日本語史研究に関する二次的情報の集約 —語誌データベースの構築—

山崎 誠（国立国語研究所）

本発表では，古辞書，言語地図，言語記事（言葉について書かれた新聞・雑誌の記事），『日本語歴史コーパス』からの頻度情報，および，日本語研究論文の情報を一度に検索できる語誌データベースの構築について説明する．これらの情報は UniDic という形態素解析用の電子化辞書の語彙素 ID をキーとして相互に結びつけられ，検索の便に供する．

Aggregation of the Secondary Information about the study on Historical Japanese: Compilation of Word History Database

Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

In this study we propose a database with which we can search the entries of Japanese old dictionaries, Japanese linguistic maps, language articles in newspapers and magazines, frequency information from Corpus of Historical Japanese, and papers of Japanese linguistics at the same time. Using lemma ID of UniDic, which is the electronic dictionary for Japanese morphological analysis, these information is combined with each other.

１．はじめに

近年，日本語史研究においてコーパスの利用がさかんになっている．その原動力となっているのは，国立国語研究所の「日本語歴史コーパス」（CHJ）[1]の充実である．

一方，日本語史研究に役立つデータであるものの，コーパス化になじまないものも存在する．例えば，ある語について調べる際に，同時代の辞書（古辞書）を参照する場合がある．日本語研究でよく参照される辞書の多くは，語釈のない，語の羅列であるものが多い．意味分類体の辞書であれば多少は語の並び順に意味があるであろうが，「日葡辞書」などの字母順の場合は，コンコーダンサーで前文脈，後文脈を示しても有益な情報とはならないであろう．

また，言語使用のレベルという観点からみると，言語作品における語の使用を一次的使用とすると，辞書の見出しとして登録された語は，二次的使用であり，いわばメタ言語的使用である．このような二次的使用は，コーパスとしてよりもデータベースの検索キーとして利用するほうがふさわしい．

次に，言語地図から得られる地理的分布の情報は，言語の歴史的分布と併せて言語変化を把握するために重要な情報となる。さらに，近代以降，新聞や雑誌で書かれた言語についての言説は，当時の人々の言語についての意識を捉える情報である．

これらの言語に関する二次的情報を集めたデータベースを作り，そこに含まれる各データを串刺しで検索できるようにすれば，個々の語をそれぞれ別個に検索せずに済み，かつ，コーパスからの頻度情報も併せて示せば，「日本語歴史コーパス」を利用した分析が進み，日本語研究のさらなる発展が期待できるだろう[2]. このようなユーザインターフェースは，とくに，研究を始めたばかりの人やコーパスに触れたばかりの人にとってコーパス利用のハードルを下げる役割を果たすことになるだろう．

本発表では，最終的に Web 上に設置するユーザインターフェースである「語誌ポータル」の構築を目指して，現在構築中である，古辞書や言語地図，言語記事，コーパスからの統計情報，研究文献の情報などの二次情報を集約したデータベースについて説明するものである．

２．「語誌データベース」の特徴

前節で述べた，二次的情報を集めたデータベースを「語誌データベース」と呼ぶことにする．語誌データベースは以下のような特徴を備えたものとして設計する．

（1）文字だけでなく元資料の画像を表示して，情報を視覚的に提供する．

（2）検索の効率化のため，形態素解析用電子化辞書である UniDic の形態論情報（語彙素，読み，語彙素 ID）を付与し，串刺し検索ができるようにする。

（3）データによっては，形態論情報ではなく，カテゴリーによる検索も併用する．

©2018 Information Processing Society of Japan

表 1 に，データの種類と検索方法を示す．

表 1　データの種類と検索方法

データ	形態論情報検索	カテゴリー検索
頻度情報	○	
古辞書	○	
言語地図	○	△
言語記事	○	○
研究文献	○	○

言語地図のカテゴリー検索が△になっているのは，文法項目として「ら抜き」「命令形」など只すねる質問があり，そのような結果を適切に検索するにはカテゴリー検索が適しているためである．

3．UniDic の利用

本研究で利用するさまざまなデータを効率よく検索するために，形態素解析用電子辞書 UniDic を利用する．UniDic は，日本語コーパスの構築の際に開発されたもので，「国立国語研究所の規定した斉一（せいいつ）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」[3]である．

UniDic を使うメリットは以下のとおりである．例えば，漢字の場合は，文字コードという形で規格が定まっていたり，あるいは『漢字字体規範データベース』（HNG）などがあり，検索のための情報整備がかなり整っている．しかし，語彙の場合は，そのような統一的なデータがないのが現状である．国語辞書の見出しも各辞書により，出入りや語形の違いなどがあり，標準的なものが存在しない．そこで，現在普及が進んでいる UniDic を使い，その語彙素 ID（語を同定するユニークな ID）を利用すれば，各データをこの語彙素 ID で結んで検索できるようになると考えた．また，『日本語歴史コーパス』『現代日本語書き言葉均衡コーパス』[4]など国立国語研究所が開発したコーパスは，解析に UniDic を使用しており，コーパスからの頻度情報を抽出するのに適している．ただし，UniDic を利用しても，見出しが存在しなかったり，あるいは，語の長さが合わなかったりする場合は，後述のように別の工夫が必要となる．

以下，語誌データベースに格納されるそれぞれの情報について順を追って説明する．

3．コーパスにおける頻度情報

コーパスにおける頻度情報は，『日本語歴史コーパス』および『現代日本語書き言葉均衡コーパス』（BCCWJ）から取得する．例えば，図 1，2 は「犬」および「猫」という語の頻度情報を示したものである．図 1 は作品ごと，図 2 は時代別に粗頻度と相対頻度（100 万語当たりの頻度である PMW）を示した．このようなグラフは，Excel のグラフ作成機能を使えば，コーパスからダウンロードしたデータを使って，描画することが可能であるが，それなりに手間がかかるものでもあるし，コンピュータリテラシーがそれほど高くないユーザにとっては有益な情報となる．

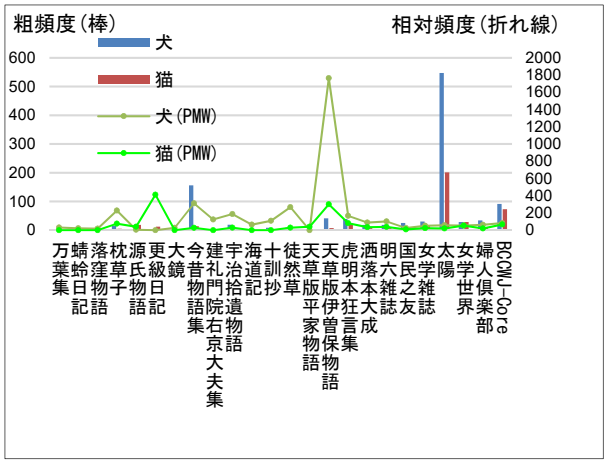


図 1　「犬」「猫」の頻度（作品別）

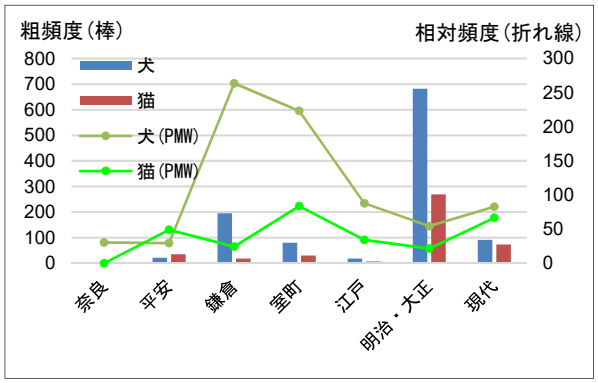


図 2　「犬」「猫」の頻度（時代別）

なお，図 3 のように『日本語歴史コーパス』のジャンル情報をもとに，ジャンル別の相対頻度のグラフを書くこともできる．図 3 は「美しい」という形容詞の例である．

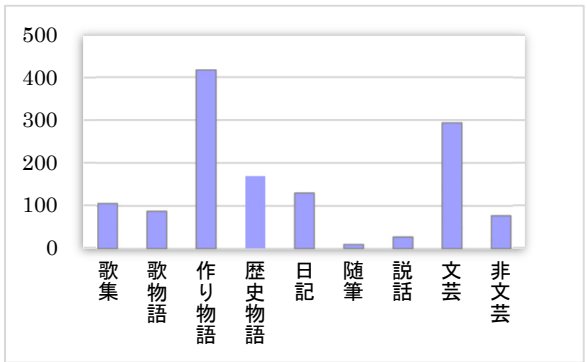


図 3　「美しい」のジャンル別相対頻度）

©2018 Information Processing Society of Japan

これらのグラフは検索の度に生成させることも可能であるが，コーパスデータはバージョンアップされるまでは固定されているため，前もって作成しておき，検索されたらすぐに表示できるようにするほうがシステムへの負担が少ない．

４．古辞書

古辞書としては，言語資料としての質的有用性，分量，入力難度，解読の難度（刊本か写本か，使われてる字種），画像の公開の有無，テキストの有無を考慮して以下のものを選定した。

①古活字版二十巻本倭名類聚鈔
底本：国立国会図書館
画像：http://dl.ndl.go.jp/info:ndljp/pid/2544216
②黒川本三巻本色葉字類抄
底本：国立国会図書館，画像：なし
③観智院本類聚名義抄
底本：天理大学図書館，画像：なし
④名語記
底本：不明，画像：なし
⑤文明本節用集（雑字類書）
底本：国立国会図書館
画像：http://dl.ndl.go.jp/info:ndljp/pid/1286982
⑥パリ本日葡辞書
底本：フランス国立図書館
画像：http://gallica.bnf.fr/ark:/12148/bpt6k852354j.r=Vocabul%C3%A1rio%20da%20l%C3%ADngoa%20de%20Japam?rk=21459;2
⑦易林本節用集
底本：国立国語研究所
画像：http://dglb01.ninjal.ac.jp/ninjaldl/bunken.php?title=ekirinbon
⑧慶長十八年版倭玉篇
底本：国立国会図書館
画像：http://dglb01.ninjal.ac.jp/ninjaldl/bunken.php?title=wagokuhen
⑨元和版下学集
底本：国立国会図書館
画像：http://dl.ndl.go.jp/info:ndljp/pid/2533531
http://dl.ndl.go.jp/info:ndljp/pid/2538644
⑩和漢音釈書言字考節用集
底本：国立国語研究所，画像：なし
⑪増補俚言集覧
底本：国立国会図書館
画像：http://dl.ndl.go.jp/info:ndljp/pid/991569
⑫改正増補和英英和語林集成（第三版）
底本：国立国語研究所
画像：http://dglb01.ninjal.ac.jp/ninjaldl/bunken.php?title=waeigorin3
⑬哲学字彙初版
底本：国立国語研究所
画像：http://dglb01.ninjal.ac.jp/ninjaldl/bunken.php?title=tetugakuzii
⑭日本辞書言海
底本：国立国会図書館
画像：http://dl.ndl.go.jp/info:ndljp/pid/992954/548

このうち，①⑬は作業をほぼ終え，現在⑥と⑫を作業中である．表 2 は，①のデータの例である．元資料からの情報に加えて，○をつけた UniDic からの形態論情報の付与している．URL は国会図書館デジタルコレクションへの該当ページへのリンクである．

表 2　『古活字版二十巻本倭名類聚鈔』のデータ構造

見出し語	潦
和訓	爾八太豆美
巻	1
部	天部
具類	雲雨類
所在	4 丁裏 7 行目
URL	http://dl.ndl.go.jp/info:ndljp/pid/2544216/13
○語彙素読み	ニワタズミ
○書字形基本形	潦
○語彙素 ID	183036

（注　○印は UniDic からの情報）

また，表 3 は，⑬のデータの例である．「形而上」という見出しは UniDic の見出しでは，複合語として扱われるので，2 単位で表される．語彙素 ID の「42889+18467」がそれに当たる．また「虚形」は UniDic に登録されていない語であるため，語彙素 ID は「*」としてある．このような語の場合は，読みや語形からの検索となる．

表 3　『哲学字彙初版』のデータ構造

見出し語	Abstract
訳語	抽象、虚形、形而上
見出し	抽象/*/形而+上
読み	チュウショウ/キョケイ/ケイジジョウ
語彙素 ID	23824/*/42889+18467
分野	-
ページ	1
割注・角書	按、易繋辞、形而上者謂之道
URL	https://dglb01.ninjal.ac.jp/ninjaldl/show.php?title=tetugakuzii&issue=1&num=9

©2018 Information Processing Society of Japan

５．言語地図

言語地図の情報として利用するのは，「国立国語研究所 言語地図データベース」[5]である．このデータベースは以下の 2 つのデータベースから構成されている．1 つは，「言語地図目録データベース」である．これは，「日本の方言学・言語地理学において 1905 年から 2006 年の間に作成された言語地図集とそこに収録された言語地図のデータベース」（[5]より引用）である．もう 1 つは，「言語地図画像データベース」である．上記「言語地図目録データベース」の中から，画像として言語地図が閲覧でき，GIS ソフトで利用できるようジオタグが付されている。

これらの言語地図は，2018 年 10 月現在 17 冊の書籍に収録されている地図を GIS 付き画像で公開している。言語地図画像データベースの例を表 4 に，その対応する画像を図 4 に示す.

表 4 「言語地図画像データベース」の構造

地図名	氷柱（つらら）
○書字形基本形	氷柱
○語彙素読み	ツララ
○語彙素 ID	60380
質問文：	−
分野	語彙
語彙（品詞）	名詞
語彙（分類）	天地
地図 ID	18963
書名	山形県方言地図
著者	矢作春樹
刊行年	1968 年 6 月 10 日
調査時期	1952−1953 年
発行所	私家版
対象地域	山形県
地図集 ID	104
URL	http://www2.ninjal.ac.jp/hogen/dp/ladp/104/18693/18693.html

2018 年 10 月現在，公開しているジオタグ付きのデータは以下の１７冊分である.

1.真田信治　「越中飛騨国境言語地図」　（真田ふみ　『越中五箇山方言語彙(5)』付．）
2.信州大学方言研究会　　『木曽及びその周辺地方の言語地図　語彙編Ｉ』
3.馬瀬良雄　『上伊那の方言』
4.馬瀬良雄　『信飛国境地帯　奈川村の方言』
5.矢作春樹　『山形県方言地図』
6.矢作春樹　『山形県方言地図』（山形県方言研究会『山形県方言概説』）
7.矢作春樹　　山形県方言地図　第２集
8.天理大学方言研究会　奈良県と三重県の境界地帯方言地図
9.学習院大学方言研究会　『千葉県夷隅川流域方言地図』
10.学習院大学方言研究会　『千葉県安房鴨川付近方言地図』
11.学習院大学方言研究会　『千葉県夷隅川流域新方言地図』
12.学習院大学方言研究会　『千葉県茂原市近傍方言地図』
13.学習院大学方言研究会 日本大学方言研究会『千葉県東金市近傍方言地図』
14.岡山大学教育学部国語研究室方言研究会『播備国境言語地図論集第１集』
15.岡山大学教育学部国語研究室方言研究会『播備国境言語地図論集第２集』
16.鏡味明克　　『兵庫岡山県境言語地図』
17 鏡味明克　　『続・兵庫岡山県境言語地図』

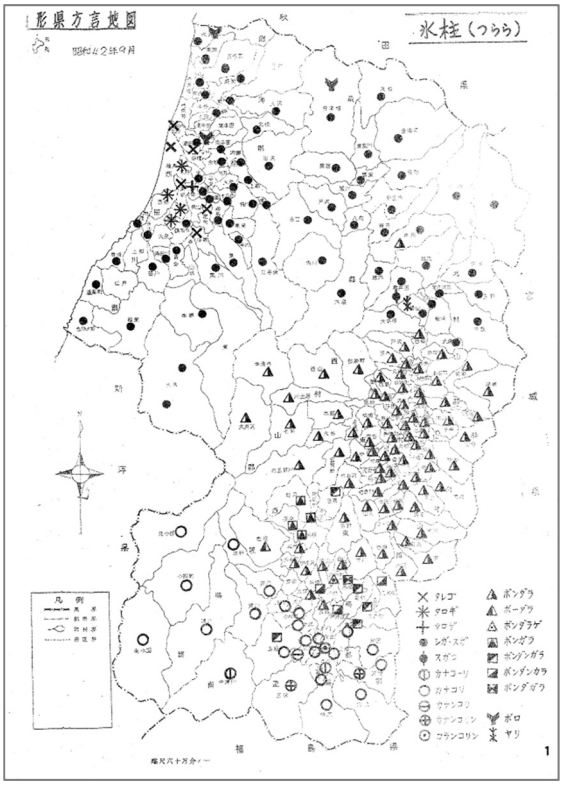


図 4「つらら」の言語地図
矢作春樹『山形県方言地図』（1968 年私家版）より

2018 年月現在，上記 17 冊の「言語地図画像データベース」のデータベースは完了（902 項目）しており，「言語地図目録データベース」のデータベース化に着手したところである.

６．言語記事

言葉について書かれた記事のデータベースである．新聞記事は，大正・昭和前期の『読売新聞』（大正 2，6，10，14 年），雑誌記事は，『文芸春秋』の大正 12 年 1 月（創刊）～昭和 20 年を対象とする．『文藝春秋』の記事には，簡単な要約を

©2018 Information Processing Society of Japan

付して，内容理解の助けとする．

2018 年 8 月現在，以下のデータがほぼ完了している．UniDic の語彙素 ID の付与はこれからである．

・新聞記事

『読売新聞』（大正 2 年：306 項目，同 6 年：310 項目）

・雑誌記事

『文藝春秋』大正 12 年 1 月（創刊）～昭和 6 年，昭和 12 年～20 年，852 項目

表 5 は読売新聞の，表 6 は文文藝春秋のそれぞれのデータ構造である．

表 5　言語記事データベース（新聞記事）のデータ構造

年	1917（大正 6）
月	2
日	22
朝夕	朝
面	7
段	4
欄名	
連載タイトル	外国固有名詞の訳し方
連載回数	
記事見出し	
執筆者	岩野泡鳴
コード	人名　発音　アクセント　外来語の表記　外国人の日本語生活
キーワード	

表 6　言語記事データベース（文藝春秋）のデータ構造

年	1938（昭和 13）
月	4
開始 P	79
終了 P	82
欄名・特集・連載名	
回数	
執筆者	上宮貴一
題名	営中雑誌終篇
内容	「名を重ず」　軍隊用語について。　「堅パン」は「かんめんぼう」、物干し場は「ぶっかんじょう」などの例を挙げ、一風変わった呼び方にしていることを指摘する。
コード	職業語　類義語
キーワード	かんめんぼう　ぶっかんじょう

7．研究文献

研究文献情報は，国立国語研究所「日本語・日本語教育文献データベース」（以下，文献 DB）の収録文献(2017 年 7 月現在）を利用する．1950 年代以降の主に日本語で書かれた研究文献約 22 万 4 千件があり，このうち，雑誌論文（学会誌，紀要などに掲載された論文）に相当する約 18 万 6 千件を対象とする。それぞれの文献に付与されているキーワードを検索のキーとする．

また，佐藤喜代治編『講座日本語の語彙 別巻 語彙研究文献語別目録』（明治書院，1983）に収められている研究文献情報も利用する予定である．

表 7 は，文献 DB のキーワード欄にある語を UniDic で形態素解析した結果，頻度順の上位語である（補助記号を除く）．

表 7 文献 DB における上位語

rank	頻度	語彙素	品詞
1	34286	の	助詞
2	16533	語	名詞
3	9324	日本	名詞
4	7938	言語	名詞
5	7042	的	接尾辞
6	5721	教育	名詞
7	4947	表現	名詞
8	4929	方言	名詞
9	4802	と	助詞
10	4354	性	接尾辞
11	3658	漢字	名詞
12	3633	国語	名詞
13	3543	学習	名詞
14	3286	研究	名詞
15	3226	学	接尾辞
16	3188	化	接尾辞
17	3124	文	名詞
18	2992	動詞	名詞
19	2822	論	名詞
20	2635	語彙	名詞

また，キーワードの品詞として多いものを表 8 に挙げた．上位には普通名詞以外に固有名詞が意外に多いことが分かる．

表 8　文献 DB における上位品詞

rank	頻度	品詞
1	12338	名詞-普通名詞-一般
2	2856	名詞-普通名詞-サ変可能
3	1596	名詞-固有名詞-地名-一般
4	1416	名詞-固有名詞-人名-名
5	1374	名詞-固有名詞-人名-姓
6	998	動詞-一般
7	978	名詞-固有名詞-人名-一般
8	488	名詞-固有名詞-一般
9	422	名詞-普通名詞-形状詞可能
10	397	接尾辞-名詞的-一般

©2018 Information Processing Society of Japan

なお，キーワードを含む文献は 70141 件（全体の 31.3%）であり，それ以外の文献については，タイトルを検索することが考えられる．

８．ユーザインターフェース

ユーザインターフェースとしては，以下のような手順で検索とその結果を表示する．

1. 検索語 OR カテゴリーを入力
　　↓
2. ヒットしたデータベース名とその件数を表示
　古辞書：a 件
　語地図：b 件
　言語記事：c 件
　研究文献：d 件

それぞれのカテゴリーをクリックして，各データの詳細表示．古辞書であれば，表２，表３，言語地図であれば，表４，言語記事であれば，表５，表６の内容が表示される．また，画像のあるものは，リンクをたどり，画像を表示できる．コーパスからの頻度情報は語彙素 ID で一意に特定されるため，ヒットした場合には即表示する．

９．課題と今後の予定

辞書の見出し語等について形態素解析を行っているが，文脈のない 1 語だけの形態素解析であるため，やや精度が低くなる，また，未知語が多い点が課題である．そもそも解析用の辞書に登録されていない語は，見出しに適宜「読み」「書字形基本形」に相当する情報を与えるとともに，元資料の文字列検も行えるようにし，検索漏れをなくす工夫が必要である．
なお，『古活字版二十巻本倭名類聚鈔』の場合，形態素解析結果，4669 項目中，UniDic の語彙素との対応がとれなかったものは 2590 項目あった。語彙素情報が津かないものは，「日本国語大辞典」などを参考に情報を付ける。

今後の予定としては，平成 30 年度内に語誌データベースを検索するポータルサイトを試験公開し，利用者からのフィードバックを得て改善していきたいと考えている．搭載するデータも順次増やしていく予定である．

謝辞

本研究は，国立国語研究所のプロジェクト「通時コーパスの構築と日本語史研究の新展開」（プロジェクトリーダー・小木曽智信）および日本学術振興会・科学研究費補助金「日本語研究用オントロジーの設計と開発」（17K18505）による成果である．

参考文献

[1]国立国語研究所：日本語歴史コーパス（http://pj.ninjal.ac.jp/corpus_center/chj/）（参照 2018-09-06）
[2]山崎誠：語誌ポータルの構築について，「国際シンポジウム「古辞書研究の射程」予稿集」44-49．（2018）
[3] 国立国語研究所 UniDic https://unidic.ninjal.ac.jp/ (参照 2018-11-01)
[4]国立国語研究所 現代日本語書き言葉均衡コーパス https://pj.ninjal.ac.jp/corpus_center/bccwj/ (参照 2018-11-01)
[5]大西拓一郎：国立国語研究所言語地図データベース（http://www2.ninjal.ac.jp/hogen/dp/ladp/ladb_index.html）（参照 2018-9-06）

©2018 Information Processing Society of Japan