「日本語歴史コーパス 平安時代編」先行公開版について

小木曽智信（国立国語研究所言語資源研究系）†
須永哲矢（国立国語研究所コーパス開発センター）
冨士池優美（国立国語研究所コーパス開発センター）
中村壮範（マンパワージャパン株式会社）
田中牧郎（国立国語研究所言語資源研究系）
近藤泰弘（青山学院大学文学部／国立国語研究所言語資源研究系）

On the Public Beta Release of *the Heian Period Series of the Corpus of Historical Japanese*

Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Tetsuya Sunaga (National Institute for Japanese Language and Linguistics)
Yumi Fujiike (National Institute for Japanese Language and Linguistics)
Takenori Nakamura (Manpower Japan Co., Ltd.)
Makiro Tanaka (National Institute for Japanese Language and Linguistics)
Yasuhiro Kondo (Aoyama Gakuin University / National Institute for Japanese Language and Linguistics)

１．はじめに

国立国語研究所では「通時コーパスの設計」プロジェクトが中心となって「日本語歴史コーパス[1]」(Corpus of Historical Japanese, CHJ) の開発準備を進めてきた（近藤 2012）。今回、このうち平安時代の仮名文学作品からなる「平安時代編」のデータ整備が進んだことから、これを先行公開版として一般公開を行うこととした。「現代日本語書き言葉均衡コーパス」（BCCWJ）の公開にも用いられているウェブインターフェイス「中納言」（小木曽ほか 2011）を利用しての公開となる。

本発表では「CHJ 平安時代編」の先行公開版の内容について説明し、「CHJ 中納言」のデモンストレーションを行う。

２．「日本語歴史コーパス 平安時代編」先行公開版の概要

「CHJ 平安時代編」は、日本の代表的な古典文学作品である、平安時代の仮名文学作品をコーパス化したものである。先行公開版では、次の 10 作品のデータが利用可能である。本文はすべて、許諾を得て小学館「新編日本古典文学全集」（新編全集）を利用している[2]。

古今和歌集、土佐日記、竹取物語、伊勢物語、落窪物語、大和物語、枕草子、源氏物語、紫式部日記、和泉式部日記

収録した本文データには「中古和文 UniDic」（小木曽ほか 2012，Ogiso et al. 2012）と MeCab を用いて形態素解析を施し、その解析結果に対して人手による修正を行った。これにより、出現するすべての語に読み・品詞・活用型・活用形・語種等の形態論情報（短単位）が付与されている。

さらに、新編全集の情報を利用して本文に「本文種別」と呼ぶ情報を付与し、当該箇所が地の文なのか会話文なのか、あるいは和歌や手紙なのかといった区別がなされている。『源氏物語』では話者も表示される。

† togiso@ninjal.ac.jp

1 これまで暫定的に「通時コーパス」と呼称されていたものの正式名称。

2 コーパス化の対象は原文のみで、現代語訳等は含まない。

テキスト量

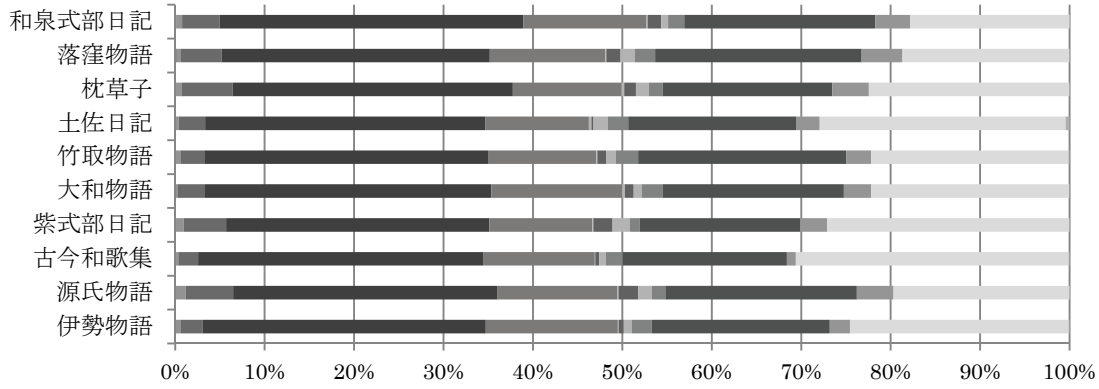
「CHJ 平安時代編」のテキストの量は、表 1 に示す通りである。全体で約 79 万語、うち 65%に近い 51 万語を源氏物語が占めている。

表 1　作品別の語数（短単位，記号を含む）

作品名	語数
伊勢物語	15894
古今和歌集	32286
和泉式部日記	12630
土佐日記	8129
大和物語	26740
枕草子	79851
源氏物語	510572
竹取物語	12583
紫式部日記	20710
落窪物語	68561
総計	**787956**

品詞・語種構成

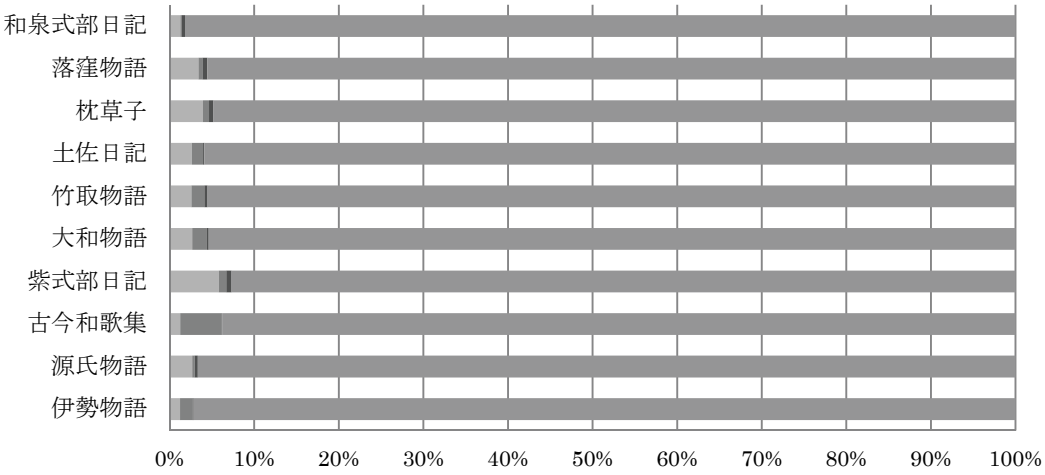
コーパスに付与された短単位の形態論情報を元に、作品ごとに品詞別の語数を集計したものが図 1 である。品詞の認定基準は『中古和文 UniDic 短単位規程集』（小椋・須永 2012）によっている。なお、以下では語数に記号を含まない。



	伊勢物語	源氏物語	古今和歌集	紫式部日記	大和物語	竹取物語	土佐日記	枕草子	落窪物語	和泉式部日記
■感動	7	334	26	15	8	7	0	51	91	16
■形状	72	4788	83	144	62	51	26	437	255	68
■形容	338	23150	688	815	692	280	195	3687	2454	452
■助詞	4325	128841	9867	5029	7327	3210	2068	20363	16175	3651
■助動	2017	58447	3875	1969	3344	1229	767	7903	6961	1476
■接続	13	449	12	19	72	12	18	184	70	21
■接頭	77	9682	142	368	224	97	16	848	862	162
■接尾	120	6675	231	332	204	112	105	950	871	84
■代名	309	6578	579	184	531	254	151	999	1209	194
■動詞	2718	93308	5687	3072	4641	2357	1239	12325	12465	2300
■副詞	308	17800	312	510	699	286	176	2657	2457	419
■名詞	3341	85827	9456	4633	5062	2238	1815	14571	10073	1916
■連体	12	14	26	3	12	11	30	9	9	0

図 1　作品別品詞構成

同様に語種別の集計を行ったものが図 2 である。どの作品でも大部分が和語であり、全体では 96%を占める。古今集は歌人の名前・官職などを含むため、固有名や漢語の割合が高くなっている。ごくわずかに現れる外来語はサンスクリット語由来の仏教語である。



	伊勢物語	源氏物語	古今和歌集	紫式部日記	大和物語	竹取物語	土佐日記	枕草子	落窪物語	和泉式部日記
■外	0	105	2	5	3	3	0	17	5	0
■漢	169	11539	388	990	607	259	174	2546	1837	136
■固	201	1205	1516	151	398	155	88	467	258	16
■混	19	1686	20	105	51	39	9	347	314	44
■和	13268	421351	29058	15841	21819	9688	6335	61607	51537	10563

図 2　作品別語種構成

形態論情報の特徴

CHJ に付与される形態論情報は、電子化辞書 UniDic（伝ほか 2007）の設計にもとづくものである。UniDic は、短単位と呼ばれる厳密な規定によって単語の区切り方が定められており、揺れが少ない斉一な単位による解析が可能になっている（小椋ほか 2011）。また、図 3 に示すように語彙素・語形・書字形・発音形という見出し語の階層構造を持っており、利用者が必要に応じて、見出し語のレベルを選択して利用することができる。図 4（次頁）は具体的な見出し語（例：何処［イズコ］）の例である。

図 3　見出し語の階層構造

「語彙素」は、異語形や異表記をまとめ上げた辞書見出し（lemma）に相当するもので、「語形」はそのうち異語形を区別したもの、「書字形」は異表記を区別したものである。「発音形」は発音を示すものであるが、CHJ においては現代における読み方を参考までに示したものに過ぎない。利用者は、表記そのものに関心があるのであれば書字形を、語形の差異に関心があるのであれば語形を、辞書見出しのレベルでまとめ上げたいのであれば語彙素を利用すればよい。

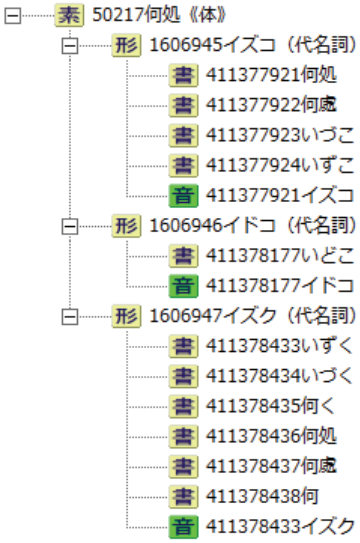


図 4　見出し語の階層構造の例（何処［イズコ］）

CHJ で利用している中古和文 UniDic の短単位は、原則として現代語と同様の基準によっており、相互に比較することができるように配慮したものである。ただし、語の歴史的変化や中古語の実態を踏まえ、時代別に異なった扱いをしている語も少なくない。たとえば、現代語では連体詞とされる「この」「その」が、中古語では代名詞「こ」「そ」と格助詞「の」に分けて数えられている。「CHJ 中納言」を用いて中古語の検索をする場合には、この短単位の規定について理解をしておく必要がある。

先行公開版データの制限

CHJ では、BCCWJ と同様、短単位だけでなく長単位の情報も付与する計画である。しかし、先行公開版で公開するデータは短単位のみである。平安時代編の完成版で、長単位のデータも公開する予定である。

また、CHJ 平安時代編が基づいている『中古和文 UniDic 短単位規程集』には完全でない部分が残されている。たとえば、複合動詞を一語と認めるか分割するかという認定基準はその例である。そのため、先行公開版では複合動詞の認定に揺れがあるなどの問題が残っている。これも完成版では統一的な基準の下に修正される予定である。

３．「日本語歴史コーパス」中納言

CHJ の公開は現在のところ、ウェブ版のコンコーダンサー「中納言」（図 5）のみで行っている。「CHJ 中納言」は、CHJ むけに若干の修正を行っているが、基本的に BCCWJ で利用されている「中納言」と同じものである。書面による申込み手続きを経ることで無償で利用できる（手続きは「日本語歴史コーパス」ホームページを参照）。

データには形態論情報が付与されているため、表層の文字列だけでなく、形態論情報を利用することで高度な検索条件の指定を行うことができる。たとえば、語彙素「読む」（終止形）を指定することで「読ま」「読み」「読む」「読め」といった各活用形を一括で検索することが可能である。また、先述の UniDic の見出し語の階層構造により、見出し語を語彙素で指定すれば、その異表記を一括検索することができる。したがって、漢字表記と仮名表記の違い、異体字や送り仮名の揺れなどを一々意識することなく検索できる。

また、たとえば品詞情報を使って「形容詞すべて」のように大きな語群を検索対象とすることもできる。形態論情報を組み合わせて、たとえば「漢語名詞」「形容詞の連体形」などの詳細な条件で検索を行うことも可能である。

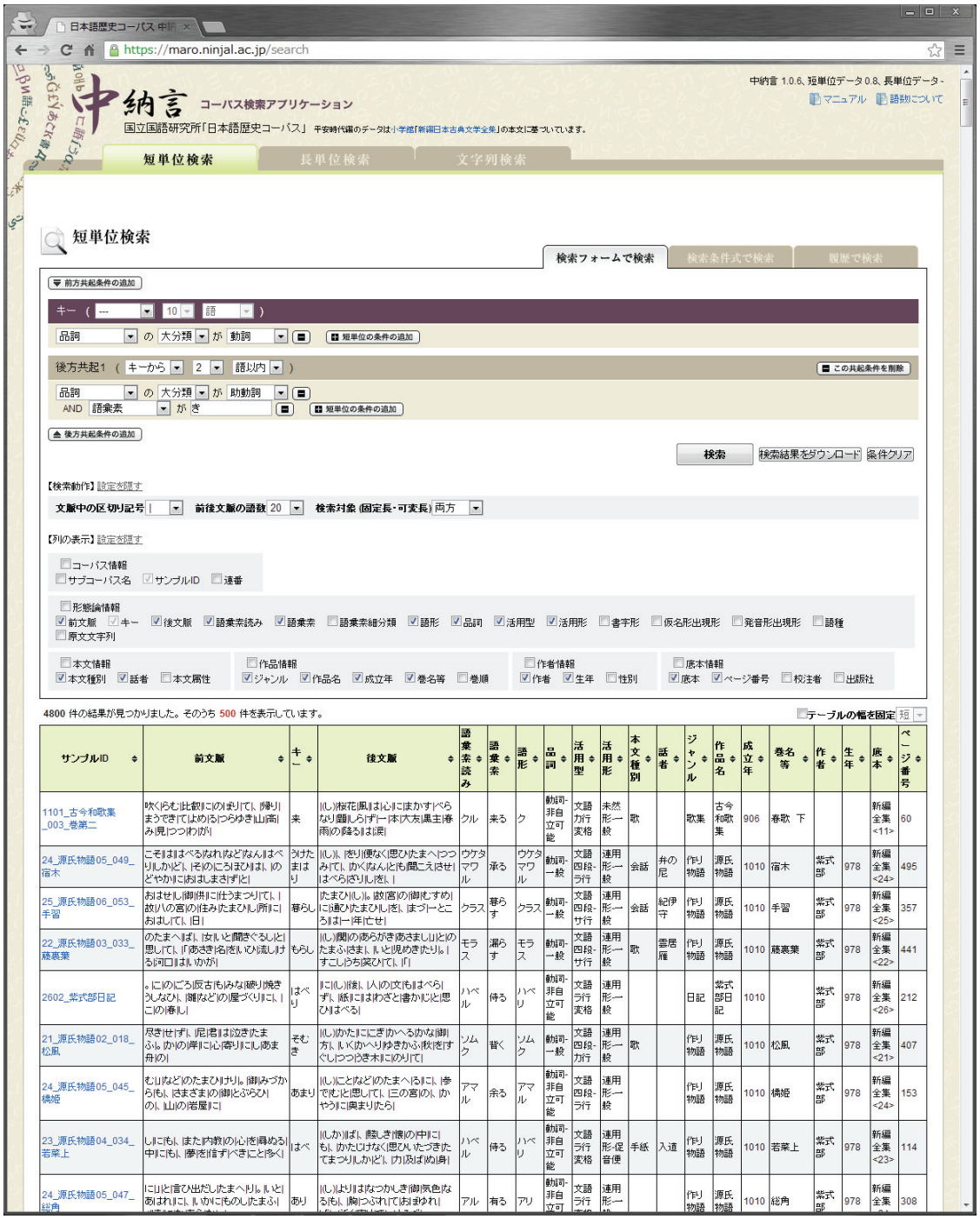


図 5　日本語歴史コーパス「中納言」検索実行画面

さらに、複数の語（最大 10 語）を組み合わせた検索も行うことができる。これにより、「特定の形容詞の連体形の後に来る名詞」であるとか、「特定の動詞に続く助動詞」、「特定の動詞の前方 5 語以内に来る“名詞＋を”」といったような、従来の索引では不可能であった検索が可能になっている。

形態論情報を使った検索以外に、「文字列検索」で表層の文字列による検索を行うこともできる。この場合にも、検索結果は形態論情報付きで表示されるため、調査したい語にどのような形態論情報が付与されているか分からない場合には、いったん文字列検索を行うことで形態論情報を確認することができる。

検索結果の項目

検索結果には、表 2 に示す項目が表示可能である。デフォルト表示が「非表示」のものは画面上のチュエックボックスをオンにすることで表示されるようになる。

「コーパス情報」は、検索結果のコーパス中の位置を示す情報である。サンプル ID と連番とで短単位の位置を一意に指定することができる。

「形態論情報」は、当該箇所の KWIC と、キーに付与されている形態論情報からなる。「キー（書字形出現形）」が実際に出現した表層形（活用変化後の形）であるのに対し「書字形」は終止形の形である。形態論情報中の「～出現形」はすべて活用変化後の形であることを示す。

表 2 検索結果表示項目

分類		順番	列名	デフォルト表示
コーパス情報		1	サブコーパス名	非表示
		2	サンプル ID	表示
		3	連番	非表示
形態論情報	KWIC	4	前文脈	表示
		5	キー（書字形出現形）	表示
		6	後文脈	表示
		7	語彙素読み	表示
		8	語彙素	表示
		9	語彙素細分類	非表示
		10	語形	表示
		11	品詞	表示
		12	活用型	表示
		13	活用形	表示
		14	書字形	非表示
		15	仮名形出現形	非表示
		16	発音形出現形	非表示
		17	語種	非表示
		18	原文文字列	非表示
本文情報		19	本文種別	表示
		20	話者	表示
		21	本文属性	非表示
作品情報		22	ジャンル	表示
		23	作品名	表示
		24	成立年	表示
		25	巻名等	表示
		26	巻順	非表示
作者情報		27	作者	表示
		28	生年	表示
		29	性別	非表示
底本情報		30	底本	表示
		31	ページ番号	表示
		32	校注者	非表示
		33	出版社	非表示

「本文情報」の「本文種別」は「会話」「手紙」「歌」「詞書」等の別である。「話者」は会話の話者表示だが、新編全集で明示されているものだけが出力され、作品によっては情報がない。「本文属性」は和歌である場合に歌番号が出力されている。

「作品情報」は当該の作品の基本的な書誌情報である。「ジャンル」には平安時代編では作り物語・日記・随筆・歌集がある。「成立年」は正確な年が不明のものは有力な説に従い、おおよその年代を記入している。「巻名等」は研究に必要と考えられる範囲で新編全集にもとづいて巻名や章段のタイトル、部立てなどを記入している。

「作者情報」は当該の作品の作者の情報である。詳細が不明のものは分かる範囲で記入している。『古今和歌集』については仮名序以外には作者情報を出力していない。

「底本情報」は CHJ 平安時代編が依拠した新編全集の情報である。「底本」は当該作品が収録された新編全集の巻数、「ページ番号」は当該箇所が現れるページ数を示す。これにより、ヒットした用例について書籍の新編全集を開いて当該箇所を確認することができる。CHJ には現代語訳や注は含まれていないため、こうした情報を確認するためには新編全集本体を参照する必要がある。

これらの情報を含む検索結果は表形式でダウンロードすることができるため、これを表計算ソフトに読み込むことで、自由に集計を行うことができる。特にピボットテーブルと呼ばれる機能を用いることで、クロス集計を自在に行うことが可能である。ダウンロードファイルには、常に表 2 の全ての項目が含まれており、さらに最終列（34 列目）に「反転前文脈」が出力される。この列は前文脈を使ってソートを行うためのもので、前文脈の文字列の並びを逆転させキーに近い文字から順に並べたものである。

先行公開版インターフェイスの制限

現在の「CHJ 中納言」では、検索対象を指定することができず、常にコーパス全体を検索することになる。したがって作品別・ジャンル別に用例数などを確認するためには、いったん検索結果をダウンロードして集計を行う必要がある。

この問題は、2013 年 3 月に予定している「中納言」のアップデートで改善される予定である。このアップデートにより、作品別・ジャンル別などの検索対象指定が可能になるほか、検索条件指定の方法などさまざまな機能が改善される予定である。

４．今後の計画とまとめ

CHJ 平安時代編の完成版は、2013 年度中の公開を予定している。完成版では、上述した制限をなくし、全ての作品について長単位を付与するほか、『更級日記』『讃岐典侍日記』等の作品を追加する予定である。

先行公開版には制限があるものの、本コーパスにより、これまでの古典語の研究手法では不可能だった検索や集計が可能になった。本コーパスが広く利用され、新しい研究成果につながることを期待したい。また、これを機にコーパス日本語学の裾野が歴史的研究の分野にまで広がり、研究がより盛んになることに期待したい。

文　献

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』22　pp.101-123

小木曽智信・中村壮範・鈴木泰山・八木豊・山崎誠・前川喜久雄（2011）「コーパス検索システム「中納言」デモンストレーション」『日本語コーパス完成記念講演会予稿集』pp.43-46

小木曽智信ほか（2012）『和文系資料を対象とした形態素解析辞書の開発』科研費 基盤研究（C）「和文系資料を対象とした形態素解析辞書の開発」（課題番号 21520492）研究成果

報告書（中古和文 UniDic ホームページからダウンロード可能）
小椋秀樹・須永哲矢（2012）『中古和文 UniDic 短単位規程集』科研費 基盤研究（C）「和文系資料を対象とした形態素解析辞書の開発」(課題番号 21520492) 研究成果報告書 2 （中古和文 UniDic ホームページからダウンロード可能）
Toshinobu Ogiso, Mamoru Komachi, Yasuharu Den and Yuji Matsumoto. (2012) UniDic for Early Middle Japanese: a Dictionary for Morphological Analysis of Classical Japanese. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), pp.911-915. Istanbul, May 2012.（http://www.lrec-conf.org/proceedings/lrec2012/pdf/906_Paper.pdf からダウンロード可能）
近藤泰弘（2012）「日本語通時コーパスの設計」『NINJAL「通時コーパス」プロジェクト・Oxford VSARPS プロジェクト合同シンポジウム 通時コーパスと日本語史研究予稿集』pp.1-10

関連 URL

日本語歴史コーパス「中納言」 http://maro.ninjal.ac.jp/
日本語歴史コーパスホームページ（国立国語研究所コーパス開発センター）
http://www.ninjal.ac.jp/corpus_center/chj
NINJAL 通時コーパスプロジェクト ホームページ http://www.historicalcorpus.jp/
中古和文 UniDic http://www2.ninjal.ac.jp/lrc/index.php?UniDic
MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://mecab.googlecode.com