NINJAL Research Digest vol.4 (2018.9)

ISSN 2432-9207
国立国語研究所 研究情報誌

国語研

ことばの波止場

NINJAL Research Digest

vol.4
2018.9

特集

言語資源の整備と研究成果発信

コーパス開発センター
研究情報発信センター

コラム　摩訶不思議な《文字》の定義　小林龍生

研究者紹介　松本 曜　金水 敏

著書紹介

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
NINJAL

コーパス開発センター

浅原正幸
ASAHARA Masayuki

あさはら まさゆき●准教授／専門は自然言語処理。奈良先端科学技術大学院大学修了、博士（工学）。2012年に本研究所着任。

コーパス開発センターのしごと

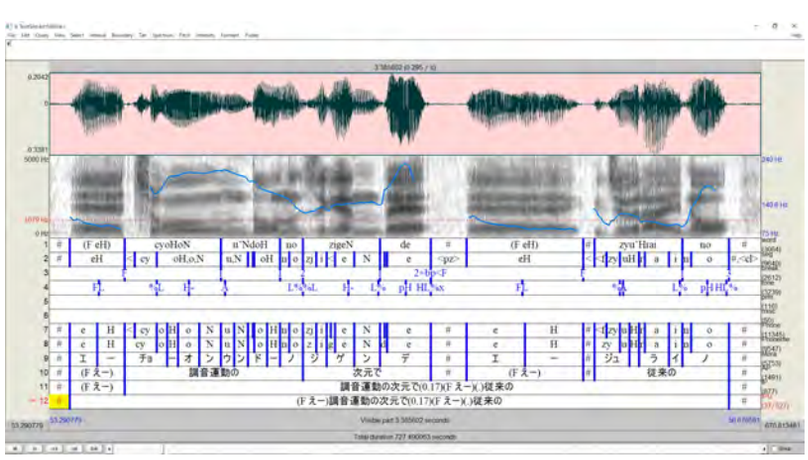
コーパスの開発には、言語学的な知識・工学的な技術・経営学的な生産管理の三つが必要になります。それぞれ特殊な技能が必要ですが、コーパス開発センターはコーパスの整備に必要な技能を持つ人員により構成されています。

コーパス開発センターは従前より国立国語研究所で整備してきたコーパスおよびツールの公開・維持・管理を行います。コーパスである『日本語話し言葉コーパス』『現代日本語書き言葉均衡コーパス』『国語研日本語ウェブコーパス』、辞書である『UniDic』『分類語彙表』、検索ツール「少納言」「中納言」「梵天」などが対象です。

コーパス開発センターは先進的なコーパスの整備を進めています。係り受けの国際的な標準であるUniversal Dependencies のツリーバンク、分類語彙表を用いた意味情報付与コーパス、視線走査装置を用いた読み時間コーパス、音声データベースなどの開発を進めています。また、研究系の各領域においても新しいコーパスの開発を行っています。これらを一括で検索可能なツールの開発もしています。それに向けて、研究所内のコーパス開発プロジェクトに対して、様々な形で支援を行うのもコーパス開発センターのしごとの一つです。

既存コーパスの維持管理

『日本語話し言葉コーパス（Corpus of Spontaneous Japanese：CSJ）』（下図）は、日本語の自発音声を集めて研究用情報を付加した話し言葉研究用のデータベースです。国立国語研究所・情報通信研究機構・東京工業大学が1999年～2004年に共同開発したもので、音声情報処理、自然言語処理、日本語学、言語学、音声学などの分野で利用されています。



『日本語話し言葉コーパス』の音声ラベルデータ

音声データ661時間、転記テキスト752万形態素と世界有数の規模です。節単位情報、分節音・イントネーションラベル、係り受け構造などが含まれた USB を有償頒布しています。

『現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese：BCCWJ）』は、現代日本語の書き言葉の全体像を把握するために構築したコーパスであり、現在、日本語について入手可能な唯一の均衡コーパスです。2006年～2010年に国立国語研究所で開発しました。書籍全般、雑誌全般、新聞、白書、ブログ、ネット掲示板、教科書、法律などのジャンルにまたがって１億430万形態素のデータを格納しており、各ジャンルについて無作為にサンプルを抽出しています。国語研が規定した短単位・長単位の2種類の単位に形態論情報が付与されています。DVD-R 4枚組のデータを有償頒布しています。

『国語研日本語ウェブコーパス（NINJAL Web Japanese Corpus：NWJC）』は、ウェブを母集団として構築した大規模テキストコーパスです。2011年～2015年に国立国語研究所で開発しました。3か月ごとに1億URL をウェブクロールすることで、250億語規模の形態素解析・係り受け解析済みデータを後に述べる検索系を介して言語研究に利用できるようにしています。また、語彙表なども無償公開しています。

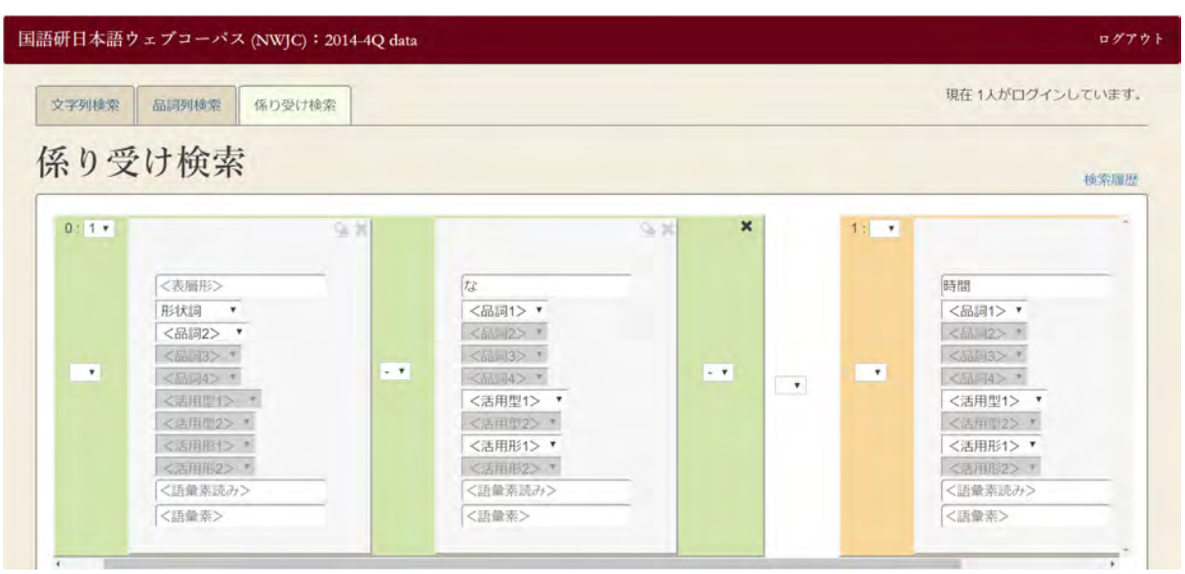
『現代日本語書き言葉均衡コーパス』を格納した検索ツール「中納言」～文字列・品詞列に基づく検索が可能

検索ツールの維持管理

コーパス開発センターは「少納言」「中納言」「梵天」と呼ばれる3種類の検索ツールを公開しています。

「少納言」はBCCWJ を公開するために開発されたウェブ上で利用可能な文字列検索ツールです。登録しなくても利用条件に承諾できる方はどなたでもご利用になれます。

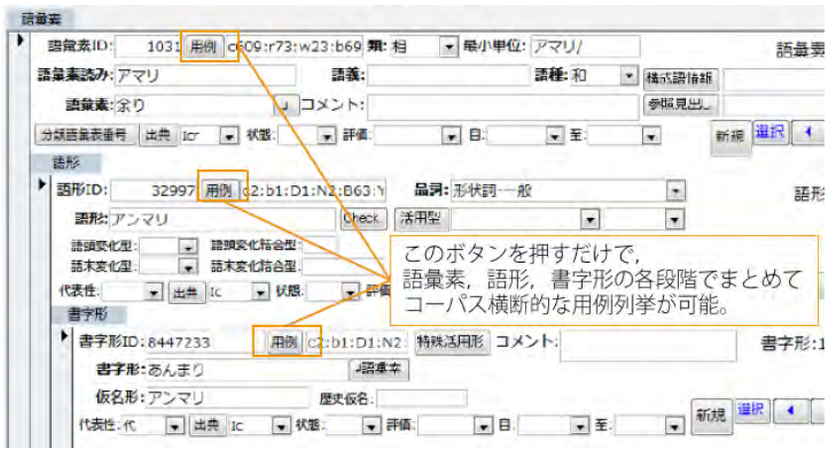
「中納言」は短単位・長単位・文字列による三つの検索が利用できるウェブアプリケーションです（右上図）。CSJ・BCCWJ の他、現在開発中の「日本語歴史コーパス（Corpus of Historical Japanese：CHJ）」「多言語母語の日本語学習者横断コーパス（International Corpus of Japanese as a Second Language: I-JAS）」が検索できます。その他、国語研に移管された「名大会話コーパス（Nagoya University Conversation Corpus：NUC）」や「現日研・職場談話コーパス（Gen-Nichi-Ken Corpus of Workplace Conversation）」が検索できます。CSJ は、有償版購入者のみ音声配信サービスが利用できます。また CHJ は、国語研所蔵の原文の画像・小学館の「ジャパンナレッジ」



『国語研日本語ウェブコーパス』を格納した検索ツール「梵天」～文字列・品詞列・係り受けに基づく検索が可能

ほか他機関の原文画像へのリンクを利用することができます。なお、中納言の利用には登録が必要です。

「梵天」はNWJCを公開するために開発された検索ツールです（前ページ下図）。一般公開版では250億語規模のテキストから高速に文字列検索できます。高機能版では形態論情報や係り受けに基づく高度な検索が可能です。背景色で係り受け関係を表したり、形態論情報がポップアップで表示されたりします。ダウンロードして、コーパス管理ツール「ChaKi.NET」で開くとより高度な分析が可能です。なお、高機能版の利用には講習会の参加が必要です。



形態論情報データベース「UnidicExplorer」

語彙資源の維持管理

コーパス開発センターではコーパスだけでなく二種類の語彙資源を整備しています。一つは『UniDic』で、もう一つは『分類語彙表』です。

『UniDic』とは、国立国語研究所の規定した斉一な言語単位（短単位）と、階層的見出し構造に基づく電子化辞書設計方針および、その実装としてのリレーショナルデータベースである UniDic データベースと、そのデータベースからエクスポートされた短単位をエントリ（見出し語）とする、形態素解析器「MeCab」用の解析用辞書解析用UniDicの総称です。コーパス開発センターでは、国語研所内で開発されるコーパスに形態論情報を付与する際に、その形態論情報に関する情報を管理するとともに「UnidicExplorer」と呼ばれるデータベースを介して UniDic データベースの情報を各プロジェクトに提供します。同データベースからエクスポートされた解析用 UniDic は『現代書き言葉 UniDic』『現代話し言葉 UniDic』『古文用UniDicS』を公開しています。解析用GUIとして Windows OS で動作する「ChaMame」と「Web 茶まめ」の2種類を無償公開しています。

『分類語彙表』とは、「語を意味によって分類・整理したシソーラス（類義語集）」です。昭和39年（1964年）に出版された初版『分類語彙表』（現在は絶版）は、現代日本語の本格的なシソーラスとして幅広く活用されてきました。その後、収録語数を増やした『分類語彙表－増補改訂版－』が刊行されましたが、研究開発用にそのデータベース版を用意しています。

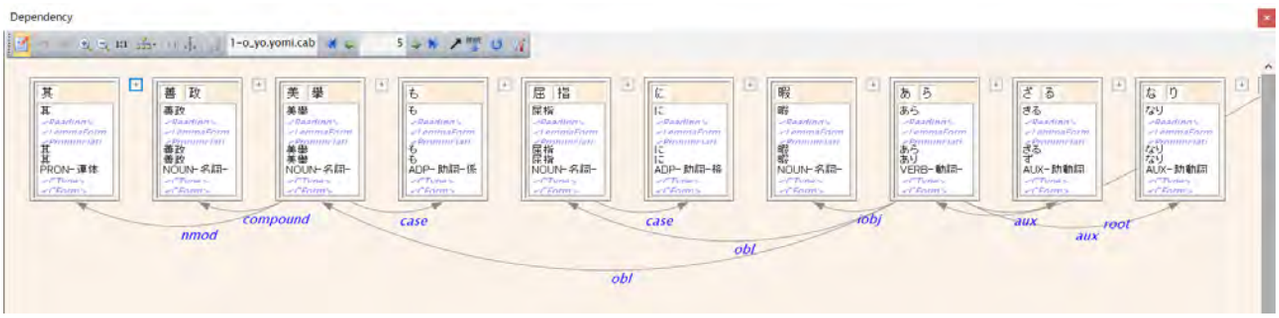
2018年2月にこの二つの語彙資源をつなぐ、新しい語彙資源『wlsp2 unidic』を公開しました。分類語彙表番号とUniDic の語彙素番号の対応表で、これを利用することで形態素解析とともに、その語彙素に対して割り当て可能な分類語彙表番号を自動展開できるようになりました。「ChaMame」のオプションを用いることで、プログラムを書かなくても分類語彙表番号付与ができるようになりました。

プロジェクト：「日本語言語資源の包括的高度共同利用環境の整備」

コーパス開発センターでは、管理業務のほかに研究も進めています。一つは国語研のコーパスの共同利用を進めるためのプロジェクト「日本語言語資源の包括的高度共同利用環境の整備」です。

検索ツール「中納言」をベースとして、さまざまな機能追加をすすめており、CSJの音声配信機能やBCCWJや I-JAS の付加的情報のダウンロードサービスなども本プロジェクトの成果です。

また 2021年度までに、「中納言」



『日本語歴史コーパス』のUniversal Dependencies～コーパス管理ツール『ChaKi.NET』による

に登録されているコーパスを横断検索するシステムを構築します。

プロジェクト:「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」

もう一つは共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」です。コーパスを用いた先進的な研究を進めるためには、付加情報が不可欠です。その中でも扱いに技術を要する統語・意味・音声の三つのアノテーションを研究対象として、他機関との共同研究を進めています。

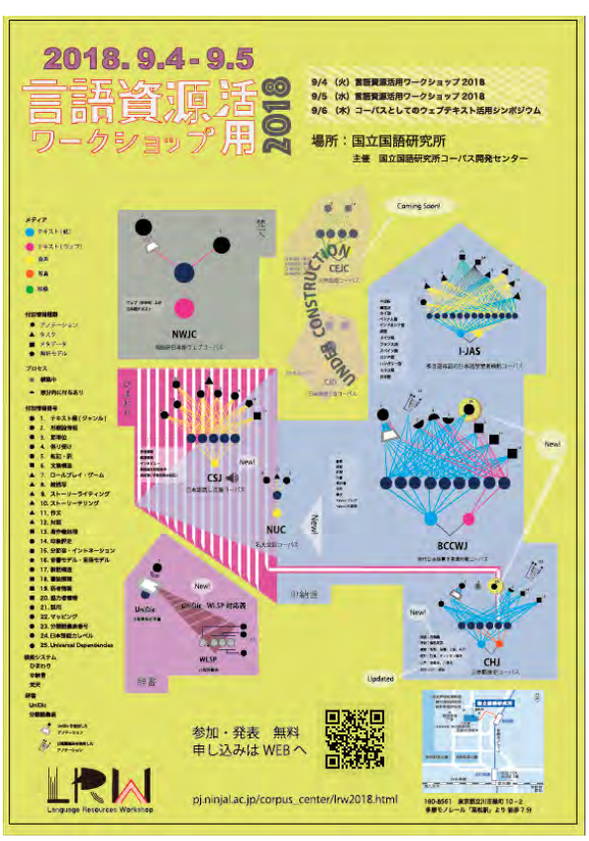
統語班（係り受け班）は、国際的な係り受けアノテーション基準 Universal Dependencies (UD) に基づく日本語の言語資源整備を進めています。UD は2014年にはじまったオープンコミュニティで、古語・危機言語を含む60言語以上のデータを公開しており、国際会議 CoNLL-2018 の多言語依存構造解析の Shared Task のデータセットとして利用されています。その他、係り受け構造付きデータを用いた基本語順の研究などを進めています。

短単位書字形	語彙素番号	短単位分類語彙表番号	短単位分類語彙表ラベル
国民	12579	1.2301	体-主体-人物-国民・住民
審査	17337	1.3065	体-活動-心-研究・試験・調査・検査など
受ける	3089	2.3770	用-活動-経済-授受
最高	13703	1.1920	体-関係-量-程度・限度
裁	13639	1.2720	体-主体-機関-公共機関
9	8929	1.1960	体-関係-量-数記号（一二三）
裁判	13810	1.3611	体-活動-待遇-裁判
官	7457	1.2411	体-主体-成員-管理的・書記的職業
司法	15710	1.3601	体-活動-待遇-国務
改革	5655	1.1500	体-関係-作用-作用・変化
アンケート	1337	1.3132	体-活動-言語-問答

『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アノテーション

意味班（語義班）は、『分類語彙表』を中心とした言語資源整備を進めています。現在、BCCWJ・CSJ・CHJ に対する分類語彙表番号アノテーションを進めています。同アノテーションデータを用いて、単語に対して意味情報を悉皆付与する all word WSD の技術について研究を進めています。これにより、コーパスが「意味」により引けるようになります。また、同データを用いた、比喩表現の調査を進めています。

音声班は、音声コーパス整備に必要な技術の研究を進めています。音声の時間情報と書き起こしとの対応関係を取る、テキスト—音声アラインメントの環境を整備して、研究所内のコーパス開発の支援を行っています。また、音声分析用のフリーソフトウェアである Praat のコーパス開発への適用などについて助言を行っています。研究としては、日本語・中国語・モンゴル語の音声データベースの整備や、調音運動データベースの構築を進めています。



言語資源活用ワークショップ2018 ポスター

言語資源活用ワークショップ

国語研で整備しているコーパス・語彙資源を用いた研究に関する情報交換をする場として、また、二つのプロジェクトの成果を発表する場として、毎年9月に「言語資源活用ワークショップ」を開催しています（初回の2016年度開催分のみ3月開催）。発表論文は、「国立国語研究所学術情報リポジトリ」に掲載されます。また2017年度開催分から、学生の発表に対して、互選に基づく優秀発表賞を設定しました。

さらにワークショップの前後に特定のテーマを取り扱うシンポジウムも開催しています。2016年度は、「語彙資源活用シンポジウム」と題し、紙の辞書・電子化辞書のそれぞれの専門家を招いて、辞書に関する様々な話題を提供してもらいました。2017年度は、国立情報学研究所データセット共同利用研究開発センターと共同で「音声資源活用シンポジウム」を開催しました。2018年度は、「コーパスとしてのウェブテキストシンポジウム」を開催しました。

さらにワークショップの前後には各種ツールの講習会も企画しています。

コーパス開発センターウェブサイト：
http://pj.ninjal.ac.jp/corpus_center/

PROJECT

言語資源の整備と研究成果発信

研究情報発信センター

高田智和

TAKADA Tomokazu

たかだ ともかず●准教授／専門は国語学。北海道大学大学院修了、博士（文学）。2005年に本研究所着任。

研究情報発信センター

研究情報発信センターは、国立国語研究所の研究成果の公表、国立国語研究所が実施した調査資料の蓄積と保存、研究文献情報の発信を行っています。研究者コミュニティーに向けて、国立国語研究所の共同利用事業を推進するためのセンターです。

機関リポジトリ

国立国語研究所の研究成果である報告書や論文は、国立国語研究所学術情報リポジトリ（https://repository.ninjal.ac.jp/）で公開しています。「国立国語研究所報告」や「国立国語研究所年報」をはじめ、「国立国語研究所論集」の収録論文や「NINJAL フォーラムシリーズ」などの研究成果を、オープンアクセスで提供しています。

データベース・データセットの公開

国立国語研究所の研究成果には、報告書や論文以外にも、日本語研究・日本語教育に関する各種データ集があります。機械可読のものは、データベース、データセットと呼ばれます。研究情報発信センターは、データ集のWeb公開も行っています。

公開データには、次のようなものがあります。

- X線映画「日本語の発音」
- 岡崎敬語調査データベース
- 沖縄語辞典 データ集
- 外来語定着度調査
- 「学校の中の敬語」アンケート調査データ
- 『日本語教育のための基本語彙調査』データ
- ことばに関する新聞記事見出しデータベース
- 雑誌『国語学』全文データベース
- 鶴岡調査データベース
- 寺村誤用例集データベース
- トピック別アイヌ語会話辞典
- 『日本言語地図』地図画像
- 日本語学習者会話ストラテジーデータ
- 日本語学習者会話データベース
- 日本語学習者会話データベース 縦断調査編
- 日本語学習者による，日本語・母語対照データベース
- 日本語観国際センサス
- 発声発語訓練例文集
- 複合動詞レキシコン
- 『方言談話資料』データ
- 『幼児・児童の連想語彙表』データ

これらのデータベースは、国立国語研究所ウェブサイト（https://www.ninjal.ac.jp/database/）で公開しています。

研究資料室

研究成果である報告書や論文、データ集の作成には、基礎調査が必

要です。方言や言語生活の研究であれば、話者にインタビューをして調査票を作り、分析のための情報カードを作成します。近年は、インタビューの録音や録画を撮ります。

書き言葉の場合も同様で、新聞や雑誌の語彙調査では、語彙カードや集計表を作成します。

こういった研究成果に至る過程の調査資料（いわば中間段階の資料）も、国立国語研究所では収集・保存をしています。紙の資料はもちろん、録音のカセットテープや録画のビデオテープも含まれます。

国立国語研究所の調査資料は、研究資料室で集中管理をし、来館利用の形で、研究者に提供しています。来館利用の方法は、国立国語研究所のホームページ（https://www.ninjal.ac.jp/info/aboutus/material-room/）を参照してください。

また、調査資料の目録はWeb公開しています（「国立国語研究所研究資料室収蔵資料」https://rmr.ninjal.ac.jp/）。現在、約240の調査資料（資料群）を保存しています。

主な収蔵資料には次のようなものがあります。

●山形県鶴岡市および附近の農村における言語生活調査

1950年に山形県鶴岡市で実施した言語生活の実態調査です。日常の言語生活と社会環境との関わりや、共通語の普及状況を把握することが目的でした。鶴岡調査はその後20年間隔で、1971年、1991年、2011年に実施され、世界最長の実時間調査となりました。（報告書：『言語生活の実態―白河市および附近の農村における―』1951年ほか）

●雑誌一般の用語の概観調査

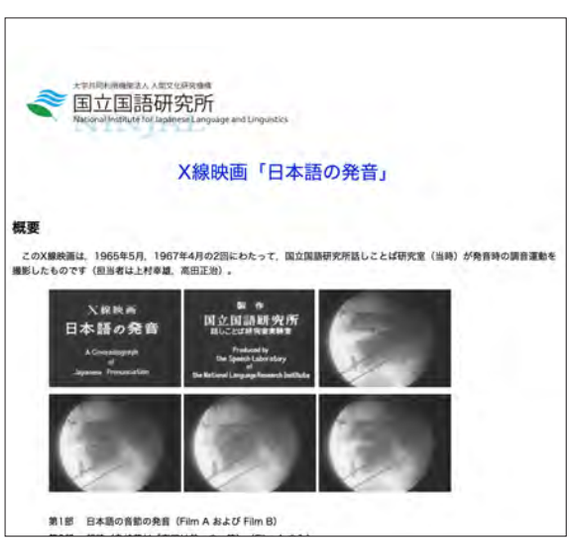
1956年発行の雑誌90種を対象とした、用語・用字の実態調査です。ランダムサンプリングの手法を導入し、言語の統計分析を開拓しました。（報告書：『現代雑誌九十種の用語用字』1962～64年）

●電子計算機による新聞の語彙調査

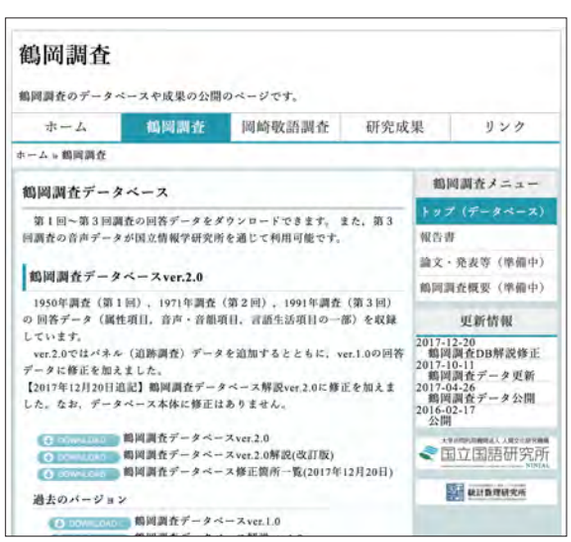
電子計算機を導入した最初の語彙調査です。1966年発行の新聞（朝夕刊1年分）を対象とし、電子計算機

国立国語研究所学術情報リポジトリ　トップページ

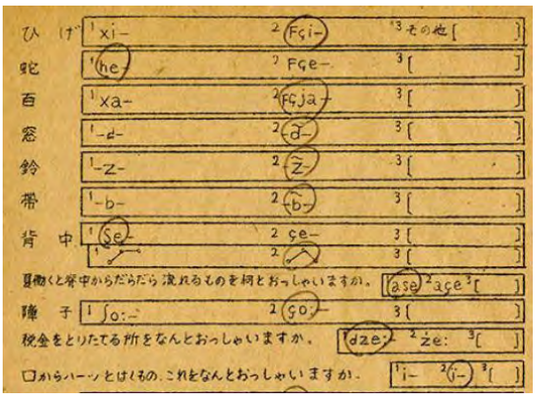
国立国語研究所　研究資料室収蔵資料



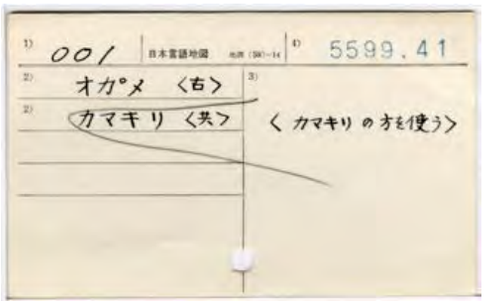
X線映画「日本語の発音」



鶴岡調査データベース



実際の調査で使用された調査票



情報カード

で日本語を分析する手法を開発しました。（報告書：『現代新聞の漢字』1976年）

●就学前児童の言語能力に関する全国調査

幼児が言語・文字をどのように習得し、どのように使用するか、またその要因が何かを明らかにするため、1967～74年に実施した調査研究です。（報告書：『幼児の読み書き能力』1972年ほか）

●日本語教育における基本文型に関する研究

日本語学習者がどのような日本語を用いて日本語母語話者とコミュニケーションを行っているかを調査し、コミュニケーション障害の要因や誤用の背景を明らかにするため、1981～84年に行った調査研究です。

●「外来語」言い換え提案

2002～06年に実施した「「外来語」言い換え提案」と、言い換え提案のための意識調査（全国調査）の資料です。（関連書籍：『分かりやすく伝える 外来語言い換え手引き』2006年、ぎょうせい）

●「病院の言葉」を分かりやすくする提案

医療従事者と患者・家族とのコミュニケーションの円滑化を目的とした、難解な医療用語の言い換え提案と、そのための意識調査、コーパス調査の資料です。（関連書籍：『病院の言葉を分かりやすく―工夫の提案―』2009年、勁草書房）

所蔵音源・映像資料

過去の調査研究で収集した録音音源と録画映像は、オープンリール、カセットテープ、8mmフィルム、ビデオテープなど、さまざまな記憶媒体で保存しています。総数はおよそ４万点です。

しかし、記憶媒体は経年劣化を起こします。また、再生用機材が生産中止になり、再生が難しくなったものもあります。そのため、録音音源と録画映像の保存と再利用のため、パソコンで視聴できるように、デジタル化を進めています。デジタル化音源・映像は、国立国語研究所内で利用できるよう、「所蔵音源・映像データベース」に蓄積しています。デジタル化音源・映像も、来館利用の形で、研究者に提供しています。

主な音源・映像資料には次のようなものがあります。

●談話語の実態

共通語による日常談話を分析するために、1952～53年に録音しました。文字起こし原稿やKWICも作成されています。（報告書：『談話語の実態』1955年）

●待遇表現の実態：松江24時間調査資料から

1963年に松江市のある市民の家庭内での一日の発話をすべて録音しました。文字化資料には文・文節・形態素の切れ目を付加し、コンピュータ処理にも利用しました。（報告書：『待遇表現の実態―松江24時

所蔵音源・映像データベース

間調査資料から―』1971年）

●談話行動の実験社会言語学的研究

1976～78年に東京と大阪で座談場面を録画しました。言語的・非言語的な観点から、言語行動様式を分析することを目的とした調査研究です。（報告書：『談話行動の諸相：座談資料の分析』1987年）

●企業の中の敬語

会社内での敬語意識と敬語使用を解明するために、1975～77年に面接調査を行い、その様子を録音しました。（報告書：『企業の中の敬語』1982年）

●方言録音文字化資料に関する研究

1977～85年度の文化庁調査「各地方言収集緊急調査」において、全国224地点の方言談話が録音・文字化されました。調査終了後、録音音源と作成資料は国立国語研究所に移管され、一部は『全国方言談話データベース日本のふるさとことば集成』（2001～08年、国書刊行会）として刊行されています。

日本語研究・日本語教育文献データベース

国立国語研究所は、創設以来、日本語研究と日本語教育に関する研究文献情報（論文や図書の書誌情報）を収集してきました。『国語年鑑』『日本語教育年鑑』として冊子を刊行してきましたが、これを引き継ぐ形で、2011年に「日本語研究・日本語教育文献データベース」（https://bibdb.ninjal.ac.jp/bunken/）を公開しました。

このデータベースには、1950年からの研究文献情報を収録し、データ件数は現在約23万件です。近年は、日本国内の学術雑誌だけでなく、韓国をはじめとして国外の研究文献情報の収集を進めています。

日本語研究・日本語教育文献データベース

また、各大学・各学協会のリポジトリで、論文本文の公開も進んでいます。「日本語研究・日本語教育文献データベース」には、これらの公開論文へのリンク情報を付与し、検索結果から論文本文へアクセスできるようにしています。

ことばの研究の「オープンサイエンス」を目指して

科学的な研究は、誰もが結果を検証できることが大切です。実態調査に基づいて研究成果を公表したとしても、研究成果に至る過程の調査資料も公表されていなくては、第三者が検証することはできません。

国立国語研究所は調査資料を保存してきましたが、調査資料の公表にはあまり熱心ではありませんでした。プライバシー保護のため、一部の調査資料に利用制限を設けるとしても、可能な範囲で検証可能な環境を整えていくことが、研究の発展に必要なことだと考えています。研究情報発信センターは、ことばの研究の「オープンサイエンス」を模索していきます。

カセットテープ・オープンリールテープなども所内に保存されている

摩訶不思議な《文字》の定義

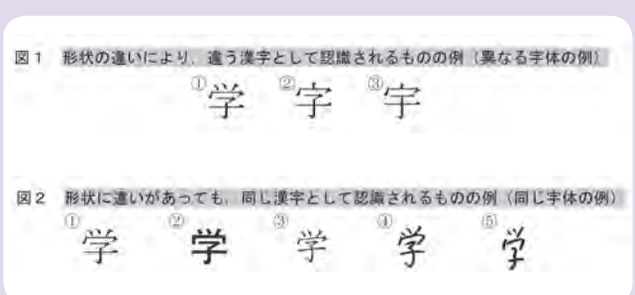
文字情報促進協議会 会長
小林龍生 KOBAYASHI Tatsuo

平成28年2月29日、文化審議会国語分科会は、報告「常用漢字表の字体・字形に関する指針」を発表した。この報告は、ISBNが付いて、三省堂から一般書籍としても発行されている。

この報告、早稲田大学教授の笹原宏之さんと文化庁国語課の武田康宏さんが中心となって纏(まと)められた渾身の力作で、単に常用漢字に留まらず、戸籍や住民基本台帳などに用いられる人名用の漢字を論じる際にも依拠するに足る貴重な指針となっている。特に、「第1章 2 常用漢字における字体・字形等の考え方」は、日本語学を専門としない一般の人びとにも分かりやすく、かつ、字体と字形の議論の層の違いが明確に述べられており、まさに白眉と言えよう。

そもそも「改訂常用漢字表」を見ると、『表の見方及び使い方』の4の項に、「字体は文字の骨組みであるが」とさらりと触れられているだけで、詳細な定義など書かれていない。この指針では、この部分を、例示とともにずいぶん丁寧に説明してくれている。

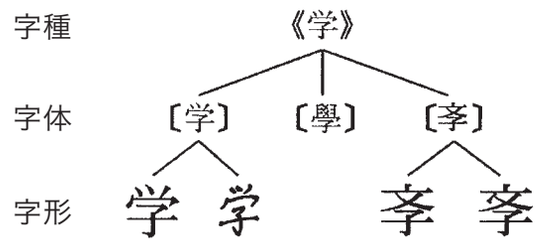
指針の本文は、文化庁のホームページにも公開されているので、そちらを参照していただくこととして、ここでは、例示されている図だけを引用しておこう。



この部分に目を通していて、ぼくは、何とも形容しがたい既視感（デジャヴ）を覚えた。う〜む、どこかで見たことがある。

しばらく黙考して、はたと思いついた。高田さんの論文だ。ずっと以前、高田智和さんから別刷りをもらった『日本語科学』23(2008年4月)95-110「行政用文字の調査研究―汎用電子情報交換環境整備プログラム―」（高田智和、井出順子、虎岩千賀子共著）に掲載されている図と同じ《学》の字が例として挙げられているのだ。

「字種」「字体」「字形」の階層構造



もしかしたら、学界では、字体と字形の違いを論ずる際、《学》の字を用いることがお作法として定着しているのかもしれないが、高田さんの図は、IPA（情報処理推進機構）の報告書などに随分と引用させてもらった。

この文化審議会の報告の元となった文化審議会の審議会資料も、渡りに船と利用させてもらって、JST（科学技術振興機構）が発行していた『情報管理』誌に「字体と字形の狭間で」という小論を書いた[1]。

そうしたら、何かの会合の後、武田さんが、えらくこの駄文を褒めてくれた。

この《字体》と《字形》の関係は、高田論文や文化審議会報告などを読んで、分かってしまえば何と言うこともないのだが、世上ではこの違いが混同され、本来《字体》レベルでなされてしかるべき議論に、《字形》レベルの相違が紛れ込んで、議論を錯綜させることがしばしばある。また、具体的な字形の相異を、同一字体内の微細な差異と捉えるか、字体レベルの差異と捉えるかは、その用途や文化的な背景によって随分と異なる。極言すると、論者が10人いたら、議論は100通りある、といった塩梅になる。

一つだけ卑近な例を挙げておくと。小学生でも知っている《次》の字。

ユニコードのIVD[2]に登録されているAdobeのAJ1 collectionと文字情報基盤事業のMoji_Joho collectionの《次》の字のところを見ると。

印刷業界ではデファクトスタンダードとして定着しているAJ1-6のcollectionでは、次のように3種類の異なる字体が掲げられている。

1 https://www.jstage.jst.go.jp/article/johokanri/58/3/58_176/_html/-char/ja

2 Ideographic Variation Database。同一の符号位置に統合される複数の字体を区別するためのメカニズムであるVS(Variation Selector)を統合漢字に適応し、基底文字とVSの組をIVS(Ideographic Variation Selector)として登録するためのデータベース。

6B21

次 E0100 Adobe-Japan1 CID+2253

次 E0101 Adobe-Japan1 CID+13799

次 E0102 Adobe-Japan1 CID+13800

それに対して、筆者も係わってきた文字情報基盤整備事業の成果物である Moji_Joho collection では、２種類の字体のみが掲げられている。

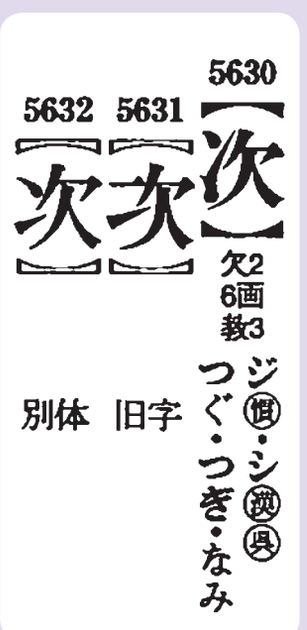
6B21

次 E0103 Moji_Joho MJ014749

次 E0104 Moji_Joho MJ014748

ここで筆者は、「３種類の字体」「２種類の字体」という書き方をしたが、内実は、「AJ1-6では次の字を３種類の字体に区別し」「MJでは《次》の字を２種類の字体に区別している」というのが正確なところであろう。

ちなみに、（漢籍ではなく）日本で用いられてきた漢字という側面に注目して編纂された新潮社の『日本語漢字字典』で《次》の項を見ると。



この辞書は、JISの符号化文字集合の策定にも係わった新潮社校閲部の小駒勝美さんの力作なのだが、《旧字》《別体》という用語を使い分けて、字体差に係わる面倒な議論からうまく逃れている。《字体》と《字形》という言葉は、面白いことに、日本の工業標準（JIS）X 0213では、下記のように、あえて対応する英語表記を避けて、ローマ字表記のみを記している。

i）字体（ZITAI）図形文字の図形表現としての形状についての抽象的概念。

h）字形（ZIKEI）字体を，手書き，印字，画面表示などによって実際に図形として表現したもの。

ぼく自身は、《字体》を glyph（文字の抽象的な図形概念）、《字形》を glyph image（個々の文字の具体的な可視化表現）に対応付けて用いているが、いずれにしても、冒頭に挙げた指針の字体、字形概念と大きく食い違っているわけではない。

ところで、いわゆる符号化文字集合の世界には、《字体》《字形》の区別どころか、《文字》というわけの分からない存在がある。英語では、character。

現在では、スマートフォンからネットワーク上の大規模データベースまで、文字情報のやりとりには、いわゆるユニコードが使われている。公的規格としては、ISO/IEC JTC1/SC2[3]が策定しているUCS[4]が相当する。

この翻訳規格である、JIS X 0221を見ると、《文字》の定義は、下記のようになっている。

「文字（character）データの構成、制御又は表現に用いる要素の集合の構成単位」

なんのことやら。

さらにやっかいなことに。

「図形記号は、文字の代表的な可視化表現とみなさなければならない。この規格群は、各文字の形を正確に規定しようとするものではない。文字の形は、採用するフォントデザインに左右されるものであり、この規格群の適用範囲外とする」（17 第2パラグラフ）

ここで、《図形記号》は graphic symbol の訳で、graphic symbol は、図形文字（graphic character）または合成列（composite sequence）の視覚表現。

これまた、なんのことやら。

蛮勇をふるってまとめると。

符号化文字集合にとって大切なのは、《文字》の具体的な形ではなく、対象となる文字集合の中で、文字集合を構成する要素（＝文字）が排他的に他の要素と区別出来ること。

規格票に印刷されている図形は、《文字》に対応付けられる《字形》の一例で、単なる参考情報。

要は、情報技術的に区別する必要があるものが区別出来ればいいわけで、社会生活上必要のない微細な差異には拘泥する必要がない、ということなのだろう。

とはいえ、この「社会生活上必要」という言葉が、また厄介者で、国や地域によっても、使われる文脈によっても、さらには個人的なコノテーションによっても異なってくる。

高田さんの論文や文化審議会報告によって、《字体》と《字形》の理論的な区別はよく分かったが、その区別がどう適用されるかは、時と場合によって異なるという、言葉を対象として議論する際に忘れてはならない要諦にまいもどってしまった。

3　国際標準化機構(ISO)と国際電気標準会議(IEC)が共同で運営している合同技術委員会(JTC1)の下で活動している第2小委員会(SC2)

4　ISO/IEC 10646 Universal Multi-Octet Coded Character Set。翻訳規格として、JIS X 0221国際符号化文字集合がある。

研究者紹介 007

松本 曜

理論・対照研究領域 教授

「泣く」「泣きすがる」「泣き落とす」～語がもつ意味を、より深く

まつもと よう●1960年札幌市出身。スタンフォード大学言語学科博士課程を修了後、神戸大学教授などを経て、2017年10月に国語研に着任。単著に『Complex predicates in Japanese』（くろしお出版）。最新刊は、『日本語語彙的複合動詞の意味と体系』（ひつじ書房、共著）。

── 先生のご研究を簡単に教えていただけますか。

専門は語彙の意味論です。単語の意味を考えるだけでなく、単語と単語がどういう関係を持っているのか、そして単語の意味がその単語の文法的な性質をどのように決めているのか、単語と単語が組み合わされてどのような単語を作るのか、そして単語の意味が私たちの世界観とどのように関係しているのか、などを研究するという分野です。

── なぜ意味論に興味を？

高校２年生の英文法の授業が非常におもしろかったのが言語に関心を持った最初でした。言葉の法則性みたいなものがすごくおもしろくなって。当時は理系志望だったのですが、その授業がきっかけになって後で文系に進みました。当時の愛読書が研究社の『英和中辞典』という辞典で、授業の合間に読んでいました。この単語とこの単語はどのように意味が違うんだろうとか、この動詞はこの構文に使えるのに、どうしてこちらの動詞は使えないんだろうとか、そういうのを考えながら辞書を読むのが趣味で。考えてみると、そのときと同じことを今している感じです（笑）。

── 先生の研究対象の言語としては何語が中心なのでしょうか。

英語圏の意味理論を通して日本語を見る研究が一番多いですね。だから対象としては、日本語が多く、全体の６割ぐらい、あとは英語が３割、他の言語が１割ぐらいです。結局、どの言語を研究しても語の意味の性質を知ることはできるわけで、そうであれば日本語のほうが研究をしやすいという現実があります。

ただ私の場合、日本語を研究していても常に言語一般がどうなっているかというところに関心があります。

── 現在の研究の中心は？

大きなプロジェクトが二つあります。一つはいろいろな言語の移動動詞（人やものが移動する表現を伴う動詞（例）「走る」「投げる」）の性質を研究して、そこから日本語を見るというものです。世界の15くらいの言語話者を対象に、同じビデオを見せて、ビデオに出てくる移動事象を、それぞれの話者がどう表現するかを統一的に比較するプロジェクト（実験）で、10年くらい行っています。

もう一つはフレーム意味論という理論に基づく動詞の研究です。

例えば「泣く」という動詞の意味は、普通は、悲しくて目から涙を出すとかそういう意味だと思いますよね。ところが複合動詞の中には「泣きすがる」とか「泣きつく」とか「泣き落とす」などがあります。どうしてこういう動詞があるかというと、泣くことによって他の人に訴えかけたり、感情をぶつけることによって相手の気持ちを変えるとか、そういうことが行われるからだと思います。

ということは、その「泣く」という動詞の背景にある知識の中には、泣くことに伴って起こる出来事についての情報も含まれているのではないかということになります。つまり泣くという行動をしたら他の人にどういう影響があるかとか、泣くことで人はどういうことをするだろうかとか。そういうのも「泣く」と関連する知識の中に含まれているんじゃないかと。だからこそ、先ほどの表現が成立しているんじゃないかなと思うんですね。

そのような背景的な知識を含めて意味を理解する理論を「百科事典的意味論」と言いますが、動詞の意味記述には、おそらく従来考えられていたよりも広い範囲の情報が必要ではないかと考えています。そう考えると、いろんなことが説明できるようになると考えています。それをコーパスなどを使って調査しています。

── 研究で大変なことはありますか？

目指していることの一つに、網羅的に研究したいということがあります。特定の動詞だけ取り上げて議論しても、その結果がどこまで広く動詞に当てはまるのか分からないからです。例えば、複合動詞の研究では、4000近くの複合動詞を確認しました。他の研究においても本当はすべての動詞を見たいのですが、その一方で一つ一つの動詞を調べるコーパス研究や実験研究はすごく時間がかかります。一つの動詞についてたくさん用例を見たい、実験調査をしたい、それでいながらすべての動詞を見たいといっても無理ですよね。バランスを取るのが難しいです。

── 神戸大学から国語研に移られてから1年が経ちました。

大学時代よりも研究ネットワークの重要性を感じるようになりました。国語研では個人の研究以上に、プロジェクトが重要な役割を果たしていて、これまで以上に他の大学の先生の研究に関わることが多くなりました。それは自分にとって、とてもよかったと感じています。

研究者紹介 008

金水 敏

客員教授・大阪大学教授

日本語史・現代日本語・役割語 多岐にわたる研究のルーツに迫る

きんすい さとし●1956年大阪府出身。東京大学助手、神戸大学助教授などを経て、1998年に大阪大学に着任。2003年に発表した「役割語」の概念は日本語研究を超えた話題に。2006年『日本語存在表現の歴史』で新村出賞受賞。日本語文法学会元会長、日本語学会現会長。

── 研究者になったきっかけは？

子どもの頃から本を読むのが好きだったんですが、高校の時に岩波新書で大野晋先生の本を読んで、「こういう世界もあるんだな」と日本語に興味を持ちました。

当時は１年生から興味関心に合わせていろんなゼミを選ぶ、少人数・演習形式の授業もあって、古田東朔（とうさく）先生のゼミで教わりました。そこですごくほめてもらったこともあり、面白いなと思ってその後、国語学に進学しました。

── そこで、今のご研究にもつながる「昔の日本語」に出会ったんですか？

古田先生のゼミでは、近世語のゼミで浮世床といった滑稽文を読んでいました。そもそも大野先生の本も『日本語の起源』ですからね。

卒論は『「は」と「が」』ですが、国語学なので国語史の授業をずっと受けていましたし、ゼミではキリシタン資料も扱いました。そういう意味では大学に入ってずっと国語史とのつながりがあり、僕自身の学問的なルーツは日本語史だと思っています。

修士の時に、山口明穂先生の脚結抄（あゆいしょう）のゼミで、存在動詞「あり」についての抄があって、それを担当して力を入れて発表して敷衍（ふえん）して修論にしました。その後、存在文についての博士論文を書き、『日本語存在表現の歴史』（ひつじ書房）になりました。そういう意味では研究者としての出発点はその存在表現の歴史で、それが博士論文までつながっていきました。

その後、ほとんど業績もないのに神戸大学の教養部に採用していただきました。この教養部がすごく面白いところで、田窪行則先生（現国語研所長）をはじめ、言語学の面白い人がたくさんいて、その人たちに影響されて生成文法や形式意味論を学びました。これらの交流は自分の研究人生にとって非常に大きかったです。

── 役割語はどのように生まれたんですか？

子ども時代から漫画やアニメ、特に『鉄腕アトム』が大好きで、お茶の水博士に憧れていました。ですから、博士語というのは結構最初から思いついていたんです。

博士が「そうじゃ」と言うイメージは自分の頭の中にもともとインプットされていましたが、それが普通の言葉遣いではないと気づくのはずっと後のことです。「〜いる」と「〜おる」の使い分けについて歴史的経緯や方言の対立を勉強していた時に、「お茶の水博士が「わしは知っておるぞ」みたいな形で「おる」を使うのはなぜだろう。これは今までの概念では、説明できないな」と思ったわけです。老人や博士になって言葉遣いが変わるなんて、現実社会で普通はないわけですから。

仮に漫画の中で特定の役割を表すのに使われているんだったら、「役割語」と呼んでみてはどうかと。フィクションの中で現実とは違った言葉遣いがあることは、江戸時代から指摘している人はいました。重要なのは「役割語」とラベルを付したことです。

いわゆる女言葉、女性語も、現実にはあまり使いません。でも「そうですわよ」「存じておりますわ」という言い方をすれば誰もがお嬢さまだと感じます。これを解決するには現実を基盤とした言語学とは違うアプローチが必要なのではと思ったんですよね。

その役割語の概念ができ、まとめた本が『ヴァーチャル日本語』（岩波書店）です。定延利之さんも似たようなことを考えていたわけですがアプローチが少し異なっていて。僕の専門はもともと日本語史ですから、歴史的なアプローチなんです。博士語や老人語も、歴史的にどういった形として進んできたかを考えたもので、それこそ大学１年の時に受けた古田先生の浮世床の影響がすごく生きてるんです。だから江戸語に博士語や老人語のルーツがあるんだというのに気づかされたのも、浮世床をやっていたからと言えるわけです。ですから幅広く研究しているように見えて、わりと全部つながっているといえばつながってるんですよね。

── いまご興味を持っている研究を教えてください。

学生さんや留学生の人で役割語に興味を持つ人がすごく多いんです。そういうこともあって、役割語を含めたキャラクターの翻訳を考えています。いまは、村上春樹翻訳調査プロジェクトを行っていて、登場人物のタイプがはっきりしていて、しゃべり方もその役割によってかなり意識して選ばれており、各国語の翻訳も多いため題材として適していると思っています。

もう一つ。存在表現から始まり意味論の勉強も結構して、指示語もやりましたので、形式意味的な枠組みを使いながら、日本語の意味論の包括的な記述研究をしたいなと。頭が動くうちに。アクティブなのは４、５年かなと思うんですけど。

神奈川大学での展示
（歴博と共同展示）

国語研での展示

弘前大学での展示

音に触れる。

音を見る。

方言を展示する国語研の初挑戦！

言語地図を作ってみよう！
傷口に貼る「絆創膏」を自分の出身地ではどう言うか？シールを貼って答えてもらう参加型地図を作りました。日本全国を席捲しているのはバンドエイドとカットバン！

方言はふつう、耳で聞き、口で伝えるものです。それを、目で見て、手で触れることのできる「展示」にするという試みを国語研が始めました。

この試みは、2017年度より始まった人間文化研究機構の「博物館・展示を活用した最先端研究の可視化・高度化事業」の一環で、国語研では、2018年7月までに既に2か所の大学（神奈川大学・弘前大学）で展示を行っています。

今後も順次日本全国の大学と協力して展示を行っていきます！

単語の違いのパターンが一目で分かる！　単語ごとに分布のパターンが違います。ユキヤケは北側、シモヤケは南側で使用されますが、他の単語は異なるパターンを示します。その違いを8枚の地図にしてみました（右の地図はそのうちの一つです）。

これであなたも琉球人！　沖縄の方言を知らない人にも沖縄の方言を見て、触って、体感してもらうことのできる展示品を作りました。右の写真にある首里方言と共通語の音の対応表を見ながら、左の写真上部の穴埋めクイズを解いて、その下のカナが書かれた駒を並べて正解すると、首里方言の発音を聞くことができます。

Book Review 著書紹介

通じない日本語
世代差・地域差からみる言葉の不思議
窪薗晴夫

平凡社新書
2017年12月

本の帯や表紙の宣伝文句は軽視されるが、出版社で何をウリにしたかが分かるし、書店で手にして買うきっかけになる。この本の帯では「パンツはいて…」「マクっていい…」というきわどい例文が書いてある。「言葉の変化／進化の裏に…法則…」は記述の態度を示す。表紙の「世代差・地域差…」は内容の２部構成を示している。これらのキャッチフレーズの効果は売行きに響く。調べてみたら、発売以来コンスタントに売れている。中身もいいからだろう。

新書だから新しい情報を期待したい。読んでみると、最近の言語変化や方言差についての多くの実例があがっている。長年の研究で丹念に集めたものがあり、国立国語研究所の研究成果『日本言語地図』『新日本言語地図』を活用した事項もある。言葉の専門家から見ると、音声に関わる部分がことに面白い。「雰囲気」が「フインキ」になるのは、「単語の末尾が長音節＋短音節だと安定するからだ」と論じる部分と、アクセントの地域差を説明する部分は、著者の本領が発揮されている。国立国語研究所の研究を分かりやすく紹介する手ごろな１冊である。

▶**井上史雄**（東京外国語大学名誉教授）

連濁の研究
国立国語研究所プロジェクト論文選集
ティモシー・J・バンス
金子恵美子　渡邊靖史　編

開拓社
2017年11月

日本語における「連濁」現象は、19世紀末のお雇い外国人ライマン氏の論文以来、未解決の問題も含めて多くの関心を集めてきた。連濁は、直接的には分節音に関する現象であるものの、語種（和語・漢語等）、音韻環境、語構造、意味、あるいはアクセントなど多くの言語事象と関っている。本書は、ライマンの法則を含む連濁の基本的諸性質の記述から始まり、研究史、生成音韻論に基づく解釈、心理言語学的アプローチなど、新たな観点からの成果が盛り込まれている。評者は、30年ほど前に一時期連濁の研究に携わっていたことがあり、通時的観点からの研究の必要性を痛感していたが、そうした研究の発展も取り上げられている。また、当時、「姫」や「紐」が何故連濁しないかについて、連濁によって唇音が連続すると発音し難いことと関連があるのではないかと思っていたが（「飛び火」の「火」が濁音化しないのと同じ）、それがOCP（必異原理）として洗練された形で説明されているのも興味深かった。本書は、連濁研究としては初めての成書であり、連濁をこれから学ぼうとする初学者にとっては好個の１冊である。また研究者にとっては、研究途上の内容も記載されているので、未解決の課題を知り、研究テーマを探るうえでも有用な書物となるであろう。

▶**佐藤大和**（東京外国語大学）

『広辞苑』第７版
新村 出 編

岩波書店
2018年1月

新しく改訂された『広辞苑』（７版）には、「ドラえもん」が載った。そのうれしさとは別に、ぼくの関心は「ことばの説明がどう変わったか」に向かう。「ナポリタン（６版）ナポリ風の料理。特にトマトソースを用いたスパゲッティ-ナポリタンをいう。⇒（７版）（「ナポリ風」の意）ゆでたスパゲッティと炒めた玉ネギ、ピーマン、ベーコンやソーセージを合わせ、トマト-ケチャップで調味した料理。」うわ、このままレシピとして使えそう。「タンタンめん（６版）辛みを利かせた挽肉やザーサイの細切りなどをのせた麺。⇒（７版）芝麻醤・醤油・ラー油などで調味し、挽肉などをのせた麺。」そうか、決め手は芝麻醤だったか！

動詞の説明は、国語研の『分類語彙表』も駆使して６千語強について再検討したそうな。「ゆ・でる（６版）①熱湯で煮る。⇒（７版）①火にかけた熱湯の中に入れ、（短時間で）加熱・調理する。」説明で「煮る」を使うのをやめたのね。「いた・める（６版）食品を少量の油を使って加熱・調理する。⇒（７版）熱した調理器具の上に少量の油をひいて、食材同士をぶつけるように動かしながら加熱・調理する。」チャーハンがパラっとしなかったのは、ぶつけ方がたりなかったからかも。辞書は読むもの、カレーは飲みもの。ページのそこかしこに潜む神を探し出そう。

▶**塩田雄大**（ＮＨＫ放送文化研究所）

編集後記

特集では、コーパス開発センターと研究情報発信センターについてご紹介しました。ふたつのセンターは、相互に、また、各研究プロジェクトと連携して、言語資源の開発整備や共同利用、研究情報・研究資料の収集や公開などに取り組んでいます。

整備・公開するデータは、新しく調査したり収集したりするものもありますが、かつての研究の一環として蓄積した資料をよみがえらせたものもあります。

表紙の写真は、国語研究所に保存されている資料のひとつで、おもに1950〜1960年代に各地の方言を録音したオープンリールテープ（「Soni」は、ソニーが東京通信工業であった時代の表記）です。

音声を記録することは、19世紀後半から欧米でおこなわれ、再生にはレコードが使われていました。ほかに、なんとワイヤーに磁気で録音する機械もあったそうです。続いて登場するオープンリールテープは、プラスチック製のフィルムに粉末状の磁性体を塗布したものですが、初期の頃には紙製の磁気テープも用いられていました。国語研究所の資料庫には、この紙製のオープンリールテープも保管されています。同じ磁気テープでも、カセットテープは今でも時々見かけることがありますね。その後、CDやMDといった光学ディスクが一般的になり、最近では、インターネットで音声ばかりか映像までも視聴できるようになりました。

過去に録音されたオープンリールテープの音声の一部は、刊行物の付属資料や、国語研究所のウェブサイトで公開され、貴重なデータの記録・保存の役目を果たすとともに、現在の研究を進めるための基礎データとしても活用されています。

国語研究所のウェブサイトには、音声のほかにもいろいろなデータベースやコーパスがありますので、一度のぞいてみていただければと思います。（井上文子）

次号予告

研究プロジェクト紹介①

言語変異と言語変化

国語研
ことばの波止場
vol.4

平成30(2018)年9月30日発行

編集　国立国語研究所研究情報誌編集委員会
発行　国立国語研究所
〒190-8561
東京都立川市緑町10−2
電話042-540-4300（代表）
協力　くろしお出版
デザイン　黒岩二三[Fomalhaut]