意味分類検索に対応したコーパス簡易検索アプリケーション「ことねり」

小木曽 智信（人間文化研究機構 国立国語研究所）
八木 豊（株式会社 ピコラボ）

概要：「ことねり」は『日本語歴史コーパス』の検索を簡単に行うことのできるツールとして開発したウェブアプリケーションである．ボタンのクリックだけでコーパスを検索ができるうえに，分類語彙表を活用して意味分類による検索も可能にし，専門家以外でもコーパスを利用することが容易にした．本稿はこのアプリケーションの開発について述べる．
キーワード：コーパス，検索ツール，ユーザーインターフェイス，分類語彙表

Cotoneri: a simple corpus search application that supports semantic classification search

Toshinobu Ogiso (National Institute for Japanese Language and Linguistics, NIHU)
Yutaka Yagi (Picolab Co., Ltd.)

Abstract: Cotoneri is a web application developed as an easy-to-use tool for searching the Corpus of Historical Japanese. The application allows users to search the corpus with the click of a few buttons, and also enables searches by semantic classification using a thesaurus, making it easier for non-specialists to use the corpus. This paper describes the development of this application.
Keywords:Corpus, Search Tool, User Interface, Thesaurus

１．まえがき

『日本語歴史コーパス』(CHJ) [1]は，デジタル時代における日本語史研究の基礎資料として，国立国語研究所で開発・公開されているコーパスである．奈良時代以前から明治・大正時代までの多様なテキストを収録している．その全てのテキストに読み・品詞などの単語情報（形態論情報）が付与されていることが特長で，これにより高度な検索や集計が可能になっている．このコーパスには中学・高校で学ばれる主要な古典のほとんどが収録されている．そのため，日本語史研究のみならず，国語（古典）教育や言語文化の教材としても利用されることが期待される．

『日本語歴史コーパス』はオンラインで利用可能なコーパス検索アプリケーション「中納言」[2]を通して一般公開されている．このツールは日本語の研究者を対象ユーザーとして設計されたものであって，高度な検索が可能となっている．しかし，その反面，一般の利用者，特に中学生・高校生を含む学生にとっては必ずしも利用しやすいものではなかった．

そこで，発表者らは「中納言」に代わってボタン操作のみでコーパス検索が可能な簡易検索ツール「ことねり」[3]の開発を行ってきた．そして今回，これに意味分類による検索機能を付与したことで，古語の語形がわからなくとも意味分類を選択していくだけでコーパス検索が利用できるように改善を行った．

２．コーパス簡易検索ツール「ことねり」

「ことねり」は「中納言」に代わって簡単に『日本語歴史コーパス』の検索を行うことのできるツールとして開発したウェブアプリケーションである．ユーザーインターフェイスは，対象ユーザーである中高生や一般の学生，コーパス初心者が戸惑うことなく検索利用できることを重視して開発した．

システムは「中納言」のラッパーとなっており，ユーザーが「中納言」のアカウントでログインした後，その権限によって「中納言」の API を通じて『日本語歴史コーパス』中の用例を取得している．ただし，見出し語別・作品別の用例数は，事前取得したものを「ことねり」のシステム側で保存しており，「中納言」側のサーバーへの負荷なく高速な表示を可能にしている．現在は『日本語歴史コーパス』にのみ対応するが，システム構成としては国立国語研究所の「中納言」で公開されている各種のコーパスに対応することが可能である．

こうした仕組みであるため，「ことねり」の利

用には『日本語歴史コーパス』の利用アカウント（「中納言」アカウント）が必要である．「ことねり」の公開 URL は https://cotoneri.ninjal.ac.jp であるが，利用時に「中納言」のログインページに誘導される．「中納言」での『日本語歴史コーパス』利用アカウントは次の URL から無料で登録できる（https://chunagon.ninjal.ac.jp/useraccount/register）．

なお，「ことねり」の名称は，「小舎人」（ロドリゲス日本大文典 Cotoneri）に由来する．

3．検索インターフェイス

「ことねり」の初期画面は「読み」で単語検索を行うものだが，この画面では銀行の ATM のインターフェイスをモデルとして，（物理的な）キーボードによる文字入力や検索ボタンの押下を不要として，画面上のひらがなのボタンを押下するごとに対象語が絞り込まれて画面下部に表示されていくインタラクティブな方式をとっている（図 1 上段）．このとき，あまりにも対象語が多くなるのを避けるため，画面右側に品詞ボタンを設けており，検索対象を絞り込むトグルスイッチとして機能する形とした．

図 1 は，実際に「あき」で始まる「名詞」「動詞」「形容詞」を表示した画面である．上から 2 段目には該当する見出し語として，名詞「秋」「秋風」「秋霧」などが表示され，その横に，それがコーパス中でどの程度よく使われる語であるのかを★印の数で示している．この印は，コーパス全体での頻度を元に品詞ごとに算出したものである．

また，見出し語の右に，ウェブ上の辞書検索サービス「weblio 古語辞典」（https://kobun.weblio.jp/，GRAS グループ株式会社）と JapanKnowledge（https://japanknowledge.com/lib/search/，ネットアドバンス社）の当該の見出し語の検索ページへリンクを表示している．後者は現状では Japan Knowledge の法人利用契約がある環境でのみ利用可能となっている．

図 1 の 2 段目の見出し語をクリックすると，3 段目に，コーパスの各作品中でその語が何回用いられているかが一覧で表示される．図 1 ではクリックし選択した「秋霧」が『古今和歌集』に 14 例，『大和物語』に 1 例，『源氏物語』に 3 例等，それぞれの作品における用例数が表示される．その右側の少数点のついた数字は，当該見出し語の当該作品中における 100 万語あたりの調整頻度（PMW:per million words）である．その上にある◆印はこの調整頻度をもとにその作品にどの程度よく現れるかを 5 段階で表示している．こうした使用頻度に関わる情報は通常の国語辞典・古語辞典では十分に示されていないものであり，コー

図 1 「ことねり」検索画面（読みで検索）

パス利用ならではの価値を有するものであると言える．

図 1 の 3 段目の作品名をクリックすると，4 段目に当該の見出し語が，その作品で使われている用例を確認することができる．「中納言」の API を通しての検索はこのタイミングで行われるため，やや時間を要し，用例取得中にはビジーカーソルが表示される．図 1 の例ではクリックし選択した『源氏物語』で，選択した「秋霧」が使われている全 3 例を表示している．この用例はコーパスから取得したもので，前後文脈付きの KWIC 形式で表示される．このとき「縦書きに切り替える」ボタンをクリックすることで用例を縦書きで表示することもできる．

コーパスからは非常に詳細な形態論情報（語彙素・語彙素読み・語形・品詞・活用型・活用形・発音形・仮名形など）が取得できるが，「ことねり」で表示する情報は利用目的を踏まえて大幅に絞り込んでおり，活用語の活用形だけを表示している（品詞はすでに単語選択の段階で限定しており，2 段目の見出し語の横に表示されている）．

4 段目にある用例の右端の「JK」ボタンは JapanKnowledge の「新編日本古典文学全集」の用例ページへのリンクとなっている．利用契約を行っていれば，これを押下することにより別ウインドウで用例の原文・頭注・現代語訳等を参照できるようになっている．

4．対象コーパスと見出し語の限定

「ことねり」では，その対象ユーザーと利用目的を考慮して，検索可能なコーパスと見出し語を制限している．

コーパスについては，中学・高校の古文の学習で取り上げられることの多い古典文学作品にしぼるため，対象を奈良時代編の「万葉集」，平安時代編と鎌倉時代編の全作品，江戸時代編の「随筆・紀行」（現在は芭蕉の紀行文のみ）に限定した．コーパス中のこれ以外の資料は、日本語史研究にとっては重要な資料であっても，古典学習という点ではほとんど利用されず，現代語訳や注釈も提供されていないものが多いためである．

見出し語については，品詞のレベルで助詞・固有名詞・連体詞・感動詞を除外した．これらの品詞は，助詞のように用例数が極めて多かったり，固有名詞のように言葉の意味を調査するという目的から外れたり，品詞認定基準が学校文法とやや異なったりするために，本システムでコーパス中の用例を確認することが有効ではないと考えたためである．しかし，これらの中にも学習上重要なものが含まれるため，今後の改良の余地を残している．なお，コーパス中で「形状詞」となっている語については，学校文法に合わせて「形容動詞」として検索可能にした．

さらに，見出し語の数が極端に多くなることを避けるため，対象コーパス全体で用例の頻度が 20 語以上の語か，または後述する意味分類の対応がとれた頻度が 4 以上の語に限った．こうした見出し語の限定は，使いやすさとレスポンスの良さを考慮して行ったものであるが，「あるはずの語が出てこない」という問題が残ることから，今後，改善を検討したい．

5．意味分類による検索

読みによる検索に加えて，今回，新たに意味分類による検索機能を付与した．検索画面上部のタブで「読み」「意味」の検索を切り替えて利用することができる．図 2 は意味分類検索で大分類「動物」から中分類「魚類」を選択した状態で，「鮎」「魚」「鮭」「鯖」などの見出し語が表示されている．見出し語をクリックして選択した後は，読みによる検索（図 1）の場合と同様に，その後が出現する資料，つづいてその用例を表示することができる．

図 2 「ことねり」検索画面（意味で検索）

５．意味分類検索の仕様

この意味分類による検索は，コーパスの見出し語と，国立国語研究所のシソーラス「分類語彙表」を関連付けることによって実現している．

『日本語歴史コーパス』は，UniDic[4][5]によって形態素解析が施されている．今回の「ことねり」が対象とする資料は，「中古和文 UniDic」や「中世文語 UniDic」「近世文語 UniDic」など各時代別の辞書[6]が用いられているが，いずれも同一の辞書データベースから時代ごとに必要な見出し語を出力した辞書であるため，見出し語に通時的な互換性がある．

『分類語彙表』は 1964 年に書籍版[7]が刊行され，その後 2004 年に増補改訂版[8]が出た後に，同データベース版[9]がインターネット上で一般公開されている．これは基本的に現代日本語の語彙を対象としたものである．一方，宮島ほか（2014）『日本古典対照分類語彙表』[10]は古典文学作品の語彙を分類語彙表の体系で分類したもので，CD-ROM のデータ付きで刊行されている．「ことねり」の対象作品の多くは『日本古典対照分類語彙表』の収録対象に含まれている．そこで，今回の意味分類検索機能の実装には，『分類語彙表－増補改訂版データベース』と『日本古典対照分類語彙表』の 2 つのデータを利用した．

分類語彙表のデータは，次のような形式であり，各見出し語について，ID と意味分類が項目名と数字で示されている．

レコード ID 番号,見出し番号,レコード種別,類,部門,中項目,分類項目,分類番号,段落番号,小段落番号,語番号,見出し,見出し本体,読み,逆読み
063576,61244,A,体,自然,動物,魚類,1.5504,01,01,01,魚（うお）,魚,うお,おう
063577,61245,A,体,自然,動物,魚類,1.5504,01,01,02,魚（さかな）,魚,さかな,なかさ
063578,61246,A,体,自然,動物,魚類,1.5504,01,01,03,－魚（ぎょ）,－魚,ぎょ,ょぎ
063579,61247,A,体,自然,動物,魚類,1.5504,01,01,04,魚類,魚類,ぎょるい,いるょぎ
063580,61248,A,体,自然,動物,魚類,1.5504,01,01,05,魚族,魚族,ぎょぞく,くぞょぎ
063581,61249,A,体,自然,動物,魚類,1.5504,01,02,01,稚魚,稚魚,ちぎょ,ょぎち
063582,61250,A,体,自然,動物,魚類,1.5504,01,02,02,幼魚,幼魚,ようぎょ,ょぎうよ
063583,61251,A,体,自然,動物,魚類,1.5504,01,02,03,成魚,成魚,せいぎょ,ょぎいせ
063584,61252,A,体,自然,動物,魚類,1.5504,01,02,04,大魚（たいぎょ）,大魚,たいぎょ,ょぎいた
063585,61253,A,体,自然,動物,魚類,1.5504,01,02,05,小魚（こざかな）,小魚,こざかな,なかざこ
063586,61254,A,体,自然,動物,魚類,1.5504,01,02,06,雑魚（ざこ）,雑魚,ざこ,こざ
063587,61255,A,体,自然,動物,魚類,1.5504,01,03,01,川魚,川魚,かわうお,おうわか

『分類語彙表』の見出し語を『日本語歴史コーパス』などのコーパスの見出し語（語彙素）と結びつけるためには，双方の ID による結合が必要となる．これを行った試みとして，近藤・田中(2020)[11]がある．その成果となるデータは『分類語彙表番号-UniDic 語彙素番号対応表』(wlsp2unidic) [12],『古典対照分類語彙表分類番号-UniDic 語彙素番号対応表』(WLSP2UniDic_historical) [13]として公開されている．[12]のデータは次のような形式となっている（分類語彙表の例と同一箇所を示す）．右端の数字がコーパスの語彙素 ID である．

分類番号,類-部門-中項目-分類項目ラベル,分類番号-段落番号-小段落番号-語番号　語彙素 ID
1.5504, 体 - 自 然 - 動 物 - 魚 類 ,1.5504-01-01-013019
1.5504, 体 - 自 然 - 動 物 - 魚 類 ,1.5504-01-01-0213912
1.5504, 体 - 自 然 - 動 物 - 魚 類 ,1.5504-01-01-039882
1.5504, 体 - 自 然 - 動 物 - 魚 類 ,1.5504-01-01-049950
1.5504, 体 - 自 然 - 動 物 - 魚 類 ,1.5504-01-01-0550443
1.5504, 体 - 自 然 - 動 物 - 魚 類 ,1.5504-01-02-0142453
1.5504, 体 - 自 然 - 動 物 - 魚 類 ,1.5504-01-02-0268832
1.5504, 体 - 自 然 - 動 物 - 魚 類 ,1.5504-01-02-0319757
1.5504, 体 - 自 然 - 動 物 - 魚 類 ,1.5504-01-02-0497979
1.5504, 体 - 自 然 - 動 物 - 魚 類 ,1.5504-01-02-0512658
1.5504, 体 - 自 然 - 動 物 - 魚 類 ,1.5504-01-02-0614820
1.5504, 体 - 自 然 - 動 物 - 魚 類 ,1.5504-01-03-01195546

今回は，これらのデータを元にコーパスに付与された見出し語（語彙素 ID）と分類語彙表番号の対応付けを行った．これにより，分類語彙表の意味分類を「部門」(5504)から「中項目」(01)，「分類項目」(01)とたどることで，自然＞動物＞魚類＞魚 という順に，意味を絞り込んでいくことが可能になる．

意味分類検索のインターフェイスでは，最初に「部門」をボタンのリストを表示し，いずれかを選択するとその配下にある「中項目」のボタンのリストを表示する．「中項目」を選択すると，分類項目は飛ばして直接見出し語を表示する形としている．分類項目をスキップしたのは，検索という目的においては中項目レベルで十分に意味の

絞り込みが可能であり，語数も十分に限定されると判断したためである．

５．意味分類検索機能の制限

このようにして実装した意味分類による検索機能だが，元となったデータの仕様に伴っていくつかの制限がある．

まず，『分類語彙表番号-UniDic 語彙素番号対応表』は，UniDic の見出し語を単位としたものであり，分類語彙表と UniDic とで見出し語のサイズが異なる場合には対応がとれないため，意味分類検索ができない場合がある．例えば，「魚市場」という語は分類語彙表では次のように 1 項目として立てられている．

022768,21626,A,体,主体,社会,事務所・市場・駅など,1.2640,06,01,03,魚市場,魚市場,うおいちば,ばちいおう

しかし，短単位を見出し語単位とする UniDic では，「魚」と「市場」に分割されるため，「魚市場」という語は意味分類検索の見出し語とならない．古典文学作品中ではこのような単位不一致の例は多いわけではないが，漢語では特に問題となることがある．

また，単純に分類語彙表には記載のない語がコーパス中に現れている場合があり，この場合は意味分類による検索ができない．

さらに，多義語については文脈上の意味で絞り込むことはできないという問題がある．例えば，分類語彙表の中で，「魚（さかな）」という語は，次のように 2 回現れる．上の例は食料としての「魚」であり，下の例は生き物としての「魚」である．

051219,49216,A,体,生産物,食料,魚・肉,1.4323,01,02,01,魚（さかな）,魚,さかな,なかさ
063577,61245,A,体,自然,動物,魚類,1.5504,01,01,02,魚（さかな）,魚,さかな,なかさ

現状では，コーパス中で，この 2 つの意味の「魚」を区別することができないため，文脈上でいずれの意味で用いられているかを問わず，見出し語として一致したものが用例として表示されるようになっている．

６．インターフェイスの評価

専門家でなくても容易にコーパスを使うことができる検索システムを目指した「ことねり」だが，現状では本格的な広報を行っていないため，幅広いユーザーからのレスポンスは得られていない．また，定量的な評価ができるだけのアンケート調査は行っていない．それでも，これまでに利用したユーザーからは，下記のような反応があった．

- 直感的に使えて，わかりやすいと思います．
- 意味検索など非常に直感的に操作できて面白かったです．
- 初めて知ったので，面白かったです．特に「意味で検索」というのは新鮮でした．

「面白い」という反応は，専門家向けの「中納言」に対するものとは全く異なるものである．コーパスを楽しみながら利用するという，新たな利用方法が開拓できていると思われ，今後の教育への応用の可能性が考えられる．

７．あとがき

従来もっぱら日本語史研究の専門家に用いられてきた『日本語歴史コーパス』だが，収録された作品や，アノテーションされた単語の情報は，広く一般にも活用されるだけの価値を秘めている．「ことねり」は，提供するインターフェイスの改良によってより多くの人達に利用してもらうための試みの一つである．

しかし，文中に記したとおり，いまだ課題も多い．対象とするコーパスと見出し語の制限については再検討の余地があるし，意味検索についても不十分な点がある．分類語彙表と UniDic 短単位の対応付けの改善や，多義語の文脈上の意味を区別するための語義曖昧性解消については，すでに一部で研究に着手しつつあるが，すぐには解決しがたい今後の大きな課題である．

また，より使いやすいものとしていくためには，ユーザーの利用目的とインターフェイスの使い勝手の評価を把握することがきわめて重要であり，今後調査を進めていきたい．

謝辞

本研究は国立国語研究所の共同研究プロジェクト「開かれた共同構築環境による通時コーパスの構築」および「多様な語彙資源を統合した研究活用基盤の共創」の研究成果の一部です．

参考文献等

[1] 国立国語研究所（2022）『日本語歴史コーパス』（バージョン 2022.3）
https://ccd.ninjal.ac.jp/chj/
[2] 国立国語研究所(2022) コーパス検索アプリケーション「中納言」
https://chunagon.ninjal.ac.jp/

[3] 国立国語研究所(2022) 日本語歴史コーパス簡易検索アプリケーション「ことねり」
https://cotoneri.ninjal.ac.jp/
[4] 伝 康晴，小木曽 智信，小椋 秀樹，山田 篤，峯松 信明，内元 清貴，小磯 花絵 (2007) 「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」，日本語科学，Vol.22，pp.101-123.
http://doi.org/10.15084/00002185
[5] 国立国語研究所(2022) UniDic：国語研短単位自動形態素解析用辞書
https://clrd.ninjal.ac.jp/unidic/
[6] 小木曽 智信，小町 守，松本 裕治 (2013) 「歴史的日本語資料を対象とした形態素解析」，自然言語処理，Vol.20，No.5，pp.727-748.
https://doi.org/10.5715/jnlp.20.727
[7] 国立国語研究所(1968)『分類語彙表』秀英出版
http://doi.org/10.15084/00002267
[8] 国立国語研究所(2004)『分類語彙表増補改訂版』大日本図書
[9] 国立国語研究所(2018)『分類語彙表 増補改訂版データベース』(ver.1.0.1)
https://clrd.ninjal.ac.jp/goihyo.html
[10] 宮島達夫ほか（2014）『日本古典対照分類語彙表』笠間書院
[11] 近藤明日子，田中牧郎 (2020) 「「分類語彙表番号－UniDic 語彙素番号対応表」の構築」，国立国語研究所論集 18: pp.77-91.
https://doi.org/10.15084/00002542
[12] 国立国語研究所(2020)『分類語彙表番号-UniDic 語彙素番号対応表』(wlsp2unidic ver.1.0.2) https://github.com/masayu-a/wlsp2unidic
[13] 国立国語研究所(2020)『古典対照分類語彙表分類番号-UniDic 語彙素番号対応表』(WLSP2UniDic_historical ver.0.8.0)
https://github.com/masayu-a/WLSP2UniDic_historical