NAIST-IS-DD1161018

博士論文

日本語通時コーパスのための
形態論情報アノテーションの研究

小木曽 智信

2014年2月6日

奈良先端科学技術大学院大学
情報科学研究科 情報処理学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
博士 (工学) 授与の要件として提出した博士論文である。

小木曽 智信

審査委員：

松本　裕治　教授　（主指導教員）
中村　哲　　教授　（副指導教員）
新保　仁　准教授　（副指導教員）
小町　守　准教授　（副指導教員，首都大学東京）

日本語通時コーパスのための形態論情報アノテーションの研究*

小木曽 智信

内容梗概

近年，コーパスを用いた日本語研究が盛んになり，国立国語研究所においては，日本語史研究のための通時コーパスを構築する準備が進められている．通時コーパスには，現代語のコーパスと同様の形態論情報を付与することが必要とされているが，従来は歴史的な日本語資料に十分な精度で形態素解析を施すことができず，形態論情報のアノテーションは困難であった．

このような中，本研究は，日本語通時コーパスのための形態論情報アノテーションを実現するために自然言語処理技術を応用して，次の貢献を行った．

1. 古文の形態素解析を実現するための言語資源として，新たに古典語の辞書と学習用のコーパスを整備し，統計的機械学習にもとづく形態素解析技術を用いて，中古和文と近代文語文について実用的な精度（見出し語認定のF値で0.96以上）が得られる形態素解析システムを実現した．

2. 上記の言語資源と通時コーパス自体の整備のために，辞書の見出し語とコーパスの出現形とを関連付け一貫性を保ちながら形態論情報の修正作業を行うことのできるデータベースシステム（国立国語研究所「形態論情報データベース」）を構築し，通時コーパス整備の基盤を整えた．

3. 通時コーパスに収録される多様なテキストに対して高い精度で形態論情報のアノテーションを行う方法を検討し，近世口語文・和漢混淆文・旧仮名遣いの口語文について，実際に形態論情報のアノテーションを行った．

4. 上記の形態素解析技術や形態論情報付きの通時コーパスを日本語研究者や人文科学系の研究者に使いやすい形で提供するために，新たなツールの作成・既存のツールの適用を行った．

*奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文，NAIST-IS-DD1161018, 2014年2月6日．

以上により，通時コーパス構築の基盤を整備し，通時コーパスを用いた日本語史研究のための環境を提供した.

キーワード

古文, 形態素解析, 言語資源, 日本語歴史コーパス

A Study on Morphological Annotation for the Japanese Diachronic Corpus*

Toshinobu Ogiso

Abstract

Recently, corpus-based study of Japanese language has become popular, and a diachronic corpus of Japanese is being developed at the National Institute for Japanese Language and Linguistics (NINJAL) to study history of Japanese language.

In order to construct a richly annotated diachronic corpus of Japanese, morphological analysis of historical Japanese text is required. However, morphological analysis of old Japanese texts with adequate accuracy was impossible by conventional means, and annotation of diachronic corpora with morphological information was difficult using existing technology.

Given this situation, this study applied natural language processing technology to carry out the annotation of morphological information for the diachronic corpus of Japanese, and made contributions as below.

1. Dictionaries and corpora of historical Japanese text were newly created as language resources to carry out the morphological analysis of historical Japanese. Using these resources and a morphological analyzer based on statistical machine learning, morphological analyses of historical Japanese texts in the literary style of the *Meiji* era and morphological analysis of literature of the *Heian* era were achieved with high accuracy (over 96% at lemmatization level).

*Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1161018, February 6, 2014.

2. For compilation and maintenance of the language resources mentioned above and the diachronic corpus itself, a database system (NINJAL morphological information database) was developed. The database system makes it possible to modify annotations in the diachronic corpus and in related dictionary entries, while maintaining consistency between the two.

3. For a variety of texts included in a diachronic corpus, methods for performing annotation of morphological information with high accuracy were studied, and actual annotation was conducted.

4. For researchers of Japanese language and scholars of humanities, some newly created tools and existing software were applied to the diachronic corpus in order to make it possible to use the morphological analysis system and the annotated diachronic corpora easily.

In this way, a basis for the compilation of diachronic corpora was constructed and an environment for studying the history of the Japanese language using diachronic corpora was provided.

Keywords:

classical Japanese, morphological analysis, language resource, corpus of historical Japanese

謝 辞

本研究をまとめるにあたり，多くの方々にお力添えを頂きました．

奈良先端科学技術大学院大学 情報科学研究科の先生方，なかでも松本裕治先生，小町守先生には，論文の細やかな点までご指導いただき，おかげにより本論文を完成させることができました．

また，千葉大学の伝康晴先生には，UniDic の設計や実装について，様々にご教授いただきました．伝先生より特定領域研究「日本語コーパス」の電子化辞書班として現代語用の UniDic を開発する中でご教示いただいたことが，本研究の出発点となりました．

国立国語研究所「通時コーパスの設計」プロジェクトのリーダーを務められた青山学院大学の近藤泰弘先生には，コーパスの利用等にあたりたいへんお世話になりました．このプロジェクトの一環としてコーパスの整備を行うことができたことで，本研究を日本語史研究にとって役立てることができました．

通時コーパス以外にも，この研究に関連する国語研究所の研究プロジェクト，科研費プロジェクトの共同研究者の皆様にご協力いただきました．特に，国語研究所の田中牧郎氏には「近代語コーパス設計のための文献言語研究」プロジェクト，科研費プロジェクトでお世話になっただけでなく，研究の多くの面で支えていただきました．私が中心となった国語研究所の「統計と機械学習による日本語史研究」プロジェクトでは，小町先生，本学大学院の岡照晃氏に共同研究に加わっていただくことで，通時コーパスへの自然言語処理技術の応用を進めることができました．

本研究の中心をなす，歴史的資料を対象とした UniDic のうち，「近代文語 UniDic」は，立命館大学の小椋秀樹氏・国語研究所の近藤明日子氏とともに取り組んだ成果を基礎とするものです．また，同じく「中古和文 UniDic」は科研費「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）の下で伝康晴氏・小椋秀樹氏・近藤明日子氏・田中牧郎氏らとともに開発され，その後，国立国語研究所共同研究プロジェクト「通時コーパスの設計」及び「統計と機械学習による日本語史研究」の下で整備を続けてきたものです．これらの辞書の開発は，科研費の共同研究者，須永哲矢氏・冨士池優美氏ら国立国語研究所のプロジェクト研

究員，技術補佐員をはじめとする多くの方々の協力によって実現したものです．また，本研究の一部である形態論情報データベースの開発・実装については，国語研究所の中村壮範氏に多大のお力添えを頂きました．

このほか，新たな科研費プロジェクト「近世口語文を対象とした形態素解析辞書の開発」においては，共同研究者である千葉大学の岡部嘉幸先生，埼玉大学の村上謙先生，国立国語研究所プロジェクト研究員の市村太郎氏，鴻野知暁氏をはじめとする方々にご協力いただきました．

私の本務である国立国語研究所では，前川喜久雄言語資源研究系長・コーパス開発センター長をはじめ皆様に，私の博士論文執筆を慮ってさまざまな形でご配慮をいただきました．

この場を借りて厚く御礼申し上げます．

最後に，傍らでいつも温かく見守ってくれた家族に感謝します．

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究の背景

日本語の歴史を研究するうえで，論証のための用例を検索することは極めて重要である．現代語の研究とは異なり，過去の言語の研究においては内省にもとづく意味や文法性の判断ができないため，残された資料における用例がほとんど唯一の立論の材料となる．したがって，用例検索の手法は日本語史研究で常に課題となってきた．

古くは調査のたびに全文を確認する必要があり，研究者にとって極めて負担の重い作業であったが，昭和になると，主要な文学作品の総索引（concordance）が整備され，日本語史研究の基礎的なツールとなった．その後，1990 年代にコンピュータの利用が一般化すると，古典作品の電子化テキストが作成され，テキストデータを文字列検索によって研究に利用する形態が広まった．

2000 年代以降になると，言語研究を目的とした大規模なコーパスの構築が行われる．歴史的な資料としては，国立国語研究所が 2005 年に公開した『太陽コーパス』[29] がある．これは明治・大正期の雑誌記事をコーパス化したもので，文語文を中心に約 1,450 万字の記事を含む大規模なものである．このコーパスには文書構造やテキストの校訂情報などがタグ付けされているものの，形態論情報は付与されていない．これは，作成当時，文語文の形態素解析を行うすべがなかったためである．そのためコーパスの利用方法も文字列検索を中心としたものに限られた．

一方，現代語のコーパスを使った研究は，2011 年に公開された『現代日本語書き言葉均衡コーパス』（BCCWJ）[34] によって大きく進展する．BCCWJ は様々なレジスターの現代日本語の書き言葉 1 億語以上を収録した大規模な日本語コーパスである．このコーパスの構築のために，言語研究に適した形態素解析辞書として新たに UniDic が開発され，収録する全てのテキストに対して短単位と長単位という二つの言語単位による形態論情報が付与された．これにより，日本語の研究にコーパス言語学の手法が本格的に導入され，複雑な共起条件を指定した用例検索や，コロケーション強度の取得，テキストごとの特徴語抽出，多変量解析

を用いた研究など，索引による語例検索やテキストの文字列検索を超えた新しい手法による日本語研究が可能になった．

BCCWJ の完成後，日本語コーパス構築の課題として，歴史的な研究を行うためのコーパスの不足が浮かび上がってきた [12]．そこで，国立国語研究所では日本語の通時コーパス[1] の構築の準備を行うこととなり，基幹型プロジェクト「通時コーパスの設計」[2] の下で江戸時代以前の日本語資料のコーパスの構築準備と試験的なコーパス作成が開始された．また，「近代語コーパス設計のための文献言語研究」[3] では『太陽コーパス』に続く近代語コーパス整備のための研究が行われた．

さきの『太陽コーパス』では文字列レベルでの利用に留まっていたが，新たな通時コーパスでは，コーパス言語学的な手法を用いた研究を可能にするために，すべてのテキストについて BCCWJ と同等の形態論情報を付けることが求められている．歴史的な日本語資料は当然に有限であって，現代語と比較すれば規模も小さいが，人手によって整備できる量は優に超えている．また，均質なタグ付けを行うためには自動処理が必須である．したがって，通時コーパスの構築のためには自動形態素解析を行わなければならない．

しかし，古文を対象とした形態素解析は長い間実現せず，コンピュータによる古文の処理を行おうとする人々から待ち望まれている状態にあった．たとえば村上（2004）[57] は，計量文献学の立場から，古典の研究資料としての価値を論じた上で次のように述べている．

> 「古典に関して計量分析で著者に関する疑問を解明できたなら，古典研究に大きな刺激を与えるにちがいない．ただ，残念なことに文章を自動的に単語に分割し，品詞情報等を付加する形態素分析のプログラムの開発が古文の場合，遅れている」(p.191)

また，近藤（2009）[66] は古典語研究の立場から次のように述べる．

> 「古典語は形態素解析の自動化がしにくいため，単語レベルの索引を作るには，すべて手作業で形態素解析を行う必要があるため，多くの資料を対象に語彙研究することは困難である」

[1]本研究では「通時コーパス」の語を特定のコーパスを指すものではなく，一般的な用語として使用する．国語研究所のプロジェクトによる通時コーパスを特に指す場合には「国語研通時コーパス」の呼称を用いる．なお，「通時コーパスプロジェクト」による成果物としてのコーパスは「日本語歴史コーパス」の名称で公開されている．

[2]国立国語研究所共同研究プロジェクト「通時コーパスの設計」（2009 年 10 月〜，プロジェクトリーダー：近藤泰弘）

[3]国立国語研究所共同研究プロジェクト「近代語コーパス設計のための文献言語研究」（2009 年 10 月〜2012 年 9 月，プロジェクトリーダー：田中牧郎）

このように，古文など歴史的な資料を対象とした形態素解析は，新しい通時コーパスを構築するために必要とされているにもかかわらず，コーパス構築や日本語研究での利用に実用的な精度での解析を行うことができず，その実現が強く望まれている状況にあった．

1.2 研究の目的

本研究の目的は，日本語の歴史を通時的に研究することのできる通時コーパスに対して形態論情報のアノテーションを行うことを可能にすることである．

そのために，通時コーパス構築の基盤として活用することのできるような，歴史的資料の形態素解析を実現しなければならない．既存の日本語形態素解析システムはもっぱら現代語のテキストの解析を目的としてきたため，これらのテキストに対応できない．しかし，今日の統計的機械学習にもとづく形態素解析技術を用いれば，歴史的な資料のための言語資源，すなわち専用の辞書と学習用のコーパスを整備することで，既存の実装を用いて形態素解析を実現にすることができる．

また，通時コーパスに収録されるテキストは多種多様であるため，その全体に対して形態論情報のアノテーションを行うには，各時代・文体に対応した形態素解析辞書が必要となる．多様なテキストを高い精度で解析するための方法を検討し，解析を実現していく．

形態素解析のための辞書とコーパスのデータを整備するためには，データベースシステムを構築し，辞書の見出し語とコーパスの出現形とを関連付けながら修正作業を行う環境を整備する必要がある．通時コーパスの構築では，残された貴重なテキストを最大限に活かすために，形態論情報を高い精度で整えなければならない．コーパス全体の形態論情報を人手で修正して，日本語史研究に利用できるだけの精度にしていくための環境を整備する必要がある．

さらに，こうして作られた古文の形態素解析システムや形態論情報が付与されたコーパスは，人文科学系の研究者に使いやすい形で提供される必要がある．形態素解析を行いやすくするツールや，古文のコーパスの検索ツール，解析結果を利用して総索引を作成するツールなどが挙げられる．こうしたユーザー向けのツールの作成も研究課題の一部である．

以上のように，自然言語処理技術を応用して，(1) 通時コーパス構築の基盤となる古文の形態素解析を実現し，(2) 解析結果を整えていく環境を整備することにより通時コーパスに形態論情報を付与し，(3) 古文の形態素解析とコーパスを活用していく環境を用意する．このことを通して，日本語史研究の今後の発展に貢献することが最終的な目的である．

1.3　本論文の構成

本論文は，次の各章によって構成される．

第 2 章では，通時コーパスと形態素解析に関連するこれまでの研究を確認したのち，形態論情報アノテーションの対象となるテキスト群を確認し，その多様性への対処方法について検討する．

第 3 章では，通時コーパスへの形態論情報付与に必要とされるデータベースシステムの設計と構築について述べる．このシステムは，はじめ BCCWJ と現代語用の形態素解析辞書の構築のために開発したものであるが，歴史コーパス用の拡張を行い，見出し語の通時的な管理を可能にし，通時コーパス構築用のシステムとして運用した．

第 4 章では，通時コーパスの中の主要なテキスト群である「中古和文」と「近代文語文」に焦点を当て，その形態素解析を実現する．既存の形態素解析器を活用し，辞書の見出し語を拡張し，学習用のコーパスを整備することで通時コーパス構築に実用的な精度での解析が可能になった．これらの辞書の構築過程で得られたデータをもとに，「中古和文」「近代文語文」以外の多様な文体のテキストの形態素解析の方法について検討する．

第 5 章では，近世の洒落本を初めとする口語文，中世の口語を反映する狂言，平安時代の和漢混淆文，通時コーパスに含まれる多様なテキストの形態素解析の方法について論ずる．4 章の結果をふまえ，少量の学習用コーパスを作成して専用のモデルを作成することで，これらのテキストでも（人手による修正を前提とした）実用的な精度での解析が可能になった．また，近代以降の旧仮名遣いで書かれた口語文については，現代語に極めて近いため，現代語コーパスと専用コーパスを用いることで対応を行った．

第 6 章では，ユーザ向けのツールの作成と提供について論じる．古文の形態素解析システムや日本語歴史コーパスの利用者のほとんどは人文科学系の研究者であり，コンピュータの利用にかならずしも習熟していない．こうしたユーザにとっても利用しやすい形で提供するために，形態素解析を行いやすくするツール，古文のコーパスの検索ツール，解析結果を利用して総索引を作成するツールの作成や提供を行った．

第2章　通時コーパスと形態論情報付与

2.1　関連研究

2.1.1　日本語史研究資料のコーパス化

1.1 で述べたとおり，日本語の歴史を研究するうえで，論証のための用例を検索することは極めて重要である．現代語の研究とは異なり，歴史的な研究では，研究者や話者の内省にもとづく意味や文法性の判断ができないため，残された資料における用例は最も重要な研究の基礎であり，用例検索の方法は日本語史研究で常に課題となってきた．時代の変化に合わせ，書籍として刊行された総索引，電子化テキスト，文書構造タグ付きのテキストコーパス，そして形態論情報付きのコーパスへと移り変わってきた．本節では，このデータの変遷を振り返り，今後求められる日本語史研究資料としてのコーパスについて論ずる．

総索引の時代（電子化以前）

用例にもとづく実証的な日本語史の研究は江戸時代に始まる．当初は，用例の検索は作品をほぼ諳んじるまでに読み込んだり，調査のたびに全文に目を通して確認したりする必要があり，極めて負担の重い作業であった[1]．やがて昭和になると，1929〜31 年にかけて刊行された『万葉集総索引』[84] を嚆矢として[2]，『対校源氏物語用語索引』[14]，『徒然草総索引』[61] などが刊行され，主要な文学作品の総索引が整備された．これによって全文を確認することなく用例の数と出現箇所を知ることが可能になり，総索引は用例検索に欠かせないものとなった．

総索引を用いた用例検索は，今日でもほとんどの日本語史研究者にとって研究の最も基礎的な作業となっている．作品に索引が整備されているかどうかは，そ

[1] 大野（1999）[54](pp.138-143）に総索引が未整備だった時代の調査の様子は描写されている．

[2] 単語を単位とした総索引ではないが，和歌の検索用の書籍としては『国歌大観』（1901-03）が早い．

の作品が日本語史資料として活用されるか否かに直結するほどの重要性を持つ．したがって総索引の作成・刊行は資料研究の一歩として今日でも行われており，日本古典作品の総索引のシリーズの一つである笠間索引叢刊は 122 巻以上を数えている．

初期のコンピュータ利用（1980 年代）

総索引の作成・利用が続く傍ら，1980 年代になると人文科学系の研究者の間でもコンピュータの利用が可能になり，文学作品の電子化やその処理の試みが始まる．組織によるものとしては，国文学研究資料館の星野（1983）[7] が万葉集・古今集のテキストデータ化に取り組んでいる．また，風間ほか（1983）[13] では，キリシタン資料の『ぎやどぺかどる』の索引作成のためにコンピュータを導入している[3]．訓点資料である『恵果和上之碑文』を扱った金水（1984）[15] も古典籍を電子化して利用した早い例である．いずれも大学に導入された大型計算機を用いた成果であった．西端幸雄（1983）[21]，西端幸雄・藤田久・成田徹（1989）[23] などのパソコンを用いた総索引作成も行われている．

この頃，『フロッピー版古典対照語い表』[72] が刊行されている．これは 1971 年に刊行された『古典対照語い表』[71] の電子版である．『古典対照語い表』は本文テキストは含まず，刊行されていた総索引を基にして，古典文学作品に出現する全ての語の頻度を作品別の一覧表にしたものであるが，古典文学作品の語彙を見渡すことを可能にした画期的なものであった．そのフロッピー版は電子的に利用可能な形で公開された古文の言語資源として最も早いものの一つである．

電子化テキストと CD-ROM（1990 年代）

1980 年代終わり頃から徐々に古典作品の電子化テキストが個人レベルでも作成され，パソコン通信のコミュニティなどを通して共有され始めた[4]．これらの多くは研究者によって作成されたデータであり，学術的な研究に必要な情報が付与されていて有用である．しかし，ファイルごとに固有の形式で記述され統一されていなかったため，利用しにくい面があった．

1990 年代に入ると，電子化テキストの作成と利用が拡大する．やはり大部分は小規模な個別作品を電子化するものであったが，1990 年に長瀬真理らによって作

[3] この成果は，後に豊島正之（1987）[60] として総索引にまとめられ刊行されている．

[4] 現在でも，「日本文学等テキストファイル」[46] や「J-TEXT 日本文学電子図書館」[55] においてこれらの公開されたテキストデータが確認できる．

成された「『源氏物語』テキストデータベース」[56] は，小学館『日本古典文学全集　源氏物語』[37] の全文を，英語の通時コーパスである「ヘルシンキコーパス」

1992 年には国語学会（現日本語学会）の春季大会テーマとして「国語研究資料の「電子化」とその利用」が立てられ（[65] など），パソコンの普及とともに一般の日本語研究者にとっても電子化テキストの利用が身近なものになった．西端幸雄・木村雅則・志甫由紀恵編（1996）[22] は平安日記文学 5 作品の語彙索引だが，データを収めたフロッピーディスクが付属しており有用であった．総索引の作成に電子化テキストが用いられ，そのソースデータがそのまま公開されたものであり，1990 年代のコンピュータを利用した日本語史研究のあり方のモデルとも言えるものである．このころの電子化テキストの利用形態は，grep による文字列検索が一般であった．

同じ頃，インターネット上でのテキスト公開も始まる．1995 年には，国文学研究資料館より岩波書店の「日本古典文学大系」のテキストデータベースが研究者向けに公開された．これは 100 巻に及ぶ古典文学作品を SGML によってマークアップしたデータ（安永 1998）[44] を Web サービスの形で公開したもので，大規模な古典のテキストデータを利用可能にした画期的なものであった．ただし，データは書籍として出版された文学作品の版面を起こしたものであり，必ずしも文書構造をマークアップしようとしたものではない．また，単語の情報は基本的に付与されておらず，文字列検索による利用となるが，古典作品特有の表記の多様性により，索引の代用としての利用には困難が伴った．インターネット上では，著作権保護期間を過ぎた文学作品の大規模なテキストアーカイブである「青空文庫」

1990 年代後半にはまた，CD-ROM によって多くの古典作品の電子データが商業的に出版された．岩波書店から 1995 年に『新日本古典文学大系』所収の勅撰和歌集を電子化した『八代集』が刊行され，角川書店からは『新編国歌大観』（1996 年），『角川古語大観　源氏物語』（1999 年）という大規模なデータベースが刊行されている．また，国文学研究資料館の「国文学研究資料館データベース　古典コレクション」が刊行され，『二十一代集（正保版本）』（1999 年），『絵入り　承応版本 源氏物語』（1999 年），『歴史物語（栄花物語・大鏡・今鏡・水鏡・増鏡）』（2003 年）等が出ている [5]．国立国語研究所からは『国定読本用語総覧』[27] が刊行された．これらは総索引を電子化して検索インターフェイスを備えた形態のもので，あくまでも従来の索引利用が PC 上で可能になったものだといえる．CD-ROM 内のテキストも自由に扱うことのできる形式ではなかったため，コンピュータ上で利用可能な言語資源としての応用は行われなかった．

[5] これらのデータは現在「古典選集本文データベース」として Web サービスの形で公開が行われている [25].

日本語研究を目的とした歴史的資料のコーパス（2000 年代）

2000 年代になると，日本語史研究を目的として組織的に作成された大規模なコーパスが作成されはじめた．「コーパス」の名称を持つ日本語研究用のデータの公開もこの頃から始まった．データの形式もプレーンテキストや独自形式のテキストではなく XML などの標準的なマークアップ言語が利用されるようになる．

日本語の歴史的な資料のコーパスは，国立国語研究所で構築された『太陽コーパス』[29] が最初である．『太陽コーパス』の完成版は 2005 年に CD-ROM で公開された．データ本体は XML 形式で，ふりがな，本文の校訂情報，記事情報，引用などの情報がタグ付けされており，XSLT を利用した形式変換や情報抽出も可能であった．全文検索システム「ひまわり」[45] により，人文科学系の研究者にも利用しやすい検索環境が用意されていた．しかしタグ付けされた情報は外形的な文書構造や文字で表現できない修飾要素に限られ，単語情報は備えていなかった．

2006 年には『近代女性雑誌コーパス』[30] が公開されているが，これも『太陽コーパス』と同様の形式である．どちらも文字列検索による利用に限られており，単語情報による検索は行えなかった．

文字列検索の限界と形態素解析の必要性

以上のように，2000 年代に至るまで，日本語の歴史的資料の用例検索は，紙や CD-ROM 版の索引によるか，電子化テキストを表層文字列で検索するかのいずれかであった．これは，現代語の日本語研究資料においても同様であった．

このような中，2005 年より『現代日本語書き言葉均衡コーパス（BCCWJ）』[34] の構築が始まる．BCCWJ はなにより 1 億語を超える規模の点で注目されるが，全文に言語研究を目的とする形態論情報が付与された点でも画期的であった．形態論情報が付けられたコーパスは，自然言語処理用の言語資源としては「RWCP コーパス [53]」「京大コーパス [35]」などですでに実現されていたが，日本語研究を目的として作られた大規模なものとしては BCCWJ が初めてである．これにより単語情報を編集した総索引と同等の情報が日本語研究のための電子化テキストに付与されることになった．

現代語の研究においても，テキストデータが盛んに用いられてきた [11] が，利用方法はもっぱら表層文字列の検索によって用例を収集するものであった．しかし分かち書きがされない日本語の文章を検索するには単純な文字列検索では全く不十分である．そのため，人文科学系の日本語研究者の間でも正規表現の利用が進んだが，それでも十分に対応しきれるものではなかった．特に，日本語学においてしばしば研究対象とされる助詞・助動詞は，短いひらがな表記であるため，

表層文字列の検索ではそれだけを検索することが極めて困難である．また，英語を対象として発達したコーパス言語学では，語を単位とした研究手法の蓄積があり，日本語学でもその応用が期待される．日本語学においても研究対象の多くは単語やその組み合わせなのであって，やはり単語を単位とした処理を行えることが望ましい．このようなことから，単語の情報がタグ付けされたコーパスの登場が待ち望まれていた．

歴史的な資料では，仮名遣いの揺れや，異体字の使用が多いなど現代語以上に表記の幅が広いため，検索に際して問題が多い．電子化テキストは，検索性，検索結果の再利用性などの点で紙の索引にまさる側面があるが，単語の情報が付けられていないため，多様な表記，予想外の語形などが多く見られる古典本文の分析用データとしては，総索引に劣る部分もあった．

表層文字列の検索では，正規表現を使うにせよ，事前に目標とする語の表記パターンを把握しておく必要があるが，内省がきかない上に表記バリエーションが極めて多いため，それが現代語以上に困難なのである．たとえば，近代語の資料における用例を見ると，文語活用の動詞「見る」と，国名の「アメリカ」の表記には次のような例が実際に用いられている．

動詞の「見る」（文語上一段活用）

みる，看る，睹る，瞰る，瞻る，矚る，見る，視る，覩る，覲る，観る，覧る，觀る，診る，賭る，閲る，覩る

国名の「アメリカ」

あめりか，アメリカ，亜墨利加，亜米利加，亞墨利加，亞米利加，阿米利加，阿美利加，阿美理駕，Ａｍｅｒｉｃａ

このように歴史的資料においては想定外の表記形が存在するため，1つの単語の検索であっても表層文字列の検索では制限が大きく，単語の情報が付与されている必要がある．さらに，現代語研究と同様の問題があり，コーパス言語学的な研究手法を日本語史研究に導入するためにも，歴史的資料についても単語の情報がタグ付けされたコーパスの登場が待ち望まれていた．

形態論情報付きの歴史的資料のコーパスへ（2010 年代）

こうした背景から，国立国語研究所では国立国語研究所の基幹型プロジェクト「通時コーパスの設計」において，日本語の通時コーパスの構築の準備を行うこととなった（近藤 2012 [67]）．最終的には江戸時代以前の日本語の全体を見渡すことのできる通時コーパスを構築しようという大規模な試みである．

このような通時コーパスの整備は，日本語史研究の発展全体にとっても大きな意義を持つものである．これまでの日本語史研究では，資料の研究上の取り扱いで高度な専門知識を要するため，その資料の専門家以外には参入が難しい場合が少なくなかった．その結果，日本語史研究があまりにも細分化し，作品・資料ごとの研究は高度化しても，全体を俯瞰する通時的な日本語史研究が行いづらい状況にあった．

専門家の監督の下で確たる調査の上にアノテーションが施された通時コーパスを作成すれば，各資料について必ずしも専門でなくとも，通時的に日本語の変化を俯瞰できることになる．これによって日本語史研究者が日本語史の全体像を把握しやすくなるだけでなく，専門外の言語研究者の参入も容易になるであろう．

2.1.2 古文の形態素解析

初期の研究（形態素解析以前）

前節で確認した歴史的日本語資料の電子化テキスト作成，コーパス化と平行するように，総索引の作成を大きな動機として，1980 年代からすでに古文のコンピュータ処理の試みが進められていた．風間ほか（1983）[13]，西端幸雄・藤田久・成田徹（1989）[23] のように，多くは手作業による境界付与を行った上で，自動処理を行おうとするものだった．伊藤（1987）[8]，伊藤（1988）[9] は「古典語文の自動分割」，すなわち自動単語分割を試みている．かな書きされる付属語と活用語尾を辞書として持ち，文節末の付属語を切り出す形で単語に分割するものであった．伊藤は，『今昔物語集』を資料として選んだ理由として，「『今昔物語集』の本文が原則として自立語は漢字表記，付属語・活用語尾は仮名表記となっている点が現代文と共通しており，現段階では古典の中で最も分割しやすいと判断したからである」と述べており，字種に依存した単語分割方法が通用しやすい作品を例として解析に取り組んだものであった．この研究はその後も伊藤（1996）[10] として発展するが『今昔物語集』以外には広がらなかった．

このほかに『源氏物語』を対象として上田裕一・上田英代・村上征勝らによる一連の研究（上田 1992 [92] など）がある．この研究はテキストの自動単語分割を行って人手による修正を施した後，総索引作成や語彙研究に活用したもので，総索引は『源氏物語語彙用例総索引』[6] 全 11 巻として結実している．

これら初期の研究は，特定の作品について総索引を作成するためにコンピュータを応用しようとするものであり，古文一般を対象とした形態素解析を志向したものではなかったと言える．

古文の形態素解析の試み

現代語の形態素解析は，1992 年に公開された JUMAN [47] 以降，実用的な解析が可能になり応用が進む．古文の形態素解析の試みも 1990 年代から行われている．

工学的な見地から古文の形態素解析を試みた早い時期の研究として，安武・吉村・首藤（1995）[87] がある．この研究は，『徒然草』を例に，古語辞典から作成した品詞の接続ルールと電子化辞書を作成し，文節数最小法によって解析を行うものであった．

また，同じ頃，山本・松本（1996）[88] は JUMAN を利用した古文の形態素解析を行っている．JUMAN はルールベースの形態素解析器であるため，古典文法を元にコストを調整してルールを作成し，辞書は『古典対照語い表』[72] のデータを利用して作成されている．

当時はまだ古語の電子的な辞書や古文のコーパスが不足していたこともあり，これらの研究はいずれも試行のレベルにとどまっており，アプリケーションの公開も行われていない．また，研究の主眼が解析手法の開発自体にあるため，コーパスや辞書も実験に必要な最小限しか整備されておらず，言語資源としての蓄積はこの間も行われてこなかった．

古典研究を行う側にも形態素解析を求める機運が熟していなかったこともあって，その後も古文の形態素解析には大きな進展がなかった．

近年では山元（2007）[16] [6] が和歌集の言語学的分析を目的として和歌の形態素解析を実現している．ただし，手法としては文字列のパターンマッチによるものであり，和歌に特化したものである．この方法は五七五七七という韻律に縛られた短いテキストである和歌においては有効であるが，散文等の古典語一般にそのまま適用できるものではない．

統計的機械学習にもとづく形態素解析と古文

1990 年代終わり頃に日本語形態素解析の手法はルールベースから機械学習ベースへと大きく転換した．

1997 年に奈良先端科学技術大学院大学松本研究室で開発され 1999 年に一般公開された ChaSen [48][7] は，隠れマルコフモデル（HMM）に基づくコーパスからのパラメータ推定（竹内・松本 1997）[19] により，JUMAN のような手作業によるコスト調整を不要にした．パラメータ学習用のコーパスとして毎日新聞の記事

[6] http://warbler.ryu.titech.ac.jp/~yamagen/kh/readme.html
[7] http://chasen.naist.jp/

に形態論情報をタグ付けしたRWCPコーパス，辞書としてIPADIC [62] を用いている．ChaSenはインターネットの利用が広がる中，フリーソフトウェアとして広く用いられた．

2006年には工藤拓によるMeCab [69][8]が発表される．これはHMMにかえて条件付き確立場（CRF）[3] を用いた汎用的な解析器 [2] である．ChaSenと同じく辞書にはIPADICが主に利用されていたが，ChaSenを上回る解析精度と速度により，今日ではその後継として事実上の標準といえる位置を占めるようになっている．

ChaSenやMeCabのような統計的機械学習にもとづく形態素解析システムの登場は，歴史的な資料の形態素解析の観点から見ると，辞書整備とコーパス作成さえ行えば，連接ルールの記述やコストの調整を行わなくとも形態素解析が実現できる環境が整ったという点で大きな前進であった．

現代人にとって歴史的な資料の文法は内省がきかないため，人手によって連接ルールを記述したりコストを調整したりすることは極めて困難である．典型的な古典に限れば，比較的整った古典文法が存在するが，通時的に見ると多くの時代のテキストが未だ研究途上にあり，各時代・各資料の文法について十分に記述された文法書は存在しない．したがって，辞書とコーパスの整備だけで形態素解析が可能になったことは，実用的な古文の形態素解析の実現へのハードルが下がったことになる．

しかしながら，古語の辞書整備とコーパス作成は，人文科学系の研究者の側からの積極的な関与がなければ困難であった．辞書やコーパスが存在しないために，機械学習ベースの解析器による古文の解析は2000年代後半まで行われてこなかった．

2.1.3 UniDicと古文

UniDicの特長

2.1.2で述べたとおり，広く用いられてきた形態素解析器であるChaSenやMeCabでは，辞書としてIPADIC [62] が事実上の標準として広く用いられてきた．

この辞書は，RWCPコーパスとともに，統計的機械学習に基づく形態素解析を実現した画期的な言語資源であり，広く活用されてきた．しかし，言語研究用の電子辞書という目的から，今日の目で見直したときには，次のような点で問題が

[8] http://code.google.com/p/mecab/

ある[9].

1. 解析単位（語）の認定基準が必ずしも明確でない
2. 学習用コーパスがもっぱら新聞記事のデータによっているため他ジャンルのテキストの解析精度が低い
3. 異表記の同語が別語と見なされるためそのままでは語彙調査に利用できない

そこで『現代日本語書き言葉均衡コーパス』(BCCWJ) [34] の構築にあたっては，国立国語研究所が中心となって，言語研究に適した新しい電子化辞書「UniDic」[10] が開発されることとなった（伝ほか 2007）[24]．UniDic は ChaSen または MeCab と組み合わせて利用する形態素解析用の辞書として公開された[11]．UniDic の特長として次の点が挙げられる．

i. 「短単位」という揺れが少ない斉一な単位を見出し語に採用している．
ii. 語彙素・語形・書字形・発音形の階層構造を持ち，表記の揺れや語形の変異にかかわらず同一の見出しを与えることができる．
iii. 語種や，アクセントや音変化の情報など，研究に有用な豊富な情報が付与できる
iv. BCCWJ に収録された多様なジャンルのテキストが高い精度で解析できる

このような特長により，先に挙げた従来の辞書が抱えていた問題を解決している．以下，この 4 点に沿って UniDic の特長を概観し，通時コーパス構築の観点からその価値を確認する．

(i) 短単位

「短単位」は国立国語研究所の語彙調査で従来から用いられていた「短い単位」，中でも『現代の語彙調査・総合雑誌の用語』[26] 『現代雑誌九十種の用語用字』

[9] IPADIC をもとにライセンス問題を解決して公開された NAIST-jdic [58] は表記ゆれ情報を持つことで 3. の問題に対処しているが，それ以外の点では同様の問題を持っている．
[10] http://download.unidic.org/
[11] 2007 年は ChaSen 版が公開され，2008 年に MeCab 用辞書が公開されたが，現在では ChaSen 版の更新は中止されている．

[32]『雑誌 200 万字言語調査』[12] で採用された「β 単位」に近いものである．『日本語話し言葉コーパス』（CSJ）[33] の構築に際してもこの短単位が採用された [13]．

短単位は，最小単位（形態素）の 1 次結合までを最大とする言語単位である．例えば，「国立国語研究所で研究している．」という文は，短単位では次の 10 単位に分割される．

/ 国立/ 国語/ 研究/ 所/で/研究/し/て/いる/．/

なお，短単位とは別に，ほぼ文節に近い長さの言語単位である長単位が規定されている．先の例は，長単位では次のように 5 単位に分割される．

/ 国立国語研究所/で/研究し/ている/．/

長単位は短単位を組み上げて作られる．BCCWJ の構築では，UniDic による解析結果である短単位列を，長単位解析器 Comainu [43] [85] によって自動処理して組み上げた後，人手による修正が行われた．両者の範囲は入れ子関係にあり，長単位の境界は必ず短単位の境界となっている．

短単位は，主として言語の形態的な側面に着目して考えた単位であり，『現代日本語書き言葉均衡コーパス』では単位の認定方法を規程集（小椋ほか 2011 [39]）に詳細に定めて揺れを防いでいる．国語研究所の語彙調査における単位の規定に際しては，「語とは何か」という本質的な議論に立ち入るのことは避け，一貫して外形的に規定可能な，操作主義的な定義によっている（小椋・冨士池 2011 [41]）．それゆえに，内省がきかない過去の言語資料を扱う際の単位としても認定しやすい面がある．

古文の形態素解析にとっても，UniDic の短単位がもつ特長は有効である．揺れの少ない斉一な単位は，テキストの解析結果を用いた語彙の比較を可能にする．

宮島（1969）[70] が夙に指摘しているように，従来の古典文学作品の総索引は単位認定や見出し付与の方針の違いにより相互の比較が難しい場合があった．しかし通時的なコーパスを構築する場合には，現代語コーパスと共通する一貫した原理に基づいた情報付与を行って，一作品・一時代にとどまらず古代から現代に至る「通時」的な観察を可能にする必要がある．UniDic をベースとすることで，作品間の比較が可能になるだけでなく，時代の違いを超え，各種のテキスト間で相互に語彙を比較することが可能になる．

[12] http://www.ninjal.ac.jp/archives/goityosa/

[13] ただし，CSJ の短単位と，UniDic や BCCWJ の短単位とでは，外来語の単位認定基準などでいくつかの違いがある．小椋・冨士池（2011）[41] 参照．

もっとも，語の歴史的変化や各時代における実態を踏まえて，時代別に異なった扱いをしなければならない場合もある．こうした処理は規程集としてまとめ，作り手とユーザーの手引きとする必要がある．

(ii) 見出し語の階層構造

UniDic の見出し語の階層構造は次のようなものである．

図 2.1: UniDic の見出し語階層

これら 4 つの見出し語の表は，語形変化表（語頭・語末変化表および活用表）と組み合わせて形態素解析辞書の語彙表にまで展開される．すなわち，辞書登録された個々の見出し語は，語頭変化・語末変化・活用変化を経て出現形（表層形）が派生されるようになっている．

(iii) 豊富な情報付与

UniDic は標準の形態素解析結果として表 2.1 に示す属性を出力する．

表 2.1: UniDic の形態論情報

階層	属性の名称	説明
語彙素	語彙素読み	語彙素見出し（カタカナ表記）
	語彙素	語彙素見出し（漢字仮名混じりの代表表記）
	語種	語種の名称（付録 B.1 参照）
語形	品詞大分類	品詞の名称（付録 B.2 参照）
	品詞中分類	
	品詞小分類	
	品詞細分類	
	活用型	活用の種類（型）（付録 B.3 参照）

	活用形	活用の形（付録 B.4 参照）
	語形基本形	異語形を区別する形（カナ）
	語頭変化型	語頭音変化の種類（型）
	語頭変化形	語頭音変化の形
	語頭変化結合型	後続要素の語頭変化形への制約の種類（型）
	語末変化型	語末音変化の種類（型）
	語末変化形	語末音変化の形
	語末変化結合型	前接要素の語末変化形への制約の種類（型）
書字形	書字形基本形	書字形見出し
	書字形出現形	書字形基本形が活用変化を受けたもの
	仮名形出現形	書字形出現形をカタカナ表記にしたもの
	仮名形基本形	書字形基本形をカタカナ表記にしたもの
発音形	発音形発音形基本形	読み上げ用の形（現代読み）
	発音形出現形	発音形基本形が活用変化を受けたもの
	アクセント型	アクセント核の位置
	アクセント結合型	前接（後続）要素との結合時のアクセント変化の種類（型）
	アクセント修飾型	活用によるアクセント変化の種類（型）

※ UniDic 1.2.13 ユーザーズマニュアル・表 1 より

表 2.1 のうち語彙素・語彙素読み・品詞・活用型・活用形・書字形・発音形（または語形）の組により，辞書中の見出し語として一意に同定することができるようになっている．

語種情報（表 B.1）を持つことは特に重要である．日本語研究にとって語種は極めて基本的な情報であり，これが付与できることは重要な価値を持っている．歴史的な資料では，中古和文や和歌ではほとんどが和語で書かれるのに対し，時代を下るにつれて漢語の使用が増えることが広く知られている．後述するとおり，現代語の UniDic でも学習素性として語種を利用することで解析精度が向上できることが分かっており（Den et al. 2008 [1]），この素性は歴史的資料の解析においても有効であると考えられる．

発音形やアクセント型，音変化情報を付与することができ，音声処理の研究への応用が可能なことも UniDic の特長の一つである．ただし，歴史的な資料については，音韻史・アクセント史研究の成果はあるものの，実際の発音やアクセントについては不明な点が多く，通時コーパスでもアノテーション対象とはしてい

ない．そのため，歴史的資料を対象としたUniDicでは，過去の書き言葉を現代人が読み上げる際の一般的な発音形を付与することとしている．

(iv) 多様なテキストへの対応

最新版のUniDicでは，現代語の様々なジャンルのテキストを98%以上の精度で解析することが可能になっている（小木曽ほか2010 [77]）．このように短単位を用いて多様なテキストが高精度に解析できることは，日本語のバリエーションとしての歴史的な資料にUniDicの見出し語を適用することが有効であることを示唆する．

また，通時コーパスの中でも現代に近いテキストを対象とする際には，BCCWJに含まれる多様なテキストのうち文体的に近いものをパラメータ学習に活用することが可能である．こうした点にも歴史的資料のための形態素解析用辞書をUniDicに基づいて開発することの有効性がある．

UniDicの見出し語階層と古文

UniDicは種々の文体に対応しているとはいえ，あくまでも現代語用の形態素解析辞書であり，そのままでは古文を解析することはできない．しかし，この辞書を元にして古文用の見出し語を追加し，学習用のコーパスを準備することにより，古文用の形態素解析辞書を作成することは可能である．UniDicがもつ斉一な単位や階層化された見出し構造は，古文の形態素解析辞書の作成時にもたいへん有用である．

UniDicでは見出し語を語彙素・語形・書字形・発音形の4段階で階層的に管理しているため，必要な語を各階層に整理して追加することができる．図2.2は，「何処（いずこ）」の例であるが，代表となる辞書見出しとして語彙素「何処」（語彙素読み「イズコ」）を立項し，その下位に異なる「語形」として「イズコ」「イドコ」「イズク」を配している（現代語とデータベースを共有するため語彙素や語形は現代仮名遣いによっている）．そして，新旧の漢字（例：「何処」「何處」）・仮名遣い（例：「いずこ」「いづこ」）や送り仮名の違いなどの異表記形が「書字形」として各語形の下に位置づけられる．

この階層に合わせて，新規に追加する語は「語彙素」のレベルで，文語活用型の語は「語形」のレベルで，旧字形などは「書字形」のレベルで追加することになる．こうすることで，現代語の語と統一的に管理することができるだけでなく，文語・口語形，異表記形がそれぞれ関係を持つものであることを示すことができる．

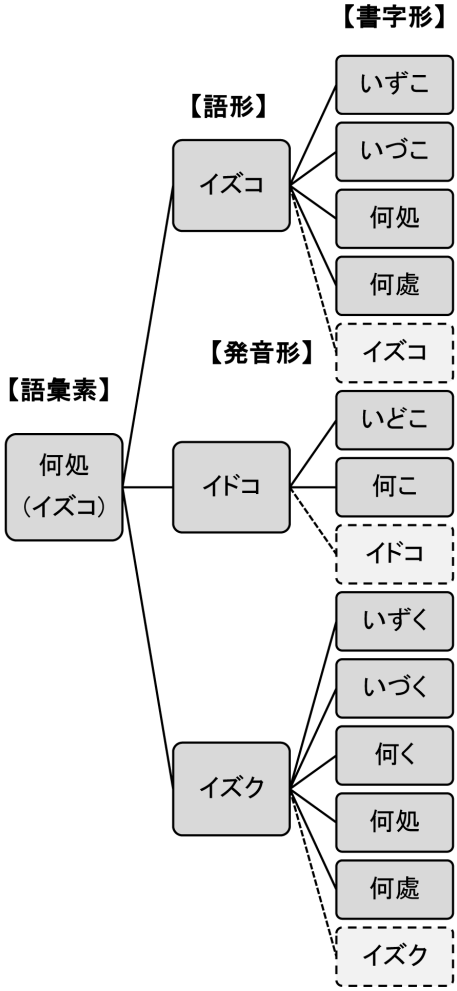


図 2.2: UniDic の階層と文語形・旧字形

辞書データに見出し語を追加登録していく際，活用語については活用表を整備する必要がある．一般的な文語の活用形をそろえたほか，歴史コーパス構築に必要な特殊な活用形の追加を行った（B.4 参照）．たとえば，「思はく」「願はく」のようなク語法は，元となる動詞の語形変異として扱った方がコーパスを利用する上で便利であるため，一活用形として扱っている．また，「読て」「読ず」のように送り仮名が省略される表記に対応するための書字形も活用表の一部として整備している．

UniDic では，形態素解析器に MeCab [2] を用いている．2.1.2 でも述べたとおり，MeCab はコーパスからのパラメータ学習器も公開されているため，古文の学

習用コーパスを用意し，UniDic の見出し語を拡充することで古文に対応することができる．後述するように，見出し語を追加するだけでは実用的な解析精度を達成することはできず，学習用コーパスを用いて形態素解析器を再学習することが重要である．

2.2 通時コーパスのテキストの多様性

2.2.1 日本語史の時代区分と主要資料

日本語の歴史は，8 世紀に成立した『万葉集』以来，文献に基づいて現代までたどることができる．国立国語研究所で計画中の通時コーパスは，この歴史的な文献資料をコーパス化し，日本語の歴史を研究する基盤として整備しようとするものである．本節では，これらのテキストを概観したうえで，形態素解析を実現するに当たってこの多様なテキストをどのように扱うべきかを検討する．

日本語の歴史は，日本語史研究においておおむね表 2.2 のように時代区分される（高山・青木 2010 [63]）．もっとも，日本語自体の変化によって規定されるべきものであるから，日本史の歴史区分と厳密に一致するものではなく，それぞれの境界も必ずしも明確なものではない．

表 2.2: 日本語の歴史的区分

区分	時代	西暦
上代語	奈良時代（及びそれ以前）	— 794 年
中古語	平安時代	794 年—1192 年
中世語	鎌倉時代・室町時代	1192 年—1603 年
近世語	江戸時代	1603 年—1868 年
近代語	明治・大正・昭和前期	1868 年—1945 年
現代語	昭和後期・平成	1945 年—

国立国語研究所の「通時コーパスの設計」プロジェクトは，このうちの上代語から近世語までを対象としており，このプロジェクトで構築されているコーパスは『日本語歴史コーパス』の名前で公開されている[14]．また「近代語コーパス設計のための文献言語研究」プロジェクトが近代語を対象としており，こちらはプロジェクトの成果は，プロジェクト開始以前に公開された『太陽コーパス』等と

[14] http://www.ninjal.ac.jp/corpus_center/chj/

あわせて，「近代語のコーパス」として公開されている[15]．本研究でいう通時コーパスには両者が含まれるが，両プロジェクトによって構築されるコーパスを「国語研通時コーパス」と呼ぶことにする．

「国語研通時コーパス」では時代幅としては8世紀から19世紀までが対象となるが，その間，継続して同種類の資料が残されているわけではない．むしろ，各時代に特有のジャンルの資料が孤立するように残されており，日本語史はそれをつなぎ合わせるようにして作られる．たとえば，上代では日本語の研究に利用できる資料は主として和歌であり，散文はほとんど残されていない．また，中世以降では，文語と口語とが乖離していくため，当時の口語を探ることの資料としては各時代に固有の資料（軍記物，抄物，キリシタン文献，狂言など）を利用することとなるため，同種の資料のみを通時的に収集することは不可能である．

表2.3に各時代の主要な資料を挙げる．

表 2.3: 時代別の主な日本語資料

時代区分	主な資料
上代	万葉集，記紀歌謡，宣命
中古	仮名文学作品（歌物語，作り物語，日記），歌集，説話集，訓点資料，古記録
中世	説話集，軍記物，抄物，キリシタン文献，狂言，擬古物語
近世	人形浄瑠璃（近松門左衛門），浮世草子（井原西鶴），噺本，洒落本，人情本，滑稽本
近代	口語小説，雑誌 その他

これらの資料のうち，「国語研通時コーパス」に収録されることが確定し，国立国語研究所においてなんらかの作業に着手しているものは表2.4の資料である[16]．

これらのテキストの文体は，文法・語彙・表記にわたって極めて多様であって，単に「古文」としてひとくくりにして済むものではない．時代差と，文語・口語という文体差を考慮に入れて各時代の日本語の文体について大まかにまとめると，図2.3のようになる．なお，それぞれの実際のテキストの例を付録Aに挙げた．

和歌の文体は平安時代に確立し，その後も大きくは変わらない文体で作られ続け，今日の短歌にまでつながっている．テキストの例として『古今和歌集』A.1.5を付録に挙げた．

[15] http://www.ninjal.ac.jp/corpus_center/cmj/

[16] なお，上代についてはこれとは別に池田幸恵らによって五国史宣命コーパスの構築が進められている（池田・須永 2013）[20].

中古和文は平安時代の口語を反映した仮名文学作品の文体で，古典の模範的文体としてその後も用いられ，中世の擬古物語や近世の擬古文，さらには明治期の雅文に至るまで続き，古文の一つの典型的な文体となっている．

歌物語の例として『伊勢物語』(A.1.1)，作り物語の例として『竹取物語』(A.1.2)と『源氏物語』(A.1.2)，日記の例として『更級日記』(A.1.3)，擬古物語の例として『恋路ゆかしき大将』(A.1.4) のそれぞれ一部を付録に挙げた．

中古和文と和歌の文体には文法や語彙において違いもあるが，和文の仮名文学作品の本文に和歌が多く含まれていることもあり，大きくは一括して扱うことができる．

漢文の訓点を読み下した形の漢文訓読文は漢語を多く含み，大部分が和語からなる和文とは語彙が大きく異なる．平安時代における文語文であったといえる．同時代の文体であっても，和文とは文法や使用される語彙（和語を含む）に差異

表 2.4: 国語研通時コーパスプロジェクトで作成中の資料

時代区分	資料	底本
中古	作り物語（竹取物語・源氏物語・落窪物語・堤中納言物語）	『新編日本古典文学全集』[38]
	歌物語（伊勢物語・大和物語・平中物語）	
	日記（土佐日記・蜻蛉日記・和泉式部日記・紫式部日記・更級日記・讃岐典侍日記）	
	随筆（枕草子）	
	和歌（古今和歌集）	
	説話集（日本霊異記・今昔物語集）	
中世	説話集（宇治拾遺物語）	『新編日本古典文学全集』
	狂言	『大蔵虎明能狂言集 翻刻註解』[18]
近世	洒落本	『洒落本大成』[93]
	人情本	『新編日本古典文学全集』
近代	近代雑誌（明六雑誌，太陽，国民之友）	（原本を使用）

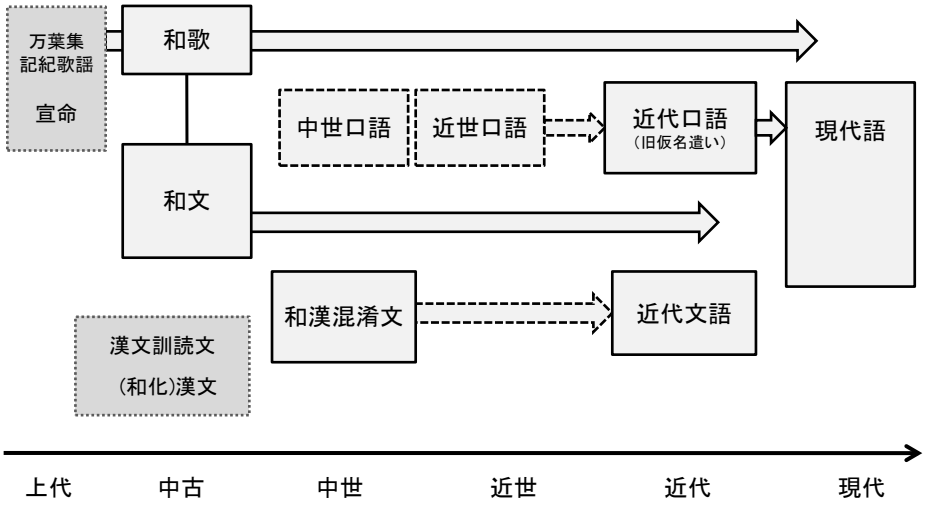


図 2.3: 各時代の資料・文体

があることが知られている [91]．この系統の文体は，和漢混淆文に大きな影響を与えつつ，その後も学者らによって用いられたほか，近代以降には論説文に広く用いられ，明治から戦前まで用いられた「普通文」と呼ばれる標準的な近代文語文の文体につながっている．

『日本霊異記』は平安初期に漢文で書かれた説話集で，読み下したものは漢文訓読文として日本語の文として読むことができ，後の『今昔物語集』につながるものである．例として，付録に「電を捉へし縁　第一」の一部（A.2.1）を挙げた．

和漢混淆文は，中世以降の説話や軍記物など，和文と漢文訓読文の混じり合った文体である．説話の例として平安時代末期成立の『今昔物語集』（A.2.1），軍記物の例として『平家物語』（A.2.2）の一部を付録に挙げた．

中世以降，口語と文語との乖離が進む．中世の口語を探ることのできる資料としては，キリシタン文献，抄物（しょうもの）[17]，狂言台本などがある．また，これに続く近世の口語は，洒落本や人情本，滑稽本などから探ることができる．当時の口語の多くは作品中の会話文として記されたものであり，書き言葉として広く用いられた

[17]室町時代の僧や学者による講義録．当時の口語が反映されている．

ものではなかった．

このうち国語研通時コーパスプロジェクトで構築がすでに行われている資料には狂言，洒落本，人情本がある．狂言の例として虎明本狂言 [18] より「あさう」（A.3.1）を，近世では，洒落本の例として『甲駅新話』（A.3.2）を，人情本の例として『春告鳥』（A.3.3）を付録に挙げた．

近代以降になると，先述の普通文と呼ばれる文語文と並んで，言文一致の流れにより口語が書き言葉に用いられるようになり，これが現代語へとつながっている．ただし，「現代かなづかい」（1946 年）・「現代仮名遣い」（1986 年）や「当用漢字表」（1946 年）・「常用漢字表」（1981 年）以前のものであり，表記法は大きく異なるものとなっている．

近代雑誌より，文語論説文の例として『明六雑誌』A.4.1 の例を挙げる．また，地の文は文語文で，会話のみ口語で書かれたテキストの例として『太陽』より「狂言娘」A.4.1 を挙げる．

さらに，文語論説文の例として『太陽』より「文学上の新事業」A.4.1，旧仮名遣いの口語小説（会話文）の例として同じく「生ける死（第三回）」A.4.1，旧仮名遣いの口語論説文の例として同じく「歴代の総理大臣（二）」A.4.1 を挙げる．

2.2.2　テキストの校訂と前処理

前節で挙げたような日本語の歴史的資料を対象としてテキスト処理を行う場合，本文校訂を避けて通ることができない．自然言語処理の対象として考える場合，校訂や前処理抜きの原文そのままを対象とすることは理想だが，こと日本語史の資料では，校訂や前処理が一切入らないテキストというものはあり得ない．例えば，近世以前の資料では，原文は変体仮名を含む崩し字で書かれているものが大部分であり[18]，これ自体をテキストとして処理することは考えられない．単に翻字（活字化）した資料をそのまま電子化テキストとしたデータであっても，この段階ですでに文字や切れ続きの認定などに一定の校訂は経ていると言える．

本文校訂には様々なレベルが考えられるが，具体的には次のような処理が挙げられる．

1. 翻字（活字化）
2. 異本との校合
3. 誤字脱字の修正

[18]例えば図 5.1 に挙げる洒落本の画像はそうしたものの一つである．

4. 句読点の追加（文境界の認定）

5. 濁点の付与

6. 表記の書き換え・統一

   (a) 仮名を漢字に書き換える

   (b) 漢字を一般的な表記に書き換える

   (c) 仮名遣いを歴史的仮名遣い（ないし現代仮名遣い）に修正

   (d) 送り仮名の整備

形態素解析を行う前提でデータを作成する場合には，それに適したレベルの校訂処理を行うべきであり，国語研通時コーパスプロジェクトで作成中の資料も，必要な校訂処理が施されている．

一方，研究資料としては原文を恣意的に改めることは不適切であり，できる限り原文を保存すべきである．国語研通時コーパスの資料では，XML によるマークアップにより，原則として底本とした資料の状態を保存しつつ，必要な情報を追加している．

ただし，国語研通時コーパスプロジェクトの底本の多くを占める『新編日本古典文学全集』[38] の本文は，文学作品として鑑賞することを前提にしたものであり，大部分はすでに高度な校訂が施されている．付録に挙げた中古和文（A.1）はそのようなものであり，本文の誤脱を校合して修正しただけでなく，全体が読みやすい漢字かな交じり文に直され，仮名遣いは歴史的仮名遣いに直されている．こうした本文は，ほとんどそのまま，形態素解析の対象として利用することができる．

一方，原本をそのままコーパスの底本とする近代雑誌や，狂言や洒落本の資料は，形態素解析以前に十分な前処理を行う必要がある．また，『新編日本古典文学全集』の本文でも，漢文を読み下している『日本霊異記』や，原文を忠実に再現している『今昔物語集』，さらに原文にあまり手を加えていない近世以降の作品については，前処理が必要である．

なお，校訂作業には仮名遣いの正規化も含まれるが，国語研通時コーパスでは仮名遣いの差は形態素解析辞書で吸収することとし，本文の修正ではなく辞書見出しの拡充によって対処している [19]．

[19] 現時点では必要となる仮名遣いのバリエーションを見出し語として追加しているが，今後は解析時に仮名遣いの揺れを考慮する方法（岡 2013b [52]）により辞書見出し拡充の負担を軽減することを検討している．

歴史的資料のコーパス化における校訂作業や前処理は，現時点ではその大部分を人手によっているが，今後自動処理が望まれるものである．

近代語資料の校訂と前処理

原資料をそのまま底本に用いる近代雑誌のコーパスでは，上述の「誤字脱字の修正」「文境界の付与」「濁点の付与」が前処理として課題となる．1 文の範囲を定めておくことは言語研究においても重要であるが，何よりも形態素解析の入力は一般に文を単位とするため，形態論情報を付与したコーパスを作る上で，文境界を付与することは欠くことのできないものである．上記の校訂作業は，『太陽コーパス』[29] の構築に際しては原則として人手によって行われた[20]．

『明六雑誌コーパス』の濁点の付与作業では，岡（2013）[51] にもとづく自動処理ツール [50] が用いられた．これは『太陽コーパス』の校訂済み濁点のデータを元にした機械学習によって実現された校訂の補助ツールである．最終的には人手による確認と修正が必要となるが，このツールの導入により大幅な省力化が可能になった．

『明六雑誌』では，多くの記事が漢字カタカナ交じりで書かれている．漢字カタカナ交じりであるか漢字ひらがな交じりであるかは，記事単位で記述しておけば十分であり，個別の表記について問題にする必要はない．したがって基本的には，漢字カタカナ交じり文と漢字ひらがな交じり文は相互に自動変換可能であるため，『明六雑誌コーパス』では記事ごとに原文がいずれであったかを属性として記述したうえで，全体を漢字ひらがな交じり文に統一している．

ただし，漢字カタカナ交じり文中に，外来語等の漢字ひらがな交じり文においてカタカナ書きされる語が現れる場合には，原文には存在しないひらがな表記の外来語をことさらに辞書登録する必要が生じる．このような表記形は，操作上生じてしまうものに過ぎず，本来は用例とすべきものでもないし，辞書登録すべきものでもない．そこで，あらかじめこうした語についてはひらがなに変換しないようにタグ付けを行った．

中近世資料の校訂と前処理

狂言や洒落本の資料でも，「濁点の付与」「文境界の付与」は大きな課題である．『大蔵虎明能狂言集』[18] や，『洒落本大成』[93] では，校訂が行われているもの

[20]形態論情報の付与を前提としていなかった『太陽コーパス』の校訂作業では，上記 3 点に加えて，仮名遣いの修正，漢字表記の修正なども行われている [28].

の，原文の表記を尊重して大きな修正は行われていない．文境界の情報も追加する必要がある．

また，近世の資料では，「ハ」「ミ」などのカタカナと字形が一致するひらがなをカタカナで翻字する習慣があり，これが無用な表記の揺れを生んでいるため，これらの仮名をひらがなに改める処理も行っている[21]．

漢文訓読系資料の校訂と前処理

漢文訓読系の資料では，漢字カタカナ交じり文で書かれているうえに，漢文の要素が交じっているため，そのままでは形態素解析を施すことができない．漢字ひらがな交じりの通常のテキストに変換し，返読を要する部分は日本語の語順に整えた後で形態素解析を行う必要がある

冨士池ほか（2013）[90] は新編日本古典文学全集の『今昔物語集』を例に，次のような問題点を指摘している．

- 返読文字：　不知（シラ）ズ，令聞（キカ）シム
- 助詞・助動詞等の省略表記：　今昔（いまはむかし）
- 捨て仮名：　候フウ，此カク
- 欠字欠文・破損
- 字種（片仮名・万葉仮名）
- 踊り字・くの字点・同の字点：　今ヤ〳〵，穴怖シ々々

これらの問題点は，形態素解析に際して，未知語を発生させたり，曖昧性を高めたりすることにつながり解析精度の低下を招く．ただそれだけでなく，次に示すような語の重複や欠損という問題にもつながる．

- 返読文字による語順の転換と形態素の重複
- 助詞・助動詞等の省略表記による形態素の不足
- 捨て仮名による形態素の重複

[21] 一般に，近世の資料の翻刻において変体仮名は現代通行の仮名字形に断りなく置き換えられている．「ハ」「ミ」の場合にのみ元の字形を残す必要は本来なく，研究資料としてひらがなに改めることに問題はない．

したがって，前処理によってこうした部分を通常の本文の形に修正しておく必要がある．

次の例は新編日本古典文学全集『今昔物語集』（A.2.1）の本文とその前処理後のテキスト例である．

（処理前）

今昔，山階寺ニ涅槃会ト云フ会有リ．此レ，二月ノ十五日ハ，釈迦如来，涅槃ニ入給ヒシ日也．然レバ，彼ノ寺ノ僧等，「昔ノ沙羅林ノ儀式ヲ思フニ，心無キ草木ソラ，皆其ノ知テ恋慕ノ形チ有キ．何況ヤ，心有リ，悟リ有ラム人ハ，釈迦大師ノ恩徳ヲ報ジ可奉シ」ト儀シ思テ，彼ノ寺ノ仏ハ釈迦如来ニ在セバ，其ノ御前ニシテ彼ノ二月ノ十五日ニ一日ノ法会ヲ行フ也ケリ．

（処理後）

今は昔，山階寺に涅槃会と云ふ会有り．此れ，二月の十五日は，釈迦如来，涅槃に入給ひし日也．然れば，彼の寺の僧等，「昔の沙羅林の儀式を思ふに，心無き草木そら，皆其の知て恋慕の形ち有き．何況むや，心有り，悟り有らむ人は，釈迦大師の恩徳を報じ奉べし」と儀し思て，彼の寺の仏は釈迦如来に在せば，其の御前にして彼の二月の十五日に一日の法会を行ふ也けり．

2.3 「通時コーパス」の形態論情報アノテーションの方針

2.3.1 多様なテキストを解析するための方針

前節で見たような種々のテキストに形態素解析を施す場合，どのような形態素解析辞書をどれだけ開発する必要があるだろうか．

テキストの文体に大きな違いがある以上，解析対象のテキストごとに最適の辞書を作成することができれば望ましいが，残された歴史的資料は有限であり，少量のテキストのために個別に辞書を作成することは現実的ではない（その手間で当該テキストをすべて人手で整備できてしまう）．したがって，文体的に近いテキストを十分な量のグループにまとめ，グループごとに適した形態素解析辞書を用意することが適切であると考えられる．

こうしたグループとしてまず考えられるのは，今から約1000年前の平安時代に書かれた仮名文学作品を中心とする和文系の資料（中古和文）である．中でも源

氏物語は日本語の古典の代表的なものであり，その文体は，中世の擬古物語や近世・近代の擬古文に至るまで模倣されながら長い期間にわたって用いられている．

もう一つのグループとしてあげられるのは，約 100 年前に広く用いられていた近代の文語文（近代文語文）である．中でも明治普通文とも呼ばれる近代の文語論説文は，明治以降，戦前にかけて広く用いられた文体であり，公文書から新聞・雑誌まで各種の資料がこれによって書かれている．

近代文語文は，平安時代以来の漢文訓読文の流れに位置づけられる文体である．漢文訓読文では漢語が多く用いられるのに対して，和文では漢語はわずかしか用いられない．また和文と漢文訓読文では使用する和語の語彙も大きく異なっていることが知られている．こうした語彙の違いからも，現代語からの時代的遠近という点からも，中古和文と近代文語文は対照的な位置にあり，この 2 つのグループについて形態素解析辞書を用意することは，通時コーパス全体の解析を目的とする上で有効であると考えられる．その方法については，4 章で取り扱う．

以上の 2 つの辞書の他に，より小規模な資料群の形態素解析の方法を検討する必要がある．例えば，中古末から現れる和漢混淆文や，近世の口語文が挙げられる．これらのテキストの解析については，5 章で取り扱う．

2.3.2　コーパスに求められる精度

形態素解析の精度は現代語を対象とした場合でも 98%程度であるが（小木曽ほか 2010 [77]），古典語研究では，現代語と比較して量が限られた資料をもとに研究を進める必要があるため，一般により高い精度でタグ付けされたコーパスが必要とされる．総索引の代替としての役割を果たすためには，古文のコーパスには 100%に限りなく近い精度が求められる．この精度は，自動形態素解析だけでは到達することは不可能であり，自動解析の後に人手による修正が必須である [22]．

こうしたコーパスを構築するためには，形態素解析の精度が高いことが望ましいことは言うまでもないが，高精度の形態素解析を実現するためのコストと，誤りを修正するためのコストの双方を考慮する必要がある．

機械学習にもとづく形態素解析を行う場合，誤りの人手修正が終わったコーパスは，そのまま学習用のコーパスとして利用可能になり，これが形態素解析の高精度化につながるという循環がある．したがって，当初は低い解析精度であってもそのコーパスを修正したうえで学習用コーパスに追加し，これにより新たなテ

[22] ただし，近代語のように大量に資料が残っている時代のテキストについては，現代語コーパスと同様に一部を高精度なものとしそれ以外は機械処理の結果のままとすることが現実的である．

キストをより高い精度で解析するという形で徐々にコーパスを構築していくことができる．

『現代日本語書き言葉均衡コーパス』の構築において必要とされた解析精度は，語彙素認定において 98%であった．通時コーパスの構築に際しては，テキスト量が限られているため，これよりも低い精度であっても修正は可能だが，修正コストを考慮すると 95%程度の解析精度は必要であると考えられる．

2.3.3　長単位のアノテーション

2.1.3 節で見たように『現代日本語書き言葉均衡コーパス』では，UniDic を用いた形態素解析で得られる短単位のほかに，文節を単位としてそこから付属語を切り離した単位である「長単位」が設定されていた．長単位は，短単位の解析・修正が終わった後，これを組み上げる形で行われた．

国語研通時コーパスでも，将来的には短単位と長単位の双方が揃うことが望ましいが，長単位が短単位を前提とするものである以上，まずは短単位の形態素解析を進める必要がある．また，漢語が少ない古典では，現代語ほどに短単位と長単位とで違いがないこともあり（冨士池 2012 [89]），長単位のアノテーションの優先度は必ずしも高くない．そこで，各時代の多様なテキストに対して，形態素解析による短単位アノテーションを行うことを優先する．

ただし，中古和文については，冨士池（2012）[89] などの検討を踏まえて，すでに長単位アノテーションに着手している[23]．中古和文の長単位アノテーションも，BCCWJ と同じく Comainu [85] によって行われた．

2.4　本章のまとめ

本章では，2.1.1 節で，歴史的な日本語資料の電子化の研究史をたどり，単語情報がアノテーションされたコーパスが求められていることを確認した．また，2.1.2 節で，資料の電子化と並行して行われてきた計算機による古文の処理，特に形態素解析の先行研究を確認した．今日では現代語の形態素解析で標準となっている統計機械学習にもとづく形態素解析が，古文の形態素解析では試みられていないことを見た．そして，2.1.3 節で，言語研究用に開発された形態素解析用の辞書である UniDic をベースとして機械学習による形態素解析を実現することが，通時コーパスの構築にとって効果的であることを確認した．

[23] 2013 年度公開の「日本語歴史コーパス 平安時代篇」完成版では，短単位と長単位の両方のデータが公開される．

そのうえで，2.2 節で通時コーパスにどのようなテキストが含まれるのかを確認し，その多様性とテキストの校訂処理の必要性を確認した．

最後に 2.3 節において通時コーパスに含まれる様々なテキストに対してどのように形態論情報をアノテーションすべきかを検討した結果，中古和文と近代文語という二つの大きな資料群を対象とした形態素解析を実現することの必要性を確認した．

第3章　形態論情報データベースの構築 [1]

3.1　形態論情報データベースの概要

通時コーパスのテキストに形態論情報のアノテーションを行うためには，データを管理するデータベースとアノテーションを支援するためのツールが必要とされる．すでに，国立国語研究所コーパス開発センターでは，『現代日本語書き言葉均衡コーパス』（BCCWJ）のコーパス構築と管理のために，UniDic 辞書データベースとコーパスとを関連付けて形態論情報の修正を行うためのデータベースシステム「形態論情報データベース」の開発を行い，運用してきた（小木曽・中村 2011 [42]）．

国語研通時コーパスのデータは辞書に UniDic を用いるなど BCCWJ と共通性が大きく，このデータベースシステムで構築作業を行うことが効果的である．新たに通時コーパスの形態論情報付与を行うにあたり，「形態論情報データベース」に改良を加え，通時コーパス用 UniDic の見出し語データやその学習用コーパスの整備を行った．本章では，このシステムの設計・実装について述べる．

3.1.1　データベースの構成

形態論情報データベースは，UniDic の見出し語を管理する部分と，コーパスを格納して修正を行う部分に分かれる．これに対応するように，データベースをインスタンスのレベルで，辞書見出しを格納する「辞書データベース」部と，コーパスを格納する「コーパスデータベース」部に分割した．そのうえで，コーパスの形態論情報と辞書の情報を同一に保つ必要があるため，それぞれのデータベースは中間に見出し語表・活用表・変化表などから生成される「語彙表」を挟んで連係させた（図 3.1）．

語彙表は辞書の見出し語を出現形レベルまで展開したもので，コーパスに出現したすべての語は，原則として語彙表のいずれかのレコードと関連付けられる．

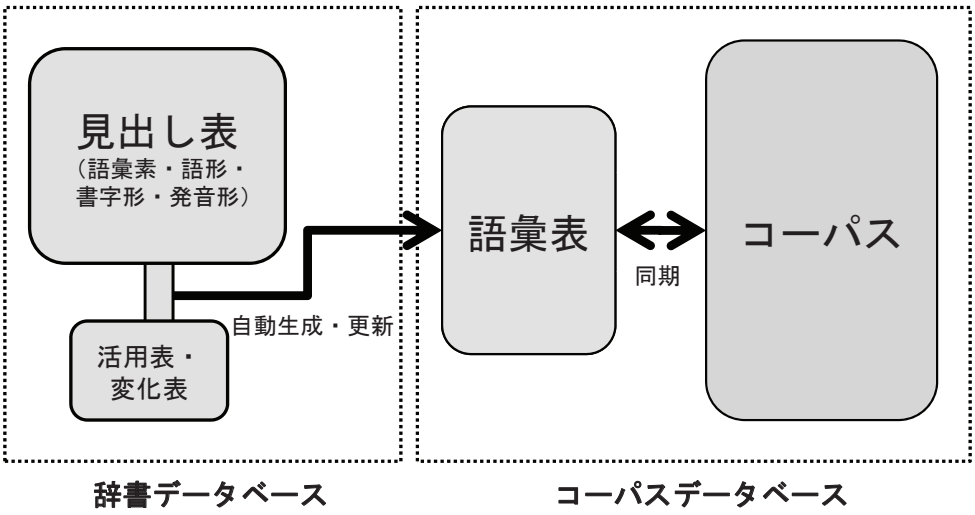


図 3.1: 形態論情報データベース全体図

重要なのは語彙表を介した両コーパスの同期である．辞書見出し語の追加時にリアルタイムで当該見出し語の語彙表レコードを自動生成・更新し，すぐにコーパス修正に利用できるようにしたほか，日次のバッチ処理で同期させる．

辞書データベースとコーパスデータベースは，語彙表を挟んだ疎結合とした．もともと，活用形を展開する必要があるため，見出し語とコーパスを直結することはできず，必ず語彙表を介することになる．また，コーパスと辞書は独立性が高く，それぞれが単独でも利用できる必要があるため，コーパスデータベースにも見出し語がもつ多様な情報を付与しておく必要がある．そのため，仮に両データベース間の結合を密として辞書側の更新をコーパスに直接反映させるとすると，パフォーマンスに大きな影響を与えてしまう．コーパスの規模が大きいため出現頻度が数万以上の見出し語は少なくないが，こうした見出し語の修正時にはコーパスのアップデートに時間を要する上，データのロックが頻発し，作業上深刻な問題となる．したがって，両データベース間は疎結合にとどめ，両者のずれは日次のバッチ処理により解消している．

辞書データベース部の詳細は 3.2 節で，コーパスデータベース部の詳細は 3.3 節で論じる．

3.1.2　利用したシステム

形態論情報データベースは，データベースサーバー（DBMS）に Microsoft SQL Server（2005 Standard Edition）を，クライアントに Microsoft Access で作成したツールを用いるクライアント・サーバ型のシステムとして構築した．開発期間が限られていたため，頻繁な仕様変更が容易に行える Access をクライアントツールの作成に採用し，Access との相互運用性が高くデータベース管理ツールが充実している SQL Sever を DBMS として採用した．

データベースは，BCCWJ や国語研通時コーパスのテキスト電子化で用いられた JIS X 0213 の文字集合（山口ほか 2012）を適切に扱える必要があった．このために SQL Server の規定の照合順序（COLLATE）として，Unicode の CJK 統合漢字拡張漢字 B 集合の文字が扱える Japanese 90 BIN2 を採用した．BCCWJ 構築開始時点で，当該文字集合が適切に扱える DBMS は少なく，このことも SQL Server を採用した理由の一つになっている．通時コーパスの構築においては，大規模文字集合が利用できることは現代語コーパス以上に重要であるが，JIS X 0213 によって十分にカバーすることが可能であった．

サーバ OS には Microsoft Windows Server[2] を採用した．十分なメモリを利用するためいずれも 64 ビット版を利用している．サーバ機のハードウェアのスペックは，メモリ：24.0GB，CPU：Intel Xeon X5355 × 2，HDD：1.0TB（RAID5）15000rpm SAS，外部増設 HDD：12TB（RAID5）である．

クライアントマシンの OS は Microsoft Windows 7 である．クライアントアプリケーションは Access で開発したが，データを全てサーバに置く ADP 形式で開発しており，クライアント側はインターフェイスとなるアプリケーションのみでデータは保持しない．クライアントアプリケーションは，マシンに Access がインストールされていない場合でも利用できるように，無償配布のランタイムと共に配布することを可能にした．

3.2　辞書データベース部の設計・実装

3.2.1　辞書データベースの概要

辞書データベースは，形態素解析辞書 UniDic の元となる見出し語のデータベースである．見出し語のテーブルのほか，活用表などの辞書作成に必要な情報からなる．

[2]バージョンは当初 2003，その後 2008R2 にアップデートした．

2.1.3 で述べたとおり，UniDic は図 2.1 のような階層化された見出し語が設定されている．「語彙素」は国語辞典の見出し語に相当するレベル，「語形」は異語形を区別するレベル，「書字形」は異表記を区別するレベル，「発音形」は発音を区別するレベルである．

辞書データベースの見出し語表は，伝ほか（2007）[24] の基本設計を踏襲し，この UniDic 見出し語階層をそのまま反映させる形で実装した（図 3.2）．辞書データベースの基本となる見出し語表を構成するテーブルは「短単位語彙素」「短単位語形」「短単位書字形」「短単位発音形」の 4 つである．各テーブル間では，レコードの生成や削除に関連するデータベース制約を設定し，不正な見出し語データの発生を防いでいる．

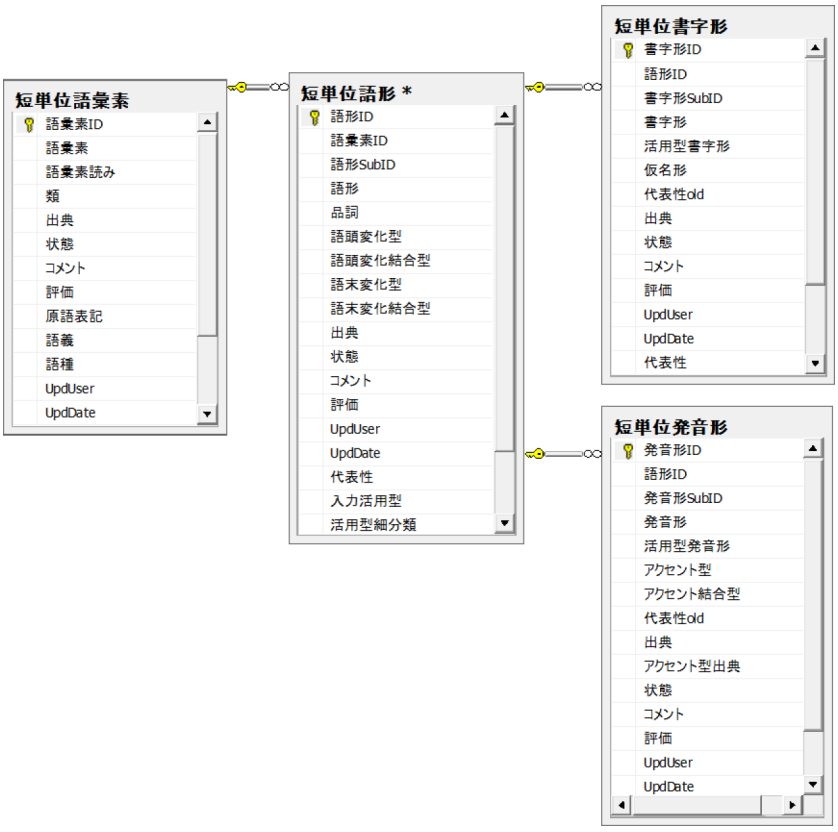


図 3.2: 辞書データベース・見出し語表のテーブル設計（短単位）

各階層の見出し語のテーブルのレコードはユニークな ID によって関連付けられており，各 ID は計算によって表の階層関係が確認できるように設計した．例えば，語形 ID は親となる語彙素の ID に 32（一つの語彙素が持ちうる語形の最大

数）を乗じたものに自身の SubID を加えたものをユニーク ID としている．

各見出しの主要な項目を表 3.1 に示す．

UniDic では，見出し語表の項目のうち，語彙素・語彙素読み・語彙素細分類・品詞・語形・活用型・書字形・発音形（表 3.1 で◎で示した）の組み合わせによって，見出し語がユニークに区別される．この関係もデータベースの外部キー制約として記述し，見出し語の二重登録を防いだ．

表 3.1 の項目に加え，各階層の見出し語のテーブルに表 3.2 の項目を持たせ，各見出し語のメタ情報を記録した．これにより，見出し語を追加・修正した際の作業者やソースのトレースを可能にし，誤った見出し語の追加・修正への対処を可能にしている．また，状態属性によりジャンル別の形態素解析辞書の作成を可能にしている．

辞書データベースでは，見出し語のテーブルのほかに，活用語を展開するための「活用表」テーブル，語頭・語末変化形を展開するための「語頭変化表」テーブル，「語末変化表」テーブルを置いた（後述）．

さらに，見出し語の書字形が含む文字レベルの情報を記述する「書字形構成漢字」テーブル，語彙素のシソーラス情報を記述する「分類語彙表番号」テーブルなどの付加情報テーブルを置き，見出し語とコーパスの研究での活用を可能にしている．この二つのテーブルはコーパス構築に必須のものではなく，網羅的に整備されているわけではないが，研究・応用に必要とされた範囲でデータ入力がなされている．

3.2.2　見出し語の使用年代情報

通時コーパスに対応するため，各見出し語階層において，見出し語に「自至情報」と呼ぶ使用年代の情報を付与した．使用年代は，短単位の規定と，『日本国語大辞典 第 2 版』[86] の初出例に基づいて決定している．

見出し語の使用年代は，語彙素のレベルで決まるものから，書字形のレベルで決まるものまで様々であるため，各レベルで使用年代を付与できるようにしている．ただし，見出し語全体の時代情報の整合性を保つため，小見出しでは原則として親見出しの辞書情報を継承しつつ狭い時代範囲に絞り込むことのみを可能にしている．

語彙素レベルで決まるものとしては，たとえば動詞「いらふ」（＝答える），副詞の「いと」（＝きわめて），形容詞「らうたし」（＝可愛らしい）のような純然たる古語があるが，このほかに語の歴史的変化に合わせて規定によって特に定めたものもある．たとえば連体詞「その」は，中古語以前においては指示詞「そ」と

表 3.1: 見出し語表の主要項目

階層	列名	説明
短単位語彙素テーブル	語彙素 ID	主キー（連番）
	語彙素◎	辞書見出しの代表表記に相当（漢字仮名混じり表記）
	語彙素読み◎	辞書見出しに相当（カタカナ表記）
	語彙素細分類◎	語彙素を語義等によって更に細分する
	類	見出し語の類（体・用・相等）による区別（品詞の上位概念）
	語種	見出し語の出自による区別
短単位語形テーブル	語形 ID	主キー
	語彙素 ID	親の語彙素の ID
	語形 SubID	同一語彙素に関連付けられる語形の連番
	語形◎	異語形を区別するレベルの見出し（カタカナ）
	品詞◎	品詞
	入力活用型◎	活用型 ※活用語の場合は必須
	活用型細分類	活用型の細分類（一部活用型で必須）
	語頭変化型	濁音化などの語頭音変化の種類（型）
	語頭変化結合型	後続要素の語頭変化形への制約の種類（型）
	語末変化型	促音化などの語末音変化の種類（型）
	語末変化結合型	前接要素の語末変化形への制約の種類（型）
	代表性	共通項目
短単位書字形テーブル	書字形 ID	主キー
	語形 ID	親となる語形の ID
	書字形 SubID	同一語形に関連付けられる書字形の連番
	書字形◎	表記を区別するレベルの見出し
	仮名形	書字形をカタカナ表記にしたもの
	活用型書字形	活用による書字形の語尾変化
	代表性	共通項目
短単位発音形テーブル	発音形 ID	主キー
	語形 ID	親となる語形の ID
	発音形 SubID	同一語形に関連付けられる発音形の連番
	発音形◎	発音を区別するレベルの見出し
	アクセント型	アクセント型（アクセント核のある位置）
	アクセント修飾型	活用によるアクセント変化の種類（型）
	アクセント結合型	前接（後続）要素との結合時のアクセント変化の種類（型）
	活用型発音形	活用による発音形の語尾変化
	代表性	共通項目

表 3.2: 見出し語表の共通項目

列名	説明
出典	当該の見出し語のソースとなった資料（新聞，Web 掲示板など）
状態	当該の見出し語の形態素解析辞書での利用状態を示す（ジャンル限定情報，辞書に出力しないなど）
代表性	当該見出し語が同階層において代表性を持つかどうか
コメント	当該の見出し語に関する情報（自由記述）
更新日時	最終更新日時
更新ユーザー名	最終更新を行ったユーザー名

格助詞「の」に分割した方が適当であると考えられるため，連体詞としての使用年代は「中世」以降に限定している．

語形レベルでは，「文語形容詞-シク」や「文語四段」などの文語活用の動詞・形容詞の活用型を「近代」以前に限定した．逆に「形容詞」や「五段」などの口語活用の動詞・形容詞の活用型は，「近世」以降に限定した．

また，語形変化によって新たな異語形が生じたものは，語形レベルでその年代に応じて使用年代を制限している．語形変化が起きた語の例として形容詞「新しい（アタラシイ)」がある．この語は上代においては「アラタシ」であったが，中古以降「アタラシ」に変化した．一方，「アタラシ」という語形は「惜しい」の意味の別語として上代から近代まで使われている．このことを UniDic の見出し語階層では図 3.3 のように表現している（書字形レベルの使用年代情報は省略した．この例では語形の情報をそのまま継承する)．

書字形のレベルで使用年代を制限したものは，仮名遣いや漢字の異体字に関するものである．現代では用いられない旧字形の「讀書（読書)」「辯明（弁明)」「聯絡（連絡)」のようなものは「近代」までとし，逆に「キレる」「イタい」のような現代語特有の表記は「現代」専用としている．

なお，活用語の場合には，一部の活用形だけが限られた時代で用いられることがある．たとえば，口語の形容詞「長い」は語形としては「近世」以降「現代」まで用いられる．この語は表記レベルで「ながひ」「なげへ」などの活用形を持つが，このような現代仮名遣いとも歴史的仮名遣いとも異なる表記形は，近世（および近代）のテキストの解析でのみ必要となるものである．そのため，活用表に使用年代情報を持たせ，この活用形の使用年代を「近世」から「近代」に限定した．「長い」の活用表全体では表 3.3 のようになる．

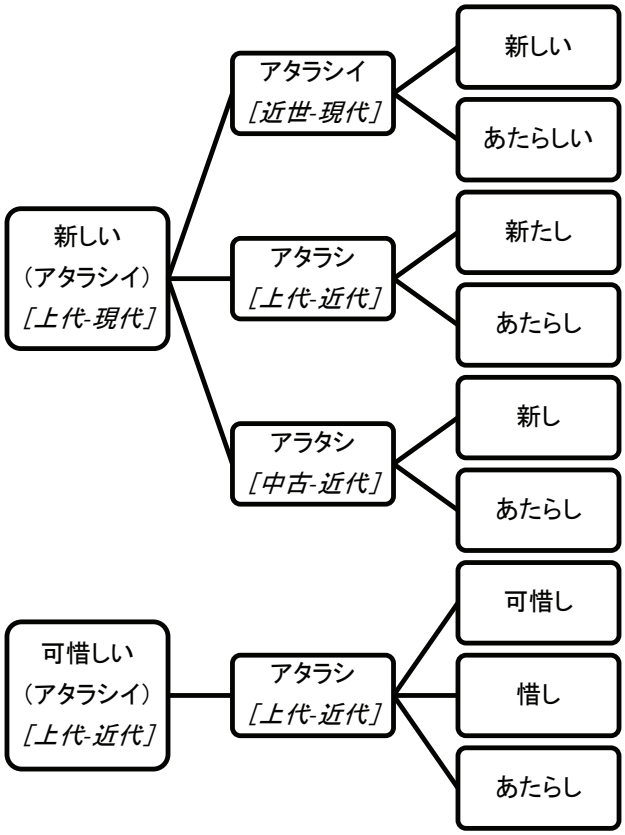


図 3.3: 見出し語の使用年代情報の例

3.2.3 語彙表の展開

辞書データベースには，見出し語表のほかに，活用語を展開するための「活用表」，語頭・語末変化形を展開するための「語頭変化表」「語末変化表」を置き，これによって辞書見出しを出現形にまで展開する．語彙表の展開は，「語頭変化」→「語末変化」→「活用形展開」の順に，見出し語を展開することで行う．以下，語彙表展開について述べる．

語頭・語末変化

語頭変化とは，「語形」が持つ「語頭変化型」に応じて，語形変化による語形を語彙表に生成する処理である．主な対象は連濁現象で，例えば「カメ（亀）」の

表 3.3: 使用年代情報を持つ活用表の例（口語形容詞「長い」）

活用形	活用形細分類	活用語尾	活用語尾書字形	活用語尾発音形	活用語尾仮名形	自	至
意志推量形	一般	ガカロウ	がかろう	ガカロー	ガカロウ	近世	現代
		ガカロウ	がかろふ	ガカロー	ガカロフ	近世	近代
	短縮	ガカロ	がかろ	ガカロ	ガカロ	近世	現代
仮定形-一般		ガケレ	がけれ	ガケレ	ガケレ	近世	現代
仮定形-融合	縮約	ガキャ	がきゃ	ガキャ	ガキャ	近世	現代
	一般	ガケリャ	がけりゃ	ガケリャ	ガケリャ	近世	現代
語幹-一般		ガ	が	ガ	ガ	近世	現代
終止形-一般	エ段	ゲエ	げえ	ゲー	ゲエ	近世	現代
		ゲエ	げへ	ゲー	ゲエ	近世	近代
	一般	ガイ	がい	ガイ	ガイ	近世	現代
		ガイ	がひ	ガイ	ガヒ	近世	近代
連体形-一般	エ段	ゲエ	げえ	ゲー	ゲエ	近世	現代
		ゲエ	げへ	ゲー	ゲエ	近世	近代
	一般	ガイ	がい	ガイ	ガイ	近世	現代
		ガイ	がひ	ガイ	ガヒ	近世	現代
連用形-ウ音便		ゴウ	ごう	ゴー	ゴウ	近世	現代
		ゴウ	ごふ	ゴー	ゴフ	近世	近代
連用形-一般		ガク	がく	ガク	ガク	近世	現代
連用形-促音便		ガカッ	がかっ	ガカッ	ガカッ	近世	現代
		ガカッ	がかツ	ガカッ	ガカツ	近世	近代
		ガカッ	がかつ	ガカッ	ガカッ	近世	近代

「語頭変化型」に「カ濁」を設定すると，語頭変化表により基本形「カメ」と，語頭文字を置き換えた濁音形「ガメ」を生成する．データベース上では，この変形は語形テーブルと語頭変化テーブルを語頭変化型で結合し，各形を生成している．書字形のレベルでは，濁音形の書字形は，漢字表記の場合には基本形と同じものが使われる（例：亀）が，ひらがな・カタカナで表記されている場合には書字形の先頭部分を変化させたもの（例：がめ・ガメ）を生成する．語末変化も語頭変化と同様で，「語形」が持つ「語末変化型」に応じて，語形変化した形を生成する．例えば「サンカク（三角）」の語末変化型に「ク促」を設定すると，語末変化表により，基本形「サンカク」と，語末文字を置き換えた促音形「サンカッ」を生成する．語頭・語末変化の例を図 3.4 に示す．

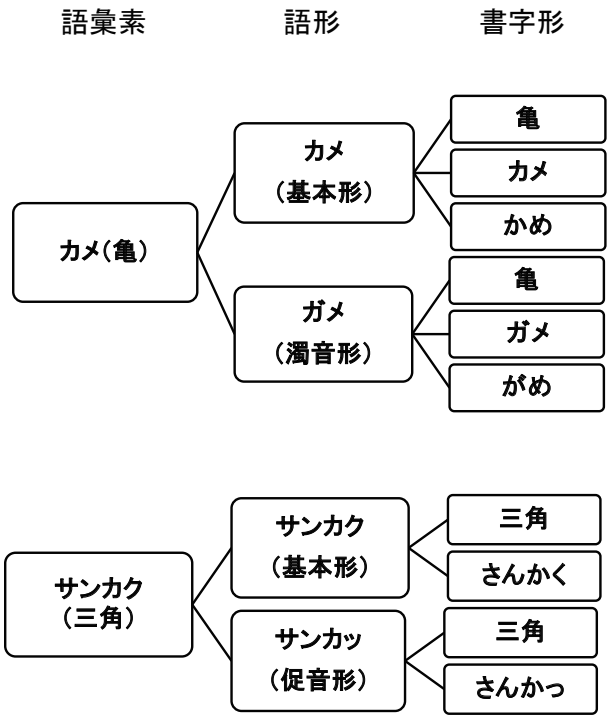


図 3.4: 語頭・語末変化の例

活用

活用は，語形が持つ活用型に応じて，活用形を展開する処理である．活用型の一覧はデータベースの「活用型」テーブルに記述されている．データベース上では短単位語形テーブルと活用表テーブルを活用型によって結合することで各活用形を生成する．

活用に際して，書字形が異なると変化する語尾の部分が異なる場合がある．たとえば，カ行変格活用の動詞「来る」では，仮名で書かれた「くる」の場合，未然形の書字形は「こ」，連用形は「き」だが，漢字で書かれた「来る」では書字形はいずれも「来」である．このように，辞書登録されている書字形ごとに変化させる部分の長さを変える必要があるため，書字形に「活用型書字形」を持たせて活用形の展開の仕方を変えている．

活用語の変化部分の長さの違いは，発音形についても起こる．たとえば，音便形の処理で語形が「オイ」でおわる形容詞は，その直前の音がオ段の場合には終止形などの発音形を長音にする必要がある（「トオイ」→「トーイ」）が，オ段以

外の場合にはその必要がない（「アオイ」→「アオイ」）．このため，発音形に「活用型発音形」を持たせて活用形の展開の仕方を変えている．

なお，通常の活用形の展開では生成できない，または特定の語においてのみ展開する特殊な活用形は，「特殊活用形テーブル」に活用した形の書字形を登録する．たとえば活用語尾までがカタカナ表記される「イイ（良い）」「デキル（出来る）」や，活用語尾のない特殊な表記「也」（助動詞「なり」の終止形），特殊な語形「ま〜す」（助動詞「ます」の終止形）などがそれにあたる．特殊活用形に登録した書字形は語彙表生成時にそのまま語彙表に追加される．

語彙表の展開

語彙表生成時には最終的な出現形のレコード一つ一つに語彙表内で一意となる語彙表 ID を割り当てる．語彙表 ID は通常 10 進数の数値として扱われるが，ビット列としてみると，発音形・語頭変化・語末変化・活用それぞれの展開処理において，各変化形の表現に十分なビット幅をフィールドとして追加したものとなっている．図 3.5 に例として形容詞「辛い」を語彙表展開して生成される出現書字形「がらかっ」の語彙表 ID（10 進数・2 進数）を示す．

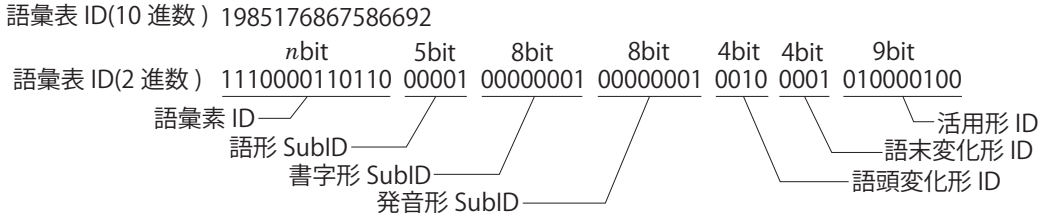


図 3.5: 語彙表 ID の例

この設計により，語彙表 ID のみから，語彙素・語形・書字形等の見出し語 ID や変化形の ID を容易に計算できるようにしている．全体として通常の整数型（32 ビット）で表現できる範囲を超えるため，bigint（64 ビット符号付き整数）型で表現する．したがって，語彙素 ID の最大数は 25 ビット分確保可能である．

図 3.6 に例として形容詞「辛い」の語彙表 ID の生成を含めた語彙表の展開を図示する．

辞書データベースでは，語彙素テーブルの主要項目のほか，語彙素・語形・書字形・発音形テーブルを結合した主要項目，語彙表テーブルにも一意制約が設定されているため，語彙素・語形・書字形・発音形が登録できてもその後語彙表展

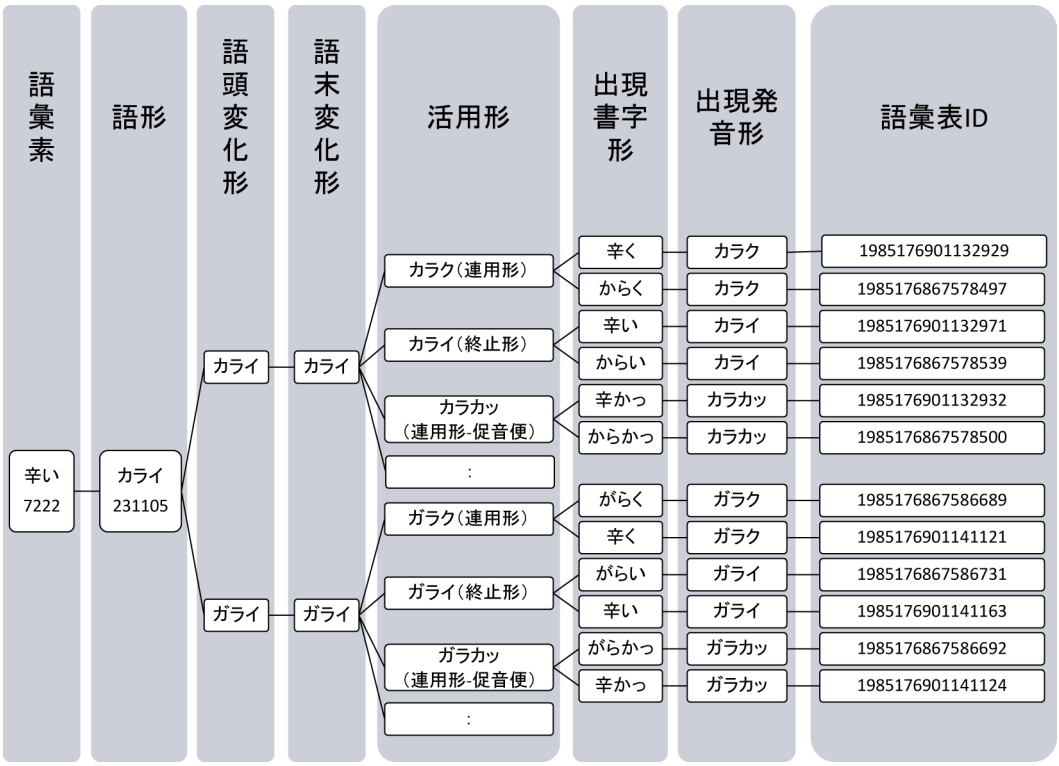


図 3.6: 語彙表展開の例

開したものが重複した場合には，語彙表展開がロールバックされ登録自体も無効となる．つまり語彙表テーブルは常にデータの重複がない状態であることが保証されている．

3.2.4　辞書データのエクスポート

辞書データベースは，形態素解析器（MeCab ）用の辞書（見出し語のリスト）を出力する役割を担っている．この辞書とコーパスデータベースから出力される人手修正済みの学習用コーパスを利用して形態素解析辞書を作成する．UniDic 1.x 系列の形態素解析辞書の作成に当たっては，辞書データベースの見出し語表・活用表・語頭語末変化表を組み合わせて語彙表展開済みの表形式テキストを出力し，これを MeCab のソースデータとして提供した．

さらに，見出し語表を結合して UniDic の階層構造を再現した XML 形式の見出し語表出力を可能にした．同時に活用表や語形変化表も XML 形式で出力する

ことで，辞書データベースの大部分を XML 形式で外部に提供することも可能になった．

3.3　コーパスデータベース部の設計・実装

3.3.1　コーパスデータベースの設計

国語研通時コーパスのテキストは，XML 文書として提供される．したがって，形態論情報のアノテーションは，この XML 文書に対して行う必要がある．また，テキストの形態論情報は，形態素解析等の自動出力結果を人手で修正した後，元の XML 文書に対するアノテーションとして出力する必要がある．

BCCWJ では短単位と長単位という階層的な関係を持つ二つの言語単位によって形態論情報がアノテーションされた（小椋・冨士池 2011 [41]）が，「日本語歴史コーパス 平安時代篇」でも同様に短単位と長単位の二つの単位でのアノテーションを予定している．

関係データベースを用いてこうしたアノテーションを施した XML 文書を扱うために，スタンドオフ・アノテーションの方法論に基づき，XML 文書が含む文字データ（CDATA）とタグをテーブルに分割し，ファイル先頭からの文字オフセット値（開始終了位置）によって関係づけて管理する設計とした．全体の整合性を保持するため，文字やタグを含む全てのデータの修正をこのデータベース上で行う．

コーパスデータベース中のテーブルは，XML 文書起源のものとして「文字」テーブル，「タグ」テーブル，「ルビ」テーブル，「文字修正」テーブルがあり，これに後述する数字処理による「数字」テーブル，形態論情報アノテーションとしての「短単位」テーブルと「長単位」テーブルが加わる．形態論情報も文字位置によってテーブルを関連付けて管理する．このほかに，全文検索用の「文」テーブルや長単位修正作業用の「長単位語彙表」テーブルを置く（図 3.7）．このうち，コーパスデータベースにとって必須のデータは文字テーブルと短単位テーブルであり，XML 文書の復元や長単位アノテーションを必要としない場合にはこれ以外のテーブルは不要となる．

コーパスデータベースと辞書データベースとは，語彙表を介して短単位テーブルが接続する．長単位は定義上，コーパスに出現したものをそのまま単位として認める形をとるため，コーパスから切り離した見出し語表としては管理しない．したがって，形態論情報データベースではコーパスデータベースの中でのみ取り扱い，辞書データベースでは管理しない．長単位語彙表テーブルは一度出現した

長単位を記録してアノテーション作業に利用するための作業用テーブルであり，辞書見出しとしての整備を意図したものではない．

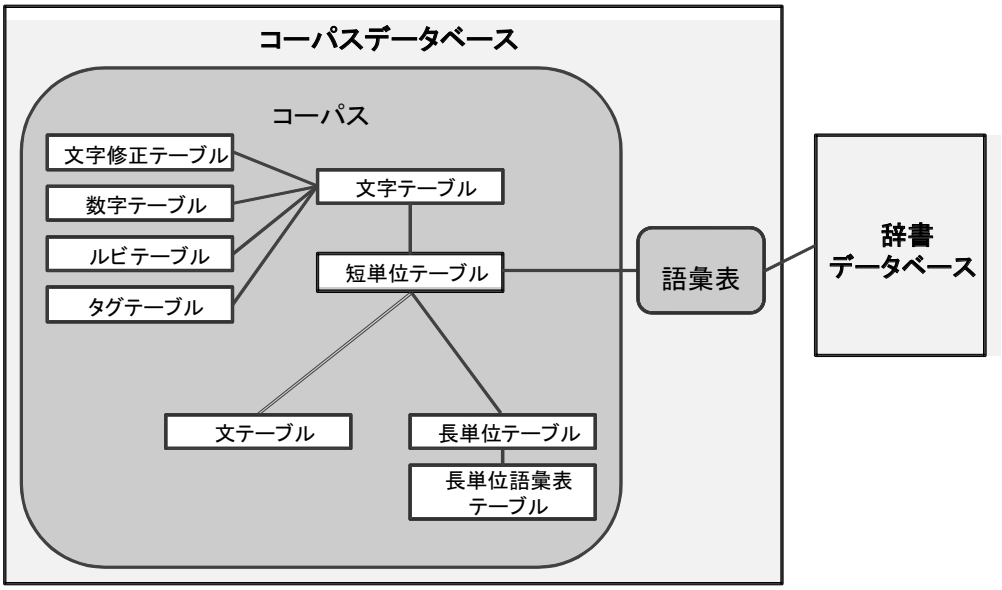


図 3.7: コーパスデータベースのテーブル関連図

コーパスデータベースの根幹である短単位テーブルの主要な項目を表 3.4 に示す．

3.3.2 XML 文書と形態論情報のインポート

コーパスデータベースに XML 形式でリリースされるデータをインポートする方法を図 3.8 のように設計・実装した．既述の通り，XML 形式のデータを表に変換し，それらの表を，文字位置（ファイル先頭からの文字オフセット値）をキーにした ID で相互に関係づける．この際，辞書登録やコーパス修正時に確認することが必要なルビタグ・数字タグ・文字修正タグのみを専用のテーブルに格納して編集可能とし，それ以外のタグについては元の形のまま「タグ表」にまとめて保存している．インポート処理の過程で形態素解析の上で妨げとなるタグの除去などの処理が加わるため，それぞれの表の情報を取り出す段階が異なっている．形態素解析は MeCab と UniDic を用いておこなった．

表 3.4: 短単位テーブルの主要項目

項目	説明	分類
コーパス名	コーパス名（レジスター）	基本となる出典情報
サンプル ID	コーパスのサンプル ID	
連番	サンプル内の並び順	
文境界	文頭（B）またはそれ以外（I）	
文字開始位置	文字テーブルの開始 ID	文字表・その他のテーブルとの接続用
文字終了位置	文字テーブルの終了 ID	
語彙素読み	当該短単位の語彙素読み	出現形をユニークに区別する形態素情報（基本 8 属性）
語彙素	当該短単位の語彙素	
語彙素細分類	当該短単位の語彙素細分類	
品詞	当該短単位の品詞	
活用型	当該短単位の活用型（簡略活用型）	
活用形	当該短単位の活用形（簡略活用形）	
出現書字形	語形変化・活用後の書字形	
出現発音形	語形変化・活用後の発音形	
語彙表 ID	語彙表展開後の語として一意な ID	語彙表との接続用
更新作業者	最終更新ユーザー名	更新情報
更新日時	最終更新日時	

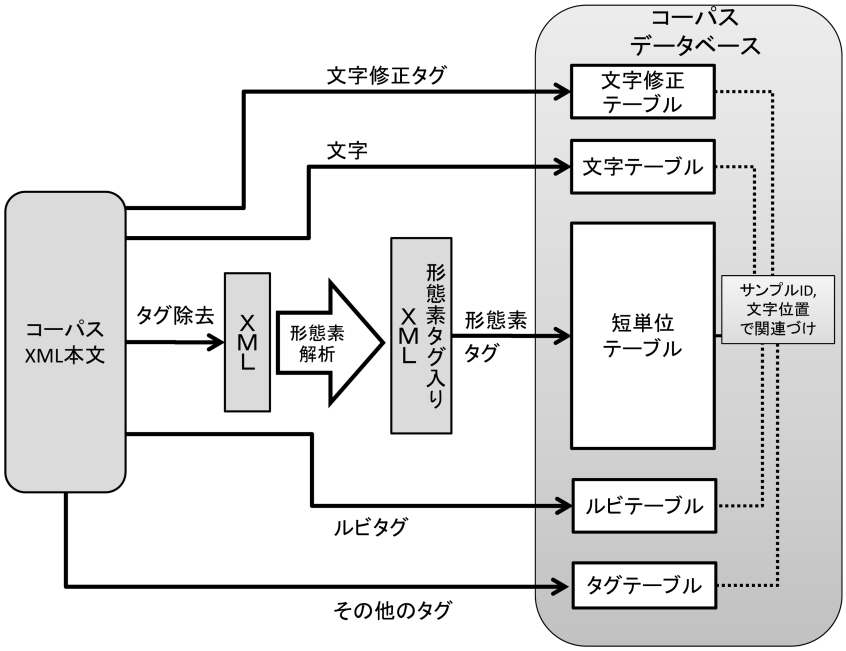


図 3.8: XML 文書の形態素解析とインポートの流れ

長単位のデータは，修正済みの短単位データをコーパスデータベースからエクスポートし，Comainu [43] によって処理を行った後，データベースの長単位テーブルにインポートする．Comainu は短単位を組み上げる形で長単位を生成するため，短単位と長単位の間で齟齬を来すことはない．インポートされた長単位テーブルからは，長単位の修正作業用に長単位語彙表テーブルを作成する．

以上のような手順でコーパスデータベースに格納されたデータは，3.4 節で説明するクライアントアプリケーション「大納言」を通して修正される．

3.3.3　辞書データベースとの関連づけと整合性の確保

辞書データベースとコーパスとの関連付けには，短単位テーブルが保持する語彙表 ID を用いるが，短単位テーブルは表 3.4 に示した属性値も保持している．語彙表 ID と属性値で二重に情報が保持されているため，語彙表 ID での接続が失われた状態でも，表 3.4 の基本 8 属性により，語彙表（及び辞書データベースの見出し語表）との接続を回復できる．これにより，辞書見出しの更新によって語彙

表 ID が変わってしまった場合や，コーパスの修正で一部の属性が一括変更された場合などに，コーパスと語彙表の関連づけが失われた際にも，語彙表 ID か基本 8 属性のいずれかをキーとして同期をとることができるようにした．

語彙表の更新は，辞書データベースの見出し語の追加時・修正時にリアルタイムで該当する語彙表レコードを自動生成・更新する．これにより辞書追加した語をすぐにコーパス修正に利用できるようにしている．また日次のバッチ処理により上述のコーパスと語彙表との同期処理を行い，語彙表 ID と基本 8 属性とのいずれによっても対応がとれないレコードが発生した場合には作業者に修正を促すことで関連付けを保っている．

3.3.4　修正済みコーパスのエクスポート

人手で修正を行った形態論情報は，元の XML 文書にタグとして埋め込んだ XML 形式でエクスポートすることができる．XML エクスポート用の SQL 文では，各テーブルを結合し，データベース内部で XML 型 のデータとして生成した後，ファイル出力している．これによりデータが整形式の XML であることが保証される．テーブルの結合時には，3.3.2 節で示したインポートの流れを逆にたどる．この際，タグテーブルを参照するが，ルビや数字などの別テーブルで管理するタグはタグテーブルからではなく，それぞれのテーブルの情報を元にタグを再構成して出力する．

当然ながら，表形式の形態論情報を出力することも可能であり，後述する Web ベースのコーパス検索アプリケーション「中納言」（6.5 節参照）のソースデータはデータベースから形態論情報を表形式で出力したものがソースとなっている．形態素解析辞書 UniDic の機械学習に用いるコーパスも，コーパスデータベースの短単位テーブルの一部を出力したものである．

3.4　クライアントアプリケーションの開発

3.4.1　辞書データベース用アプリケーション「UniDic Explorer」

辞書管理ツール「UniDic Explorer」は辞書データベースに見出し語を追加・修正するために開発したクライアントツールである．アプリケーション上に UniDic の見出し語表の階層をそのまま可視化しており，階層構造を意識した辞書管理を可能にしている（図 3.9）．

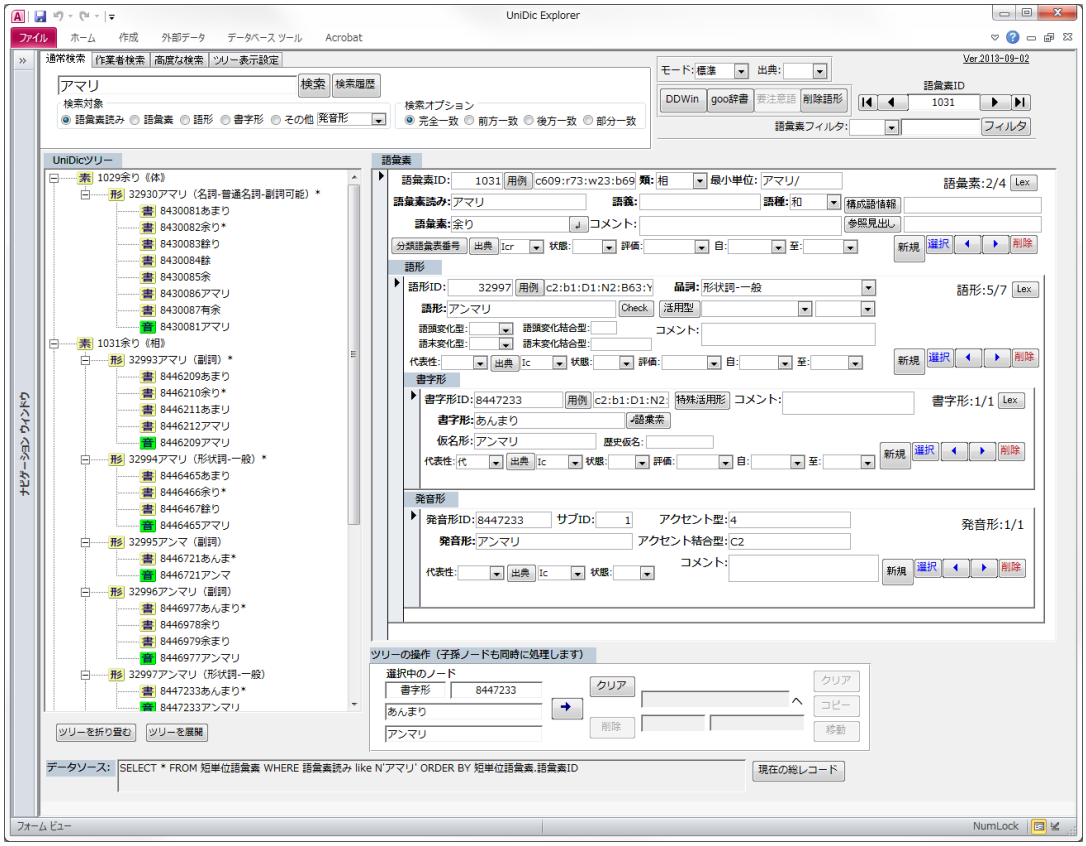


図 3.9: 「UniDic Explorer」実行画面

上段左の検索用コントロールで，各階層の見出し語の情報（語彙素・語彙素読み・語形・書字形・その他）を対象に見出し語表を検索すると，左ペインにマッチした語が UniDic の階層を反映したツリー構造で表示される．右ペインには各階層の見出し語が，階層構造を反映した重層的なフォームの形で表示される．

見出し語の追加は，各見出し階層画面の「新規」ボタンによって行う．見出し語表のデータベース制約により，見出し語は必ず親となる見出し語に追加する形で入力するよう制限されており，逆に見出し語を削除する場合には，その見出し語の子となっている見出し語をあらかじめ削除しておかなければならない．これによって見出し語表の階層構造の整合性を確保している．画面下部の「ツリーの操作」では，見出し語の移動・コピー・削除を行うことができる．この画面では，当該見出し語だけでなく，子や孫となる見出し語ツリー全体をまとめて処理することができる．

見出し語は語彙表を介してコーパスと接続されているため，当該見出し語の実際の用例をこのアプリケーションから確認することができる．当該語のコーパス中の頻度は右ペインの各階層の見出し語の部分に常に表示されている．頻度情報の横の「用例」ボタンを押下することで，当該語のコーパス中の用例を文脈付きで全て表示することができる．

3.4.2　コーパスデータベース用アプリケーション「大納言」

現代語において，短単位の自動解析精度はおおむね98%程度であり，歴史的資料においてはおおむね90%–97%程度である（4章，5章参照）．長単位解析の精度も（短単位データが全て正解であることを前提として）99%ほどであり，人手による修正が必要であった．こうした形態論情報アノテーションの人手修正を行うためのツールが，「大納言」（図3.10）である．「大納言」の中心となる機能は形態論情報の修正であるが，それ以外にも多くの機能を持つため，画面上段のタブによってモードや機能を切り替えて利用する形になっている．

形態論情報アノテーションの修正

多くの修正作業は，形態論情報を使った検索の結果に対して行うことになるが，その検索条件の指定では，「語彙素」「書字形」などの単純な形態論情報の検索だけでなく，形態論情報を前後5グラムまで自由に組み合わせた高度な検索が可能である．また，単位境界を意識しない全文検索を行って，検索結果に形態論情報を表示させることもできる．

短単位アノテーションの修正作業は，短単位の「分割結合」モードで行う．検索結果から修正対象を選択し，当該箇所の短単位境界を文字単位で分割・結合して正しい境界を指定する．境界が直ったところで語彙表を参照して，辞書データベースに登録された語の出現形を当てはめる．この際，該当する短単位がなければ，「UniDic Explorer」で新規の見出しを追加した後，新たに語彙表に追加された出現形を使用する．

長単位の修正時には「長単位」モードで短単位の情報を閲覧しながら，短単位を基本単位として長単位を分割・結合して正しい長単位境界を指定する．長単位境界が直ったところで長単位語彙表を参照して適切な長単位を選択する．この際，該当する長単位がなければ，選択箇所の短単位から自動構成される長単位をもとにして長単位語彙表に新しい語彙を追加してこれを当てはめる．

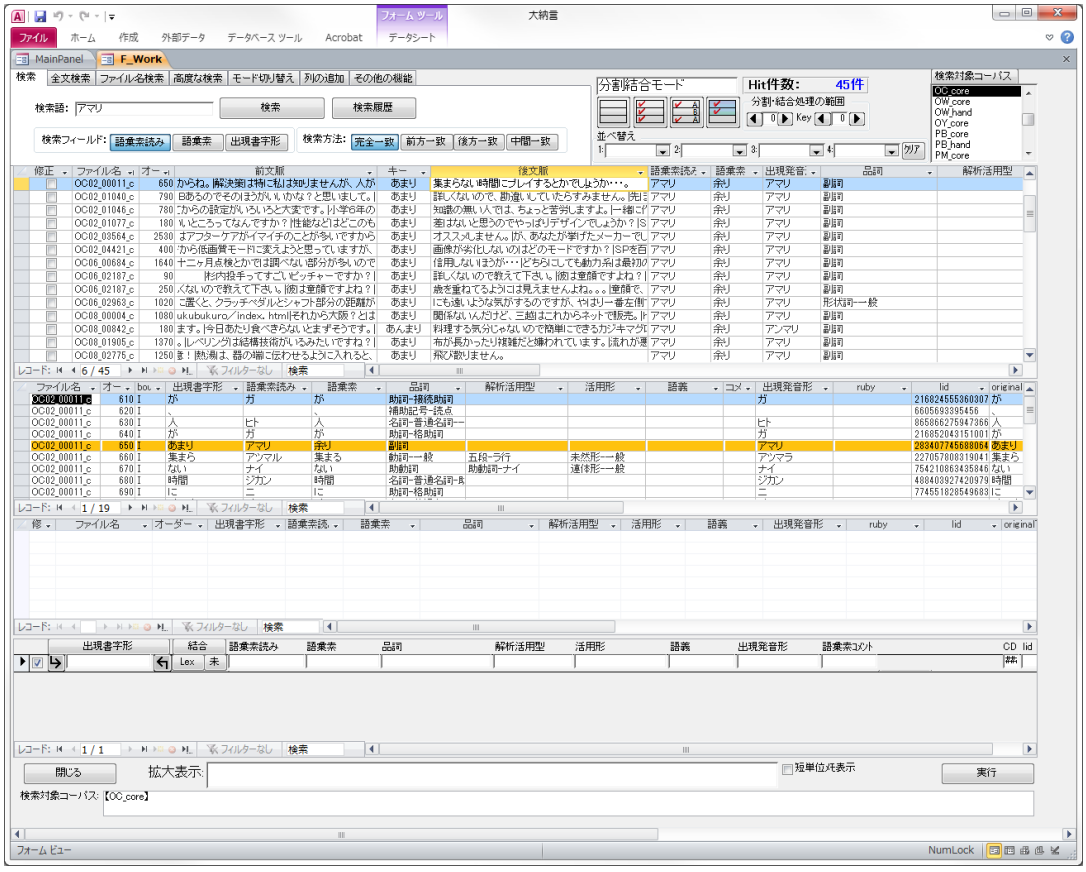


図 3.10: 「大納言」実行画面

こうした形態論情報の修正処理は，修正箇所と同一の形態論情報の組み合わせを持つもの全てを対象にして一括で行ったり，必要なものだけを作業者が選択して一括で行ったりすることが可能で，これによって効率的な修正作業を実現している．

文字とタグの修正

「大納言」では形態論情報の修正作業のほかに，原テキストの文字修正，ルビの文字修正を行うことも可能にしている．3.3 節で示したとおり，コーパスデータベースは文字ベースの開始終了 ID で全体が関連付けられている．そのため，これらのタグの修正時にも ID の整合性を保持し，最終的に XML として出力可能な状態を保たなければならない．「大納言」を通してこれらの修正を行うことで，

作業者が ID を意識することなく全体の整合性を保てるようにした．

コーパス中の文字の修正が必要な場合には，文字テーブルを修正した後，文字修正テーブルに修正内容を記録する．（文字起こしエラーではなく）原文を修正する場合には，タグとして XML 文書に出力する必要があるため，こうした管理を行っている．

このほか，タグテーブルについても，XML 文書を極力整形式に保ったまま，直接修正できる機能を実装している．

3.5 本章のまとめ

以上に述べた「形態論情報データベース」を開発し，歴史的な資料のための改良を加えることで，通時コーパスのテキストに形態論情報アノテーションを施し，コーパス全体に人手による修正を施すことを可能にした．また，本システムによって歴史的な資料を対象とした UniDic の見出し語データの整備を支援し，見出し語データと対応付けられた学習用コーパスを提供したことで通時コーパスのための形態素解析の実現に貢献した．

このデータベースシステムは，現在，国語研通時コーパスの構築に利用されているほか，BCCWJ のタグ修正や新形式のデータ出力などメンテナンス作業の基盤としても活用されている．今後もコーパスの構築を支えるシステムとして活用される予定である．

第4章　「近代文語UniDic」と「中古和文UniDic」の開発 [1]

4.1　中古和文と近代文語文

2.3節で述べたとおり，通時コーパスの形態論情報アノテーションを実現する上で，中古和文と近代文語文の自動形態素解析を実現することが最初の課題となる．

最初に，現代語用の辞書による解析精度を確認した後，辞書への見出し語の追加，学習用のコーパス整備を行って中古和文と近代文語文それぞれの専用の辞書を作成し，精度評価・エラー分析を行う．

4.2　現代語用のUniDicによる古文の解析精度：ベースライン

一般に公開されている現代語用の形態素解析辞書はこれによって古文を解析することは考慮されていない．古文では，特に助動詞などの機能語の用いられ方が大きく異なるため，現代語用の辞書で古文を解析することは困難であると考えられる．現代語用に作られたUniDicも同様であるが，新たに作成する古文用辞書と比較するためのベースラインとして，手始めに現代語用のUniDicで近代文語文と中古和文を解析した場合の精度を調査する．

評価に用いる現代語用のUniDicは一般公開されているunidic-mecab-2.1.0である．UniDicの学習にはコーパスとしてBCCWJのコアデータのほか日本語話し言葉コーパスとRWCPコーパスの一部が用いられており，学習素性としては語彙素・語彙素読み・語種・品詞・活用型・活用形・書字形が用いられている[1]．UniDicの見出し語は，語彙素・語形・書字形・発音形の階層構造を持っている．そのため，解析結果の精度評価もこの階層ごとに行った．最も基礎的なレベルとして，単語境界の認定が正しく行われているかを見る「Lv.1境界」を設定した．またこれに加えて品詞・活用型・活用形の認定が正しいかどうかを見る「Lv.2品

詞」，Lv.1・Lv.2 に加えて語彙素（辞書見出し）としての認定も正しかったかどうかを見る「Lv.3 語彙素」を設定した．Lv.3 は，たとえば「金」が「キン」でなく「カネ」と正しく解析されているかどうかを見ることになる．さらに，Lv.1〜Lv.3 が正しいことに加え，読み方が正しいかどうかを見る「Lv.4 発音形」を設定した．Lv.4 は，古文の場合には発音というよりは語形の違いが正しく認定されているかどうかを評価するものである．たとえば，「所」が文脈にあわせて「トコロ」ではなく「ドコロ」と正しく解析されているかどうかを見ることになる．

ところで，現代語の辞書で古文を解析した場合には，活用型が文語であるか口語であるかの違いによって誤りとされる例が多い．たとえば動詞「書く」は口語ではカ行五段活用（「五段-カ行」）だが，文語ではカ行四段活用（「文語四段-カ行」）で定義されているため両者が一致しないと品詞レベルで誤りと見なされる．しかし両者は本質的には同語であるといってよく，相互に容易に変換することができる．そこで，こうした口語・文語の活用型の対についてはいずれを出力した場合にも正解と見なした場合の精度についても調査した．具体的には，「文語形容詞-ク」と「文語形容詞-シク」を「形容詞」と同一視し，「文語四段」は「五段」，「文語サ行変格」は「サ行変格」，「文語下二段」は「下一段」，「文語上二段」は「上一段」と同一視した．さらに文語の「-ハ行」「-ワ行」を「-ア行」と同一視した．

以上の観点でまとめた解析精度の調査結果を表 4.1 に示す．評価コーパスは，後述する人手修正済みのコーパスから約 10 万語を文単位でランダムサンプリングしたものである．評価項目は次に示す Precision（精度），Recall（再現率），F 値（Precison と Recall の調和平均）である．

$$\text{Precision} = \frac{\text{正解語数}}{\text{システムの出力語数}}$$

$$\text{Recall} = \frac{\text{正解語数}}{\text{評価コーパスの語数}}$$

$$\text{F 値} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

活用型の補正なしの場合の語彙素認定の F 値でみると，近代語では 0.6775，中古和文では 0.5432 となっており，予想通り通時コーパス構築の実用に耐える精度ではない．中古和文と比べ近代文語の方が比較的精度が良いが，これは現代語との年代差が少ないため使用される語彙が近いことによるものと考えられる．補正後は，近代語では 0.7323，中古和文では 0.5939 となっている．補正による上昇は

表 4.1: 現代語用の UniDic による近代文語・中古和文の解析精度

評価レベル	評価項目	近代文語		中古和文	
		補正なし	補正あり	補正なし	補正あり
Lv.1 境界	Recall	0.9154	—	0.8471	—
	Precision	0.8757	—	0.7920	—
	F 値	0.8951	—	0.8186	—
Lv.2 品詞	Recall	0.7198	0.7804	0.6018	0.6624
	Precision	0.6886	0.7466	0.5627	0.6193
	F 値	0.7038	0.7631	0.5816	0.6401
Lv.3 語彙素	Recall	0.6928	0.7489	0.5621	0.6145
	Precision	0.6628	0.7164	0.5256	0.5746
	F 値	0.6775	0.7323	0.5432	0.5939
Lv.4 発音形	Recall	0.6882	0.7440	0.5560	0.6083
	Precision	0.6584	0.7117	0.5199	0.5688
	F 値	0.6729	0.7275	0.5374	0.5879

※ 活用型変換による補正の有無による精度の比較を含む

0.05 ポイント程度であり，単純な活用型変換を行ってもさほど精度は向上しないことが分かる．

このように古文の形態素解析のためには辞書への古文の見出し語追加が必須であり，また古文のコーパスで再学習を行うことで解析精度の向上が期待できる．以下，4.3 節で見出し語の追加と学習用コーパスの構築について説明する．4.4 節では見出し語追加と再学習を行った提案手法による解析精度を他の手法と比較して確認し，その後この辞書による各種テキストの解析精度について議論する．

4.3 見出し語の追加と学習用コーパスの作成

現代語用の UniDic をもとにした，最初の古文用形態素解析辞書として「近代文語 UniDic」を開発，公開した（小木曽・小椋・近藤 2008）．これは主として近代の文語論説文（明治普通文）を対象とした解析辞書であり，文語の活用・旧仮名遣い・旧漢字などに対応し，文語文を正しく解析することが可能になっている．解析精度は，現代語版の UniDic には及ばないものの，おおむね 96%以上を達成している．これにより，「太陽コーパス」（国立国語研究所 2005）の文語記事な

ど，近代文語文で書かれたテキストを解析して研究に利用することができるようになった．

古文用の形態素解析を行うために，現代語用の辞書に見出し語の追加を行った．UniDic では見出し語を語彙素・語形・書字形・発音形の 4 段階で階層的に管理しているため，近代語解析に必要な語を各階層に整理して追加することができる．現代語としては使われなくなっている語は「語彙素」のレベルで，文語活用型の語は「語形」のレベルで，異体字や旧字形など表記の違いは「書字形」のレベルで追加することになる（2.1.3 節）．

これまでに近代文語文のために追加した語彙は，語彙素レベルで 10,814 語，語形レベルで 12,417 語，書字形レベルで 25,224 語であった．また，中古和文のために追加した語彙は，語彙素レベルで 5,939 語，語形レベルで 7,351 語，書字形レベルで 13,763 語であった．両者には共通の語彙も多いが近代文語文の語彙追加を先に行ったため，中古和文のための追加数が少なくなっている．もともとあった現代語の見出し語とあわせ，全体の見出し語数は，語彙素 225,588 語，語形 253,061 語，書字形 413,897 語となっている．

見出し語の追加は，現代語形から派生させた文語形や旧字形を追加するところからはじめ，既存の古語辞典やデータ集の見出し語からも追加を行った．しかし，形態素解析辞書の見出し語としては，UniDic 体系に基づく詳細な品詞を付与し，実際に出現する表記形を入力する必要があるため，単なる辞典の見出し語リストは多くの場合，登録用のソースとして不十分である．そのため，大部分の語彙は，後述する学習用コーパスを整備する過程で不足するものを追加する形で行った．

見出し語の単位認定については，通時的な比較ができるようにするため，可能な限り現代語と共通の枠組みで処理を行った．しかし，語の歴史的変化や古文における使用実態を踏まえ，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，指示詞について，現代語の UniDic では「この」「その」などは一語の連体詞として扱っているが，「こ」「そ」が単独で指示代名詞として使われる中古語では，これらは代名詞＋格助詞として扱った方が適切である．このように，歴史的資料向けの辞書見出しの追加は単純な作業ではなく，通時的な共通性に配慮しつつ，各時代の言語の実態を反映させるかたちで見出し語を認定するという高度な判断にもとづくものである．その積み重ねによって作られた見出し語リストは，日本語の通時的な処理を行う上で基礎となる重要なデータであると言える．

また，中古和文 UniDic の見出し語認定基準は規程集としてまとめ公開している [40][2] が，これは通時的な比較を考慮した歴史的資料の処理にとって利用価値が高い資料である．中古和文用の規定はそのまま他の時代の資料に適用できるわ

[2]http://www2.ninjal.ac.jp/lrc/index.php?UniDic

けではないが，日本語の古典文法は中古和文を基準として作られたものであるため，この規定を中核として追加・修正を行うことで各時代向けの辞書を作成していくことが可能である．

語彙の追加と平行して活用表の整備も行った．UniDic はもともと文語の活用表を一部備えていたが，これを整備して網羅的なものとするとともに，通時コーパス構築に必要な活用形の追加を行った．

UniDic は現代語用に整備されてきたため，古文では用いられない語彙を多く含んでいる．しかし，基礎語彙の多くは中古和文でも共通であり，どの語が不要であるかを事前に判断することは必ずしも容易ではない．また，古文の形態素解析辞書にとって見出し語の肥大化は大きな問題ではない上，不要語があることによる解析精度への悪影響は認められなかったため，現代語用の見出しも原則としてそのままとした．同様の理由から，近代文語 UniDic と中古和文 UniDic の間でも同一の語彙表を用いている．

以上のような見出し語の追加だけであれば，学習用コーパスの作成に比べて低コストで行うことができる．そこで，見出し語の追加だけで十分な精度向上が見られるのかどうか確認するため，古文用の見出し語を追加した現代語用の UniDic を作成し，4.2 節の表 4.1 と同様に解析精度を調査した．その結果を表 4.2 に示す．見出し語の追加によって，近代文語では，語彙素認定の F 値で約 0.06 ポイント向上し 0.7363，補正後の数字で 0.7970 となっている．また，精度の低かった中古和文では，語彙素認定の F 値で約 0.07 ポイント向上し 0.6190，補正後の数字で 0.6640 となっている．しかし，やはりこの精度は不十分であり，見出し語の追加だけでは通時コーパスの構築にとって十分なだけの精度は得られなかった．

4.3.1 学習用コーパスの準備

前節で確認したように，古文の形態素解析のためには，見出し語の追加だけでなく学習用コーパスの整備が必要となる．

近代文語では，主たる解析対象の明治期の文語論説文を中心に，表 4.3 の約 64 万語の人手修正済みのコーパスを作成した．近代詩・小説・法令・論説文の大部分は「青空文庫」[3] 所収のテキストを利用し，論説文としては他に上田修一氏作成の「文明論之概略」テキストデータ [4] を利用した．また，雑誌の本文は国立国語研究所の『太陽コーパス』の一部のテキストを利用した．以上のテキストに対して独自に UniDic ベースの形態論情報をタグ付けしたデータに加え，国立国語

[3] http://www.aozora.gr.jp/

[4] http://web.keio.jp/~uedas/bunmei.html

表 4.2: 古文用の見出し語を追加した現代語用の UniDic による解析精度

評価	評価	近代文語		中古和文	
レベル	項目	補正なし	補正あり	補正なし	補正あり
Lv.1	Recall	0.9565	—	0.9103	—
境界	Precision	0.9637	—	0.9154	—
	F 値	0.9601	—	0.9129	—
Lv.2	Recall	0.7655	0.8313	0.6600	0.7149
品詞	Precision	0.7713	0.8375	0.6637	0.7189
	F 値	0.7684	0.8344	0.6619	0.7169
Lv.3	Recall	0.7336	0.7941	0.6173	0.6621
語彙素	Precision	0.7391	0.8000	0.6207	0.6658
	F 値	0.7363	0.7970	0.6190	0.6640
Lv.4	Recall	0.7288	0.7891	0.6102	0.6550
発音形	Precision	0.7343	0.7950	0.6136	0.6586
	F 値	0.7316	0.7920	0.6119	0.6568

※ 活用型変換による補正の有無による精度の比較を含む

研究所で公開された形態論情報付き『明六雑誌コーパス』[31] を学習用のデータとして利用した.

表 4.3: 近代文語の学習・評価用コーパス

ジャンル	語数
近代詩	58,375
小説	39,293
法令	30,868
論説文	233,145
『太陽』	100,873
『明六雑誌』	180,598
総計	643,152

中古和文では，2012 年に国立国語研究所によって公開された「日本語歴史コーパス 平安時代編 先行公開版」[5] のデータを学習に利用した．このコーパスは，表

[5]http://www.ninjal.ac.jp/corpus_center/chj/

4.4 に示す平安時代の仮名文学作品を中心とした約 82 万語の人手修正済みの形態論情報を含んでいる．

表 4.4: 中古和文の学習・評価用コーパス

作品名	語数
伊勢物語	15,894
源氏物語	510,600
古今和歌集	32,256
更級日記	16,774
讃岐典侍日記	18,541
紫式部日記	20,707
大和物語	26,742
竹取物語	12,584
土佐日記	8,129
枕草子	79,850
落窪物語	68,564
和泉式部日記	12,633
総計	823,274

学習用のコーパスは段落や改行，振り仮名などがタグ付けされたテキストに形態論情報を付与したもので，これを 3 章で述べた「形態論情報データベース」上で辞書データと紐付けて管理している．近代語コーパスでは，さらに濁点が付されていない部分に濁点を付与するなど表記上の問題に対処するためのタグ付けを行い，また文末を認定してセンテンスタグを付与している．日本語の歴史的資料のコーパスを形態素解析辞書の機械学習に利用可能な形で整備したのはこれが初めての試みである．

近代文語文と中古和文は，同じ古文と言っても大きく性質の異なる文体である．近代文語文では低頻度語が多く，上記のコーパス中，書字形（基本形）で集計した場合，頻度 1 の語が 18,120 語，頻度 2 の語が 5,943 語含まれていた．一方，中古和文では，頻度 1 の語が 6,198 語，頻度 2 の語が 2,134 語にすぎない．近代文語文は，明治維新後に西洋の新たな文物を吸収していく時代において，新しい書き言葉を確立していく過程にあった文体であるため，新たに作られながらも定着しなかった語などの低頻度語が目立つ．また，書かれる内容が多様で文体差が大きい．一方，中古和文は，古代の宮廷における限られたコミュニティの中で当時の

話し言葉に基づいて書かれた文章であり，内容的な幅も狭いため，語彙の広がりが小さい．こうした違いは，形態素解析の精度にも影響を及ぼしてくると考えられる．

4.3.2 MeCab を用いたコーパスからのパラメータ学習

MeCab でコーパスからのパラメータを学習する場合，設定ファイル rewrite.def と feature.def によって，学習に用いる内部素性のマッピングと，内部素性から CRF 素性を抽出するためのテンプレートを書き換えることができる．近代文語 UniDic と中古和文 UniDic の学習にあたっては，CRF 素性を抽出するテンプレート（feature.def）は，どちらの辞書でも現代語用のものをそのまま用いている（付録 C.2 参照）．利用している素性は，語彙素・語彙素読み・語種・品詞（大分類・中分類・小分類）・活用型・活用形・書字形（基本形と出現形）とその組み合わせである．UniDic では語種を学習素性として利用しているのが特徴となっており [1]，この素性は古文の解析にも大きく寄与している．

一方，内部素性のマッピング（rewrite.def）は，現代語用の UniDic のものを修正して，語彙化する見出し語を文語の助動詞・接辞に置き換えた（付録 C.1 参照）．たとえば，次の助動詞については品詞ではなく，語彙のレベルで連接コストを計算している．

> き，けむ，けらし，けり，こす，ごとし，ざます，ざんす，じ，ず，
> たり，つ，なり，ぬ，べし，べらなり，まし，まじ，む，むず，めり，
> らし，らむ，り，んす，んなり，んめり，非ず

助詞・助動詞などの語彙化すべき見出し語については近代文語文と中古和文とで共通する部分が多いため，rewrite.def は近代文語 UniDic と中古和文 UniDic とで共通のものを利用した．

4.4 解析精度の評価

4.4.1 解析精度

上述の方法で作成した「近代文語 UniDic」「中古和文 UniDic」の解析精度を調査した．評価コーパスは，4.3.1 節で示した人手による修正済みのコーパスから文単位でランダムサンプリングした 10 万語分とし，その残りを訓練コーパスとし

た．評価コーパスは，辞書ごとに固定して，以後の精度評価でも同一のものを用いている．4.2 節での調査と同様に，単位境界・品詞・語彙素・発音形の 4 つのレベルで調査を行った結果を表 4.5 に示す．なお，評価コーパスはもともと学習用に整備したものの一部を転用したものであるため，当初含まれていた未知語は辞書に登録されている．そのため，解釈不能語などを除いて原則として未知語を含んでいない．

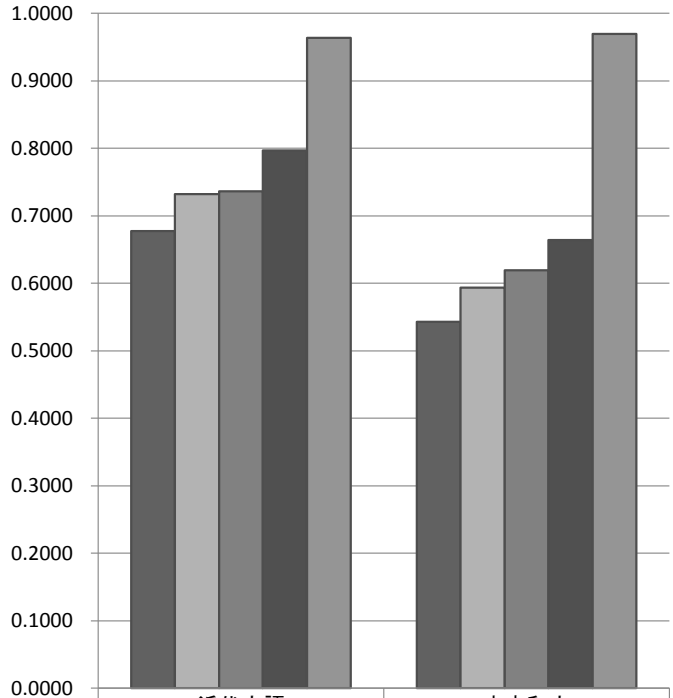
表 4.5: 「近代文語 UniDic」「中古和文 UniDic」の解析精度

評価レベル	評価項目	近代文語	中古和文
	評価語数	100,016	100,074
	出力語数	99,909	100,078
Lv.1 境界	正解数	99,104	99,428
	Recall	0.9909	0.9935
	Precision	0.9919	0.9935
	F 値	0.9914	0.9935
Lv.2 品詞	正解数	97,071	97,919
	Recall	0.9706	0.9784
	Precision	0.9716	0.9785
	F 値	0.9711	0.9784
Lv.3 語彙素	正解数	96,376	97,076
	Recall	0.9636	0.9700
	Precision	0.9646	0.9700
	F 値	0.9641	0.9700
Lv.4 発音形	正解数	96,004	96,834
	Recall	0.9599	0.9676
	Precision	0.9609	0.9676
	F 値	0.9604	0.9676

語彙素認定の F 値で，近代文語は 0.9641，中古和文では 0.9700 となっており，ベースライン（表 4.1）や見出し語のみを追加した場合（表 4.2）と比較して大幅に精度が向上している．近代文語 UniDic と中古和文 UniDic を比較すると，すべてのレベルで中古和文の方が良い精度となっている．これには中古和文の方が訓練コーパスの量が多いことも影響しているが，それよりも 4.3.1 節で見たようなテキストの性質の違いによる影響が大きいと考えられる（4.4.5 節参照）．

図 4.1 に，再学習を行った提案手法と，4.2・4.4 節で確認した各種手法（現代語

辞書によるベースライン，再学習を伴わず見出し語だけを追加した辞書）の解析精度を比較した結果を示す．精度は語彙素認定の F 値である．図中の「補正」とは活用型の文語形への変換を行った場合の精度である．



	近代文語	中古和文
ベースライン	0.6775	0.5432
ベースライン（補正）	0.7323	0.5939
見出し語追加	0.7363	0.6190
見出し語追加（補正）	0.7970	0.6640
見出し語追加＋再学習	0.9641	0.9700

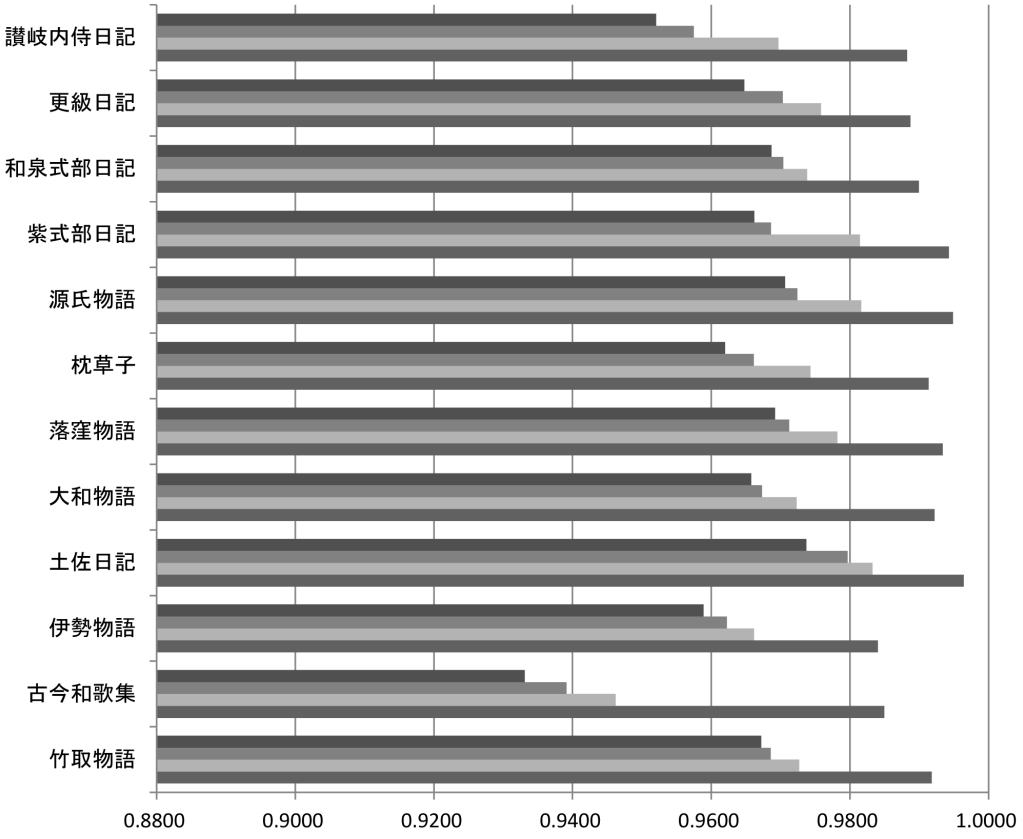
図 4.1: 各種方法による解析精度の比較（語彙素レベル・F 値）

このように，コーパスによる再学習によって初めて実用的な精度での解析が可能になる．BCCWJ の構築に利用された現代語の UniDic の解析精度が語彙素認定の F 値で約 0.98 であり，ジャンルによっては 0.96 程度に留まることと比較しても，コーパス構築に利用するために十分な精度が出ていると言える．

4.4.2　作品・ジャンル別の解析精度

中古の文学作品別の解析精度

4.4 節でみた中古和文の解析精度を，作品別にまとめ直した結果を図 4.2 に示す．表 4.1・表 4.2 と同一の評価用データを対象に，作品別に集計している．



	竹取物語	古今和歌集	伊勢物語	土佐日記	大和物語	落窪物語	枕草子	源氏物語	紫式部日記	和泉式部日記	更級日記	讃岐内侍日記
■発音形	0.9672	0.9331	0.9589	0.9737	0.9658	0.9692	0.9620	0.9706	0.9662	0.9687	0.9648	0.9521
■語彙素	0.9686	0.9391	0.9622	0.9797	0.9673	0.9712	0.9661	0.9724	0.9686	0.9704	0.9703	0.9575
■品詞	0.9727	0.9462	0.9662	0.9833	0.9723	0.9782	0.9743	0.9816	0.9814	0.9738	0.9758	0.9697
■境界	0.9918	0.9849	0.9841	0.9964	0.9922	0.9934	0.9914	0.9949	0.9943	0.9899	0.9887	0.9882

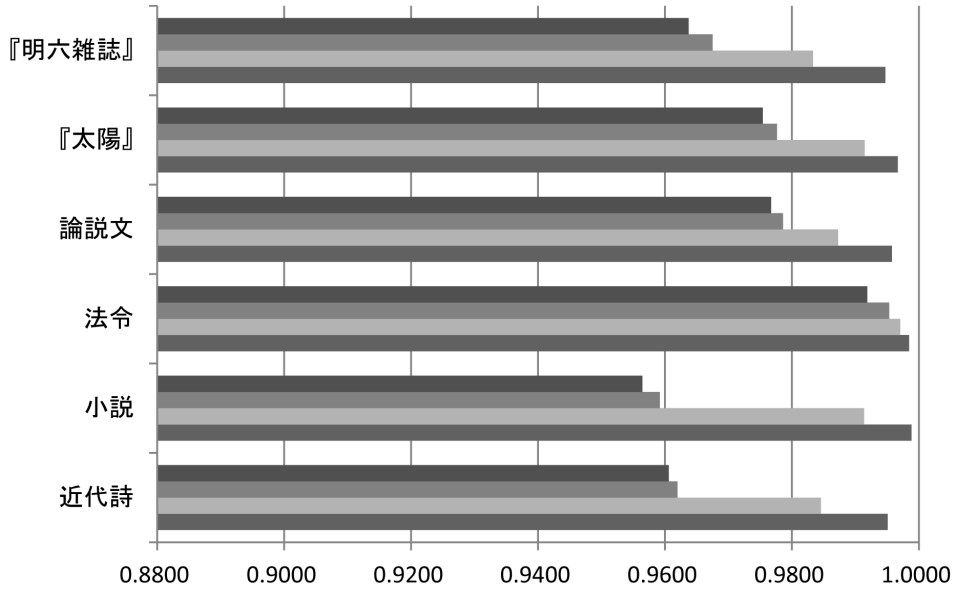
図 4.2: 中古の文学作品別の解析精度

『土佐日記』の解析精度が最も高く語彙素認定で 0.9797，『古今和歌集』が最も低く 0.9391 であった．『土佐日記』は，簡潔な文で明快な記述が多く，平均文

長を見ると全 12 作品中最も短い（平均文長は『源氏物語』30.7 語に対し『土佐日記』では 14.5 語）．こうしたことが高い解析精度につながっていると考えられる．一方，『古今和歌集』はこれだけが和歌集であり他の散文とは異質であるうえ，コーパスには詞書や人名などが多く含まれる（和歌本体より多い）ため，一般的な中古和文とは違いが大きく，解析精度の低下につながっている．『古今和歌集』以外では，最も成立年代が下る『讃岐内侍日記』で精度がやや低い．学習用コーパスの大部分を『源氏物語』が占めるため，全体として『源氏物語』に文体（語彙・語法）が近い作品は高い精度で解析ができる傾向がある．

近代のジャンル別の解析精度

同様に，近代文語文の解析精度を，ジャンル・媒体別にまとめ直した結果を図 4.3 に示す．表 4.3 に示したジャンル・媒体ごとに，評価用データ約 1 割を文単位でランダムサンプリングして集計したものである．



	近代詩	小説	法令	論説文	『太陽』	『明六雑誌』
■発音形	0.9606	0.9564	0.9919	0.9767	0.9754	0.9637
■語彙素	0.9620	0.9592	0.9953	0.9786	0.9777	0.9675
■品詞	0.9846	0.9914	0.9971	0.9873	0.9914	0.9833
■境界	0.9951	0.9988	0.9984	0.9957	0.9967	0.9947

図 4.3: 近代のジャンル別の解析精度

語彙素認定で法令が 0.9953 と突出して精度が高い．これは，法律の文体が極めて人工的に整えられたものであり，語彙が限定され，固定的な言い回しが繰り返されることによるものである．コーパスの大部分を占める『太陽』と論説文は，それぞれ 0.9777，0.9786 と高い精度で解析ができている．『明六雑誌』も 0.9675 とこれに次ぐ．一方，近代詩は 0.9620，小説は 0.9592 であり，文学作品は他のジャンルに比較して精度が低い．これは，多くの語彙が含まれ，多彩な表現が用いられていることによるものであると考えられる．

4.4.3 未知語を考慮した解析精度

表 4.5 の精度は，基本的に未知語が存在しない状態のコーパスを評価対象とした場合のものであった．しかし，実際の解析対象には未知語が含まれているのが通常である．そこで，未知語を含んだテキストを解析した際の精度を検証した．

評価コーパスには同一のものを用いて，評価コーパスのみに現れ，それ以外の人手修正済みコーパス（＝学習用コーパス）には一度も出現しない語を，近代文語 UniDic・中古和文 UniDic それぞれの辞書から削除して未知語を発生させた．削除した語数は，近代文語 UniDic では 2,089 語（評価コーパス中の出現回数 3,128），中古和文 UniDic では 795 語（評価コーパス中の出現回数 824）であった．4.3.1 節で見たように，近代文語文には低頻度の語が多く含まれるため，中古和文に比べて未知語が多く発生することになる．この条件で作成し直した辞書の解析精度を表 4.6 に示す．

Precision が特に低下しており，語彙素認定の F 値で見ると，近代文語文では 0.0376 ポイント低下して 0.9265，中古和文では 0.0089 ポイント低下して 0.9611 となっている．未知語が多い近代文語文では影響が大きいが，それでも十分に実用的な精度が得られている．

4.4.4 未知の資料の解析精度

前節で見たように学習用コーパスと同一の作品では十分な精度が得られたが，学習用コーパスとは完全に無関係な資料の解析精度を確認する必要がある．未知の資料には，当該辞書の適用対象といえるテキストと，文体差があり必ずしも適切な対象であるとはいえないテキストがある．ここでは，中古和文 UniDic を例に，その対象内の文体で書かれたテキストであるといえる擬古物語『恋路ゆかしき大将』と，時代的には中古和文に近いが和漢混淆文と呼ばれる別種の文体である『今昔物語集』の一部の解析精度を調査する．調査対象のテキストはいずれも

表 4.6: 未知語の有無による解析精度比較

評価レベル	評価項目	近代文語		中古和文	
		未知語なし	未知語あり	未知語なし	未知語あり
Lv.1 境界	Recall	0.9909	0.9685	0.9935	0.9886
	Precision	0.9919	0.9497	0.9935	0.9834
	F 値	0.9914	0.9590	0.9935	0.9860
Lv.2 品詞	Recall	0.9706	0.9433	0.9784	0.9725
	Precision	0.9716	0.9250	0.9785	0.9673
	F 値	0.9711	0.9340	0.9784	0.9699
Lv.3 語彙素	Recall	0.9636	0.9356	0.9700	0.9637
	Precision	0.9646	0.9175	0.9700	0.9586
	F 値	0.9641	0.9265	0.9700	0.9611
Lv.4 発音形	Recall	0.9599	0.9318	0.9676	0.9612
	Precision	0.9609	0.9137	0.9676	0.9561
	F 値	0.9604	0.9227	0.9676	0.9586

未知語を一部含んだ状態である．それぞれ，付録 A.1.4，A.2.1 に示すような文体である．

中古和文 UniDic による擬古物語と和漢混淆文の解析結果を表 4.7 に示す．擬古物語では，語彙素認定の F 値で 0.95 以上の精度を確保している一方，和漢混淆文では 0.85 程度となっている．ここから，中古和文 UniDic がターゲットとしていた文体であれば未知の資料であっても十分な解析が可能であること，一方ターゲットとしていない文体では十分な精度が得られず，再学習など新たな取り組みが必要であることが分かる．

4.4.5 学習に用いるコーパスの量の解析精度への影響

つづいて，学習に用いるコーパスの量の解析精度への影響を確認するために，コーパスの量を変化させて解析精度を評価した．評価コーパスは辞書ごとに固定した 10 万語で，4.4 節と同一のものである．学習用のコーパスは，評価コーパス以外の人手修正済みデータを文単位でランダムに並び替えた後，先頭から指定語数分取得している．語数は，2 万語までは 5000 語ごと，10 万語までは 2 万語ごと，それ以上は 10 万語ごとに学習用コーパスを増やし，近代文語 UniDic では 50 万語，

表 4.7: 「中古和文 UniDic」による擬古物語・和漢混淆文の解析精度

評価項目		擬古物語	和漢混淆文
語数	評価語数	44,842	10,715
	出力語数	44,088	10,530
Lv.1 境界	正解数	44,285	10,372
	Recall	0.9876	0.9680
	Precision	0.9892	0.9613
	F 値	0.9884	0.9646
Lv.2 品詞	正解数	43,405	9,532
	Recall	0.9680	0.8896
	Precision	0.9696	0.8834
	F 値	0.9688	0.8865
Lv.3 語彙素	正解数	42,822	9,188
	Recall	0.9550	0.8575
	Precision	0.9565	0.8515
	F 値	0.9557	0.8545
Lv.4 発音形	正解数	42,685	9,051
	Recall	0.9519	0.8447
	Precision	0.9535	0.8388
	F 値	0.9527	0.8418

中古和文 UniDic では 70 万語まで評価している．比較のため現代語用の UniDic についても同様の方法で 100 万語まで評価した．現代語の学習・評価用コーパスには BCCWJ のコアデータを利用した．

この結果を図 4.4 に示す．縦軸が語彙素認定の F 値，横軸がコーパス量である．

現代語 > 中古和文 > 近代文語の順に解析精度が低くなるが，この傾向はコーパス量が同じであればどの段階においても同じであり，この差は各辞書が対象とするテキストの（短単位による形態素解析という観点での）難易度を反映したものだといえそうである．ただし，近代文語文の精度低下には，口語による表現が部分的に挿入される場合があることも影響している [6]．口語表現を除外するようにコーパスを整備したり，口語表現の品詞認定基準を改めたりすることで，他の辞書との差は小さくなるものと思われる．

[6]近代文語文のコーパス中，口語表現が占める割合は，抜き取り調査にもとづく概算で 1%程度である．

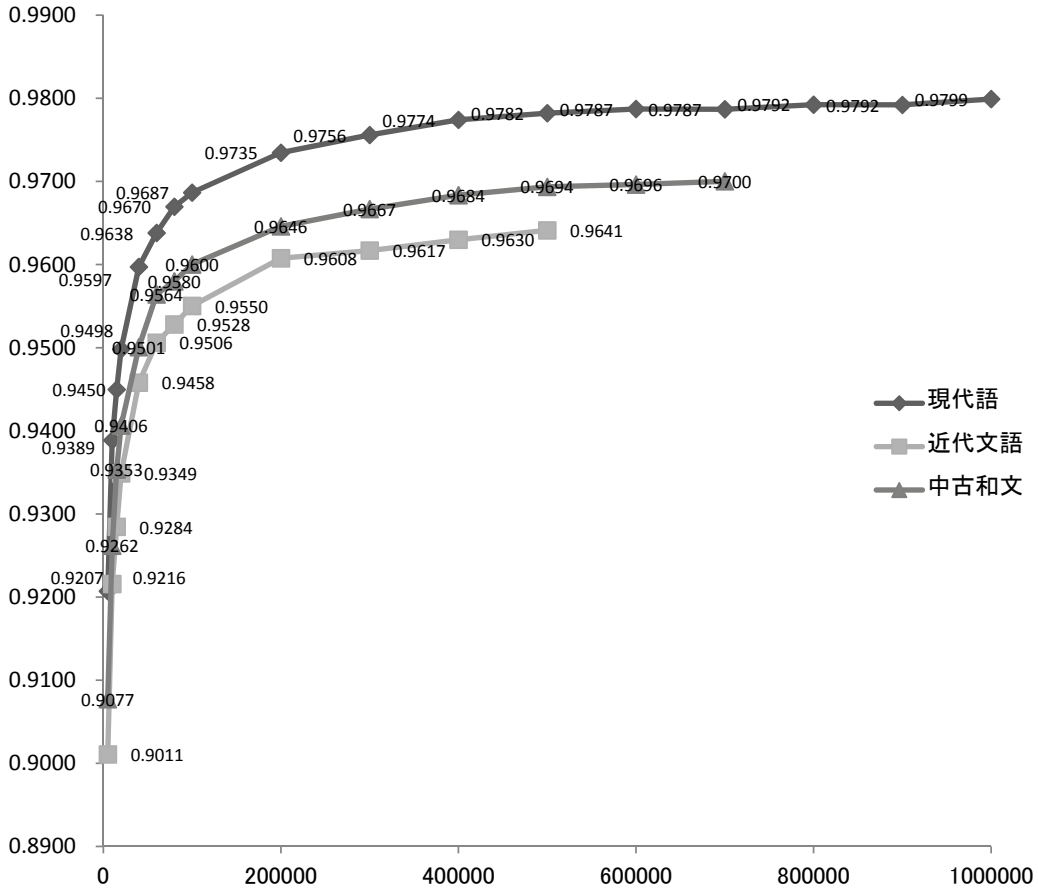


図 4.4: 各種 UniDic の学習曲線（語彙素レベル・F 値）

訓練コーパス量が 5000 語でも 0.9 以上の F 値が得られているが，これは現代語の UniDic で解析した表 4.1 や表 4.2 の結果よりも遙かに良い数値である．古文の形態素解析では，少ない量であっても専用のコーパスを使って辞書を作成することが効果的であることがわかる．また，約 5 万語のコーパスで 95%の精度に達しており，どの辞書でも約 10 万語を境に精度向上が大幅に鈍化し飽和していく．短単位の形態素解析辞書を新たに作成するのに必要な学習用のコーパスは約 5〜10 万語というのが一つの目安であるといえる．

4.5 エラー分析

4.5.1 高頻度の解析エラー

4節 (表4.5) の精度調査におけるエラーから，近代文語UniDic，中古和文UniDicのエラーの傾向を分析する．表4.8に，境界認定・品詞認定・語彙素認定の各レベルにおいて特に高頻度のエラーをまとめた．表中括弧内の数字はエラー数である．以下，これらのエラーについてレベル別に確認する．

表 4.8: 高頻度の解析エラー

		中古和文	近代文語
境界認定 Lv.1		「に/て」(48), 「御/前」(5), 「と/も」(4)	「然れ/ども」 (18), 「論/派」(10), 「に/て」 (9), 「と/も」(6), 「異/なる」 (6), 「彼/の」(5)
品詞認定 Lv.2	品詞	「に」助動詞/格助詞/接続助詞 (342), 「を」格助詞/接続助詞/終助詞 (129), 「また」接続詞/副詞 (17)	「に」助動詞/格助詞/接続助詞 (182), 「も」接続助詞/係助詞 (33), 「で」格助詞/助動詞/接続助詞 (20)
	活用形	助動詞「む」終止形/連体形 (124), 動詞「給ふ」終止形/連体形 (30), 助動詞「ず」終止形/連用形 (17)	助動詞「ず」終止形/連用形 (62), 動詞「言ふ」終止形/連体形 (32), 助動詞「む」終止形/連体形 (23), 動詞「有り」終止形/連用形 (22)
語彙素認定 Lv.3		「御」ミ/オオン (399), 「音」オト/ネ (27), 「こと」異/事 (24), 「中」ウチ/ナカ (18), 「夕」ユウベ/ユウ (13)	「人」ジン/ニン (48), 「等」トウ/ラ (18), 「空」クウ/ソラ (13)

4.5.2 境界認定レベルのエラー

境界認定のレベルでは，近代文語UniDicは結果として過分割となっているものが272例，同数となるもの74例，過結合となるものが258例であった．中古和

文 UniDic は過分割となっているものが 189 例，同数となるものが 25 例，過結合となるものが 176 例であった．

中古和文・近代文語でともにエラーが多い「にて」「とも」は，語源にさかのぼると「に/て」「と/も」であり，複合してできた語を語源的に見て 2 語と扱うか，新たにできた 1 語として扱うかという認定基準の立て方の問題と関わる．歴史的な言語変化によって一語化が進展していくわけだが，もともと連続して出てきやすい語の連続であり，また全ての「に/て」「と/も」連続が一語化するわけではないため判別が難しい．

近代文語の「然れども」は，現在の規程では，「しかれども」と読む場合には「然り」と「ども」に分割し，「されども」と読む場合には一語の接続詞として扱っている．したがって問題は「然れ」を「され」と読むか「しかれ」と読むかという点にあり，実質上は語彙素認定のエラーであるとも言える．「彼の」も同様で，「かれ（の)」と読めば 2 語になり，「か（の)」と読めば連体詞として一語と見なされる．このように近代語では漢字表記語が多く読みに曖昧さがあり，そこに語の歴史的変化により一語化が進展していることが複合して問題となっている．

一方，近代文語の「論派」は，「〇〇論派」とある場合に，UniDic の短単位規定で「((〇〇）論）派」という語構成を考えて「-論」「-派」がいずれも接尾辞となって切り出されるのに対し，「論派」という一語の名詞も存在するためにこちらが優先されることによっている．漢語の多い近代語で目立っているが，これは現代語でも同種の現象が起きる問題であり，漢語の単位認定について形態素解析では扱いきれない語構成の問題が短単位認定基準に取り込まれていることが要因であるといえる．

4.5.3　品詞認定レベルのエラー

品詞認定のレベルでは，品詞そのものの認定エラーと，活用形の認定のエラーが区別される．

品詞そのものの認定で中古和文・近代文語ともに最多のものは「に」の認定が助動詞・格助詞・接続助詞の間で揺れるものである．「に」の判別は現代語においても助動詞「だ」連用形と格助詞「に」の間などで問題になるが，古文では接続助詞の「に」が高い頻度で用いられるため曖昧性が高い．接続助詞「に」は連体形接続であるため，接続の上でも他と区別が付かない．このほか，中古和文では「を」の判別エラーが多い．現代語では格助詞以外の用法を持たないが，中古語では接続助詞・終助詞があるためエラーにつながっている．また，近代文語では「も」「で」の判別エラーが目立つが，いずれも同形の接続助詞があることで

現代語よりも曖昧性が増している．さらに「で」は近代においては口語的な助動詞「だ」連用形としての「で」が用いられることがあるため，現代語でも発生する助動詞「だ」連用形と格助詞「で」の判別と同じ問題が生じている．

このように品詞の認定では，コピュラ助動詞（「なり」「だ」）の連用形と格助詞との判別という現代語でも見られるエラーがあることに加えて，古文では同形の接続助詞が存在するためにより曖昧性が高くエラーにつながっている．

活用形の認定では，高頻度の助動詞や動詞の終止形・連体形の判別と，助動詞「ず」・動詞「有り」の終止形と連用形の判別でエラーが多く発生している．

終止形・連体形の判別エラーは，係り結びに起因する古文特有のものである．古文では，文中に「ぞ」「か」「や」の係助詞や疑問詞（不定語）が存在する場合，係り結びの法則によって文末が終止形ではなく連体形になるという現象がある．ところが，四段活用では終止形と連体形が同形であるため，文末に位置する場合には文中に上述の要素（係り）が存在するかどうかによって同形の動詞を判別しなければならず，この困難がエラーにつながっている．係り結びは中古和文で特に多いため終止・連体形の判別エラーも中古和文に多く，活用形間の誤り全体 363 例のうち 214 例がこのエラーである．

助動詞「ず」・動詞「有り」の終止形と連用形の判別は，文末の認定に関わるものである．助動詞「ず」やラ行変格活用の語では終止形と連用形が同形であるが，終止形と連用形の違いは多くの場合，文が中止しているのかそこで終わっているのかの違いに相当する．ところが，近代文語文では，文末が句点として必ずしも明示されない[7]．句点と読点が区別されず，ともに「,」で表されている文が多く，こうした場合には中止か終止かの区別が極めて難しい．このことが終止・連用形の選択エラーにつながっている．これも古文のテキスト特有のエラーである．

4.5.4　語彙素認定レベルのエラー

語彙素認定では，中古和文における接頭辞「御」が「ミ」「オオン」の間で揺れる例が極めて多かった．これを含め中古和文における高頻度の語彙素認定エラーは，品詞が一致する上に語種までが同じ語の間で生じている．近代文語で最大の「人」が「ジン」「ニン」で揺れる例も語種が同じものである．UniDic の見出し語中には，同一の漢字表記語が音読する漢語と訓読する和語に区別される例が多いが，学習素性に語種を利用していることもあり，語種をまたいだ誤りは比較的少なかった．語彙素認定のエラーは現代語でも生じうるタイプの誤りである．ただ

[7]このため，近代語のコーパスでは人手によって文境界をタグ付けしている．古文の解析にとって文末の自動認定は残された重要な課題の一つである．

し，現代語では接頭辞「御」は漢語「ゴ」と和語「オ」でほぼ区別が付くが，中古和文では和語に複数の読みがあるため語種では判別ができないといった違いがある．

4.5.5　エラーのまとめ

以上のエラーのうち古文特有といえる問題は，次の 2 点である．

- 係り結びに起因する文末活用語の終止形・連体形の判別
- 文末表示の曖昧さに起因する文末活用語の終止形・連用形の判別

特に係り結びの問題に対処するには文中の離れた要素を考慮する必要があるが，提案手法では局所的な形態論情報だけを素性として利用しているため対応できていない．しかし，これらのエラーは比較的簡単なルールによって自動修正できるため，形態素解析後の後処理で対応することが考えられる．

その他のエラーは同種の問題が現代語でも発生しうるものである．しかし，「に/て」「と/も」，「に」「も」「を」などの助詞に関する判別は，古文の方が同音異義となる語彙が多いため曖昧性が増していた（「を」は現代語では曖昧性がない）．また，特に近代文語では漢字表記語の割合が大きく，その 読みの曖昧性がエラーにつながる例が現代語よりも多い．中古和文の接頭辞「御」も同様で和語に絞っても現代語より多様な読みがあるため判別が困難である．

以上のように，エラーの原因には文法現象から語彙，表記法の違いまで，古文特有の現象が関わっているものが見られた．

4.6　本章のまとめ

UniDic の見出し語を増補し，学習用のコーパスを整備することによって，「中古和文 UniDic」[5] と「近代文語 UniDic」[78] の 2 つの形態素解析辞書を作成した．この辞書により，語彙素認定の F 値で，近代文語は 0.9641，中古和文では 0.9700 という高い精度で解析することが可能になった．これにより，通時コーパス構築の基盤となる形態素解析システムが整ったといえる．これらの形態素解析辞書はすでに Web 上で一般公開を行っている [8]．

開発過程で，古文の形態素解析には見出し語の追加だけでは十分な精度が得られないこと，5000 語程度の少量であっても専用の学習用コーパスを用意すること

[8] http://www2.ninjal.ac.jp/lrc/index.php?UniDic

が効果的であることが確認された．他分野の辞書による解析精度が低いこととあわせ，このことは，古文の形態素解析では，他分野のコーパスによって学習したパラメータの転用を図ることは有効ではないことを示唆している．また，短単位にもとづく形態素解析辞書の学習には，5〜10 万語の学習用コーパスを用意すれば歴史的日本語コーパスの構築にとって十分であることが確認された．さらに，エラーの分析から，残されたエラーの多くは，現状の解析器と学習可能な素性では対処の難しいものであることが確認された．

近代文語と中古和文を比較すると，近代文語の解析精度が低かったが，その理由の一つは近代文語の中身が多様で，ドメインの分割がうまくできていないことにあるものと思われる．比較的少量の学習用コーパスで効果が見込まれることが確認されたことから，近代文語文をより小さなドメインに分割することで全体として精度を向上させられる可能性がある．同様に，会話文と地の文とで別の辞書を作成することでも精度の向上が期待できる．今後の課題としたい．

通時コーパスの構築のためには，今後，様々なタイプのテキストの解析を行っていく必要がある．今回の調査でも，中古和文と同時代の資料であっても和漢混淆文は中古和文 UniDic では十分な精度で解析できないことが確認された．今後，和漢混淆文をはじめとする多様なジャンルのテキストを対象とした形態素解析辞書を作成していく必要がある．その中では，仮名遣いのバリエーションへの対処や送り仮名の大幅な省略などの表記揺れへの対処も必要となる．今回得られた情報をもとに必要なコーパスを整備するとともに，新たな解析器も活用しつつ，通時コーパスのための形態素解析を行っていきたい．

第5章　多様な古文テキストへの対応

5.1　中古和文・近代文語文以外のテキスト

4章では日本語の歴史的資料のうち，「中古和文」「近代文語文」という二つの主要な文体のテキストの形態素解析について論じた．しかし，通時コーパスで扱う必要のある資料の中には，これ以外にもいくつもの重要な資料群が存在する．2.2節で見たように，こうした資料の例として，上代の万葉集，中古末の『今昔物語集』，中世の狂言台本，近世の洒落本などがある．これらの資料は，もともと残された量が比較的小さいため，それだけのために「中古和文UniDic」や「近代文語UniDic」と同レベルでの辞書やコーパスの整備を行うことは困難である．

このような場合，比較的質が近いUniDicのコーパスを用いて転移学習によって当該資料群の解析を可能にしていく方法が考えられる．しかし，4章で見たように，歴史的日本語資料はそれぞれが大きく性格が異なる文体であるため，他の辞書を流用するより，少量であっても専用のコーパスを用意して再学習することが有効であることが確認された．5,000語程度の学習用コーパスであっても比較的高い精度が得られることから，これら小規模な資料群でも，学習用のコーパスを整備することが十分に可能である．見出し語の追加とコーパスの整備というオーソドックスな方法が最も簡易でかつ効果的であると考えられる．

本章では，国語研通時コーパスの構築作業がすでに始まっている資料群（2.4）のうち，コーパスの整備が進んでいるものから順に，次の3種類のテキストについてその形態素解析の方法について論ずる．

- 中世・近世口語文（狂言と洒落本）
- 旧仮名遣いによる近代以降の口語文
- 漢文訓読文と和漢混淆文の説話集

5.2　中世・近世の口語文 [1]

5.2.1　洒落本・狂言の性格

国語研通時コーパスでは，中世・近世の口語を反映した資料群をコーパス化するためにて，狂言と洒落本の一部について電子化・形態論情報の付与に着手している．狂言は『虎明本狂言集』を，洒落本は『洒落本大成』収録作品の一部を対象としている．

洒落本と狂言は本来，かなり異なる性格の資料であるが，現代語とも「近代文語」「中古和文」とも大きく異なる文体である．洒落本も，さらに細分するのであれば，前期の上方語と，後期の江戸語とで別の辞書を用意することも考えられる．一方で，いずれも中世から近世にかけての口語を反映する資料という点では共通性も持つ．

歴史的な資料の形態素解析を行う場合，資料群を細分化しすぎると，多種類の学習用コーパスと辞書を作成する必要が生じる上に，解析対象の資料が十分に残らないため，効果的ではない．そこで，本節ではこれらの資料を解析することのできる共通の辞書の開発を視野に，二つの資料をあわせて取り扱う．それぞれ，次のような性質の資料群である．

狂言

狂言は，中世から近世にかけての言語資料として重要な位置を占めている．登場人物が多くその身分関係が明確であること，対話劇の形で進行し場面・状況が明確であることから，口語資料としての価値は極めて高い．

その中でも『虎明本』は，寛永 19（1642）年に大蔵流十三世宗家大蔵虎明の手によって書かれた大蔵流の祖本である．本狂言 237 曲を収めており，狂言の類別や詞章の整備された台本として，質・量とも第一級の資料である．その詞章には，中世の言葉を伝承していると見られる点，書写当時である近世初期の日常語の影響を受けたと思われる点，舞台言語として整理され固定化・類型化する兆候が見られる点がある．狂言史上の位置を踏まえ，他の台本との比較ということが不可欠であるが，注釈書や総索引が整備され，中世から近世の言語資料として広く利用されてきた（小林・市村 2013 [59]）．

虎明本では，狂言台本としての性格上，書写の手間を省くために通常であれば漢字表記される語が仮名書きされることが多いという特徴があり，これにより辞

[1] 本節の内容は [75] にもとづく．

書未登録の表記が発生しやすく，形態素解析を難しくしている．一方，同一人による写本であるため，全体として均質性も持つ．

虎明本のテキストの例として大名狂言「あさう」の一部を付録 A.3.1 に掲げた．

洒落本

洒落本は，江戸時代の遊里を舞台とした小説の一種で，明和（1764–1772）から天明（1781–1789）の頃を中心に，文政 (1818 – 30) のころまでに多く刊行された．登場人物の会話部分に当時の話し言葉が反映されているとされ，日本語史研究上，近世後期の口語の実態を探る上での重要資料である．大きく分けて江戸版と上方版があり，その口語体の会話部分はそれぞれの地域の言葉を反映する場合も多い．また年代も 18C 後半から 19C 前半までと幅広く，近・現代語への過渡的状況を伺うのに適している．方言や中央語の形成を知る上でも，不可欠な資料である（市村ほか 2013 [64]）．

洒落本は，作品ごとに内容が大きく異なるだけでなく，江戸・上方で言語そのものが大きく異なっている．作者・形式も多様で，全体としてテキストの均質性は低い．さらに言葉遊び的な要素をしばしば含むため，形態素解析やコーパス化には課題が多い．

以下に洒落本の一つ，1757 年（宝暦 7）刊の『聖遊廓』の一部を掲げた．また，付録に「甲駅新話」（A.3.2）と「陽台遺編・雛閣秘言」（A.3.2）の例を挙げた．

洒落本のテキスト例「聖遊廓」

> 爰に聖人のかよひたまへる郭(くるわ)あり 揚屋(あげや)の亭主(ていす)は李白(りはく)とかや 中にも孔子はくるわにてすい(はで)といはれて端手ならず　ゑちご縮(ちゞみ)のかたびらにもんろの羽織(はおり)すそながく深(ふか)あみがさにあわざうり古金買(ふるかねかい)の目利(めきゝ)にも太夫かいとは見へざりし　李白がかたへ御入りあれば　▲亭主李白 是は仁さまおめづらしい さあ﹀﹀おくへ ともてはやす　▲孔子 なんと李す此中は久しいの 無事で珍重﹀﹀　と座敷へ行　▲李白女房滝(たき)　是はおめづらしいおかほ．おうはさばつかり申ておりました　▲中居なつ もし仁さま此中横堀でお見うけ申ましたゆへ大かたおよりなさるであろふとぞんじましたに．よふまたせなさつたの　▲孔子 ヲヽよりたかつたけれども行時(ゆくとき)に径(こみち)によらず．

なお，洒落本のコーパスの一部については，コーパスの検索結果から本文画像を呼び出すことができるように原文の画像ファイルと関連付けを行っている．上

掲のテキストに対応する箇所を，「国立国語研究所研究図書室所蔵 日本語史研究資料」として公開されている『聖遊廓』[2] 画像ファイル[3] から図 5.1 に示す．

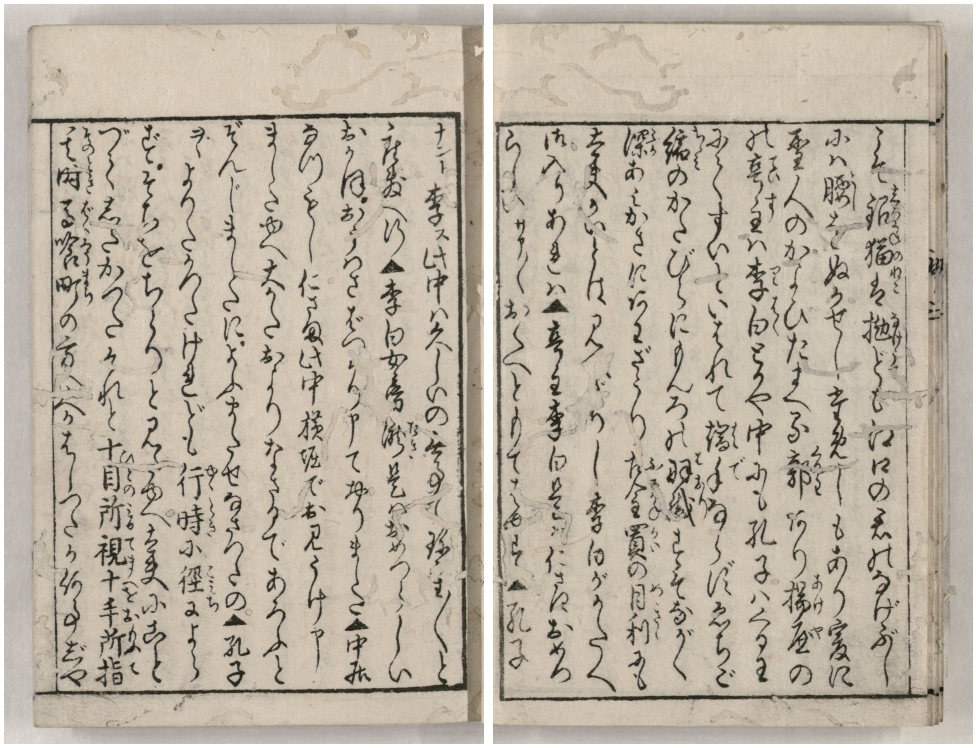


図 5.1: 洒落本「聖遊廓」原文画像

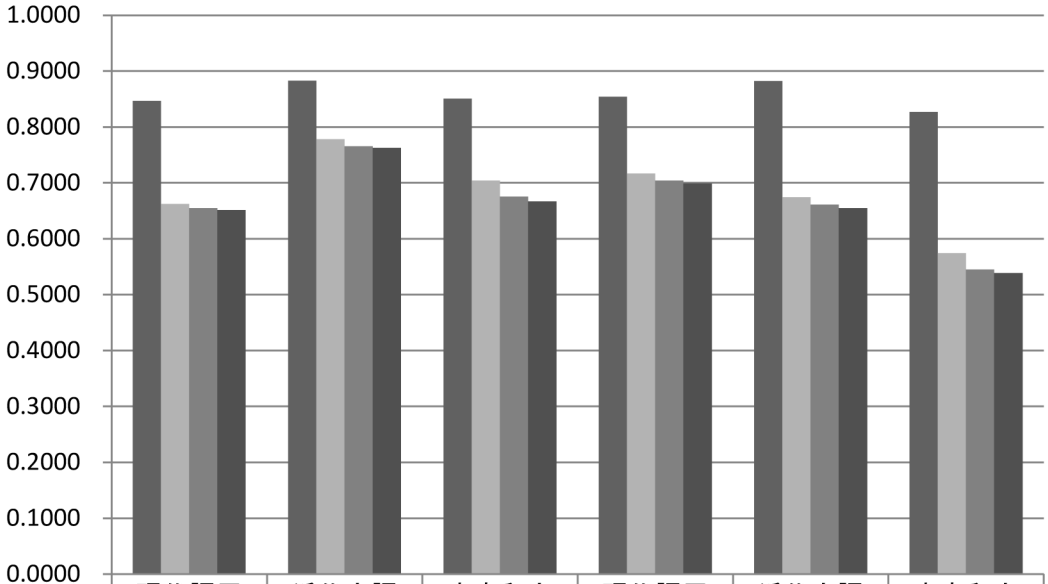
5.2.2 既存の UniDic による解析精度

狂言・洒落本のテキストは，いずれも現代語とは大幅に異なる上に，典型的な古文である平安和文などとも大きく異なる文体で書かれている．4 章でみたように，これまでに歴史的な資料を対象とした形態素解析辞書として「中古和文 UniDic」「近代文語 UniDic」を開発・公開してきたが，このいずれも狂言・洒落本の解析には適していない．

既存の辞書での解析精度を確認するために，形態素解析器に MeCab[2] を用い，現代語用の UniDic と中古和文 UniDic，近代文語 UniDic のそれぞれで狂言・洒落

[2]表紙に貼られた題簽（だいせん）には「雪月花」という書名が書かれている．
[3]http://db3.ninjal.ac.jp/ninjaldl/bunken.php?title=hiziriyukaku

本の評価用コーパスを解析し，精度を評価した．結果を図 5.2 に示す．評価データは，後述する学習用に整備したコーパス（表 5.1・表 5.2）の約 10%を文単位でランダムサンプリングしたものである．数値は語彙素認定の F 値である．



	狂言			洒落本		
	現代語用	近代文語	中古和文	現代語用	近代文語	中古和文
lv.1	0.8471	0.8829	0.8506	0.8543	0.8822	0.8269
lv.2	0.6626	0.7778	0.7040	0.7168	0.6744	0.5740
lv.3	0.6551	0.7653	0.6753	0.7044	0.6610	0.5450
lv.4	0.6513	0.7624	0.6667	0.6988	0.6549	0.5385

図 5.2: 既存の UniDic による狂言・洒落本テキストの解析精度

2.3.2 節で見たとおり，現代語のコーパス構築において必要とされた解析精度が，語彙素認定で 98%であった．歴史コーパスの構築においても，おおむね 95%以上の解析精度が必要とされる．しかし，図 5.2 にみるように，既存の解析辞書では近世口語資料の解析は難しく，狂言のテキストが最高で 76.5%，洒落本のテキストが最高で 70%程度の解析精度に留まっている．本格的なコーパス構築に用いるには大幅に精度が不足しており，新たな解析辞書を作成する必要がある．

5.2.3　学習・評価用コーパス

狂言は表 5.1 に示す約 5.3 万語分が，単語情報付きのコーパスとして整備済みである．本文は全て『大蔵虎明能狂言集 翻刻註解』[18] によっている．電子化テキストに，文書構造のアノテーション，濁点付与・文境界付与等の本文整備を施した後，既存の形態素解析辞書を用いて形態素解析を行った．形態素解析は，図 5.2 の結果に基づき，狂言は近代文語 UniDic で，洒落本は現代語用の UniDic で解析することから始めた．この結果を「形態論情報データベース」に格納した後，人手によって解析の誤りを修正している．

表 5.1: 狂言の学習・評価用コーパス

種類	曲名	語数
脇	ゑびす大黒	1173
	連歌毗沙門	1090
	福の神	599
	大黒連歌	543
	びしやもん	355
	餅酒	1749
	かくすい	684
	昆布柿	876
	鴈かりがね	1036
	三人夫	901
	つくしのおく	1037
	松ゆづり葉	395
	すゑひろがり	2048
	よろい	1250
	はりだこ	1567
	隠笠（宝之笠）	540
	財のつち	1058
	目近籠骨	2569
	三本の柱	1132
	松やに	688
	せんじ物	1197
	牛馬	1933
	なべやつばち	2060
	たうずまふ	586
	はちたたき	630
	連歌十徳（天神）	462
	祇園	315
	三国之百姓	85

	三人の長者	134
大名	あさう	2001
	入間川	1929
	鴈盗人	1624
	鬼がわら	596
	ふずまふ	402
	かずまふ	1358
	じせんせき	1729
	ふたり大名	941
	しんばい	274
聟・山伏	ゑびす毗沙門	1535
	鶏聟	1700
	ひつしき聟	1360
	はうちやう聟	1333
	おか太夫	1497
鬼・小名	あさいな	1862
女	わかな	752
	つりばり	822
出家座頭	腹不立	1799
集	ながみつ	1016
計		53222

洒落本は表 5.2 に示す約 4.4 万語分が，単語情報付きのコーパスとして整備済みである．本文は，「跖婦人伝」と「遊子方言」は小学館『新編日本古典文学全集』[36]，それ以外の 6 作品は『洒落本大成』[93] によっている．

人情本と滑稽本

狂言と洒落本の他に，江戸時代の口語資料として，滑稽本『浮世床』と人情本『春告鳥』についてもコーパスの整備作業を行った．これは，将来的に洒落本以外の江戸語資料を作成することを念頭に，先行して整備しているものである．いずれの本文も小学館『新編日本古典文学全集』[36] によっている．

人情本とは，江戸時代末期から明治初期にかけて流行した小説で，庶民の恋愛や人情を扱ったものである．為永春水の『春色梅児誉美』『春色辰巳園』が特に著名であるが，付録 A.3.3 に挙げる『春告鳥』もその春水の作である．

また，滑稽本とは，江戸時代後期の小説の一種で，庶民の日常生活における滑稽な様を描いたものである．十返舎一九の『東海道中膝栗毛』や，式亭三馬の『浮

表 5.2: 洒落本の学習・評価用コーパス

底本	書名	出版地	語数
新編全集	跖婦人伝	江戸	5700
	遊子方言	江戸	8986
洒落本大成	聖遊廓	大坂	4417
	甲駅新話	江戸	8613
	興斗月	京都	3622
	陽台遺編・妣閣秘言	大坂	4165
	風流裸人形	京都	3075
	箱まくら　序・上	京都	5319
計			43897

世風呂』が特に著名であるが，付録 A.3.4 に挙げる『浮世床』も『浮世風呂』に続く三馬の代表作の一つである．

人情本と滑稽本は，いずれもその判型から中本と呼ばれ，洒落本の影響の下に成立した近世小説であり，その会話文は近世江戸語の口語を反映した資料として貴重である．

洒落本と滑稽本・人情本は，書かれた時代も近く，特に江戸を舞台とするものは言葉の上での共通性が高い．洒落本に限定せず，これらの江戸時代の口語資料を解析することができれば，辞書の応用範囲が広がる．そこで，洒落本の解析精度評価とあわせて，滑稽本・人情本についても解析精度の評価を行った．

表 5.3: 滑稽本・人情本の学習・評価用コーパス

底本	種類	書名	語数
新編全集	滑稽本	浮世床　初編	4869
	人情本	春告鳥　初編	14805
		春告鳥　二編	9601
計			29275

5.2.4　近世口語共通辞書の解析精度

狂言と洒落本とに共通の辞書を作成するために，表 5.1・表 5.2 のコーパスのうち，評価用を除く全てのコーパスを利用して学習を行い，近世口語用の形態素解析辞書を作成した．

見出し語は，従来の中古和文 UniDic・近代文語 UniDic で用いていたものに，表 5.1・表 5.2 のコーパスで出現した語を追加したものを利用した．見出し語は，活用形展開後で総計 134 万に上る．近世口語では利用されない語彙を含むが，(1) どの語が不要であるかを事前に判断することは必ずしも容易ではないこと，(2) 古文の形態素解析辞書にとって見出し語の肥大化は大きな問題ではないこと，(3) 不要語があることによる解析精度への悪影響は特に認められなかったこと，によりそのまま利用している．

なお，MeCab の設定は 4.3.2 節と同様である．

表 5.4 にこの近世口語共通辞書の解析精度を示す．各レベルの意味は図 5.2 と同様，数値は F 値である．

表 5.4: 近世口語共通辞書の解析精度

	狂言	洒落本	人情本	滑稽本
Lv.1	0.9835	0.9613	0.9640	0.9592
Lv.2	0.9379	0.8617	0.8708	0.8660
Lv.3	0.9298	0.8524	0.8636	0.8427
Lv.4	0.9270	0.8473	0.8591	0.8369

既存の辞書と比較すると精度は向上しているが，特に洒落本（人情本・滑稽本）の精度が低く，コーパス構築に十分な性能とはいえない．これは，一つには洒落本等近世の口語資料となる版本の多くが，会話は口語文，地の文（ト書き，序）は文語文で書かれるというテキストの混質性が原因になっていると考えられる．

また，現状の学習用コーパスの量は全体で約 12.6 万語だが，4 章で見たとおり，近代文語 UniDic では約 64 万語，中古和文 UniDic では約 82 万語を用いており，現在の学習用コーパスの量は，同等の解析精度を得るには不足している．

この問題とは別に，狂言と洒落本という質的にかなり異なるテキストを近世口語として一括していることにも原因があると考えられる．

5.2.5　狂言・洒落本専用辞書の解析精度

4 章での実験により，歴史的資料の形態素解析を行う際には，異分野のテキストによる学習結果を流用するより，少量であっても専用コーパスによる学習が効果的であることが分かっている．そこで，狂言と洒落本を分割し，それぞれの専用辞書を作成して近世口語共通の辞書と解析精度を比較することにする．分割により，もともと十分ではない学習用コーパスの量はほぼ半減することになるが，専用の学習用コーパスのみを利用することによるメリットがそれを上回る可能性がある．

狂言の学習は狂言のコーパスだけを学習に利用し，洒落本は，滑稽本・人情本に時代的にも内容的にも比較的近いため，これらを利用する「洒落本・人情本・滑稽本用」と純粋に洒落本だけを利用するもの「洒落本専用」の 2 通りを作成した．それぞれの辞書によるターゲット資料の解析精度を表 5.5 に示す．見方は表 5.4 と同様であるが，表 5.5 では狂言と洒落本が，それぞれ別の辞書による評価結果となっていることに注意されたい．表 5.5 の数値は，学習用コーパスの量を大幅に減らしたものであるにもかかわらず，表 5.4 の共通辞書による解析精度よりも向上している．したがって，狂言と洒落本とでは別の解析辞書を用意すべきであることが分かる．

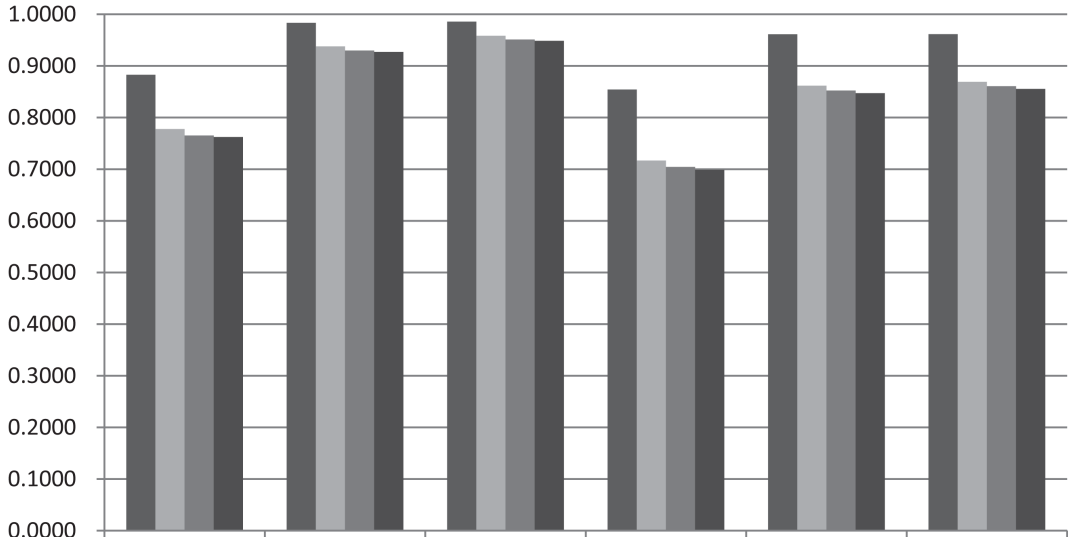
一方，洒落本については，洒落本だけで学習した「洒落本専用」よりも，滑稽本・人情本をまじえてコーパスサイズを増やした「洒落本・人情本・滑稽本用」のほうがよい精度となっており，これら近世後期の資料群はまとめて取り扱うことが有効であると考えられる．

表 5.5: 狂言・洒落本専用辞書の解析精度

辞書	狂言専用	洒落本専用	洒落本・人情本・滑稽本用
解析対象	狂言	洒落本	洒落本
Lv.1	0.9859	0.9614	0.9614
Lv.2	0.9583	0.8677	0.8691
Lv.3	0.9511	0.8574	0.8607
Lv.4	0.9486	0.8521	0.8554

図 5.3 は，本節でこれまでに精度を確認してきた辞書による解析精度を，比較のためにグラフにまとめたものである．既存の辞書による解析結果のうち最良の

もの（狂言は近代文語 UniDic，洒落本は現代語用の UniDic）と，近世口語共通辞書，専用辞書の解析結果を比較した．



	既存最良（近代文語）	共通辞書	専用辞書	既存最良（現代語）	共通辞書	専用辞書
	―狂言―			―洒落本―		
lv.1	0.8829	0.9835	0.9859	0.8543	0.9613	0.9614
lv.2	0.7778	0.9379	0.9583	0.7168	0.8617	0.8691
lv.3	0.7653	0.9298	0.9511	0.7044	0.8524	0.8607
lv.4	0.7624	0.9270	0.9486	0.6988	0.8473	0.8554

図 5.3: 各辞書による狂言・洒落本テキストの解析精度比較

このように提案手法により精度はかなり向上し，狂言については語彙素認定の F 値で 0.9511 と，コーパス構築のために十分利用できるものとなった．

しかし洒落本については語彙素認定の F 値で 0.8607 と大幅に低い数値となっている．これには，会話は口語文，地の文（ト書き，序）は文語文で書かれるという文体の混質性が影響している．エラーを確認すると 4.2 節で見た文語活用と口語活用の判定の問題が関わっており，活用型変換による補正を行うことで精度は大きく上昇する．したがって実質的には，見かけの数値ほどに解析精度が低いわけではない．

しかし，実際のコーパス構築に際しては対策が必要であり，そのためには会話部分とト書き部分で辞書を切り替えて解析を行うことが効果的であると考えられる．こうした方法は，A.4.1 に示すような近代の文章でも有効である．

5.3　旧仮名遣いの口語文 [4]

現代語の書き言葉は，短歌や俳句などの例外を除き，ほぼ全てが現代仮名遣いによっている．しかし，やや時代をさかのぼると，文法的には現代語と変わらない内容であっても，歴史的仮名遣いなどの旧仮名遣いで書かれている資料が少なくない．たとえば，現行の日本国憲法でさえ原文は歴史的仮名遣いで書かれている．明治期の言文一致よりのち戦後の国語改革の定着までの間に書かれたテキストの多くはこのような旧仮名遣いの口語文でかかれている．こうした資料を電子化してコンピュータ上で利用していく場合，表記の違いは処理上の問題を引き起こす．日本語の自然言語処理の基礎であるところの形態素解析においても，思わぬ解析エラーを引き起こし，全体の精度を低下させることになる．文字単位で対応表を用意すればことたりる旧漢字については比較的容易に常用漢字に置き換えることができる．しかし，仮名遣いについては，語を単位とするものであるため，このような単純な置換によるわけにはいかない．

5.3.1　旧仮名遣いの口語文の性格

明治時代後半の言文一致により文法的には現代語に近い口語文体が成立した．しかし，当時のテキストは歴史的仮名遣いか旧来の慣用（旧仮名遣い）によっており，今日の仮名遣いとは異なっている．このような旧仮名遣いの口語文体（以下，旧仮名口語）のテキストは，戦後の現代仮名遣い（「現代かなづかい」）の定着までの期間にかなりの量が残されており，資料の質の面でも，先述した日本国憲法などの法令・公文書から近代の文学作品，新聞等まで幅広いジャンルの重要なテキストが残されている．

これらの旧仮名口語テキストは，再出版されたり電子化されたりする機会に現代仮名遣いに直されるものも少なくないが，旧仮名遣いのままで残されているものも多い．また，できる限り一次資料に基づくべきであるという点からも，旧仮名口語テキストを直接扱わなければならない場合は多い．

すでに電子化されたデータのうち，大規模なものとしては，国立国語研究所で構築された『太陽コーパス』[29] に含まれる口語記事がある．『太陽コーパス』は約 3,400 記事，約 1445 万文字という大規模なテキストからなるが，このうちの約半分が旧仮名遣いの口語記事である（『太陽』の口語記事の例を付録 A.4.1，A.4.1 に挙げた）．このほか，文学作品の多くも原典は旧仮名口語で書かれており，旧

[4] 本節の内容は [73] にもとづく．

仮名遣いのまま電子化されたものも少なくない．「青空文庫」[5] に収録された作品の中でも，たとえば芥川龍之介の作品では 374 作品中 209 作品が旧仮名遣いであり，青空文庫における近代の作家のテキストのかなりの部分を旧仮名口語テキストが占めている．

一般的な現代語の文章は現代仮名遣いの口語文であるから，旧仮名口語との違いは，基本的には仮名遣いという表記法の一部分に過ぎないはずである．しかし，旧仮名遣いが主に利用された時代は，すでに 60 年以上前のことである．そのため，対象となる旧仮名口語テキストは，現代語テキストと比べた場合，語彙や文法など表記以外の点でも違いを生じている．特に，『太陽コーパス』の口語文のように明治期の口語資料となると，語彙的には近代文語 UniDic で利用されるような，現代語では用いられない古い見出し語が必要とされる．

また，表記の違いにおいても，仮名遣いだけでなく，使用される漢字の面でも大きな違いがある．旧仮名口語文では，単に対応する旧漢字が利用されるだけでなく，今日では用いられない幅広い漢字表記が利用される傾向にある．

旧仮名口語テキストに形態素解析を施す場合，現代語用の辞書を用いると仮名遣いが異なる平仮名が現れる部分で解析に失敗することになる．一方，近代語の文語文を対象として開発された「近代文語 UniDic」を利用すれば，こうした旧仮名遣いや古い時代特有の語彙には比較的良く対応できる．しかし，近代文語 UniDic は文語文を対象として開発されたものであるため，口語特有の表現の解析に失敗する場合が少なくない．たとえば，活用語の場合には文語活用の語として登録されているために，口語の音便形に対応できない場合があるほか，口語の助動詞などが正しく解析できない場合がある．また，UniDic では口語文法と文語文法に対応させる形で動詞の活用型を区別しているが，近代文語 UniDic では文語形を優先して利用するため，現代語の解析結果とは異なる結果を返すという問題もある．

このような問題に対処するため，新たに開発した「旧仮名口語 UniDic」では，現代語用の UniDic と近代文語 UniDic の見出し語を全て利用し，さらに不足する表記（書字形）を追加した．そして，現代語用の UniDic と近代文語 UniDic の学習用コーパスから必要と考えられるコーパスを流用し，これに新たなコーパスを加えて学習を行った．

5.3.2 見出し語の拡充

見出し語には，現代語用の最新の UniDic の見出し語に加え，近代文語 UniDic の見出し語を全て利用した．先述したとおり，旧仮名口語文は単に仮名遣いが異

[5] http://www.aozora.gr.jp/

なるだけではなく，書かれた年代の違いから近代語特有の表記・語彙を含むためである．これに加えて，次に示す活用形の問題に対処するために，活用表を整備して見出し語の拡充を行った．

UniDic において，たとえば動詞「買う」の場合には，活用型が口語では「五段-ワア行」，文語では「四段-ハ行」のように区別されている．そして，口語では終止形「買う」，文語では終止形「買ふ」の形だけが見出し語として登録されていた．そのため，旧仮名口語文で「買ふ」が現れる場合，文語形である「四段-ハ行」として解析されてしまう．しかし，旧仮名遣いであるからといって口語形であることには変わりないわけであるから，他と同じ「五段」活用として認定されることが望ましい．

そこで，「買ふ」「買ひ」「買へ」（「かふ」「かひ」「かへ」）の活用形（書字形）を，口語の「五段-ワア行」の動詞から派生させるように活用表を拡充した．このとき，「買ふ」のようにハ行に活用しているものを「ハ行」とせず「ワア行」とすることの是非が問題になるが，ここでは「買う」が仮名遣いにかかわらず同語として容易に抽出できることと，現代語コーパスとの互換性を優先して，これらも「ワア行」の一活用形とした．

また，旧仮名口語文が必ずしも歴史的仮名遣いによらないことを踏まえ，連用形ウ音便では「買う」（「かう」「こう」）だけでなく，「買ふ」（「かふ」「こふ」）も派生させている．特にウ音便で「ふ」表記がなされる頻度が高いためである．さらに，促音の「っ」が小書きされないことに対応するため「買つ」（「かつ」）を促音便形として派生させた．

これにより，口語活用の「買う」全体で表 5.6 のように多数の活用形表記形を持つこととなった．「新規追加分」に○印を付けた行が，新たに追加した活用形である．ここでいう意志推量形とは，未然形に助動詞「う」が付いた形を指す．音変化で「う」を切り出せない場合が少なくないことから UniDic では活用形の一つとして扱っている．

5.3.3　学習用コーパス

先述したとおり，旧仮名口語は，現代語と比べて単に表記が違うだけでなく，語彙や文法の面でも違いが見られる．したがって，本来であれば専用の学習用コーパスを大量に用意して，コスト学習を行うことが望ましい．しかし，コーパスを人手で修正して整備するためには多大なコストを要するため，旧仮名口語の大量の学習用コーパスを用意することは困難である．そこで，現代語用の UniDic のために整備されたコーパスと近代文語 UniDic 用に整備されたコーパスを一部用

表 5.6: 拡張した活用表の例（動詞：五段-ワア行）

活用形	活用語形	活用書字形	新規追加分
未然形-一般	カワ	買は	○
	カワ	買わ	
連用形-一般	カイ	買ひ	○
	カイ	買い	
連用形-ウ音便	コウ	買う	
	コウ	買ふ	○
連用形-促音便	カッ	買っ	
	カッ	買つ	○
終止形-一般	カウ	買ふ	○
	カウ	買う	
連体形-一般	カウ	買ふ	○
	カウ	買う	
仮定形-一般	カエ	買へ	○
	カエ	買え	
命令形	カエ	買へ	○
	カエ	買え	
意志推量形	カオウ	買はう	○
	カオウ	買はふ	
	カオウ	買おう	
	カオッ	買おっ	
	カオッ	買おつ	○
	カオ	買お	

いつつ，これに旧仮名口語専用の学習用コーパスを加えて MeCab による機械学習を行うこととした．

このために，後述する評価用データとあわせて約 81,400 語分の旧仮名口語テキストに対して UniDic による形態素解析を施したのちこれに人手による修正を加えて，正解となるコーパスを作成した．このうち，表 5.7 に示す約 58,000 語のコーパスを学習に利用した[6]．テキストは，「青空文庫」と『太陽コーパス』から選定した．

現代語のコーパスは，全てのコーパスを学習に利用するのではなく，ターゲットとなる文体と違いが大きく，不要と思われるものは除くこととした．そのため，Web のブログと掲示板のデータは利用していない．その結果，書籍を中心に約 2,116,400 語を学習に利用することとなった．近代文語 UniDic 用のコーパスから

[6]表中，『太陽コーパス』を出典とするデータのタイトルの数字は，たとえば「192501-07_」の場合，1925 年の第 1 号にある 7 番目の記事であることを示す．

は，文体的に口語に近いものを中心に約 144,500 語分を選定して学習に利用した．

表 5.7: 旧仮名口語文の学習用コーパス

出典	タイトル（著者）	語数
青空文庫	蟲の聲（永井荷風）	2283
	計畫（平出修）	9455
	運動会の風景（葉山嘉樹）	1231
太陽コーパス	192501-07_近代兵器の進歩並に将来の趨勢	5753
	192501-114_日本の記念切手と時価	3878
	192501-116_漫画小説 握り損ねた玉	1658
	192501-122_卓上私語	649
	192501-12_鼻で見，指で聞く少女	2700
	192501-13_政界太平記	2066
	192501-15_歴代の総理大臣（一）	1826
	192501-20_最近に於ける飛行機の発達	6130
	192501-28_貸金庫とはドンなものか	2369
	192501-40_予の実験したる熱湯浴若返法	2367
	192501-57_最近X光線療法の進歩	5574
	192501-65_長篇小説　蛇人（第一回）	4786
	192501-72_世界的の大発明として推称すべきゴ氏の汚物焼却炉	2745
	192501-78_放送無線電話の沿革と現状及其将来	2568
計		58038

5.3.4　旧仮名口語文用辞書の解析精度

このようにして作成した「旧仮名口語 UniDic」と MeCab で旧仮名口語テキストを解析した場合の精度を調査した．比較対象の UniDic は現時点での最新の公開版を利用している [7]．

解析精度の評価には，表 5.8 に示す約 23,400 語の人手修正済みコーパスを利用した．学習用コーパスと同様，評価用コーパスも「青空文庫」と『太陽コーパス』から選定したものである．

[7]現代語用の UniDic は ver.1.3.12，近代文語は ver.1.2 を用いた．

表 5.8: 旧仮名口語文の評価用コーパス

出典	タイトル（著者）	語数
青空文庫	井戸の底に埃の溜つた話（葉山嘉樹）	1331
	幽霊の足（相馬御風）	647
	硯友社の沿革（尾崎紅葉）	7470
太陽コーパス	192501-02_近代文明と発明	4989
	192501-16_現代の女性美	2296
	192501-55_帝都の復興に際して偉人星亨氏を想ふ	2364
	192501-75_麻疹の予防の急務	4342
計		23439

解析精度

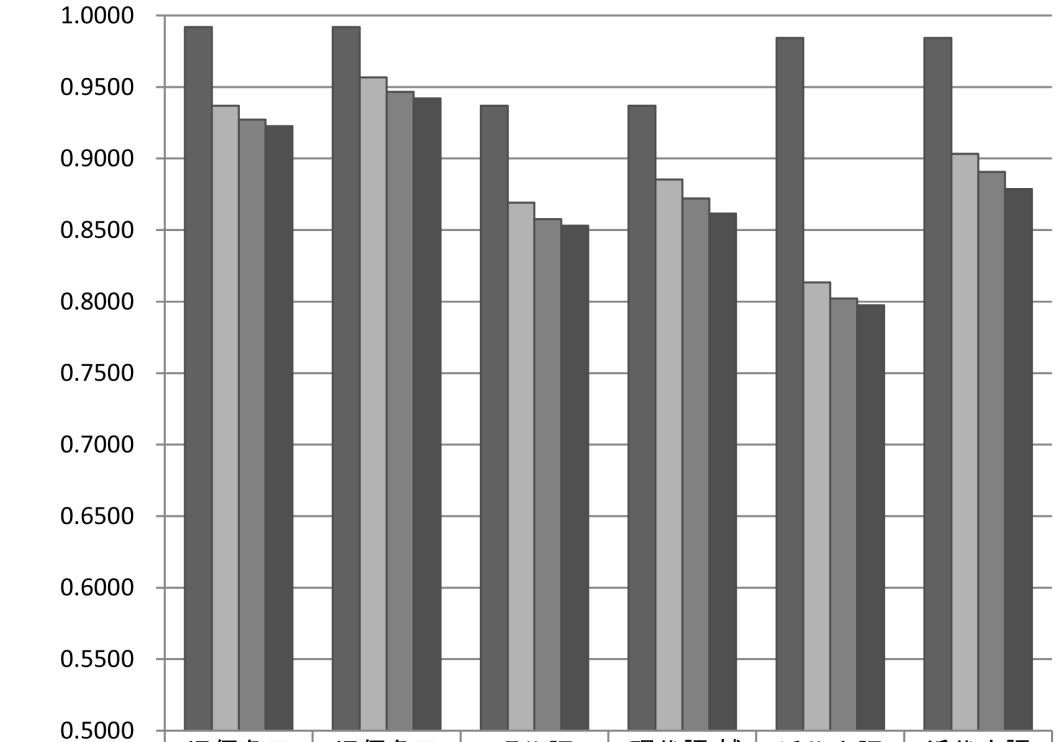
図 5.4 は，評価用コーパス全体を対象に各種の UniDic で解析した結果の精度をグラフにしたものである．評価方法は 4.4 節と同様である．

それぞれの辞書による通常の解析結果に加え，今回は 4.2 節での調査と同様に，活用型を補正した結果の評価も調査した．

図 5.4 から分かるとおり，補正の有無にかかわらず，旧仮名口語 UniDic の精度が他の辞書を大きく上回っている．UniDic の評価精度で基準として用いてきた語彙素認定レベルでは，補正なしの場合，現代語用の UniDic に対して 0.07 の差，近代文語 UniDic に対しては約 0.12 以上の差に及ぶ．補正を行った場合でも，現代語用の UniDic に対して約 0.07 以上，近代文語 UniDic に対しては約 0.05 以上の差を付けている．補正した場合にも他の辞書を大きく上回っていることから，単に活用型の文語・口語選択の選択によって精度が向上しているだけではなく，全体として解析がうまくいっていることが分かる．

旧仮名口語 UniDic の解析精度は，語彙素認定では，補正なしで 0.9273，補正ありで 0.9467 となっており，この精度は現代語用の UniDic や近代文語 UniDic 等と比べるとやや低い．活用型以外における口語文法・文語文法による揺れが存在することが影響している可能性がある．また，このような基準の難しさもあり，コーパスの人手修正が万全でないために，コーパスの方に誤りが残っていたため誤りと誤判定された可能性も否定できない．

現代語と近代文語を比較すると，補正なしでは現代語用の UniDic が近代文語 UniDic を上回っているのに対し，補正を行うことによって近代文語 UniDic が現代語用の UniDic を上回るようになることが注目される．



	旧仮名口語UniDic	旧仮名口語補正	(現代語)UniDic	(現代語)補正	近代文語UniDic	近代文語補正
Lv.1境界	0.9920	0.9920	0.9371	0.9371	0.9844	0.9844
Lv.2品詞	0.9370	0.9567	0.8693	0.8854	0.8134	0.9032
Lv.3語彙素	0.9273	0.9467	0.8576	0.8721	0.8023	0.8907
Lv.4発音形	0.9228	0.9421	0.8531	0.8616	0.7975	0.8787

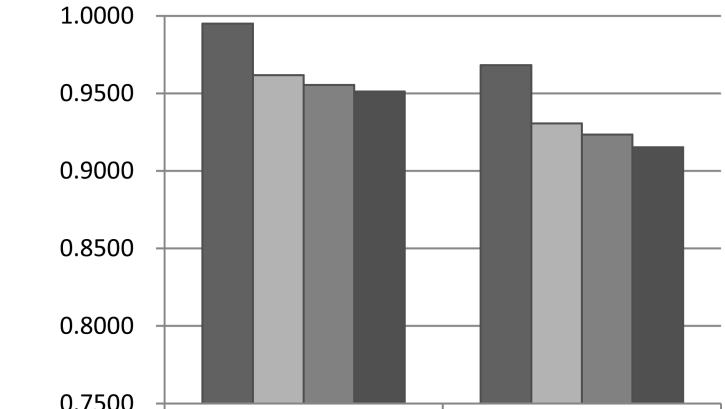
図 5.4: 既存の UniDic による旧仮名遣いテキストの解析精度

現代仮名遣いと旧仮名遣いが混在するテキストの解析精度

「旧仮名口語 UniDic」の実際の利用を考えた場合，青空文庫の多数のテキストを一括して形態素解析する場合など，現代仮名遣いと旧仮名遣いのテキストが混在しているデータに対して適用することが少なくないと考えられる．そこで，現代語用の UniDic と旧仮名口語 UniDic の二つの辞書について，現代語のテキストと旧仮名口語のテキストをほぼ同じ分量含めたテキストを対象として，仮名遣いが混在する場合の精度調査を行った．

全ての人手修正済みの現代語コーパスを現代語用の UniDic のコスト学習に利

用しているため，評価専用の現代語コーパスが入手できないため，今回は，辞書の学習に利用した，BCCWJ の書籍コアデータから約 25,000 語を評価に使用した．このデータは現代語用の UniDic と旧仮名口語 UniDic の両方で学習に利用したものである．評価コーパスは，旧仮名口語と現代語の書籍コアデータを合わせ，全体で約 48,400 語である．結果を図 5.5 に示す．数値はいずれも補正なしのものである．



	旧仮名口語UniDic	(現代語) UniDic
Lv.1境界	0.9949	0.9682
Lv.2品詞	0.9617	0.9307
Lv.3語彙素	0.9554	0.9234
Lv.4発音形	0.9513	0.9152

図 5.5: 現代仮名遣いと旧仮名遣いが混在するテキストの解析精度

図 5.5 から明らかなとおり，現代仮名遣いと旧仮名遣いが混在するテキストの場合であっても旧仮名口語 UniDic が現代語用の UniDic の解析精度を上回っている．実は，現代語のテキストについても旧仮名口語 UniDic は高い解析精度を示し，書籍コアデータの解析精度では現代語用の UniDic とほぼ同等の解析精度を示している．そのため，混在する場合にも旧仮名口語におけるアドバンテージがほぼそのまま反映される結果となっている．

現代仮名遣いテキストの評価データが，学習に利用されたデータであるため，今後より精確な調査が望まれるが，旧仮名口語 UniDic が，仮名遣いが混在する場合にも有効であることは確かである．

エラー分析

精度評価を踏まえ，旧仮名口語 UniDic ではどのような場合に誤りが多いのかエラーを調査した．

境界認定（Lv.1）のエラーでは，促音便が「つ」で表記される場合が目立つ．これは，活用形の整備不足により起きている．また，連体詞「その」「この」などを代名詞「そ」「こ」と格助詞「の」に誤ったものが多い．これは，近代文語と現代語とで単語認定基準にずれがあるために起きていることであり，学習用コーパスを事前に現代語側の基準にあわせておくことで避けられる問題である．その他は同表記の語が現れやすいという各見出し語固有の事情によるものが多い．このように境界認定のエラーは，今後の対応で改善が見込めるものが多かった．品詞認定（Lv.2）で問題となるエラーは，連体形と終止形を誤る例，未然形と連用形を誤る例が非常に多かった．この中には評価コーパス側に問題があった例も含まれている可能性がある．語彙素認定（Lv.3）で問題となるエラーは，同表記となる語がある「今日（きょう)」と「今日（こんにち)」,「昨日（きのう)」と「昨日（さくじつ)」,「門（もん)」と「門（かど)」などの対が目立った．発音形認定（Lv.4）で問題となるエラーは,「一（いち)」と「一（いっ)」などの数詞の誤りが目立つ．

全体として，語彙素認定と発音形認定のエラーは避けがたいものが多いが，境界認定と品詞認定のエラーは今後の対策で改善できる可能性がある．特に品詞認定では活用形の違いで大きく精度を落としていることから，活用形の問題を中心に対策を行う必要がある．

5.4　漢文訓読文と和漢混淆文の説話集

5.4.1　『日本霊異記』と『今昔物語集』の性格

中古末から現れる和漢混淆文は，漢語を多く含み漢文訓読由来の語法が目立つ点では近代文語に近い．しかし近代文語文は文法が簡略化され固定的な言い回しが多くなっている上に，両者には語彙的にも非常に大きな違いがあるため，別途辞書を作成する必要があると考えられる．

和漢混淆文の資料の中で大きな位置を占めるものとして，中古末から中世の間に書かれた説話集がある．その最大のものは平安時代末期に成立した『今昔物語集』（A.2.1 に例を示した）で，全 31 巻（ただし巻八，巻十八，巻二十一を欠く），1059 話という大冊である．

説話集は，和漢混淆文以外の文体で書かれたものもあり，平安時代初期に景戒によって書かれた仏教説話集『日本霊異記』（日本現報善悪霊異記）は，変則的な漢文で書かれている．また，鎌倉時代初期に成立した『宇治拾遺物語』（A.2.1 に例を示した）は，和文で書かれている．今日では，漢文で書かれた『日本霊異記』は，読み下した漢文訓読体の文章の形で古典文学作品として読まれており，小学館『新編日本古典文学全集』の本文も漢文訓読体の文章が載せられている（A.2.1 に例を示した）．国語研通時コーパスにも，この読み下し文が収録される．

『宇治拾遺物語』のような和文体の説話は「中古和文 UniDic」によって十分な精度で解析を行うことができるが，『今昔物語集』や『日本霊異記』は，既存の辞書では適切に解析を行うことができない．そこで新たな辞書が必要とされるが，『今昔物語集』と『日本霊異記』はともに仏教説話が中心で内容的に重なる部分が多いため，語彙的にも共通性が高い．また『今昔物語集』は先行する様々な説話集を原拠として成立しているが，元の文体を反映して漢文訓読体に近い文体で書かれた部分があるため『日本霊異記』の読み下し文とも共通性がある．こうした点から，『今昔物語集』と『日本霊異記』とを共通の辞書で扱いうると考えられる．

5.4.2　学習・評価用コーパス

『今昔物語集』と『日本霊異記』のテキストのうち，表 5.9 の説話について，中古和文 UniDic で解析したものを元にして，人手による修正を施したコーパスを約 1.4 万語整備した．このコーパスを学習・評価に利用して，『今昔物語集』と『日本霊異記』の解析のための辞書を作成することを考える．

なお，これらのコーパスは，2.2.2 節でみたように，漢字カタカナ交じり文を漢字ひらがな交じり文に直し，漢文の語順のままの部分を修正するなどの前処理を行ったうえで解析を行っている．

表 5.9: 『日本霊異記』『今昔物語集』の学習・評価用コーパス

作品名	出典	語数
日本霊異記	上巻 第十三	285
	上巻 第三十三	258
	中巻 第六	294
	中巻 第九	301
	中巻 第十六	907
	中巻 第二十二	661
	中巻 第三十一	349
	中巻 第三十八	198
	下巻 第二十五	612
	小計	3865
今昔物語集	巻第十二 於山階寺行涅槃会語第六	901
	巻第十二 和泉国尽恵寺銅像為盗人被壊語第十三	807
	巻第十二 紀伊国人漂海依仏助存命語第十四	1128
	巻第十二 河内国八多寺仏不焼火語第十八	373
	巻第十二 薬師寺食堂焼不焼金堂語第二十	1029
	巻第十二 関寺駈牛化迦葉仏語第二十四	1928
	巻第十二 奉入法華経筥自然延語第二十六	433
	巻第十二 書写山性空聖人語第三十四	2972
	巻第十二 天台円久於葛木山聞仙人誦経語第三十八	571
	小計	10142

5.4.3 見出し語の拡充

『今昔物語集』と『日本霊異記』の解析のためには，コーパスとともに，見出し語についても追加整備する必要がある．そのためにまず，それぞれのテキストを中古和文 UniDic で解析し「未知語」とされたものを中心に見出し語を追加した．

さらに，漢文訓読体のテキスト特有の送り仮名が省略された活用形を整備した．漢文訓読体では，次のように送り仮名が表記されない例が多いためである．

- 思ズ（おもはず） 未然形「思」
- 思テ（おもひて） 連用形「思」

- 思ベシ（おもふべし） 終止形「思」

これに対応して，表 5.10 のように未然形・連用形・終止形等の送り仮名省略形を，活用形展開で出力するようにした．

表 5.10: 追加した活用形の例（動詞：文語四段-ハ行-一般）

活用形	活用語形	活用書字形	新規追加分
未然形-一般	オモワ	思は	
	オモワ	思	○
連用形-一般	オモイ	思ひ	
	オモイ	思	○
連用形-ウ音便	オモウ	思ふ	
	オモウ	思	○
連用形-促音便	オモッ	思つ	
	オモッ	思ツ	
終止形-一般	オモウ	思ふ	
	オモウ	思	○
連体形-一般	オモウ	思ふ	
	オモウ	思	○
已然形-一般	オモエ	思へ	
命令形	オモエ	思へ	
意志推量形	オモオウ	思はう	
	オモオウ	思はふ	
ク語法	オモワク	思はく	

このような活用形を出力することは，「思」という表層形が，未然形「オモワ」にも，連用形「オモイ」にも，終止形「オモウ」にもなり得ることを意味し，活用語の曖昧性を高めることになる．ただし，こうした活用形は，パラメータ学習時に語彙化される助詞・助動詞が必ず後接することもあり，形態素解析において必ずしも多数のエラーにつながるわけではない．しかし，特に人手による修正時に混乱を招くことから，「形態論情報データベース」上で，使用年代・位相の情報を付与し，他の時代向けの辞書では原則として使わないこととした．

以上のように，従来の中古和文 UniDic・近代文語 UniDic で用いていた見出し語に，『今昔物語集』『日本霊異記』で新出の語と，表 5.10 の送り仮名省略活用形を追加したものを日本霊異記・今昔物語集のための辞書見出しとして整備した．以下，本節で述べる辞書ではすべてこの見出し語を共通で利用している．

5.4.4　漢文訓読文と和漢混淆文用の辞書の作成

共通辞書の作成

『今昔物語集』と『日本霊異記』の共通性に着目して，共通の辞書を作成するために，表 5.9 のコーパスのうち，評価用を除く全てのコーパスを利用して学習を行い，「霊異記・今昔共通辞書」を作成した．コーパスサイズは，霊異記・今昔の学習用コーパスを混ぜた約 14,000 語である．

表 5.11 にこの共通辞書の解析精度を示す．評価コーパスは表 5.9 のうち 1 割を文単位でランダムサンプリングしたものである．各レベルの意味は図 5.2 と同様，数値は F 値である．

表 5.11: 霊異記・今昔共通辞書の解析精度

	日本霊異記	今昔物語集
Lv.1	0.9795	0.9862
Lv.2	0.9328	0.9376
Lv.3	0.8974	0.9224
Lv.4	0.8787	0.9110

専用の辞書作成

前節では『今昔物語集』と『日本霊異記』の共通性に着目して，共通辞書を作成したが，両者はもともと和漢混淆文と漢文（を訓読したもの）という文体の違いをもち，成立年代にも大きな開きがある．

そこで，それぞれ別に専用の辞書を作成して，先の共通辞書と精度を比較することにする．見出し語は共通で，学習用コーパスを，それぞれの作品の一部からのみ取って，「今昔専用辞書」と「霊異記専用辞書」を作成した．コーパスサイズは，それぞれ表 5.9 に示したように『日本霊異記』が 3,865 語，『今昔物語集』が 10,142 語の 9 割であり，共通辞書と比較しても極めて小さい．

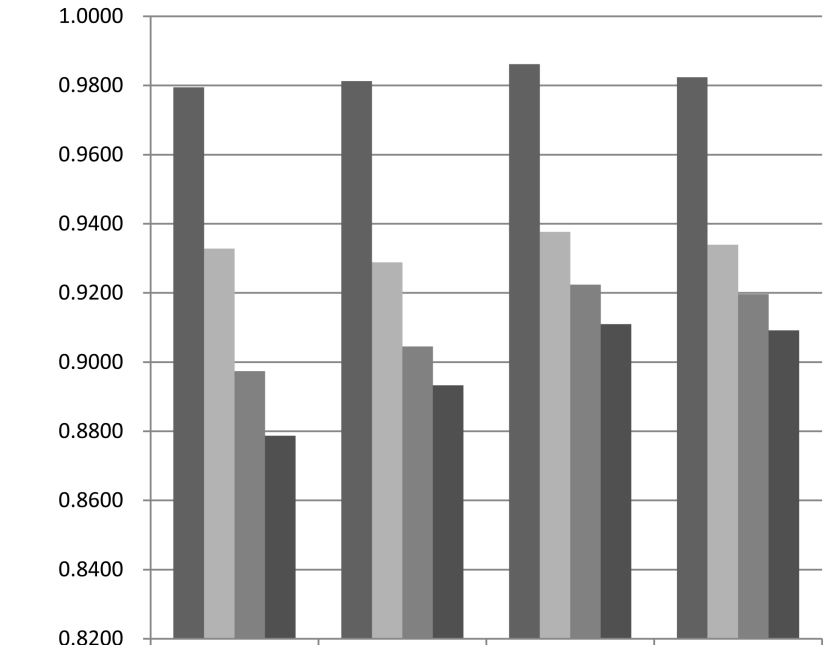
表 5.12 にこの共通辞書の解析精度を示す．評価コーパスは前節の表 5.11 と同一の学習対象外のものを対象としている．各レベルの意味は図 5.2 と同様，数値は F 値である．

表 5.12: 霊異記専用辞書・今昔専用辞書の解析精度

	日本霊異記	今昔物語集
Lv.1	0.9813	0.9824
Lv.2	0.9288	0.9339
Lv.3	0.9045	0.9196
Lv.4	0.8933	0.9092

5.4.5 各方法による辞書の解析精度

上記の「今昔・霊異記共通辞書」と「今昔専用辞書」「霊異記専用辞書」の解析精度を比較したものが図 5.6 である.



	共通辞書	専用辞書	共通辞書	専用辞書
	日本霊異記		今昔物語集	
Lv.1境界	0.9795	0.9813	0.9862	0.9824
Lv.2 品詞	0.9328	0.9288	0.9376	0.9339
Lv.3 語彙素	0.8974	0.9045	0.9224	0.9196
Lv.4 発音形	0.8787	0.8933	0.9110	0.9092

図 5.6: 各辞書による『日本霊異記』『今昔物語集』の解析精度比較

整備したコーパスは全体で約 14,000 語に過ぎないが，語彙素認定の F 値で 0.90 を超えており，今後コーパスの整備を行うことで十分な精度が達成できるものとが考えられる

『日本霊異記』については，専用のコーパスのみを用いることで精度が上がっているが，『今昔物語集』についてはわずかながら専用辞書で精度が低下している．学習・評価用コーパスのサイズが小さいため目安に留まるが，両者をまとめて一つの辞書で扱うことは妥当であると考えられる．

5.5 本章のまとめ

本章では，4 章で扱った中古和文・近代文語文のように大掛かりなコーパス整備を行うことのできない資料群について，形態素解析の実現に向けた取り組みを行った．

5.1 節で扱った，中世から近世にかけての口語文については，新たにコーパスからの学習を行って狂言用と洒落本用の形態素解析辞書を作成し，既存の辞書を上回る精度で解析を行うことが可能になった．また，近世口語を一括した共通辞書を作成する場合と，対象分野を分割した専用辞書を作成する場合とで精度の比較を行うことにより，狂言と洒落本とは別に扱うことが良いことが確認された．現在，通時コーパスプロジェクトの一環として，狂言コーパスと洒落本コーパスの構築を行っており，5.1 節の成果はこの構築に活かされている．

5.4 節で扱った『日本霊異記』と『今昔物語集』でも，5.1 節と同様に共通辞書と専用辞書を作成して，解析精度の比較を行った．学習用のコーパスサイズは 1.4 万語と小さいものの，語彙素認定の F 値で 0.90 を超える精度が出ており，人手修正を前提としたコーパス整備の出発点となった．今後，修正済みデータが蓄積されることにより，4.4.5 節で見たように，より高い精度での解析が可能になると考えられる．

5.3 節で扱った旧仮名遣いの口語文は，BCCWJ の現代語コーパスの一部と少量の専用コーパスをパラメータ学習に用いることによって，既存の辞書を上回る解析精度を実現した．今後，近代語コーパスの整備のために活用される．

以上のように，4 章の場合と比較して簡易な方法によって，多様なテキストの形態素解析を，通時コーパス構築のために必要となる精度で行うことが可能になった．

第6章　ユーザー向けツール

6.1　形態素解析を活用するためのツール

本章では，人文科学系の研究者を中心とするユーザに対し，歴史的資料の形態素解析やその結果の活用を促すためのツール類について論じる．

6.2 節では，形態素解析辞書を容易にインストールすることができるパッケージにまとめて配布し，グラフィカルな解析補助インターフェイスによる解析を可能にしたことについて述べる．

6.3 節では，形態素解析結果を総索引の形で出力するツールを開発し，総索引を容易に作成することを可能にしたことについて述べる．これにより，これまでの研究方法との連続性を保った形で形態素解析を応用した研究を行うことができる．

6.4 節では，コーパス管理ツール「茶器」に形態論情報を付与した古文のテキストをインポートし，個人レベルで検索や集計，データ修正を可能にする方法について述べる．

6.5 節では，形態論情報が付与された通時コーパスである「日本語歴史コーパス 平安時代編」と，これを Web 上で利用可能にしたコンコーダンサー「中納言」について述べる．「日本語歴史コーパス 平安時代編」は「通時コーパスの設計」プロジェクトで開発・公開されたコーパスの最初のものである．

6.2　形態素解析辞書の配布と解析補助 GUI

4 章で扱った中古和文 UniDic・近代文語 UniDic は，すでに Web 上で一般公開を行っている [1]．公開にあたっては，人文科学系の研究者にも利用しやすいように，インストーラーを付与したパッケージにまとめ，画面の指示に従うだけで容易にインストールすることができるように配慮した．

人文科学系の研究者への形態素解析技術の普及が遅れた原因の一つに，当初の形態素解析システムがコマンドラインの CUI でしか扱うことができず，彼らに

[1] http://www2.ninjal.ac.jp/lrc/index.php?UniDic

とって利用しづらい状況にあったことが挙げられる．たとえば，人文科学系の研究者の間では，今日でも 2000 年に公開された古い解析用 GUI アプリケーションである WinCha（ChaSen 2.1 の Windows 用パッケージ）が利用されている状況にあるが，これはコマンドプロンプト画面による CUI 操作を避けるためである．

そこで，Windows 用のパッケージでは，人文科学系の研究者が容易に解析を行うことができるようにインターフェイスを工夫した解析補助アプリケーション「茶まめ」を添付した（図 6.1）．「茶まめ」は，Javascript で実装された簡易なアプリケーションで，入力されたテキスト（またはテキストファイル）を前処理した後，内部的に MeCab を呼び出して UniDic を用いて形態素解析を行い，結果をファイルやアプリケーションに出力する．

このツールは, 文字コードの変換にも対応しており, Shift_JIS, UTF-16, UTF-8 などでエンコーディングされたテキストを辞書にあわせて UTF-8（ないし Shift_JIS）に自動変換する．これにより，テキストの貼り付けとマウスクリックだけで形態素解析を行うことが可能になっている．また，形態素解析の前処理として，漢字カタカナ交じり文の漢字ひらがな交じり文への変換など，簡単な前処理も行うことができる．

また，形態素解析結果の出力先として，その場に表示したりファイルに出力したりすることができるだけでなく，直接表計算ソフト（Excel）に書き出すことができるようにして利便性を高めている．

UniDic は，2.1.3 節の表 2.1 に示したように，単語に対して非常に多くの属性を出力することができるが，茶まめを用いた解析結果の出力は，このうち，表 6.1 に示した情報から成る表形式のテキストである．このうち「語種」を除く 8 属性[2] によって，UniDic の見出し語を一意に特定することができる．

「茶まめ」のような GUI は，人文科学系の研究者にとって形態素解析を身近なものにする価値を持っている．日本語史研究者の間においてはコンピュータ利用はこれまで盛んではなかったため，歴史的な資料の形態素解析を普及していくためには，こうした補助ツールの存在は欠かせないと考えられる．

6.3 総索引作成ツール [3]

1.1 節で述べたように，日本語・日本文学の研究，なかでも日本語の歴史的な研究において，用例検索のための総索引はたいへん重要な位置を占めている．歴史的資料を対象とした形態素解析が実用化されたことにより，かつては膨大な手作

[2]「語彙素」はハイフン区切りにより「語彙素細分類」を含む

[3]本節の内容は [81] にもとづく．

図 6.1: 解析補助アプリケーション「茶まめ」

表 6.1: 「茶まめ」が出力する形態論情報

属性名	説明
語彙素	辞書見出しの代表表記
語彙素読み	辞書見出しの読み（カナ）
語形	異語形を区別する形（カナ）
品詞	品詞（大-中-小分類）
活用型	活用型（活用語のみ）
活用形	活用形（活用語のみ）
書字形	テキストに出現した表記形
発音形	読み上げ用の形（現代読み）
語種	和語・漢語・外来語等の別

業を必要とした総索引の作成を半自動的に行うことが可能になった．しかし，一般的な日本語研究者にとって既存のソフトウェアを用いて総索引を作成することは容易ではない．そこで，形態素解析結果から容易に紙ベースの総索引を作成することのできるソフトウェアの開発を行った．対象資料の電子化テキストから，「近代文語 UniDic」「中古和文 UniDic」を利用した形態素解析（及び必要に応じた人手修正）を経て，文脈付き総索引の生成を可能にするものである．

6.3.1 日本語学と総索引

1.1 節および 2.1.1 節でも見たとおり，語の用例を見つけ出し，それに基づいて議論を行うことが研究の出発点となる日本語学において，語の検索のための総索引はたいへん重要である．特に文法性判断などで内省がきかない歴史的研究において，用例の実態調査は研究の基礎となる重要な役割を果たすものである．そのため，多くの文学作品や日本語史資料について総索引が作られ，研究に利用されてきた．日本文学研究においても総索引は研究に欠くことのできないものとして広く用いられている．

今日では主立った文学作品の総索引は出そろった感があるが，比較的マイナーな作品を対象としたものや，別系統の伝本にもとづくものなどはいまだ不十分であり，現在でも新たな総索引が作成・刊行されている．たとえば，6.3.3 節で取り上げる『鎌倉時代物語集成』（市古・三角 1988）[83] 所収の擬古物語の総索引の整備状況は，表 6.2 に示す通りであり，半数以上の索引については索引がない状況である．

表 6.2: 『鎌倉時代物語集成』所収作品の総索引の有無

作品名	総索引の有無
あきぎり	なし
あさぢが露	自立語索引あり
あまのかるも	総索引あり
在明の別	なし
石清水物語	総索引あり
いはでしのぶ	なし
風につれなき物語	なし
風に紅葉	語句索引あり
苔の衣	なし
木幡の時雨	総索引あり
恋路ゆかしき大将	なし

小夜衣	総索引あり
雫に濁る	なし
しのびね物語	なし
白露	なし
住吉物語	総索引あり
とりかへばや	総索引あり
兵部卿物語	なし
松陰中納言物語	なし
松浦宮物語	総索引あり
むぐらの宿	自立語索引あり
無名草子	総索引あり
八重葎	なし
(別本) 八重葎	なし
山路の露	総索引あり
夢の通ひ路物語	なし
夜寝覚物語	総索引あり
我身にたどる姫君	なし
雲隠六帖	なし
下燃物語	なし
豊明絵草子	なし
なよ竹物語	総索引あり
掃墨物語	なし
葉月物語	なし

資料の電子化・データベース化が進む中でも，これまでと同じ使い勝手を求める声は少なくなく，紙ベースの書籍形態の総索引の需要は大きいようである．近年における日本語研究分野の科研費採択課題から見ても，総索引を新たに作ろうとする試みが続けられていることがわかる[4].

初期に作成された総索引の多くは，文脈なしの自立語索引であったが，日本語史研究においては付属語も重要な調査対象である．そのため，今日では付属語までを含み，KWIC 形式の文脈を付与するものが一般化している．このような総索引を作成するに当たってはコンピュータを用いた何らかの手助けが必要とされる．文脈生成や見出し語のソート，整形については，レポート出力可能なデータベースソフトなどを用いることで対応できるが，一般的な日本語・日本文学の研究者にとってはハードルが高い作業である．また，単語に分割して品詞や読みなどを

[4]「科学研究費助成事業データベース」(http://kaken.nii.ac.jp/) によると，2009 年以降の日本語学・日本文学の研究課題だけでも索引作成を伴うものが 10 件以上確認できある

付与する作業は，人手によらなければならずたいへんな手間を要する．しかし，歴史的資料を対象とした形態素解析の実現により，その大部分を自動化することが可能になった．

中古和文 UniDic や近代文語 UniDic を用いることにより，単語分割や品詞・読みの付与といった，総索引作成の手間を大幅に減らすことが可能になる．4 章で見たとおり，歴史的資料を対象とした形態素解析システムの解析精度は 96%程度であり，そのまま総索引の元データとするには問題が残る．しかし，これに人手によるチェックと修正を加えることで，総索引作成用として十分な精度に高めることが可能である．

6.3.2 総索引作成ツール

近代文語 UniDic や中古和文 UniDic を利用して，日本語学の研究者が容易に総索引を作成することができるようにするために，UniDic の出力形式を入力として，最終的に PDF（ないし HTML）形式の総索引を出力とする総索引作成ツールを開発した．総索引作成の全体の流れは図 6.2 のようになる．

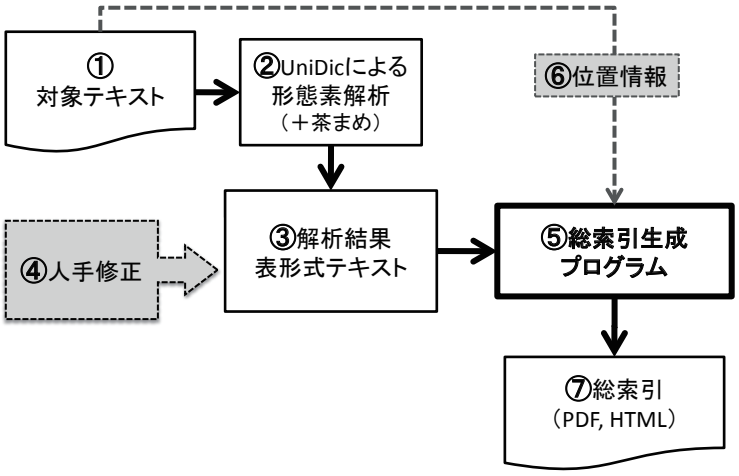


図 6.2: 総索引作成の流れ

入力となるテキストは，プレーンテキストまたは XML 形式のファイルである．XML 形式の場合も，タグ付けされた情報は後述する位置情報以外には用いない．タグを除去してプレーンテキスト化し，改行と句点を基準にして文を区切ったものを形態素解析への入力とする．

形態素解析そのものは，UniDic 標準の解析方法による．6.2 節で述べたとおり，UniDic には形態素解析を手助けするインターフェイスプログラム「茶まめ」が付属しており，コンピュータに不慣れなユーザでも容易に形態素解析を行うことが可能になっている．そこで，この「茶まめ」による出力を総索引作成ツールの入力とすることとした．

茶まめを用いた解析結果の出力は表 6.1 に示した情報が含まれるタブ区切りテキスト形式であるが，情報が多すぎて煩瑣になることを避けるため，このうち「語彙素」「語彙素読み」「語形」「品詞」「活用型」「書字形」「語種」を総索引に出力することとした．また印刷される結果には出力しなかったものの「活用形」を索引のソートに利用した．

先述したとおり，形態素解析結果をそのまま総索引に用いることは，精度の上で問題がある．したがって，本格的な索引を作成するためには，多くの場合には人手による修正を行う必要があるが，その際にも，上記の形式で出力すれば，このツールの入力とすることができる．したがって，解析結果をそのまま用いて全自動で総索引を作成するだけでなく，人手修正後のデータを用いて総索引を作成することも可能である．

形態素解析結果の修正

総索引作成ツールの入力である形態素解析結果はタブ区切りテキスト形式であり，必要に応じて表計算ソフトなどを用いて修正することができる．しかし，形態素解析結果だけを見て，テキストの他の箇所との整合性を保ちながら解析結果の修正作業を行うことは必ずしも容易ではない．テキストの形態素解析結果（コーパス）と形態素解析辞書の見出し語をデータベース上で関連づけ，辞書を参照しつつテキスト解析結果の修正を行うことができれば，比較的容易に修正作業を行うことができる．

次節（6.3.3）での作成例では，修正作業用のシステムとして 3 章で述べた「形態論情報データベース」を用いて実際に総索引を作成した．このデータベースシステムは国語研究所内部のコーパス構築用の環境であり，一般に利用可能なものではないが，後述する「茶器」（6.4 節）を用いることで，一般ユーザにもコーパス整備が可能な環境を構築することができる．

形態論情報の簡略化

UniDic における品詞は，たとえば普通名詞は「名詞-普通名詞-一般」のように，「大分類-中分類-小分類」の形で表される（B.2 参照）．しかし，一般の総索引利用

者にとっては，この長い形式は煩雑である．また，「大分類」や「中分類」だけを取り出した場合にも，品詞によっては，簡略に過ぎたり細かすぎたりといった問題が生ずる．そこで，索引を作成する段階で，この長い品詞を適切に簡略化させる機能を用意した．

UniDic の品詞と簡略品詞の対応表を設定ファイルとして用意し，これを読み込んでパターンマッチによって置き換えるものである．品詞対応表は単純なテキストファイルであり，索引の作成者が自由にカスタマイズすることが可能になっている．

このような簡略化は，活用型・活用形でも必要となる．UniDic の出力では，活用型は「文語四段-カ行」，活用形は「終止形-一般」などと，これも総索引の利用者から見た場合にはやや煩雑な長い形で出力されている（B.3 参照）．そこで，これらも品詞と同様の対応表によって簡略化を行っている．

本文中の位置情報

索引において，語の出現箇所は原文のテキストに戻って本文や注釈などを確認するために，非常に重要な情報である．したがって，総索引作成システムでは位置情報を出力できるようにする必要がある．

プレーンテキストを入力とした場合には，すべてのテキストが形態素解析の対象となるため，位置情報をタグ付けして残すことができない．テキストファイルから取得可能な位置情報としては，ファイル先頭からの文字数・文数・語数などが考えられる．文字数は数字が大きくなりすぎ扱いにくく，語数は形態素解析結果の修正によって変化してしまうため，文数を採用することとした．

XML ファイルの場合には位置情報をタグ付けしておくことができる．そこで，総索引作成システムのオプションで形態素解析対象となった XML ファイルを指定することにより，位置情報を索引に埋め込むことができるようにした．タグの形式としては，

```
<info position="位置情報" />
```

という簡単な空要素タグを参照して利用することとした．位置情報は，終了位置を表すものとし，索引には，おのおのの語から見て後方に最初に現れる位置情報を出力している．位置情報と語との関連づけは，ファイル中の文字位置によっている．

なお，位置情報については，ページ区切り，行区切りなどを階層化して入力することも考えられるが，現在のところ，簡潔さを優先して一つのタグの情報をそのまま出力するのみとなっている．

総索引の出力

総索引作成ツールは，整形した文脈付き総索引を出力する．文脈は，形態素解析結果の書字形（表層の出現文字列）を出現順に組み上げて生成する．見出し語は，読み（UniDic の「語彙素読み」）を基本に代表表記（同「語彙素」）・品詞・活用型をキーとして並び替え・グループ化し，アイウエオ順に出力する．

出力形式は，オプションにより，LaTeX（PDF）と HTML に対応する．LaTeX（PDF）形式は高品質な組み版・印刷に供するためのものであるが，LaTeX ファイルをコンパイルして PDF を出力するために Unicode に対応した LaTeX 環境が必要となる．一方，HTML 形式は特別な環境なしに容易に利用可能にするためのもので，CSS による簡易なデザインを施している．CSS の修正により，デザインを変更することも可能である．

総索引作成ツール GUI

以上の処理を行うための GUI ツールを作成した．その実行画面を図 6.3 に示す．これにより，コマンドラインを使用することなく総索引の作成処理が行えるようになった．

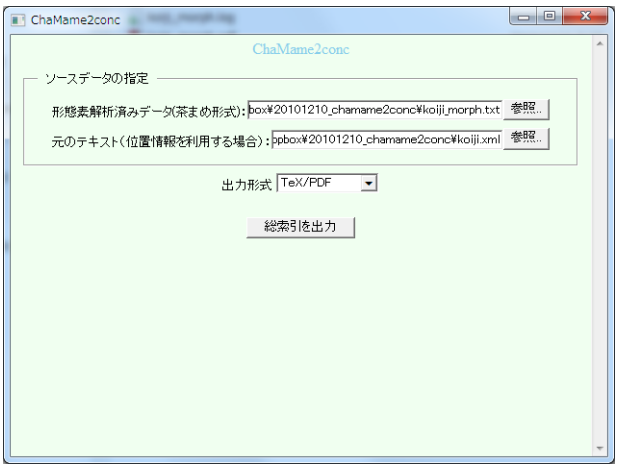


図 6.3: 総索引作成ツール

6.3.3　総索引の作成例 ―『恋路ゆかしき大将』文脈つき総索引

このツールを利用して，実際に索引サンプルを作成した．対象とした作品は，鎌倉期の擬古物語の一つ，『恋路ゆかしき大将』[5] である．この作品を含め，中世の物語作品は，総索引が作られていないものが多く，本システムの活用が期待されるジャンルのひとつであるといえる．

『恋路ゆかしき大将』総索引の作成は，本文を OCR によってテキストデータ化するとことから始めた．テキストデータ化の対象ページ数は計 117 ページで，二段組みの上段に本文，下段に現代語訳があるうちの上段のみをテキスト化した．文字数約 7.5 万，短単位語数 4.5 万（記号・句読点含む）である．

作業手順とそれに要した日数は表 6.3 に示したとおりである．作業者 1 名，1 日あたりの作業時間はおおむね 6 時間程度であった．こうして出来上がった総索引のサンプルを，図 6.4，および図 6.5 に示す．

表 6.3: 『恋路ゆかしき大将』総索引作成の作業日数

作業手順	作業日数
1. 書籍をスキャン	1 日
2. OCR による本文のテキストデータ化と校正作業	3 日
3. タグ付け，XML データ化	3 日
4. 「中古和文 UniDic」による形態素解析	（自動）
5. 形態素解析結果人手修正	16 日
6. 総索引作成ツールによる処理	（自動）
（計）	23 日

今回の総索引作成にあたっては，書籍形態の本文しか存在しない状態から開始した．実際には，本文の OCR 結果に対する校正をどの程度行うか，形態素解析結果にどの程度細かく修正を施すかなどによって作業時間は変わってくるが，およそ 3 週間で，約 4.5 万語のテキストの総索引が作成できたことになる．なお，今回は本文校正，形態素解析結果ともに，2 回以上の十分な確認作業を行っている．この作業時間は，手作業を多く必要とした従来の総索引作成に比較して大幅な短縮を実現しており，形態素解析により大きく省力化が進んだことになる．

形態素解析に用いた中古和文 UniDic は，主として平安時代の和文資料を対象としたものであるが，4.4.4 節で確認したとおり，鎌倉期以降の作品であっても，

[5] 作者不詳．本文は宮田 2004 [17] によった．

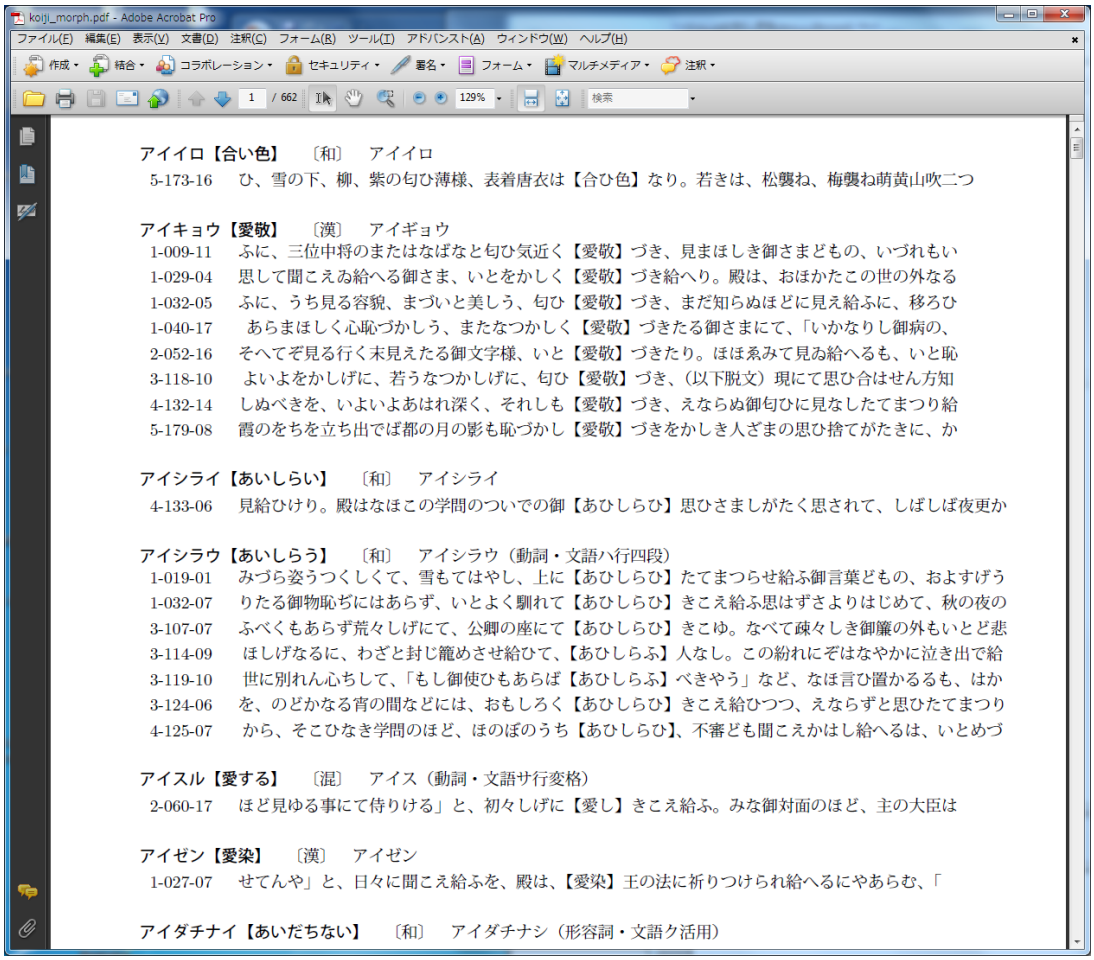


図 6.4: 総索引の例（PDF 版）

図 6.5: 総索引の例（HTML 版）

中古和文に類する文体の資料に対しては，ほぼ同程度の解析精度が期待できる．実際の作業においても，平安時代作品を解析した場合に比して不都合は感じられなかった．

歴史的資料を対象とした形態素解析の実現により，古典テキストの高度な利用が可能になった．しかし，日本語・日本文学の一般的な研究者にとって，形態素解析やデータベースの操作といった技術はいまだに敷居の高いものであって，なかなか利用が進んでいない．本ツールのように，新しい技術による成果を従来からある研究用の資源の形式で提供することによって，技術への理解とその普及を図ることも重要である．

6.4 「茶器」による通時コーパスの利用 [6]

4 章で見たように通時コーパス構築のために，歴史的な日本語を対象とした形態素解析が実用化されたが，一般の日本語史研究者が自ら資料を作成して形態素解析を行い，形態論情報付きのコーパスを作成するためには，その解析結果を研究者が利用するための環境が整備されなければならない．形態素解析によって単語ごとに分割され形態論情報が付与された巨大な表形式データは，多くの人文科学系の研究者にとってはそのままでは利用が困難であり，検索や集計のためのツールが必要とされる．しかし，3 章で見た「形態論情報データベース」は国語研究所内部のコーパス開発用システムであって，一般の研究者が利用できるものではなかった．コーパスを用いた日本語の歴史研究の発展のためには，日本語学の研究者が個人で容易に使うことのできるコーパス利用ツールが求められる．

2.3.2 節で見たように，古典語研究では，現代語と比較して量が限られた資料をもとに研究を進める必要があるため，高い精度でタグ付けされたコーパスが必要とされる．そのためには，人手による形態素解析結果の修正が柔軟に行えるプログラムが求められる．その一方で，多くの人文科学系研究者が利用可能なように手軽にパソコンにインストールして利用できるものである必要がある．

このようなニーズを満たすツールとして奈良先端科学技術大学院大学で開発されたコーパス管理ツール「茶器」[4] [7] がある．本節ではこの「茶器」に形態素解析済みの古典語のデータを格納して，日本語史研究で活用する方法について論じる．

[6] 本節の内容は [79] にもとづく．

[7] http://sourceforge.jp/projects/chaki/

6.4.1　コーパス管理ツール「茶器」

「茶器」は，上述の問題点を解消することが可能な汎用コーパス管理ツールであり，次のような特徴を備えている．

> 「茶器」は，タグ付きコーパスの検索および管理を支援する目的で作成されたツールである．文字列，単語列，および，係り受け関係による検索機能を備えている．単語列による検索では，単語の表層形以外に，読み，品詞や活用形などの文法情報を指定して検索を行うことができる．係り受け関係による検索では，文節内の単語列の指定と文節間の係り受け関係を指定した文検索が可能である．また，コーパス内の単語の頻度や前後文脈における単語の頻度など，簡単な統計処理を行うことができる．茶器は，タグ付きコーパスを関係データベースシステム（MySQL を使用）に格納し，検索要求を記述し結果を表示するためのインタフェースを提供する．対象言語は，多言語を目指しており，日本語，英語，中国語のデータを取り扱うことが可能である．
>
> （「茶器」使用説明書 version 2.1[49]）

「茶器」は，近年「ChaKi.NET」としてシステムが一新され，SQLite などの簡易なデータベースに対応したことによって，いっそう利用のしやすさを増している．SQLite の可搬なデータベースファイルを利用することで，タグ付きの古典語コーパスを広く配布して，研究者のパソコンでローカルに利用することができるようになった．

6.4.2　形態素解析済み古文コーパスのインポート

対応するデータ形式は MeCab ないし ChaSen による形態素解析結果と CaboCha [68] による係り受け解析結果であるが，標準の「Text2Corpus」機能により，指定したテキストファイルに対して形態素解析・係り受け解析を施し，インポートすることが可能になっている．文字コードは Unicode に対応しているため，古文で用いられる一般的には使用頻度の低い文字も取り扱うことが可能である．

Text2Corpus では，形態素解析に用いる辞書として UniDic を指定することができる．現代語用の UniDic の利用を想定した機能だが，4 章で開発した「中古和文 UniDic」の形態論情報は現代語用の UniDic と互換性があるため，辞書ファイルを差し替えることで，そのまま古文のテキストを解析し，インポートすることができる．

なお，茶器は CaboCha による係り受け解析結果を利用することが可能であるが，歴史的資料のためのモデルは存在しない．しかし，京大コーパスにもとづく標準の現代語用モデルを利用した解析結果であっても，人手による修正のベースとして利用する事は可能である．

図 6.6 は「茶器」に形態素解析済みの『土佐日記』をインポートし，タグ付けを行っている画面イメージである．

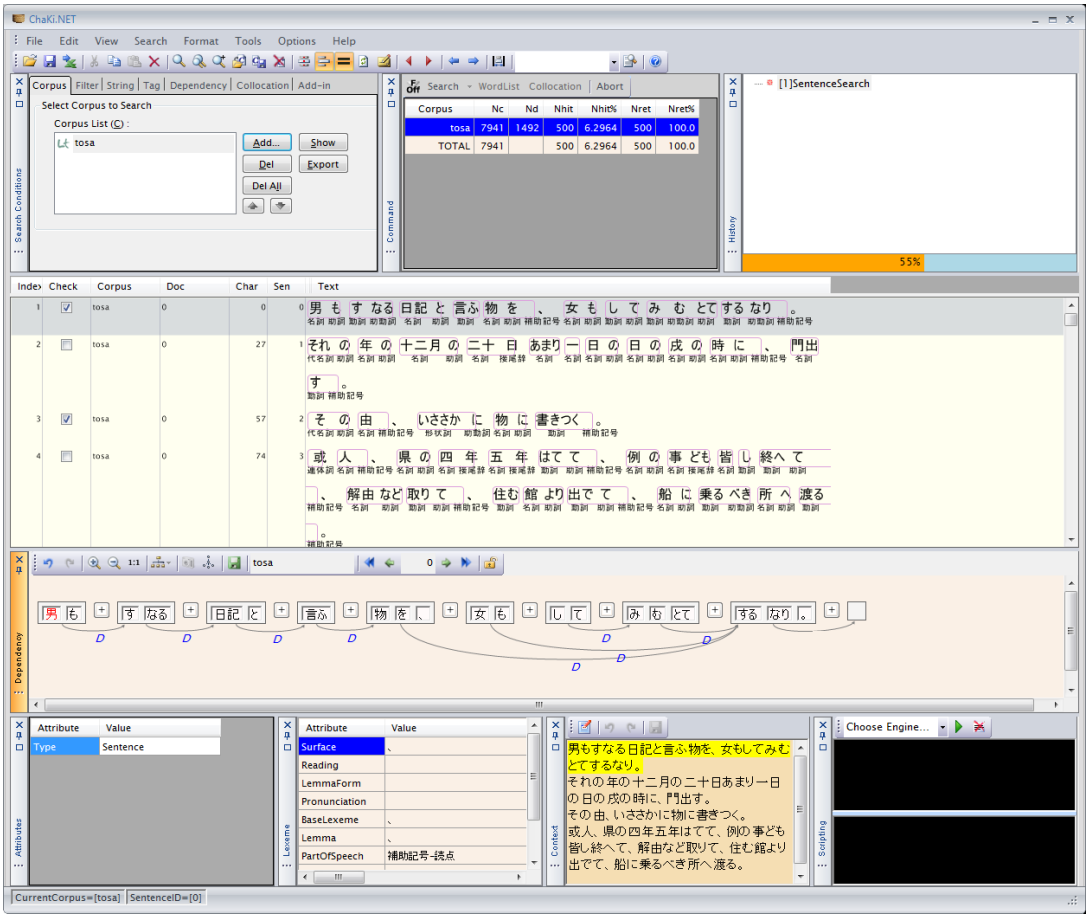


図 6.6: 「茶器」による歴史的資料の利用例（『土佐日記』）

歴史的資料に対応した UniDic と「茶器」により，日本語研究者が容易に形態素解析済みの古典語コーパスを利用する環境が整ったといえる．

6.4.3 「茶器」と形態論情報

「茶器」は，形態論情報として次の 9 属性を取り扱うことができる．括弧内は UniDic での対応する用語である（同一名称の場合は省略した）．

- Surface ＝表層形（書字形）
- Reading ＝読み（仮名形）
- LemmaForm ＝（語彙素読み）
- Pronunciation ＝発音（発音形）
- BaseLexeme ＝基本形の表層形（書字形基本形）
- Lemma ＝（語彙素）
- ParOfSpeech ＝品詞
- CType ＝活用型
- CForm ＝活用形

UniDic は，語種やアクセント型などの多様な情報を付与することができ，その属性数は合計 20 以上に上る（2.1.3 節・表 2.1）ため，上記の 9 属性に対応しない情報については，次のカスタムフィールドにすべてをまとめて次のように格納している．

- custom ＝ orth pronBase goshu iType iForm fType fForm kanaBase form formBase iConType fConType aType aConType aModType

6.4.4 古典語研究用ツールとしての利用

「茶器」の検索機能によって，単純な文字列検索はもちろんのこと，タグ付けされた形態論情報の品詞や活用型を検索キーとしたり，複数の検索語を組み合わせたりした自由度の高い高度な検索を行うことができる．検索結果は KWIC 形式で表示したり，外部にエクスポートしたりすることができる．さらにコロケーション[8] や語彙頻度表などの統計情報を出力することが可能になっている．

[8]「茶器」のコロケーション強度は，検索結果として表示中の KWIC データをコーパスの語数と見なして計算が行われるため，一般にコーパス言語学で用いられる，コーパス全体の語数を用いるものとは異なる点に注意が必要である．

電子化され検索可能な古典語資料はこれまでにもあったが，多くは総索引の電子版としての利用方法に限られ，「茶器」のような高度な検索や集計，統計情報の取得は行えなかった．古典語研究においてこのような処理が可能になったことで新たな発見が期待される．そのための主要な機能は，正規表現を利用した文字列検索，タグ情報検索，共起検索，ワードリスト，そして係り受け検索である．

文字列検索（StringSearch）

「茶器」画面右上の検索条件指定パネル（SearchCondition パネル）で，さまざまな方法での検索を行うことができる．文字列検索は中でももっとも単純なものだが，「茶器」では正規表現を利用した検索を行うことができる．一般にコーパス検索ツールでは使用できる正規表現に制限があることが多いが，「茶器」では Perl 5 互換の強力な正規表現が利用できる．

語検索（TagSearch）

語検索は形態素解析によって付与された語のタグ情報を利用して検索を行うものである．先述の 9 属性を自由に組み合わせて，検索に利用することができる．それぞれの項目で正規表現による指定が利用可能である．UniDic の見出し語は，語彙素・語形・書字形・発音形の四つのレベルに階層化されているため，調査対象に合わせて選択することで，有効な検索ができる．さらに，複数の語を組み合わせ，共起条件を設定した検索も可能である．

ワードリスト検索（WordList）

検索条件を指定した後に，コマンドパネルの WordList コマンドを利用することで，条件を満たした語の集計を行うことができる．表 6.4 は，ワードリスト検索を使って完了の助動詞「ぬ」「つ」の前 2 語以内に来る動詞を用例数の多いものから順にそれぞれ 28 位までリストアップしたものである．

助動詞「つ」「ぬ」の使い分けに動詞の種類が関わっていることは広く知られているが，その違いは「茶器」の検索ですぐに得られるリストによって確認することができる．

表 6.4: 助動詞「つ」「ぬ」の上接動詞

ぬ		つ	
給う	657	給う	260
成る	476	侍る	105
侍る	223	為る	82
有る	112	思う	80
過ぎる	111	見る	78
出でる	92	聞こえる	73
止む	77	有る	63
果てる	71	奉る	62
経る	63	過ぐす	39
参る	44	見える	25
為る	43	思す	24
罷出づ	34	宣う	20
思す	32	言う	20
返る	30	初める	19
泣く	29	果てる	19
出で来る	29	成す	17
更ける	25	取る	15
思う	24	聞く	15
おわします	22	遣る	14
居る	22	申す	14
おわす	21	置く	13
止まる	20	許す	13
思し成る	20	捨てる	13
奉る	19	来る	13
隠れる	19	変える	12
然り	18	止す	12
初める	17	渡る	12
思い成る	17	渡す	12

係り受け検索（DependencySearch）

検索条件パネルの Dependency タブにより，文節の係り受け関係を条件に指定した検索を行うことができる．先述したとおり，現在は古典語の係り受け解析は開発途上であるため，現時点では十分なデータが利用できないが，将来的にはこの機能を用いることで，単に隣接しているだけではなく，係り受け関係にある語を検索することが可能になる．日本語研究でコーパスを利用する場合，意図しない不要な用例が含まれるのを承知で検索し，その検索結果を研究者が選別する「ゴ

ミ取り」作業が大きな負担になっていたが，係り受けまで整備されたコーパスが用意できれば，こうした手間が軽減できる．

以上のようなコーパスの検索や集計では，データのサイズと処理速度が問題になるが，古典語のコーパスのサイズは限られているため，快適に利用することが可能である．検証のために用意した 58.8 万語のコーパスでは，上記のうちもっとも時間のかかる検索であっても 15 秒以内に取得することができた[9]．

このように「茶器」の機能を用いることで，これまでには行えなかった新しい視点からの古典語研究が可能になったと言える．

6.4.5　タグ付けツールとしての利用

2.3.2 節で述べたとおり，古典語コーパスの研究では，コーパスのタグ付け精度は現代語以上に高い精度が求められる．しかし，自動形態素解析だけでその精度を実現するには困難であるため，自動解析結果を人手で修正し，高精度なデータを用意する必要がある．

「茶器」を用いることで，コーパスのタグ付け・修正を行うことができるため，このような形態素解析の誤り修正のために利用することができる．修正用の辞書見出し語は，インポートしたコーパスから自動生成されているので，既出の語であれば正しい見出し語を選択するだけで解析結果の修正を行うことができる．

また，現状では係り受けまでタグ付けされた古典語のコーパスは存在しないが，「茶器」を用いることで，古典語コーパスに対する係り受けのタグ付けを行うことができる．図 6.7 は，「茶器」の文節係り受けのアノテーション画面（Dependency パネル）で，実際に『源氏物語』桐壺巻の冒頭の 1 文について，文節係り受けのアノテーションを行ったものである[10]．

古典語のコーパスへの係り受けのタグ付けが実現すれば，より高度なコーパス利用が可能になり，古典語研究において新しい発見を導く可能性がある．

文節係り受けのタグ付けは，文法判断に内省がきかない古典語の場合には，現代語と比べて遙かに難しい．しかし，現代語とは異なり，古典語のデータは量が限られているため，十分な時間をかけることができれば，今後，コーパス全体に対してタグ付けを行うことも可能だと思われる．

[9] Windows7 x64, Core i7 2.53GHz, 8GB RAM の環境で確認した．

[10] この文節係り受けアノテーション例は，冨士池（2012）[89] の試行結果によるものである．

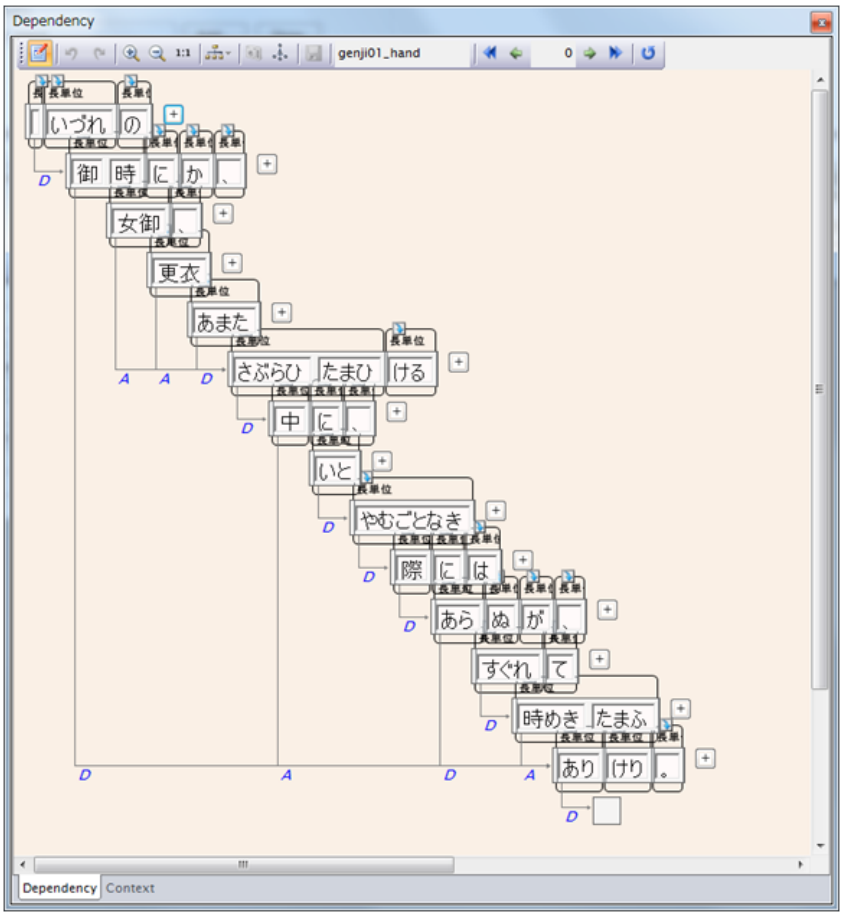


図 6.7: 文節係り受けのアノテーション（『源氏物語』冒頭）

6.5 「日本語歴史コーパス」と「中納言」[11]

6.5.1 「日本語歴史コーパス」先行公開版の概要

「日本語歴史コーパス」（Corpus of Historical Japanese, 以下 CHJ）は，国立国語研究所「通時コーパスの設計」プロジェクトで構築が進められている日本語史研究を目的としたコーパスである．そのうち「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．2012 年に公開された先行公開版では，次の 10 作品のデータが利用可能である[12]．

> 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

本文は，許諾を得て小学館「新編日本古典文学全集」（新編全集）[38] を利用している．ただし，コーパス化の対象は原文のみで，現代語訳等は含まない．

収録した本文データには，4 章で述べた「中古和文 UniDic」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ている．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．中でも『源氏物語』については発話者の情報も付与されている．

「CHJ 平安時代編」のテキストの量は，表 6.5 に示す通りである．全体で約 79 万語であり，うち 65%に近い 51 万語を『源氏物語』が占めている．

CHJ には中古和文 UniDic にもとづく形態論情報が付与されているため，2.1.3 節で述べた短単位の長所を活かした利用を行うことができる．CHJ 先行公開版で公開中のデータは短単位のみであるが，今後，BCCWJ と同様に長単位の情報も付与したうえで本公開が行われる予定である．

6.5.2 「日本語歴史コーパス」中納言

CHJ の公開は現在，インターネット上で利用するウェブ版のコンコーダンサー「中納言」（図 6.8）で行っている．「CHJ 中納言」は，CHJ むけに若干の修正を

[11] 本節の内容は [82] にもとづく．

[12] 2013 年度末までに，CHJ 平安時代篇の完成版が公開される予定である．完成版では，先行公開版の 10 作品に「平中物語」「堤中納言物語」「更級日記」「讃岐典侍日記」を加えた 14 作品の短単位と長単位のデータが利用可能になる．

表 6.5: 「日本語歴史コーパス 平安時代編」先行公開版の作品別語数

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

※短単位による．記号を含む．

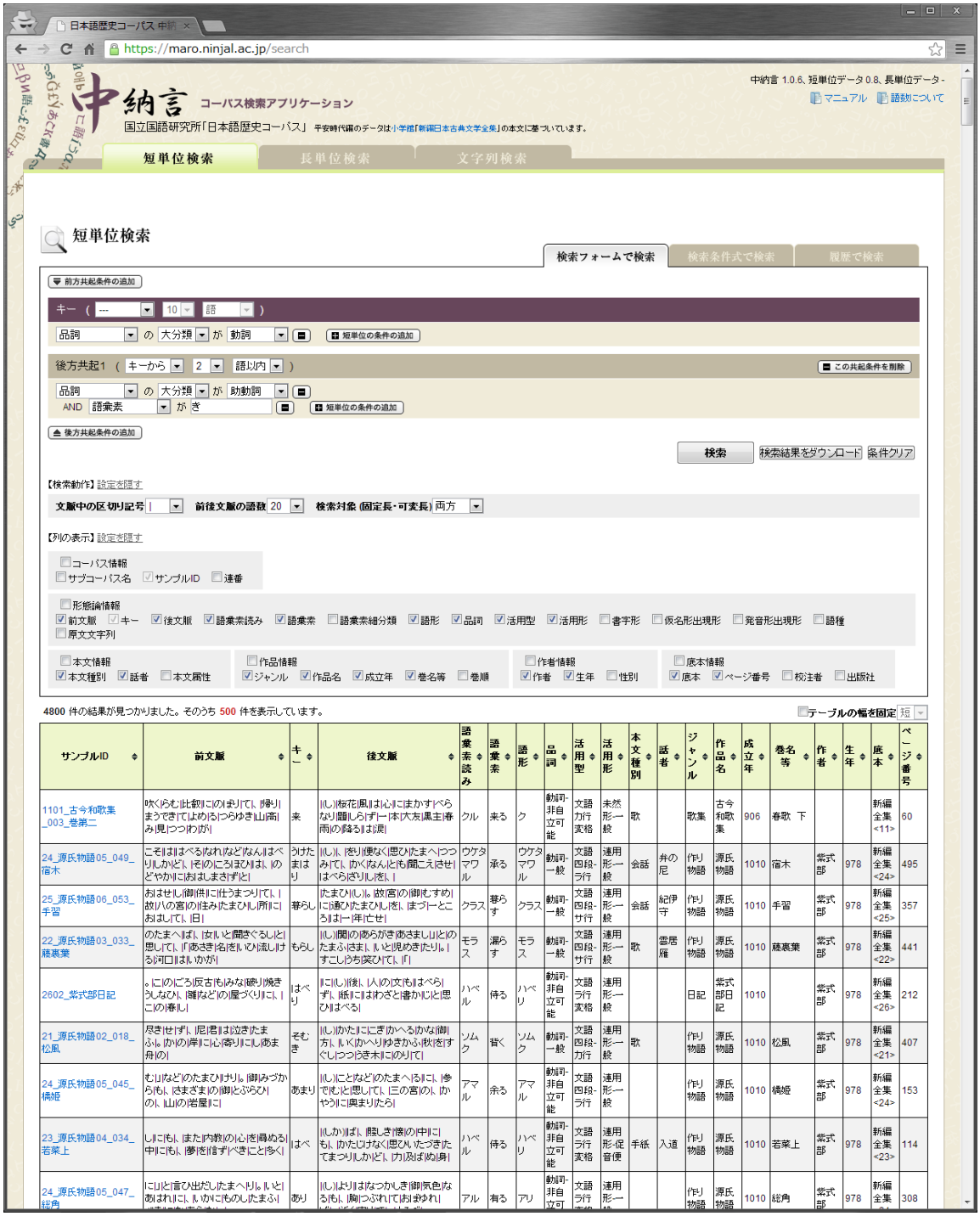
行っているが，基本的にBCCWJで利用されている「中納言」と同じものである．書面による申込み手続きを経ることで無償で利用できる[13]．

検索条件の指定

「中納言」では，上述した形態論情報を用いることで，従来の索引と同様に活用形や異表記にとらわれない見出し語による検索が可能であるが，単にそれだけでなく，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．また，他の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．さらに，複数の語（最大10語）を組み合わせた検索を行うことができため，「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．図6.9は，こうした検索条件指定を行う画面で，助動詞「ぬ」の直前2語以内に来る動詞を検索している例である．

画面上で行った検索条件は，内部的に「検索条件式」と呼ぶ式に変換された後に検索が行われる．検索条件式は検索履歴として保存されており，再利用するこ

[13] http://www.ninjal.ac.jp/corpus_center/chj/



サンプルID	前文脈	キー	後文脈	語彙素読み	語彙素	語形	品詞	活用型	活用形	本文種別	話者	ジャンル	作品名	成立年	巻名等	作者	生年	底本	ページ番号
1101_古今和歌集_003_巻第二	吹くからむ|比叡|に|の|ほり|て|、|帰り|まうできて|よめ|る|つらゆき|山|高|み|見|つつ|わが|	来	|ん|桜花|風|は|心|にまかすべからなり|題しらず|ー|本|六友|黒主|春雨の|降る|は|涙|	クル	来る	ク	動詞-非自立可能	文語カ行変格	未然形一般	歌		歌集	古今和歌集	906	春歌 下			新編全集<11>	60
24_源氏物語05_049_宿木	こそ|は|べるなれ|など|な|ん|は|べり|し|か|ど|、|その|ころ|ほど|り|は|、|の|どやか|に|おはし|まさ|ず|と|	うけたまはり	|ん|、|を|り|便なく|思ひたまへ|つつみ|て|、|かく|なん|とも|聞こえさせ|は|べら|ざり|し|を|、|	ウケタマワル	承る	ウケタマワル	動詞-一般	文語四段-ラ行	連用形一般	会話	弁の尼	作り物語	源氏物語	1010	宿木	紫式部	978	新編全集<24>	495
25_源氏物語06_053_手習	おはせし|御|供|に|仕うまつり|て|、|故|八の宮|の|住みたまひし|所|に|おはし|て|、|旧|	尋らし		たまひ|ん|し|ん 故|宮|の|御|むすめ|に|通ひたまひ|し|を|、|まづ|ー|とこ|ろ|は|ー|年|亡せ|	クラス	尋ぬ	クラス	動詞-一般	文語四段-サ行	連用形一般	会話	紀伊守	作り物語	源氏物語	1010	手習	紫式部	978	新編全集<25>	357
22_源氏物語03_033_藤裏葉	のたまへ|ば|、|女|いと|聞きぐるし|と|思し|て|、|『|あさき|名|を|いひ|流しけ|る|河口|は|い|かが|	もらし		ん|、|関|の|あらがき|あさまし|山|と|のたまふ|さま|、|いと|児めきたり|。| すこし|うち|笑ひ|て|、|『|	モラス	漏らす	モラス	動詞-一般	文語四段-サ行	連用形一般	歌	雲居雁	作り物語	源氏物語	1010	藤裏葉	紫式部	978	新編全集<22>	441
2602_紫式部日記	。|に|の|ごろ|反古|も|みな|破り|焼き|うしなひ|、|雛|など|の|屋づくり|に|、|こ|の|春し|	はべり	|に|し|、|後|、|人|の|文|も|はべら|ず|、|紙|に|は|わざと|書か|に|と|思ひ|は|べる|	ハベル	侍る	ハベリ	動詞-非自立可能	文語ラ行変格	連用形一般			日記	紫式部日記	1010		紫式部	978	新編全集<26>	212
21_源氏物語02_018_松風	尽き|せ|ず|、|尼|君|は|泣きたまふ|。|か|の|岸|に|心|寄り|に|し|あま舟|の|	そむき		し|、|か|た|に|に|ぎ|か|へる|かな|御|方|、|く|か|へり|ゆき|かふ|秋|を|すぐし|つつ|ら|き|木|に|の|り|	ソムク	背く	ソムク	動詞-一般	文語四段-カ行	連用形一般	歌		作り物語	源氏物語	1010	松風	紫式部	978	新編全集<21>	407
24_源氏物語05_045_橋姫	む|山|などのたまひ|け|り|。|御|みづから|も|、|はま|さ|の|御|と|ぶらひ|の|、|山|の|岩屋|に|	あまり		し|、|こと|などのたまへ|る|に|、|参でむ|と|思し|て|、|三の宮|の|、|かやう|に|奥まり|たら|	アマル	余る	アマル	動詞-非自立可能	文語四段-ラ行	連用形一般			作り物語	源氏物語	1010	橋姫	紫式部	978	新編全集<24>	153
23_源氏物語04_034_若菜上	し|に|も|、|また|内|教|の|心|を|尋ぬる|中|に|も|、|夢|を|だ|ず|べき|こと|多く|	はべ		し|か|ば|、|限|し|き|懐|の|中|に|も|、|かたじけなく|思ひ|、|たづきたてまつり|し|か|ど|、|内|及ば|ぬ|身|	ハベル	侍る	ハベリ	動詞-非自立可能	文語ラ行変格	連用形-促音便	手紙	入道	作り物語	源氏物語	1010	若菜上	紫式部	978	新編全集<23>	114
24_源氏物語05_047_総角	に|し|と|言ひ出だしたまへ|り|。|い|と|あはれ|に|、|い|かに|も|のしたまふ|	あり		し|より|は|なつかしき|御|気色|なる|も|、|胸|つぶれ|て|おし|ゆれ|	アル	有る	アリ	動詞-非自立可	文語ラ行	連用形一			作り物語	源氏物語	1010	総角	紫式部	978	新編全集	308

図 6.8: 「日本語歴史コーパス」中納言

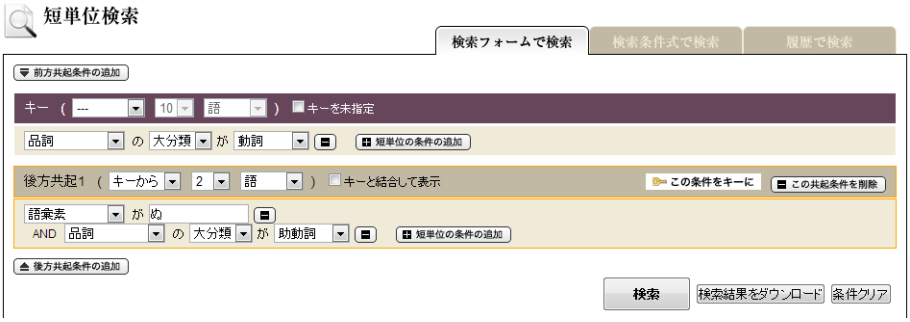


図 6.9: 形態論情報を使った検索条件指定（中納言）

とができる．また，検索条件式を論文等に記述することにより研究の再現性を高めることができる．次の例は，上述した助動詞「ぬ」の直前 2 語以内に来る動詞を検索する検索条件式の例である．

```
キー: 品詞 LIKE "動詞%"
AND 後方共起: (語彙素 = "ぬ" AND 品詞 LIKE "助動詞%")
ON 2 WORDS FROM キー  WITH OPTIONS unit="1"
AND tglWords="20" AND tglKugiri="|" AND tglFixVariable="2"
```

なお，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って検索することも可能である．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 6.6 に示す項目が表示可能である [14]．

「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．

[14] 表中の※は標準では非表示になっているもので，画面上のチェ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．

表 6.6: 検索結果表示項目（中納言）

コーパス情報	レジスター (コーパス名) ※, サンプル ID, 連番※
KWIC	前文脈, キー (書字形出現形), 後文脈
形態論情報	語彙素読み, 語彙素, 語彙素細分類※, 語形, 品詞, 活用型, 活用形, 書字形※, 仮名形出現形※, 発音形出現形※, 語種※, 原文文字列※
本文情報	本文種別, 話者, 本文属性※
作品情報	ジャンル, 作品名, 成立年, 巻名等, 巻順※
作者情報	作者, 生年, 性別※
底本情報	底本, ページ番号, 校注者※, 出版社※

「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．

検索結果表示画面の一部を図 6.10 に示す．

検索結果のダウンロード

検索結果は，画面上では 500 例までしか表示されないが，全例（最大 10 万例）をダウンロードすることができる．この検索結果を表計算ソフトに読み込むこと

サンプルID	前文脈	キー	後文脈	語彙素読み	語彙素	語形	品詞	活用型	活用形	本文種別	話者	ジャンル	作品名	成立年	巻名等	作者	生年	底本	ページ番号
22_源氏物語03_031_真木柱	、出だしたてまつるいでくもあらげ。」男君たち、「十なるは殿上したまふ、いと	うつくし	。」人にほめられて、宿直などようはあらねど、いと心うううつう、ものの	ウツクシイ	美しい	ウツクシ	形容詞一般	文語形容詞シク	終止形一般			作り物語	源氏物語	1010	真木柱	紫式部	978	新編全集<22>	379
18_枕草子_145_一四五_うつくしきもの	池より取りあげたる。」雛のいと小さき。」何も何も、小さきものは、みな	うつくし	。」いみじう泣く児えたる、ちごの、二つばかりなるが、二藍の直衣など、	ウツクシイ	美しい	ウツクシ	形容詞一般	文語形容詞シク	終止形一般			随筆	枕草子	1001	うつくしきもの	清少納言	966	新編全集<18>	271
21_源氏物語02_018_松風	やうやう心ちとけて、もの言ひ笑ひなどして暗れたまふを見るままに、「ほまれさりて	うつくし	。」抱きてはおはするほまし、見るかひありて、宿世によなしと見えたり。」 またの	ウツクシイ	美しい	ウツクシ	形容詞一般	文語形容詞シク	終止形一般			作り物語	源氏物語	1010	松風	紫式部	978	新編全集<21>	415
1701_落窪物語_003_巻之三	も、勝時たまへるかともみよ」、いとなまめかしういらへて、いと白う、	うつくし	げなる子の、三つばかりなるを、懸けにすゑて、うつくしがり遊たまへれ	ウツクシイ	美しい	ウツクシ	形容詞一般	文語形容詞シク	語幹-一般			作り物語	落窪物語	986	巻之三			新編全集<17>	225

図 6.10: 検索結果表示画面の一部（中納言）

で，自由に集計を行うことができる．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 6.6 の全ての項目が含まれている．さらに最終列に「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．

6.6 本章のまとめ

本章では，人文科学系の研究者を中心とするユーザに対し，歴史的資料の形態素解析の活用を促すためのツール類について述べた．具体的には，以下の 4 点である．

- 形態素解析辞書を容易にインストールすることができるパッケージにまとめ，グラフィカルなインターフェイスと共に配布すること
- これまでの研究方法との連続性を保つため，形態素解析結果を総索引の形で出力すること
- 個人で利用可能なコーパス管理ツールにインポートして高度な検索や集計，データ修正を可能にすること
- コンコーダンサーを研究者向けのサービスとして展開すること

いずれも，本研究によって実現した歴史的資料の形態素解析を応用して，コーパスに基づく新しい日本語史研究を進めていくために必要とされるものである．さらなる普及のためには，今後，こうしたツール類を解説するチュートリアルなどを通して研究者に広めていく必要がある．

第7章　結論

7.1　本論文の成果

本研究は，日本語通時コーパスのための形態論情報アノテーションを実現するために自然言語処理技術を応用して，次の貢献を行った．

1. 古文の形態素解析を実現するための言語資源として，新たに辞書と学習用のコーパスを整備し，統計的機械学習にもとづく形態素解析技術を用いて，中古和文と近代文語文について実用的な精度（見出し語認定のF値で0.96以上）が得られる形態素解析システムを実現した．

2. 上記の言語資源と国語研通時コーパス自体の整備のために，辞書の見出し語とコーパスの出現形とを関連付けながら形態論情報の修正作業を行うことのできるデータベースシステム（国語研究所「形態論情報データベース」）を構築し，通時コーパス整備の基盤を整えた．

3. 国語研通時コーパスに収録される多様なテキストに対して高い精度で形態論情報のアノテーションを行う方法を検討し，近世口語文，和漢混淆文，旧仮名遣いの口語文について，実際に形態論情報のアノテーションを行った

4. 上記の形態素解析技術や形態論情報付きの通時コーパスを人文科学系の研究者に使いやすい形で提供するために，新たなツールの作成・既存のツールの適用を行った．

以上により，通時コーパス構築の基盤を整備し，通時コーパスを用いた日本語史研究のための環境を提供した．

日本語史研究というこれまでに自然言語処理の応用が少なかった分野において，今日の技術を応用して言語資源を構築することにより，コーパスを用いた日本語史研究を発展させるための基礎を作り，自然言語処理の応用の場を広げるという形で，二つの分野において裨益するものになったものと考える．

7.2 展望

本研究により，通時コーパスに必要な形態論情報アノテーションのうち，短単位についてはほぼ実現の目処が立ったと言える．ただし，テキストによっては改善の余地が残る．例えば，会話文と地の文で解析用の辞書を切り替えるなどの対応により解析精度を向上させられる可能性がある．

また，本研究を通して資料のタイプに応じたたくさんの辞書を作成することになった．現段階では，どの辞書がどの資料に有効であるかが比較的はっきりしているが，今後多様な資料を扱っていく中では，性格が不明確な資料に対して，適切な辞書をいかに選択するかという問題が生じる可能性がある．また，複数の辞書の中間的な文体の資料に対して，中間的な辞書で対処することが望ましい場合が想定される．こうした問題に対処するためには，これまでに用意してきた辞書見出しと修正済みのコーパスと時代文体情報を統合して，柔軟な対応を行えるシステムを構築することが有効であろう．今後の課題としたい．

また，原文が万葉仮名で書かれている上代語の資料や，ローマ字で表記された中世のキリシタン文献では，形態素解析の前処理やデータベースシステムで新たな対応が必要となる．このためには，原表記を保存した上で，本文を一般的な仮名表記に置き換えて解析を施し，「形態論情報データベース」上に新たに原表記の文字のテーブルを用意して，原表記と形態素解析結果とを関連付けるといった対処が必要になると考えられる．これも今後の課題である．

本研究では直接扱わなかった長単位・文節境界のアノテーションについては，中古和文ですでに解析が実現しているが，近代文語文をはじめとする漢語を多く含む文体では未着手であり，規定の整備を含め，これから新たに取り組む必要がある．

形態論情報アノテーション以外における，自然言語処理技術の通時コーパスへの応用としては，文節係り受けのアノテーションや文脈に即した意味情報のアノテーションなどが考えられる．こうした高度なアノテーションには様々な困難があり，大きな労力が必要となると予想されるが，資料が有限である通時コーパスでは不可能な課題ではない．今後の日本語史研究の発展のために検討されるべき価値があるだろう．

付録 A　通時コーパスのテキスト例

2.2 節で示した，通時コーパスに含まれるテキストの例を挙げる．

A.1　中古和文

A.1.1　歌物語

伊勢物語「東下り」（新編全集 12 巻 p.122）
〈校訂済み，振り仮名付き〉

富士の山を見れば、五月(さつき)のつごもりに、雪いと白うふれり。

時しらぬ山は富士の嶺(ね)いつとてか鹿子(かのこ)まだらに雪のふるらむ

その山は、ここにたとへば、比叡(ひえ)の山を二十(はたち)ばかり重(かさ)ねあげたらむほどして、なりは塩尻(しほじり)のやうになむありける。

なほゆきゆきて、武蔵(むさし)の国と下(しも)つ総(ふさ)の国とのなかにいと大(おほ)きなる河あり。それをすみだ河といふ。その河のほとりにむれゐて、思ひやれば、かぎりなく遠くも来(き)にけるかな、とわびあへるに、渡守(わたしもり)、「はや船に乗れ、日も暮れぬ」といふに、乗りて渡らむとするに、みな人ものわびしくて、京に思ふ人なきにしもあらず。さるをりしも、白き鳥の、はしとあしと赤き、鴫(しぎ)の大きさなる、水の上に遊びつつ魚(いを)を食ふ。京には見えぬ鳥なれば、みな人見しらず。渡守に問ひければ、「これなむ都鳥(みやこどり)」といふを聞きて、

名にしおはばいざ言問(ことと)はむみやこどりわが思ふ人はありやなしやと

とよめりければ、船こぞりて泣きにけり。

A.1.2　作り物語

竹取物語「石上の中納言と燕の子安貝」（新編全集 12 巻 p.49）

〈校訂済み，振り仮名付き〉

中納言石上麿足の、家に使はるる男どものもとに、「燕の、巣くひたらば告げよ」とのたまふを、うけたまはりて、「何の用にかあらむ」と申す。

男ども答へて申す、「燕をあまた殺して見るだにも、腹になき物なり。ただし、子をうむ時なむ、いかでかいだすらむ、侍んなる」と申す。「人だに見れば、失せぬ」と申す。

また、人の申すやう、「大炊寮の飯炊く屋の棟に、つかの穴ごとに、燕は巣をくひはべる。それに、まめならむ男ども率てまかりて、足座を結ひあげて、うかがはせむに、そこらの燕子うまざらむやは。さてこそ取らしめたまはめ」と申す。中納言よろこびたまひて、「をかしきことにもあるかな。もつともえ知らざりけり。興あること申したり」とのたまひて、まめなる男ども二十人ばかりつかはして、麻柱にあげ据ゑられたり。

殿より、使ひまなく賜はせて、「子安の貝取りたるか」と問はせたまふ。燕も、人のあまたのぼりゐたるに怖ぢて巣にものぼり来ず。かかる由の返りごとを申したれば、聞きたまひて、「いかがすべき」と思しわづらふに、かの寮の官人くらつまろと申す翁申すやう、「子安貝取らむと思しめさば、たばかりまうさむ」とて、御前に参りたれば、中納言、額を合せて向ひたまへり。

源氏物語「若紫」（新編全集 20 巻 p.205）

〈校訂済み，振り仮名付き・発話者表示付き〉

日もいと長きにつれづれなれば、夕暮のいたう霞みたるにまぎれて、かの小柴垣のもとに立ち出でたまふ。人々は帰したまひて、惟光朝臣とのぞきたまへば、ただこの西面にしも、持仏すゑたてまつりて行ふ尼なりけり。簾すこし上げて、花奉るめり。中の柱に寄りゐて、脇息の上に経を置きて、いとなやましげに読みゐたる尼君、ただ人と見えず。四十余ばかりにて、いと白うあてに痩せたれど、つらつきふくらかに、まみのほど、髪のうつくしげにそがれたる末も、なかなか長きよりもこよなういまめかしきものかな、とあはれに見たまふ。

きよげなる大人二人ばかり、さては童べぞ出で入り遊ぶ。中に、十ばかりやあらむと見えて、白き衣、山吹などの萎えたる着て走り来たる女子、あまた見えつ

る子どもに似るべうもあらず、いみじく生ひ先見えてうつくしげなる容貌なり。髪は扇をひろげたるやうにゆらゆらとして、顔はいと赤くすりなして立てり。

尼君「何ごとぞや。童べと腹立ちたまへるか」とて、尼君の見上げたるに、すこしおぼえたるところあれば、子なめりと見たまふ。紫「雀の子を犬君が逃がしつる、伏籠の中に籠めたりつるものを」とて、いと口惜しと思へり。

A.1.3　日記

更級日記「〔一二〕土忌みの宿、くしき猫のおとずれ」（新編全集 26 巻 p.300）
〈校訂済み，振り仮名付き〉

三月つごもりがた、土忌みに人のもとに渡りたるに、桜さかりにおもしろく、今まで散らぬもあり。かへりてまたの日、あかざりし宿の桜を春くれて散りがたにしも一目みしかなといひにやる。

花の咲き散るをりごとに、乳母亡くなりしをりぞかし、とのみあはれなるに、同じをり亡くなりたまひし侍従の大納言の御むすめの手を見つつ、すずろにあはれなるに、五月ばかり、夜ふくるまで物語をよみて起きゐたれば、来つらむ方も見えぬに、猫のいとなごう鳴いたるを、おどろきて見れば、いみじうをかしげなる猫あり。いづくより来つる猫ぞと見るに、姉なる人、「あなかま、人に聞かすな。いとをかしげなる猫なり。飼はむ」とあるに、いみじう人なれつつ、かたはらにうち臥したり。尋ぬる人やあると、これを隠して飼ふに、すべて下衆のあたりにも寄らず、つと前にのみありて、物もきたなげなるは、ほかざまに顔をむけて食はず。

A.1.4　擬古物語

擬古物語 恋路ゆかしき大将・巻一（中世王朝物語全集 8[17]）
〈校訂済み〉

暁方になるままに、おびたたしう吹きまさりたる風の紛れに、いと疾う内裏へ参り給ひぬ。「今宵は中宮の御宿直なりけるが、下りさせ給ひけるままに、上は藤壷にわたらせ給ふ」と聞こゆれば、そなたざまへ参り給ふに、立て蔀など、よろづの所あらはに、例ならず見わたされて、姫宮の御方の御小壺の叢に、童べ下りて、虫屋ども手ごとに持たり。御覧ずるとて、二宮、御簾を高くもたげさせ給

へるに、十二二ばかりにやと見ゆる御丈立ちにて、うつぶきて立ち給へれば、前へ靡き掛かれる御髪の削ぎめふさやかに、絵に描きたらん心地して、まみ・額・髪ざし、かの雪の朝の御面影なるものから、なほけしき異にて気高う、匂ひも光も類なき御さまは、姫宮にこそはおはしますめれ。よろづのことに騒がず鎮まる御心も、ただ今はいかがはあらん、深く心騒ぎして、おどろかれ給ふ。我が上の空にもの憂く浮きたつ心は、この御さまなどを朝夕見奉らんには慰めなんかし、さりとて当時、世の常に思ひ寄るべき御年のほどならねど、ただまぼり奉らまほしきに、「あはれ、雛屋に虫のゐよかし。一つにあらば、いかに嬉しからん」とのたまへば、二宮、「あらわろや。苔や露も入れさせ給はば、雛のため、いかにうつくからん」と笑ひ聞こえ給へば、げにと思したるさまにて、まめだち給へる御まみのわたり、見る我もうち笑まれて、幾千代まぼるとも飽く世あるまじきに、おとなしき人参りて引き直しつれば、口惜しうて歩み過ぎ給ふ。

A.1.5　歌集

古今和歌集「春歌 上」（新編全集 11 巻 p.31）
〈校訂済み，振り仮名付き〉

ふる年に春立ちける日よめる　　在原元方（ありはらのもとかた）
年のうちに春は来（き）にけりひととせを去年（こぞ）とやいはむ今年（ことし）とやいはむ

春立ちける日よめる　　紀貫之（きのつらゆき）
袖（そで）ひちてむすびし水のこほれるを春立つけふの風やとくらむ

題しらず　　読人しらず（よみひと）
春霞（はるがすみ）たてるやいづこみよしのの吉野（よしの）の山に雪はふりつつ

二条の后（にでう きさき）の春のはじめの御歌（おほんうた）
雪のうちに春は来にけり鶯（うぐひす）のこほれる涙（なみだ）今やとくらむ

雪の木に降りかかれるをよめる　　素性法師（そせいほふし）
春たてば花とや見らむ白雪（しらゆき）のかかれる枝（えだ）にうぐひすの鳴く

A.2　漢文訓読文・和漢混淆文

A.2.1　説話集

日本霊異記（日本国現報善悪霊異記）「電を捉へし縁　第一」（新編全集 10 巻 p.24）

〈校訂済み（漢字カタカナ），振り仮名付き〉

少子部の栖軽は、泊瀬の朝倉の宮に、二十三年天の下治めたまひし雄略天皇大泊瀬稚武の天皇と謂す。の随身にして、肺脯の侍者なりき。天皇、磐余の宮に住みたまひし時に、天皇、后と大安殿に寐テ婚合したまへる時に、栖軽知らずして参ゐ入りき。天皇恥ぢて輟ミヌ。

時に当りて、空に電鳴りき。即ち天皇、栖軽に勅して詔はく、「汝、鳴雷を請け奉らむや」とのたまふ。答へて白さく、「請けまつらむ」とまうす。天皇詔言はく、「爾らば汝請け奉れ」とのたまふ。栖軽勅を奉りて宮より罷り出づ。緋の縵を額に著け、赤き幡桙を擎ゲテ、馬に乗り、阿倍の山田の前の道と豊浦寺の前の路とより走り往きぬ。軽の諸越の衢に至り、叫囁びて請けて言さく、「天の鳴電神、天皇請け呼び奉る云々」とまうす。然して此より馬を還して走りて言さく、「電神と雖も、何の故にか天皇の請けを聞かざらむ」とまうす。走り還る時に、豊浦寺と飯岡との間に、鳴電落ちて在り。栖軽見て神司を呼び、輦籠に入れて大宮に持ち向ひ、天皇に奏して言さく、「電神を請け奉れり」とまうす。時に電、光を放ち明り炫ケリ。天皇見て恐り、偉シク幣帛を進り、落ちし処に返さしめたまひきと者へり。今に電の岡と呼ぶ。古京の少治田の宮の北に在りと者へり。

今昔物語集「巻第十二・於山階寺行涅槃　会語第六」（新編全集 35 巻 p.167）

〈校訂済み（漢字カタカナ），振り仮名付き〉

今昔、山階寺ニ涅槃会ト云フ会有リ。此レ、二月ノ十五日ハ、釈迦如来、涅槃ニ入給ヒシ日也。然レバ、彼ノ寺ノ僧等、「昔ノ沙羅林ノ儀式ヲ思フニ、心無キ草木ソラ、皆其ノ知テ恋慕ノ形チ有キ。何況ヤ、心有リ、悟リ有ラム人ハ、釈迦大師ノ恩徳ヲ報ジ可奉シ」ト儀シ思テ、彼ノ寺ノ仏ハ釈迦如来ニ在セバ、其ノ御前ニシテ彼ノ二月ノ十五日ニ一日ノ法会ヲ行フ也ケリ。四色ノ羅漢ハ威儀ヲ調ヘ、三部ノ伎楽ハ音ヲ発ス。

而ルニ、此ノ会ノ儀式、初メハ少シ愚也ケルヲ、尾張ノ国ノ書生ナル者有ケリ、国司ノ政ノ枉レル事ヲ見テ、心ヲ仏法ニ係テ、頭ヲ剃テ、本国ヲ去ナム

ト思ヒケル間、山階寺ノ僧善殊僧正ト云フ人、請ヲ得テ彼ノ国ニ至ルニ、此ノ書生、本意有ルニ依テ、彼ノ僧正ニ伴ヒテ、本国ヲ棄テ、山階寺ニ行テ、頭ヲ剃リ衣ヲ染テ、彼ノ僧正ノ弟子ト成ヌ。名ヲ寿広ト云フ。本ヨリ心浄クシテ悟リ賢カリケレバ、正教ノ道ヲ学ビ、音楽ノ方ヲ知レリ。然レバ、世ノ人皆、此ノ寿広ヲ敬ヒ貴ビテ、和尚ノ名ヲ得タリ。而ル間、此ノ寿広、更ニ此ノ涅槃会ノ儀式ヲ造テ、色衆ヲ調へ、楽器ヲ副ヘテ、改メテ厳重ニ行ヘリ。

宇治拾遺物語　「十七　小野篁広才の事」（新編全集 50 巻 p.137）

〈校訂済み，振り仮名付き〉

今は昔、小野篁といふ人おはしけり。嵯峨帝の御時に、内裏に札を立てたりけるに、「無悪善」と書きたりけり。帝、篁に、「読め」と仰せられたりければ、「読みは読み候ひなん。されど恐れにて候へば、え申し候はじ」と奏しければ、「ただ申せ」とたびたび仰せられければ、「さがなくてよからんと申して候ふぞ。されば君を呪ひ参らせて候ふなり」と申しければ、「おのれ放ちては誰か書かん」と仰せられければ、「さればこそ、申し候はじとは申して候ひつれ」と申すに、御門、「さて何も書きたらん物は読みてんや」と仰せられければ、「何にても読み候ひなん」と申しければ、片仮名の子文字を十二書かせて給ひて、「読め」と仰せられければ、「ねこの子のこねこ、ししの子のこじし」と読みたりければ、御門ほほゑませ給ひて、事なくてやみにけり。

A.2.2　軍記物

平家物語「猫間」（新編全集 46 巻 p.125）

〈校訂済み，振り仮名付き〉

康定都へのぼり院参して、御坪の内にして、関東のやうつぶさに奏聞しければ、法皇も御感ありけり。公卿殿上人も皆ゑつぼにいり給へり。兵衛佐はかうこそゆゆしくおはしけるに、木曾の左馬頭、都の守護してありけるが、たちゐの振舞の無骨さ、物いふ詞つづきのかたくなる事かぎりなし。理かな、二歳より信濃国木曾といふ山里に三十まで住みなれたりしかば、争でか知るべき。

或時猫間中納言光隆卿といふ人、木曾に宣ひあはすべき事あッておはしたりけり。郎等ども、「猫間殿の見参にいり、申すべき事ありとて、いらせ給ひて

候」と申しければ、木曾大きにわらッて、「猫は人にげんざうするか」。「是は猫間の中納言殿と申す公卿でわたらせ給ふ。御宿所の名とおぼえ候」と申しければ、木曾、「さらば」とて対面す。猶も猫間殿とはえいはで、「猫殿のまれ〳〵わいたるに物よそへ」とぞ宣ひける。中納言是を聞いて、「ただいまあるべうもなし」と宣へば、「いかがけどきにわいたるにさてはあるべき」。何もあたらしき物を無塩といふと心えて、「ここに無塩の平茸あり。とう〳〵」といそがす。根井の小弥太陪膳す。田舎合子のきはめて大きにくぼかりけるに、飯うづたかくよそひ、御菜三種して、平茸の汁で参らせたり。木曾がまへにも同じ体にてすゑたりけり。木曾箸とッて食す。猫間殿は合子のいぶせさに召さざりければ、「それは義仲が精進合子ぞ」。中納言召さでもさすがあしかるべければ、箸とッて召すよししけり。木曾是を見て、「猫殿は小食におはしけるや。きこゆる猫おろし給ひたり。かい給へ」とぞせめたりける。中納言かやうの事に興さめて、宣ひあはすべきことも一言もいださず、軈ていそぎ帰られけり。

A.3　中世・近世口語資料

A.3.1　狂言

虎明本狂言　あさう（大蔵虎明能狂言集 翻刻註解 [18]p.161）

〈原文にはない濁点を補い，その部分を下線で示した〉

麻生の何某〽信濃の国の住人、あさうのなにがしです、そせうの子細あつて、在京仕る処に、安堵の御教書を下され、新地を拝領いたし、あまつさへおいとまを下された、のさ者をよび出し、よろこばせうとぞんずる、藤六あるかやい　藤六〽お前に　麻生〽下六もよべ　藤六〽やい下六めすハやい　下六〽何とめすといふか　藤六〽あふ　下六〽とういふてくれひで、お前に　麻生〽両人ながらはやかつた、やいなんぢらがよろこぶ事があるハ　藤六・下六〽そハまづめでたひ事で御ざあるが、何事で御ざあるぞ　麻生〽永々在京いたす程にと有て、あんどの御教書を下され、新地を拝領して、おいとままで下されたが、かたじけなひ事でハなひか　藤六〽扨も〳〵それハ思召まゝなお仕合で御ざある、此やうなめでたひ事ハ御ざるまひ　下六〽藤六が申ごとく、又両人の者どもハ、さやうの事をこそまつていて御ざあるに、さて〳〵めでたい事でござる　麻生〽さうよ〳〵、国本を出る時ハ、われも〳〵と供をしてのぼつたれども、永々ざい京なれハ、今

までおらひで、こと〳〵く国へくだつてあるに、なんぢら二人ハ今までつめていた程に、くだつたらハくわつとふち(扶持)をせうぞ　藤六・下六〽それハかたじけなふござる　麻生「馬にのせ(乗)ふぞ　藤六・下六〽なを〳〵うれしう御ざる、　麻生〽さりながら、はじめからのり付ぬ馬にのせたらハ、おつる(落)事があらふ程に、馬にのるまで牛にのせう　藤六・下六〽それハともかくも御意次第で御ざる

A.3.2　洒落本

甲駅新話

大木戸(おゝきど)の塵(ちり)は水売(みづうり)の雫(しづく)にしめり天竜寺(てんりうじ)の鐘(かね)は蜩(ひぐらし)の声(こへ)にひゞく[クツワノヲト]ちやんらん〳〵[馬士二人歌]おゝれへとなア引いかぬアう ソレそうだになア引[アトの馬士]かみ村(むら)のウ江五右衛門(ゑごゑむ)がアよめ女(じよう)ナア産月(うみづき)だアといつけがどふだア まだひり出(だ)さねへかなア[サキノ馬士]大キナ声ニテ あんだかハアよんべも夜(よ)ふてへ疝積(せんしやく)のウいてへとつておれらアも張番(はりばん)のしたがががらら出(で)そくねたアよ何がはあ蚊(か)にはおじめられるしたゞもいられねへからおゝめ小(こ)めのウしてひどヲ引いやつを二本(ほん)とられた事よ[アト]四文銭(おゝぜに)でか[サキ]おゝよ[アト]そりやアはあたけへかん病(びよう)のしたナア[又歌]おうらのウせゑどへなアしのびなあ小(こ)ヲざくらの枝(ゑだア)おゝりになああゝさへ引[谷粋(ヤスイ)]

藍さびちゞみのかたびら紅麻に白ぬめゑりのじゆばん帯は黒びろうどにあさぎ小伯を合せたちうや帯呂の山まひ染に桐の三ツ紋付た羽折ひもは駿河打のほそ色はむらさきなれどもさめて藤色かとうたがふ茶つかの少シよごれた脇さし一本おとし指にしかまぼこ形のすげ笠に白キ麻のひもを付てかむり笠の裏に小サキ風車二本さしたり

陽台遺編・娼閣秘言

[女郎歌夕]まづたばこのむ 中居くめ茶持テ来ルまた切炭のいけてある火入と下地の火入といれかへる[中居くめ] 夕 さんなんにも御用はないかへ[女郎]いゝへ源七がたば粉もつてきたら。くだんせ。こなんはおきばんかさむかろふナ。[くめ]アイおきばんしやわいな。何なと用が有ならいひなんせ[女郎]その屏風もつとこちらへ引てくだんせ。[くめ]アイ 歌川の瀬もなる夜半もなるにといひ〳〵勝手へ行。[女郎]まづねる。ふとんきる。扨たばこ吸付ル申ゝ。これいな。又いなまたたぬきねいりなんするといふてこそぐる。[客]ウヽヽヽアヽついねたそふな。もうなん時じや[女郎]しらんわいな[客]しらんほどなら。もふよいいにぢぶんじや[女郎]客のかいなへ喰つく[客]あいたゝこりやめつたにかぶるまいほうれんそうの汁がついて有ぞ[女郎]だんないわいな。いつそしんだがましじやわいな。[客]ハヽア世をくはんじさつしやつたの。此上に桜のつぼみがちつ

たらよいていはつものじや女郎わたしがていはつしたら。よろこびなんしよな申といふて。又きせるをとり。エヽ心いきの通らぬきせるじや。おまへはせんど桔梗屋へおいなんして井筒さんにあいなんしたげなようきくと思ひなんせ。どふでまた。井筒さんのよふに。わたしらはないはづじやわいな客これはめいわく成程以中や祇童と付合で桔梗屋へいたが井筒とやら井げたとやらはしらぬ。女郎へエさよふで御座りましよ。わたしやそのあけすぐに井筒さんへは付届して置キましてこざりますアイきつとゝいふてやりまして御座りますといふてつめる

A.3.3　人情本

春告鳥

さまゝのこと思ひ出す桜かな、その桜節、憂ことを、わすれさせんと勧められ、迎嶋の別荘よりうしやの土留木を二人連、おとなしき風俗の息子と桜川新孝
新「若旦那マア私にだまされたと思つて、例の所へ往て御覧じまし。おめへさんにやア、急度お気に入るにやア相違ごぜへません。此間も若竹の兼八が貴君さんのお噂を申て、今度出来た薄雲さんを是非あなたに出会し申てへと、くれゝ左様申て居ました 息「そりやア有難へが相方で出るか何だか知れるものかナ
新「ナニゝ大丈夫でごぜへます　トいふとき乗切の船頭小舟のうすべりをひろげ 舟「サア お乗んなさいまし 新「ヲイゝこれは御苦労サア 若旦那、お乗なさいまし 息「アイ汐があるから洲を乗り切るに楽だノト二人はのりてむかふへわたる 息「エヽコウ此洲へ犬が幾疋も来て居て汐の下る間平気であそんでゐるからおかしいノウ 新「左様サ、江戸の方の犬は斯いふ世界は知りますめへ

A.3.4　滑稽本

浮世床

「をぢさん、あさりはどうだ。あさりむきん。「ぼうはぽつぼにゐて、あつたでいゝぞ。ヲヽあぶいゝ。「かうぢ、アイよびなすつたかエ。ホイ「かうぢ。「おとよさん、かうぜいがみをかひにいくから、おまへもおいで。「ヲイはゐへの。今おめざめか。「ちよびと、やつつけばゝあとやらかしてへ。「伝さん、けふはしごとか。「べらぼうにとほいちやうばよ。「はやくおいでよ。「喜八さん、かほみせ

のひやうばんをきゝなすつたか。「アイサ、三げんながら、いゝさうさ。きついもんだネ。「ぶちこい〳〵、エヽおゝしき〳〵、いせやのよつとうらのめんかぶりとけしかけてやらう。ありやりやんりうとい。「どうだ八公、すてきとさむいぢやアねへか。ヲヤ〳〵がうけつにたまつた。こいつはをさまらねへ。「ヲツトあぶねへ〳〵。へンちげへねへ。「又あくびだ。あアあアあア引。にやん妙ひやうれんじやぶつ。「コレ〳〵こゝのもんくがありがてへ。こゝを一ばんきかつし。「まづうけちんからさきさ。

A.4　近代語資料

A.4.1　近代雑誌

明六雑誌 1874 年 1 号・洋字を以て國語を書するの論・西周

〈明六雑誌コーパスの一部，文語論説文〉

吾輩日常二三朋友ノ盍簪ニ於テ偶當時治亂盛衰ノ故政治得失ノ跡ナド凡テ世故ニ就テ談論爰ニ及ブ時ハ動モスレバカノ歐洲諸國ト比較スルㇵノ多カル中ニ終ニハ彼ノ文明ヲ羨ミ我ガ不開化ヲ歎ジ果テ果テハ人民ノ愚如何トモスルナシト云フㇵニ歸シテ亦欷歔長大息ニ堪ザル者アリ　夫維新以來賢材モ輩出シ百度モ更張シ官省寮司ヨリ六十餘縣ニ至ルマデ既ニ昔日ノ日本ニ非ズ　其善政美擧モ屈指ニ暇アラザルナリ　然ルニ退テ熟々之ヲ考フレバ百端未ダ脱垢ノ地ニ至ラザル事ノミニシテ善政アレドモ民其澤ヲ蒙ラズ美擧アレドモ得失相償ハザル等ノ事多シ　是何トナレバ維新以來日タル未ダ久シカラザレバ外面ノ規模ハ如何ニ盛大ニモアレ衷情未ダ洽浹セザレバナリ　是殆猿ニ衣裳爨婦ニ舞衣ヲ被セタル如シ　故ニ上旨ハ下達セズ下情ハ上伸セズシテ全身不遂ノ人ノ如シ　是ヲ以テ間ニ一二賢明英傑ノ人有テ之ヲ鼓舞シ之ヲ振起セント欲スルモ猶眠リヲ貪ルノ兒ヲ醒起シ醉倒シタル夫ヲ扶助スルガ如シ　手倦ミ力竭キ己亦從テ倒レントス

太陽 1895 年 11 号・狂言娘

〈太陽コーパスの一部，地の文は文語，会話は口語（旧仮名遣い）〉

娘（むすめ）は少時（しばし）女（をんな）を睨（にら）みしが、軈（やが）て一歩進（ひとあしすゝ）んで、『お前（まへ）、また來（き）たのかい。何遍（なんべん）來（き）たつて行（い）かないから、さう思（おも）つてるが能（い）い。お玉（たま）の使（つかひ）に來（き）たんだらう。』と、問掛（とひかけ）し眼（め）には光（ひかり）あり。

女は豫て娘を見知れるにや、又かと思へる氣色にて、眉を顰め、『お前さんは、又此樣處を歩いて居なさるんだね。暗くなツて來たから、早く歸りなさるが能いよ。』

『又歸れツて、新富町にかえ。可厭な事ツだ、家爺さん處なんか。』

『新富町。其樣處ぢやないよ、直ぐ其處のお前さんの宅にさ。今日も令母さんのお墓參だね。』

娘は莞爾と笑つて、左も馴々しく女の傍へ進み寄りぬ。

『誰かと思つたら、婆さんだツたね。今日も能く出てお出でだね。私しやね、お前を見ると、可哀想で仕樣がないの。だから、何時でも何か遣りたいの。お前は母親さんのお墓を、何時でも掃除しといてお呉れだもの。今日も爲といてお呉れか。今ね、お墓にお參りを爲て、歸途に又與るから、其處に待てゝお呉れよ。』

娘は墓に行んとてか、川沿に上手へ行んとすれば、『もう暮るのに、早く歸んなさるが能いよ。其樣方に行ツちやア……。』と、女聲を掛れば、娘は振返つて、『だツてお前、母親さんのお墓參を爲なけりや惡いもの。母親さんが毎日私しを待てゝね、能くお參を爲てお呉れだツて、大層お喜びなの。母親さんは可哀想なんだよ、お前にも話したツけね。誰もお參りを爲る者はないし、私がお參に行かなきや、淋しくツて爲樣がないとお云ひだもの。お墓に行つてね、母親さんに逢つて來るんだよ。お前も一處にお出でゞないか。』と、尚ほとぼ〳〵と歩み行く。

太陽 1895 年 03 号・文学上の新事業

〈太陽コーパスの一部，文語論説文〉

我が社會の事、不整頓なるもの少なからぬが中に、文學の如きは、恐くはその最も亂雜なるものゝ一ならん。我が文界は今尚ほ過渡の時代にありと云はざるべからず。從來の和漢文學及び其混和に成りし文學を總稱して我が國文學と名くるに妨なしとするも、未だ其國文學は世界の最も進歩せる偉大なる文學に見る所の雄篇大作と相比すべきものを有すと云ふを得ず。固より或は戲曲に、或は小説に、或は神話に、或は詩歌に頗る見るべき者なきにあらず、又その各々一種特別の美を具ふる所あるを否むべからずと雖も、未だ以て世界の大文學と稱するには足らざるなり。こは多くの論者の既にしば〳〵痛言し以て我が文學をして將來に大成を期せしめんと欲したる所なるが、啻にその如く雄篇大作に乏しきの憾あるのみならず、文章に用うる言語の形に於いても大に吾人をして不滿足を感ぜしむるも

のあり。爰には我が文學の發揮する想の價値如何を云はず、又全躰文の想と形とは如何なる關係を有するものなるかをも論ぜず、唯だ想を傳ふるの機關たる文章の形式に於いて甚しく不完全なる所あるを云はんと欲す。

1925 年 03 号・長篇科學小説　生ける死『第三回』

〈太陽コーパスの一部，旧仮名遣いの口語会話文〉

『『夢ぢやないか？』とトムスンは云ひました。

『いや、夢ぢやない、抓つてみたら痛かつたから。

『極樂だ。』とハムデンは呟きました。

『さうですよ、リヴイングストン大佐、極樂を發見したんですよ、あなたは――ああ、こんな凍つた大陸の眞中で！』

『全く私達皆の悦びと云つたらありませんでしたね。現代に於ける地理學上の最大の發見だと思ひましたね。いや全く氣が狂やしないかとさへ思ひましたよ、私は。どうしても、涙が流れるのを止めることが出來ませんでしたよ。

『「犬はどうしてるかしら。」とトムスンが云ひました。「あいつ等があんなに嗅いでばかりゐたわけがわかつたよ。」

『犬共はしきりに四邊を見廻し、この大發見を祝ふやうに、頻りに尾を振つてゐました。

『「極樂だ。」と私も叫びました。「その通りだよ、ハムデン。極樂と命名しよう――極樂園と。」

『全く極樂の様に見えたのですよ。』

『空氣はまだすつかり澄み切つてはゐませんでした。最初は、さう見えたのです、が、謂はばあらゆるものの上に不思議な夢のやうなものが搖曳してゐるのだといふ事に間もなく氣が附きました。靄ではなかつたんです。何だかわかりませんがね。この光景を見て、第一に思出したのはターナーの傑作ですな。しかし、全くその通りともいへませんでした。それにまた、何處へ行つても、空氣はちつとも動かないやうに見えました。それから雲の様子は！　成程壯大とか壯麗とかいふのではない、が、どんな畫家も未だ嘗て畫布の上に描き得なかつたやうな美しさがありました。まるでお伽の國に浮んでゐる雲のやうでしたよ。

1925 年 02 号・歴代の総理大臣（二）

〈太陽コーパスの一部，旧仮名遣いの口語論説文〉

黒田内閣は、大隈外相が實權を握り、攻撃も之に集つた。前の順序で、大隈が大久保の後を繼ぐべきであり、薩長聯合で排斥せられたのが、恰も大久保が大隈を用ゐた如く、黒田が之を用ゐようとし、大久保程押が利かなかつたのである。大隈が權力を振ひ、内閣瓦解の餘儀なきに終つた後、山縣が内閣を組織することになつたが、明治三十一年大隈内閣の成るまで、薩長代表者が首相となり、十七年を經て再び、大隈内閣の成るまで、准藩閥から首相を出した。

大隈放逐の際、山縣が少なからぬ力を出し、次いで黒田内閣といふ事實的大隈内閣の後に山縣が出で、後の大隈内閣瓦解にも山縣が出で、大隈の掻き廻した跡始末をするが爲とし、何でも大隈に反對するの態度に出でた。伊藤も大隈と相容れない仲になつても、互に相諒解する所があり、山縣の頑なゝのを困り者とするに一致した。それでも山縣が大隈を斥けて止まず、大隈の跡始末に困ると明言するを憚らぬ。後に山縣が大隈を首相とするに力を添へたのは、准藩閥の頼みにならなくなつたのに伴つて居る。

付録 B　UniDicの品詞・活用表等

B.1　語種

表 B.1: UniDic の語種

値	説明
和	和語
漢	漢語
外	外来語
混	混種語
固	固有名
記	記号
不明	語種不明

B.2　品詞

表 B.2: UniDic 品詞一覧

品詞	大分類	中分類	小分類	細分類	類
名詞-普通名詞-一般	名詞	普通名詞	一般		体
名詞-普通名詞-サ変可能			サ変可能		体
名詞-普通名詞-形状詞可能			形状詞可能		体
名詞-普通名詞-サ変形状詞可能			サ変形状詞可能		体
名詞-普通名詞-副詞可能			副詞可能		体
名詞-普通名詞-連用可能			連用可能		体
名詞-固有名詞-一般		固有名詞	一般		固有名
名詞-固有名詞-人名-一般			人名	一般	人名
名詞-固有名詞-人名-姓			人名	姓	姓

名詞-固有名詞-人名-名			人名	名	名
名詞-固有名詞-地名-一般			地名	一般	地名
名詞-固有名詞-地名-国			地名	国	国
名詞-数詞		数詞			数
名詞-助動詞語幹		助動詞語幹			体
代名詞	代名詞				体
形状詞-一般	形状詞	一般			相
形状詞-タリ		タリ			相
形状詞-助動詞語幹		助動詞語幹			助動
連体詞	連体詞				相
副詞	副詞				相
接続詞	接続詞				他
感動詞-一般	感動詞	一般			他
感動詞-フィラー		フィラー			他
動詞-一般	動詞	一般			用
動詞-非自立可能		非自立可能			用
形容詞-一般	形容詞	一般			相
形容詞-非自立可能		非自立可能			相
助動詞	助動詞				助動
助詞-格助詞	助詞	格助詞			格助
助詞-副助詞		副助詞			副助
助詞-係助詞		係助詞			係助
助詞-接続助詞		接続助詞			接助
助詞-終助詞		終助詞			終助
助詞-準体助詞		準体助詞			準助
接頭辞	接頭辞				接頭
接尾辞-名詞的-一般	接尾辞	名詞的	一般		接尾体
接尾辞-名詞的-サ変可能			サ変可能		接尾体
接尾辞-名詞的-形状詞可能			形状詞可能		接尾体
接尾辞-名詞的-サ変形状詞可能			サ変形状詞可能		接尾体
接尾辞-名詞的-副詞可能			副詞可能		接尾体
接尾辞-名詞的-助数詞			助数詞		助数
接尾辞-形状詞的		形状詞的			接尾相
接尾辞-動詞的		動詞的			接尾用
接尾辞-形容詞的		形容詞的			接尾相
記号-一般	記号	一般			記号
記号-文字		文字			記号
補助記号-一般	補助記号	一般			補助
空白					補助
補助記号-句点		句点			補助
補助記号-読点		読点			補助
補助記号-括弧開		括弧開			補助

補助記号-括弧閉	括弧閉			補助
補助記号-ＡＡ-一般	ＡＡ	一般		補助
補助記号-ＡＡ-顔文字		顔文字		補助

B.3　活用型

動詞（文語）

表 B.3: UniDic 文語動詞活用型一覧

活用型	補足説明
文語カ行変格	
文語サ行変格	
文語ザ行変格	「-ず」型の一字漢語サ変動詞
文語ナ行変格	
文語ラ行変格	
文語上一段-カ行	
文語上一段-ナ行	
文語上一段-マ行	
文語上一段-ヤ行	
文語上一段-ワ行	
文語上二段-タ行	
文語上二段-ダ行	
文語上二段-ハ行	
文語上二段-バ行	
文語上二段-ヤ行	
文語下二段-ア行	
文語下二段-カ行	
文語下二段-ガ行	
文語下二段-サ行	
文語下二段-ザ行	
文語下二段-タ行	
文語下二段-ダ行	
文語下二段-ナ行	
文語下二段-ハ行	
文語下二段-バ行	
文語下二段-マ行	

文語下二段-ヤ行	
文語下二段-ラ行	
文語四段-カ行	
文語四段-ガ行	
文語四段-サ行	
文語四段-タ行	
文語四段-ハ行	
文語四段-バ行	
文語四段-マ行	
文語四段-ラ行	

動詞（口語）

表 B.4: UniDic 口語動詞活用型一覧

活用型	補足説明
カ行変格	
サ行変格	
ザ行変格	「-ずる」型の一字漢語サ変動詞
上一段-ア行	
上一段-カ行	
上一段-ガ行	
上一段-ザ行	
上一段-タ行	
上一段-ナ行	
上一段-ハ行	
上一段-バ行	
上一段-マ行	
上一段-ラ行	
下一段-ア行	
下一段-カ行	
下一段-ガ行	
下一段-サ行	
下一段-ザ行	
下一段-タ行	
下一段-ダ行	
下一段-ナ行	
下一段-ハ行	

下一段-バ行	
下一段-マ行	
下一段-ラ行-一般	
下一段-ラ行-呉レル	「呉れる」（命令形「くれ」）
五段-カ行-イク	「行く（イク）」（連用形促音便あり）
五段-カ行-ユク	「行く（ユク）」（連用形に音便なし）
五段-カ行-一般	
五段-ガ行	
五段-サ行	
五段-タ行	
五段-ナ行	
五段-バ行	
五段-マ行	
五段-ラ行-アル	
五段-ラ行-一般	
五段-ワア行-イウ	「言う」（イーマス／ユー）
五段-ワア行-一般	

助動詞（文語）

表 B.5: UniDic 文語助動詞活用型一覧

活用型	補足説明
文語助動詞-キ	
文語助動詞-ケム	
文語助動詞-ケリ	
文語助動詞-コス	
文語助動詞-ゴトシ	
文語助動詞-ザマス	
文語助動詞-ザンス	
文語助動詞-ジ	
文語助動詞-ズ	
文語助動詞-タリ-完了	
文語助動詞-タリ-断定	
文語助動詞-ツ	
文語助動詞-ナリ-伝聞	
文語助動詞-ナリ-断定	
文語助動詞-ヌ	

文語助動詞-ベシ	
文語助動詞-マシ	
文語助動詞-マジ	
文語助動詞-ム	
文語助動詞-ムズ	
文語助動詞-メリ	
文語助動詞-ラシ	
文語助動詞-ラム	
文語助動詞-リ	
文語助動詞-ンス	近世上方語
無変化型	

助動詞（口語）

表 B.6: UniDic 口語助動詞活用型一覧

活用型	補足説明
助動詞-ゲナ	近世
助動詞-ジャ	
助動詞-タ	
助動詞-タイ	
助動詞-ダ	
助動詞-デス	
助動詞-ドス	関西（京都）方言
助動詞-ナイ	
助動詞-ヌ	
助動詞-ヘン	関西方言
助動詞-マイ	
助動詞-マス	
助動詞-ヤ	
助動詞-ヤス	「〜でやす」
助動詞-ラシイ	
助動詞-レル	

形容詞（文語）

表 B.7: UniDic 文語形容詞活用型一覧

活用型	細分類	補足説明
文語形容詞-ク	一般	
	多シ	「多し」（終止「多かり」）
文語形容詞-シク	シク	
	ジク	「いみじ」など

形容詞（口語）

表 B.8: UniDic 口語形容詞活用型一覧

活用型	補足説明
形容詞	一般

B.4 活用形

表 B.9: UniDic 活用形一覧

大分類	活用形	補足説明
語幹	語幹-サ	形容詞「無い」「良い」に，様態の助動詞「そうだ」が接続するときの形（「無さ-そうだ」「良さ-そうだ」）
	語幹-一般	
未然形	未然形-サ	サ変（ザ変）に，助動詞「せる」「れる」が接続するときの形（「さ-せる」「さ-れる」）
	未然形-セ	サ変（ザ変）に，助動詞「ず」が接続するときの形（せ-ず）
	未然形-一般	
	未然形-撥音便	ラ行五段活用動詞の一部で起こる撥音便（「知ん-ない」）
	未然形-補助	形容詞カリ活用未然形（「少なから-ず」）
意志推量形	意志推量形	意志・推量の助動詞「う」「よう」が接続した形全体（「行こう」「見よう」）
連用形	連用形-イ音便	
	連用形-ウ音便	

	連用形-ト	断定の文語助動詞「たり」の連用形「と」
	連用形-ニ	断定の助動詞「だ」・文語助動詞「なり」の連用形「に」
	連用形-一般	
	連用形-促音便	
	連用形-撥音便	
	連用形-省略	関西方言などで形容詞連用形が省略された形をとることがある（「欲し-ない」）
	連用形-融合	断定の助動詞「だ」の連用形に後続する係助詞「は」が融合した形（「じゃ」）
	連用形-補助	文語形容詞・文語助動詞「ず」のカリ活用連用形（「無かり」「ざり」）
終止形	終止形-ウ音便	文語ハ行四段活用動詞の終止形がウ音便化することがある（「給う［タモー］」「候［ソーロー］」）
	終止形-一般	
	終止形-促音便	形容詞の「高っ」「痛っ」などの形
	終止形-撥音便	助動詞「ず」の終止形に撥音便形がある（「（しませ）ん」）また関西方言などで撥音便形になることがある（「てん（な）」）
	終止形-融合	断定の助動詞「だ」の終止形に前接する「と」の音と融合した形（「（何のこっ）ちゃ」）
	終止形-補助	文語形容詞「多し」のカリ活用終止形（「多かり」）
連体形	連体形-ウ音便	文語ハ行四段活用動詞の連体形がウ音便化することがある（「給う［タモー］」「候［ソーロー］」）
	連体形-一般	
	連体形-撥音便	助動詞「ず」の連体形がしばしば「ん」となるほか，動詞でも「すん（の）」のように準体助詞「の」の前で撥音になる．また文語助動詞「む」「けむ」の連体形が「ん」となる
	連体形-補助	
已然形	已然形	
	已然形-一般	
	已然形-補助	
仮定形	仮定形-一般	
	仮定形-融合	
命令形	命令形	
	命令形-一般	
ク語法	ク語法	文語専用

付録 C　MeCab用の設定ファイル

歴史的資料を対象としたUniDicのためのMeCab用の設定ファイルのうち，素性テンプレートrewrite.def, feature.defを掲げる[1].

C.1　rewrite.def

素性列から内部状態素生列に変換するマッピングの定義ファイル．定義ファイル中の$\$n$は表C.1に示すUniDicの要素に対応する．各要素の説明は表2.1参照.

表 C.1: rewrite.defの素性番号

番号	辞書の素性
1	品詞大分類（pos1）
2	品詞中分類（pos2）
4	品詞小分類（pos3）
5	品詞細分類（pos4）
6	活用型（cType）
7	活用形（cForm）
8	語彙素読み（lForm）
9	語彙素（lemma）
10	書字形（orth）
11	発音形（pron）
12	仮名形（kana）
13	語種（goshu）
14	書字形基本形（orthBase）
15	発音形基本形（pronBase）
16	仮名形基本形（kanaBase）
17	語形基本形（formBase）
18	語頭変化型（iType）

[1]設定ファイルの詳細は `http://mecab.googlecode.com/svn/trunk/mecab/doc/learn.html` 参照.

19	語頭変化形（iForm）
20	語頭変化結合型（iConType）
21	語末変化型（fType）
22	語末変化形（fForm）
23	語末変化結合型（fConType）
24	アクセント型（aType）
25	アクセント結合型（aConType）
26	アクセント修飾型（aModType）

リスト C.1: rewrite.def

```
[unigram rewrite]
*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,$6,$7,$8,$9,$13
    ,$10,$14,$12
*,*,*,*,*,* $1,$2,$3,$4,$5,$6,*,*,*,*,*,*,*

[left rewrite]
BOS/EOS,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,*,*,*,*,*,BOS/EOS,BOS
    /EOS,BOS/EOS,*,*,*
助詞,*,*,*,*,*,*,(の|に|を|をば|て|は|と|が|も|ぞ|こそ|にて|のみ|とて|の|から|
    か|が|か|など|ど|と|ば|や|まで|へ|から|より|よ|し|ほど|し
    か),*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*  $1,$2,$3,$4,$5,$6,$9,$13,$12,$20,$21,
    $19
助動詞,*,*,*,*,*,*,(き|けむ|けらし|けり|こす|ごとし|ざます|ざんす|じ|ず|たり|つ|
    なり|ぬ|べし|べらなり|ましまじ|む|むず|めり|らし|らむ|り|んす|んなり|んめ
    り|非ず),*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*  $1,$2,$3,$4,$5,$6,$9,$13,$12,$20
    ,$21,$19
名詞,助動詞語幹,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,$6,
    $9,$13,$12,$20,$21,$19
形状詞,助動詞語幹,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,
    $6,$9,$13,$12,$20,$21,$19
動詞,非自立可能,*,*,*,*,*,(詫びる|来る|来たる|遊ばす|有る|忘れる|奉る|聞こえる|
    付ける|付く|抜ける|抜く|能う|得る|渡る|通す|直す|頂く|続ける|続く|成る|申す|
    上げる|出だす|習わす|終わる|終える|持て行く|侍る|止す|始める|参る|参らす|坐
    す|込める|込む|合う|行く|候う|遣る|見る|見せる|居る|給ぶ|給う|願う|間違う|敢
    う|慣わす|慣れる|掛ける|掛かる|回る|過ぎる|果てる|果たす|ものす|おわ
    す),*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,$6,$9,$13,$12,$20,$21,
    $19
形容詞,非自立可能,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,
    $6,$9,$13,$12,$20,$21,$19
接頭辞,*,*,*,*,*,*,(さ|み|めく|位|君|兼ねる|勝ち|上|人|中|殿|等|難い|品|
    様),*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,$6,$9,$13,$12,$20,$21,
    $19
接頭辞,*,*,*,*,*,*,(御|先|真|相|大),*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,
    $4,$5,$6,$9,$13,$12,$20,$21,$19
名詞,数詞,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,$6,$9,$13
    ,$12,$20,$21,$19
*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,$6,*,*,$12,$20,
    $21,$19
*,*,*,*,*,* $1,$2,$3,$4,$5,$6,*,*,*,*,*,*,*
```

```
18
19 [right rewrite]
20 BOS/EOS,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,*,*,*,*,*,BOS/EOS,BOS
       /EOS,BOS/EOS,*,*,*
21 助詞,*,*,*,*,*,*,(の|に|を|をば|て|は|と|が|も|ぞ|こそ|にて|のみ|とて|の|から|
       か|が|か|など|ど|と|ば|や|まで|へ|から|より|よ|し|ほど|し
       か),*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*  $1,$2,$3,$4,$5,$6,$9,$13,$12,$17,$18,
       $22
22 助動詞,*,*,*,*,*,*,(き|けむ|けらし|けり|こす|ごとし|ざます|ざんす|じ|ず|たり|つ|
       なり|ぬ|べし|べらなり|ましまじ|む|むず|めり|らし|らむ|り|んす|んなり|んめ
       り|非ず),*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*  $1,$2,$3,$4,$5,$6,$9,$13,$12,$17
       ,$18,$22
23 名詞,助動詞語幹,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,$6,
       $9,$13,$12,$17,$18,$22
24 形状詞,助動詞語幹,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,
       $6,$9,$13,$12,$17,$18,$22
25 動詞,非自立可能,*,*,*,*,*,(詫びる|来る|来たる|遊ばす|有る|忘れる|奉る|聞こえる|
       付ける|付く|抜ける|抜く|能う|得る|渡る|通す|直す|頂く|続ける|続く|成る|申す|
       上げる|出だす|習わす|終わる|終える|持て行く|侍る|止す|始める|参る|参らす|坐
       す|込める|込む|合う|行く|候う|遣る|見る|見せる|居る|給ぶ|給う|願う|間違う|敢
       う|慣わす|慣れる|掛ける|掛かる|回る|過ぎる|果てる|果たす|ものす|おわ
       す),*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,$6,$9,$13,$12,$20,$21,
       $19
26 形容詞,非自立可能,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,
       $6,$9,$13,$12,$17,$18,$22
27 接尾辞,*,*,*,*,*,*,*,(御|先|真|相|大),*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,
       $4,$5,$6,$9,$13,$12,$20,$21,$19
28 接頭辞,*,*,*,*,*,*,*,(さ|み|めく|位|君|兼ねる|勝ち|上|人|中|殿|等|難い|品|
       様),*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,$6,$9,$13,$12,$20,$21,
       $19
29 名詞,数詞,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,$6,$9,$13
       ,$12,$17,$18,$22
30 *,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,*,* $1,$2,$3,$4,$5,$6,*,*,$12,$17,
       $18,$22
31 *,*,*,*,*,* $1,$2,$3,$4,$5,$6,*,*,*,*,*,*
```

C.2 feature.def

素性テンプレート．定義ファイル中の F[n] は unigram 素性で，表 C.2 の番号の素性に展開される．L[n] は左文脈素性，R[n] は右文脈素性で，番号は F[n] と同様である．%t は文字種情報．

表 C.2: feature.def の素性番号

番号	辞書の素性
0	品詞大分類（pos1）
1	品詞中分類（pos2）
2	品詞小分類（pos3）

3	品詞細分類（pos4）
4	活用型（cType）
5	活用形（cForm）
6	語彙素読み（lForm）
7	語彙素（lemma）
8	書字形（orth）
9	書字形基本形（orthBase）
10	発音形（pron）
11	発音形基本形（pronBase）
12	語種（goshu）

リスト C.2: feature.def

```
1
2  UNIGRAM G01:%F[0]
3  UNIGRAM G02:%F[0],%F?[1]
4  UNIGRAM G03:%F[0],%F[1],%F?[2]
5  UNIGRAM G04:%F[0],%F[1],%F[2],%F?[3]
6  UNIGRAM C01:%F?[4]
7  UNIGRAM C02:%F?[5]
8  UNIGRAM C03:%F?[4],%F?[5]
9  UNIGRAM GC01:%F[0],%F?[4],%F?[5]
10 UNIGRAM GC02:%F[0],%F?[1],%F?[4],%F?[5]
11 UNIGRAM T01:%t
12 UNIGRAM GT01:%F[0],%t
13 UNIGRAM GT02:%F[0],%F?[1],%t
14 UNIGRAM GT03:%F[0],%F[1],%F?[2],%t
15 UNIGRAM GT04:%F[0],%F[1],%F[2],%F?[3],%t
16 UNIGRAM GCT01:%F[0],%F?[4],%F?[5],%t
17 UNIGRAM GCT02:%F[0],%F?[1],%F?[4],%F?[5],%t
18 UNIGRAM O01:%F[8]
19 UNIGRAM O02:%F[9]
20 UNIGRAM O03:%F[8],%F[9]
21 UNIGRAM GO01:%F[0],%F[9]
22 UNIGRAM GO02:%F[0],%F?[1],%F[9]
23 UNIGRAM GO03:%F[0],%F[1],%F?[2],%F[9]
24 UNIGRAM GO04:%F[0],%F[1],%F[2],%F?[3],%F[9]
25 UNIGRAM GCO01:%F[0],%F?[4],%F?[5],%F[8]
26 UNIGRAM GCO02:%F[0],%F?[1],%F?[4],%F?[5],%F[8]
27 UNIGRAM GL01:%F[0],%F[6],%F[7]
28 UNIGRAM GL02:%F[0],%F?[1],%F[6],%F[7]
29 UNIGRAM GL03:%F[0],%F[1],%F?[2],%F[6],%F[7]
30 UNIGRAM GL04:%F[0],%F[1],%F[2],%F?[3],%F[6],%F[7]
31 UNIGRAM CL01:%F?[4],%F?[5],%F[6],%F[7]
32 UNIGRAM GCL01:%F[0],%F?[4],%F?[5],%F[6],%F[7]
33 UNIGRAM GCL02:%F[0],%F?[1],%F?[4],%F?[5],%F[6],%F[7]
34 UNIGRAM LO01:%F[6],%F[7],%F[9]
35 UNIGRAM GLO01:%F[0],%F[6],%F[7],%F[9]
36 UNIGRAM GLO02:%F[0],%F?[1],%F[6],%F[7],%F[9]
37 UNIGRAM GLO03:%F[0],%F[1],%F?[2],%F[6],%F[7],%F[9]
38 UNIGRAM GLO04:%F[0],%F[1],%F[2],%F?[3],%F[6],%F[7],%F[9]
39 UNIGRAM GCLO01:%F[0],%F?[4],%F?[5],%F[6],%F[7],%F[8]
```

```
UNIGRAM GCLO02:%F[0],%F?[1],%F?[4],%F?[5],%F[6],%F[7],%F[8]
UNIGRAM W01:%F[12]
UNIGRAM GW01:%F[0],%F[12]
UNIGRAM GW02:%F[0],%F?[1],%F[12]
UNIGRAM GW03:%F[0],%F[1],%F?[2],%F[12]
UNIGRAM GW04:%F[0],%F[1],%F[2],%F?[3],%F[12]
UNIGRAM GCW01:%F[0],%F?[4],%F?[5],%F[12]
UNIGRAM GCW02:%F[0],%F?[1],%F?[4],%F?[5],%F[12]
UNIGRAM OW01:%F[9],%F[12]
UNIGRAM LW01:%F[6],%F[7],%F[12]
UNIGRAM GCLOW01:%F[0],%F?[4],%F?[5],%F[6],%F[7],%F[8],%F[12]
UNIGRAM GCLOW02:%F[0],%F?[1],%F?[4],%F?[5],%F[6],%F[7],%F[8],%F[12]
UNIGRAM GLOP01:%F[0],%F[6],%F[7],%F[9],%F[11]
UNIGRAM GLOP02:%F[0],%F?[1],%F[6],%F[7],%F[9],%F[11]
UNIGRAM GLOP03:%F[0],%F[1],%F?[2],%F[6],%F[7],%F[9],%F[11]
UNIGRAM GLOP04:%F[0],%F[1],%F[2],%F?[3],%F[6],%F[7],%F[9],%F[11]
UNIGRAM GCLOP01:%F[0],%F?[4],%F?[5],%F[6],%F[7],%F[8],%F[10]
UNIGRAM GCLOP02:%F[0],%F?[1],%F?[4],%F?[5],%F[6],%F[7],%F[8],%F[10]
BIGRAM G_G01:%L[0]/%R[0]
BIGRAM G_G02:%L[0]/%R[0],%R?[1]
BIGRAM G_G03:%L[0]/%R[0],%R[1],%R?[2]
BIGRAM G_G04:%L[0]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM G_G05:%L[0],%L?[1]/%R[0]
BIGRAM G_G06:%L[0],%L?[1]/%R[0],%R?[1]
BIGRAM G_G07:%L[0],%L?[1]/%R[0],%R[1],%R?[2]
BIGRAM G_G08:%L[0],%L?[1]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM G_G09:%L[0],%L[1],%L?[2]/%R[0]
BIGRAM G_G10:%L[0],%L[1],%L?[2]/%R[0],%R?[1]
BIGRAM G_G11:%L[0],%L[1],%L?[2]/%R[0],%R[1],%R?[2]
BIGRAM G_G12:%L[0],%L[1],%L?[2]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM G_G13:%L[0],%L[1],%L[2],%L?[3]/%R[0]
BIGRAM G_G14:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R?[1]
BIGRAM G_G15:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R[1],%R?[2]
BIGRAM G_G16:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM C_C09:%L?[4],%L?[5]/%R?[4],%R?[5]
BIGRAM G_C01:%L[0]/%R?[4],%R?[5]
BIGRAM G_C02:%L[0],%L?[1]/%R?[4],%R?[5]
BIGRAM G_C03:%L[0],%L[1],%L?[2]/%R?[4],%R?[5]
BIGRAM G_C04:%L[0],%L[1],%L[2],%L?[3]/%R?[4],%R?[5]
BIGRAM C_G01:%L?[4],%L?[5]/%R[0]
BIGRAM C_G02:%L?[4],%L?[5]/%R[0],%R?[1]
BIGRAM C_G03:%L?[4],%L?[5]/%R[0],%R[1],%R?[2]
BIGRAM C_G04:%L?[4],%L?[5]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM G_GC01:%L[0]/%R[0],%R?[4],%R?[5]
BIGRAM G_GC02:%L[0]/%R[0],%R?[1],%R?[4],%R?[5]
BIGRAM G_GC05:%L[0],%L?[1]/%R[0],%R?[4],%R?[5]
BIGRAM G_GC06:%L[0],%L?[1]/%R[0],%R?[1],%R?[4],%R?[5]
BIGRAM G_GC09:%L[0],%L[1],%L?[2]/%R[0],%R?[4],%R?[5]
BIGRAM G_GC10:%L[0],%L[1],%L?[2]/%R[0],%R?[1],%R?[4],%R?[5]
BIGRAM G_GC13:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R?[4],%R?[5]
BIGRAM G_GC14:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R?[1],%R?[4],%R?[5]
BIGRAM GC_G01:%L[0],%L?[4],%L?[5]/%R[0]
BIGRAM GC_G02:%L[0],%L?[4],%L?[5]/%R[0],%R?[1]
BIGRAM GC_G03:%L[0],%L?[4],%L?[5]/%R[0],%R[1],%R?[2]
BIGRAM GC_G04:%L[0],%L?[4],%L?[5]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM GC_G05:%L[0],%L?[1],%L?[4],%L?[5]/%R[0]
BIGRAM GC_G06:%L[0],%L?[1],%L?[4],%L?[5]/%R[0],%R?[1]
BIGRAM GC_G07:%L[0],%L?[1],%L?[4],%L?[5]/%R[0],%R[1],%R?[2]
BIGRAM GC_G08:%L[0],%L?[1],%L?[4],%L?[5]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM C_GC01:%L?[4],%L?[5]/%R[0],%R?[4],%R?[5]
BIGRAM C_GC02:%L?[4],%L?[5]/%R[0],%R?[1],%R?[4],%R?[5]
```

```
BIGRAM GC_C01:%L[0],%L?[4],%L?[5]/%R?[4],%R?[5]
BIGRAM GC_C02:%L[0],%L?[1],%L?[4],%L?[5]/%R?[4],%R?[5]
BIGRAM GC_GC01:%L[0],%L?[4],%L?[5]/%R[0],%R?[4],%R?[5]
BIGRAM GC_GC02:%L[0],%L?[4],%L?[5]/%R[0],%R?[1],%R?[4],%R?[5]
BIGRAM GC_GC05:%L[0],%L?[1],%L?[4],%L?[5]/%R[0],%R?[4],%R?[5]
BIGRAM GC_GC06:%L[0],%L?[1],%L?[4],%L?[5]/%R[0],%R?[1],%R?[4],%R?[5]
BIGRAM O_O05:%L?[7]/%R?[7]
BIGRAM G_O01:%L[0]/%R?[7]
BIGRAM G_O02:%L[0],%L?[1]/%R?[7]
BIGRAM G_O03:%L[0],%L[1],%L?[2]/%R?[7]
BIGRAM G_O04:%L[0],%L[1],%L[2],%L?[3]/%R?[7]
BIGRAM GO_O01:%L[0],%L?[7]/%R?[7]
BIGRAM GO_O02:%L[0],%L?[1],%L?[7]/%R?[7]
BIGRAM GO_O03:%L[0],%L[1],%L?[2],%L?[7]/%R?[7]
BIGRAM GO_O04:%L[0],%L[1],%L[2],%L?[3],%L?[7]/%R?[7]
BIGRAM O_G01:%L?[7]/%R[0]
BIGRAM O_G02:%L?[7]/%R[0],%R?[1]
BIGRAM O_G03:%L?[7]/%R[0],%R[1],%R?[2]
BIGRAM O_G04:%L?[7]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM O_GO01:%L?[7]/%R[0],%R?[7]
BIGRAM O_GO02:%L?[7]/%R[0],%R?[1],%R?[7]
BIGRAM O_GO03:%L?[7]/%R[0],%R[1],%R?[2],%R?[7]
BIGRAM O_GO04:%L?[7]/%R[0],%R[1],%R[2],%R?[3],%R?[7]
BIGRAM C_O01:%L?[4],%L?[5]/%R?[7]
BIGRAM O_C01:%L?[7]/%R?[4],%R?[5]
BIGRAM O_CO01:%L?[7]/%R?[4],%R?[5],%R?[6]
BIGRAM GC_O01:%L[0],%L?[4],%L?[5]/%R?[7]
BIGRAM GC_O02:%L[0],%L?[1],%L?[4],%L?[5]/%R?[7]
BIGRAM O_GC01:%L?[7]/%R[0],%R?[4],%R?[5]
BIGRAM O_GC02:%L?[7]/%R[0],%R?[1],%R?[4],%R?[5]
BIGRAM G_GO01:%L[0]/%R[0],%R?[7]
BIGRAM G_GO02:%L[0]/%R[0],%R?[1],%R?[7]
BIGRAM G_GO03:%L[0]/%R[0],%R[1],%R?[2],%R?[7]
BIGRAM G_GO04:%L[0]/%R[0],%R[1],%R[2],%R?[3],%R?[7]
BIGRAM G_GO05:%L[0],%L?[1]/%R[0],%R?[7]
BIGRAM G_GO06:%L[0],%L?[1]/%R[0],%R?[1],%R?[7]
BIGRAM G_GO07:%L[0],%L?[1]/%R[0],%R[1],%R?[2],%R?[7]
BIGRAM G_GO08:%L[0],%L?[1]/%R[0],%R[1],%R[2],%R?[3],%R?[7]
BIGRAM G_GO09:%L[0],%L[1],%L?[2]/%R[0],%R?[7]
BIGRAM G_GO10:%L[0],%L[1],%L?[2]/%R[0],%R?[1],%R?[7]
BIGRAM G_GO11:%L[0],%L[1],%L?[2]/%R[0],%R[1],%R?[2],%R?[7]
BIGRAM G_GO12:%L[0],%L[1],%L?[2]/%R[0],%R[1],%R[2],%R?[3],%R?[7]
BIGRAM G_GO13:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R?[7]
BIGRAM G_GO14:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R?[1],%R?[7]
BIGRAM G_GO15:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R[1],%R?[2],%R?[7]
BIGRAM G_GO16:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R[1],%R[2],%R?[3],%R?[7]
BIGRAM GO_G01:%L[0],%L?[7]/%R[0]
BIGRAM GO_G02:%L[0],%L?[7]/%R[0],%R?[1]
BIGRAM GO_G03:%L[0],%L?[7]/%R[0],%R[1],%R?[2]
BIGRAM GO_G04:%L[0],%L?[7]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM GO_G05:%L[0],%L?[1],%L?[7]/%R[0]
BIGRAM GO_G06:%L[0],%L?[1],%L?[7]/%R[0],%R?[1]
BIGRAM GO_G07:%L[0],%L?[1],%L?[7]/%R[0],%R[1],%R?[2]
BIGRAM GO_G08:%L[0],%L?[1],%L?[7]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM GO_G09:%L[0],%L[1],%L?[2],%L?[7]/%R[0]
BIGRAM GO_G10:%L[0],%L[1],%L?[2],%L?[7]/%R[0],%R?[1]
BIGRAM GO_G11:%L[0],%L[1],%L?[2],%L?[7]/%R[0],%R[1],%R?[2]
BIGRAM GO_G12:%L[0],%L[1],%L?[2],%L?[7]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM GO_G13:%L[0],%L[1],%L[2],%L?[3],%L?[7]/%R[0]
BIGRAM GO_G14:%L[0],%L[1],%L[2],%L?[3],%L?[7]/%R[0],%R?[1]
BIGRAM GO_G15:%L[0],%L[1],%L[2],%L?[3],%L?[7]/%R[0],%R[1],%R?[2]
```

```
BIGRAM GO_G16:%L[0],%L[1],%L[2],%L?[3],%L?[7]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM GO_GO01:%L[0],%L?[7]/%R[0],%R?[7]
BIGRAM GO_GO02:%L[0],%L?[7]/%R[0],%R?[1],%R?[7]
BIGRAM GO_GO03:%L[0],%L?[7]/%R[0],%R[1],%R?[2],%R?[7]
BIGRAM GO_GO04:%L[0],%L?[7]/%R[0],%R[1],%R[2],%R?[3],%R?[7]
BIGRAM GO_GO05:%L[0],%L?[1],%L?[7]/%R[0],%R?[7]
BIGRAM GO_GO06:%L[0],%L?[1],%L?[7]/%R[0],%R?[1],%R?[7]
BIGRAM GO_GO07:%L[0],%L?[1],%L?[7]/%R[0],%R[1],%R?[2],%R?[7]
BIGRAM GO_GO08:%L[0],%L?[1],%L?[7]/%R[0],%R[1],%R[2],%R?[3],%R?[7]
BIGRAM GO_GO09:%L[0],%L[1],%L?[2],%L?[7]/%R[0],%R?[7]
BIGRAM GO_GO10:%L[0],%L[1],%L?[2],%L?[7]/%R[0],%R?[1],%R?[7]
BIGRAM GO_GO11:%L[0],%L[1],%L?[2],%L?[7]/%R[0],%R[1],%R?[2],%R?[7]
BIGRAM GO_GO12:%L[0],%L[1],%L?[2],%L?[7]/%R[0],%R[1],%R[2],%R?[3],%R?[7]
BIGRAM GO_GO13:%L[0],%L[1],%L[2],%L?[3],%L?[7]/%R[0],%R?[7]
BIGRAM GO_GO14:%L[0],%L[1],%L[2],%L?[3],%L?[7]/%R[0],%R?[1],%R?[7]
BIGRAM GO_GO15:%L[0],%L[1],%L[2],%L?[3],%L?[7]/%R[0],%R[1],%R?[2],%R?[7]
BIGRAM GO_GO16:%L[0],%L[1],%L[2],%L?[3],%L?[7]/%R[0],%R[1],%R[2],%R?[3],%R?[7]
BIGRAM C_CO01:%L?[4],%L?[5]/%R?[4],%R?[5],%R?[6]
BIGRAM CO_C01:%L?[4],%L?[5],%L?[6]/%R?[4],%R?[5]
BIGRAM CO_CO01:%L?[4],%L?[5],%L?[6]/%R?[4],%R?[5],%R?[6]
BIGRAM G_CO01:%L[0]/%R?[4],%R?[5],%R?[6]
BIGRAM G_CO02:%L[0],%L?[1]/%R?[4],%R?[5],%R?[6]
BIGRAM G_CO03:%L[0],%L[1],%L?[2]/%R?[4],%R?[5],%R?[6]
BIGRAM G_CO04:%L[0],%L[1],%L[2],%L?[3]/%R?[4],%R?[5],%R?[6]
BIGRAM GO_C01:%L[0],%L?[7]/%R?[4],%R?[5]
BIGRAM GO_C02:%L[0],%L?[1],%L?[7]/%R?[4],%R?[5]
BIGRAM GO_C03:%L[0],%L[1],%L?[2],%L?[7]/%R?[4],%R?[5]
BIGRAM GO_C04:%L[0],%L[1],%L[2],%L?[3],%L?[7]/%R?[4],%R?[5]
BIGRAM GO_CO01:%L[0],%L?[7]/%R?[4],%R?[5],%R?[6]
BIGRAM GO_CO02:%L[0],%L?[1],%L?[7]/%R?[4],%R?[5],%R?[6]
BIGRAM GO_CO03:%L[0],%L[1],%L?[2],%L?[7]/%R?[4],%R?[5],%R?[6]
BIGRAM GO_CO04:%L[0],%L[1],%L[2],%L?[3],%L?[7]/%R?[4],%R?[5],%R?[6]
BIGRAM C_GO01:%L?[4],%L?[5]/%R[0],%R?[7]
BIGRAM C_GO02:%L?[4],%L?[5]/%R[0],%R?[1],%R?[7]
BIGRAM C_GO03:%L?[4],%L?[5]/%R[0],%R[1],%R?[2],%R?[7]
BIGRAM C_GO04:%L?[4],%L?[5]/%R[0],%R[1],%R[2],%R?[3],%R?[7]
BIGRAM CO_G01:%L?[4],%L?[5],%L?[6]/%R[0]
BIGRAM CO_G02:%L?[4],%L?[5],%L?[6]/%R[0],%R?[1]
BIGRAM CO_G03:%L?[4],%L?[5],%L?[6]/%R[0],%R[1],%R?[2]
BIGRAM CO_G04:%L?[4],%L?[5],%L?[6]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM CO_GO01:%L?[4],%L?[5],%L?[6]/%R[0],%R?[7]
BIGRAM CO_GO02:%L?[4],%L?[5],%L?[6]/%R[0],%R?[1],%R?[7]
BIGRAM CO_GO03:%L?[4],%L?[5],%L?[6]/%R[0],%R[1],%R?[2],%R?[7]
BIGRAM CO_GO04:%L?[4],%L?[5],%L?[6]/%R[0],%R[1],%R[2],%R?[3],%R?[7]
BIGRAM GCO_GCO01:%L[0],%L?[4],%L?[5],%L?[6]/%R[0],%R?[4],%R?[5],%R?[6]
BIGRAM GCO_GCO02:%L[0],%L?[4],%L?[5],%L?[6]/%R[0],%R?[1],%R?[4],%R?[5],%R?[6]
BIGRAM GCO_GCO05:%L[0],%L?[1],%L?[4],%L?[5],%L?[6]/%R[0],%R?[4],%R?[5],%R?[6]
BIGRAM GCO_GCO06:%L[0],%L?[1],%L?[4],%L?[5],%L?[6]/%R[0],%R?[1],%R?[4],%R?[5],%R
    ?[6]
BIGRAM W_W01:%L?[8]/%R?[8]
BIGRAM G_W01:%L[0]/%R?[8]
BIGRAM G_W02:%L[0],%L?[1]/%R?[8]
BIGRAM G_W03:%L[0],%L[1],%L?[2]/%R?[8]
BIGRAM G_W04:%L[0],%L[1],%L[2],%L?[3]/%R?[8]
BIGRAM W_G01:%L?[8]/%R[0]
BIGRAM W_G02:%L?[8]/%R[0],%R?[1]
BIGRAM W_G03:%L?[8]/%R[0],%R[1],%R?[2]
BIGRAM W_G04:%L?[8]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM GW_GW01:%L[0],%L?[8]/%R[0],%R?[8]
BIGRAM GW_GW02:%L[0],%L?[8]/%R[0],%R?[1],%R?[8]
BIGRAM GW_GW03:%L[0],%L?[8]/%R[0],%R[1],%R?[2],%R?[8]
```

```
BIGRAM GW_GW04:%L[0],%L?[8]/%R[0],%R[1],%R[2],%R?[3],%R?[8]
BIGRAM GW_GW05:%L[0],%L?[1],%L?[8]/%R[0],%R?[8]
BIGRAM GW_GW06:%L[0],%L?[1],%L?[8]/%R[0],%R?[1],%R?[8]
BIGRAM GW_GW07:%L[0],%L?[1],%L?[8]/%R[0],%R[1],%R?[2],%R?[8]
BIGRAM GW_GW08:%L[0],%L?[1],%L?[8]/%R[0],%R[1],%R[2],%R?[3],%R?[8]
BIGRAM GW_GW09:%L[0],%L[1],%L?[2],%L?[8]/%R[0],%R?[8]
BIGRAM GW_GW10:%L[0],%L[1],%L?[2],%L?[8]/%R[0],%R?[1],%R?[8]
BIGRAM GW_GW11:%L[0],%L[1],%L?[2],%L?[8]/%R[0],%R[1],%R?[2],%R?[8]
BIGRAM GW_GW12:%L[0],%L[1],%L?[2],%L?[8]/%R[0],%R[1],%R[2],%R?[3],%R?[8]
BIGRAM GW_GW13:%L[0],%L[1],%L[2],%L?[3],%L?[8]/%R[0],%R?[8]
BIGRAM GW_GW14:%L[0],%L[1],%L[2],%L?[3],%L?[8]/%R[0],%R?[1],%R?[8]
BIGRAM GW_GW15:%L[0],%L[1],%L[2],%L?[3],%L?[8]/%R[0],%R[1],%R?[2],%R?[8]
BIGRAM GW_GW16:%L[0],%L[1],%L[2],%L?[3],%L?[8]/%R[0],%R[1],%R[2],%R?[3],%R?[8]
BIGRAM GCW_GCW01:%L[0],%L?[4],%L?[5],%L?[8]/%R[0],%R?[4],%R?[5],%R?[8]
BIGRAM GCW_GCW02:%L[0],%L?[4],%L?[5],%L?[8]/%R[0],%R?[1],%R?[4],%R?[5],%R?[8]
BIGRAM GCW_GCW05:%L[0],%L?[1],%L?[4],%L?[5],%L?[8]/%R[0],%R?[4],%R?[5],%R?[8]
BIGRAM GCW_GCW06:%L[0],%L?[1],%L?[4],%L?[5],%L?[8]/%R[0],%R?[1],%R?[4],%R?[5],%R
    ?[8]
BIGRAM OW_OW01:%L?[7],%L?[8]/%R?[7],%R?[8]
BIGRAM GOW_GOW01:%L[0],%L?[7],%L?[8]/%R[0],%R?[7],%R?[8]
BIGRAM GOW_GOW02:%L[0],%L?[7],%L?[8]/%R[0],%R?[1],%R?[7],%R?[8]
BIGRAM GOW_GOW03:%L[0],%L?[7],%L?[8]/%R[0],%R[1],%R?[2],%R?[7],%R?[8]
BIGRAM GOW_GOW04:%L[0],%L?[7],%L?[8]/%R[0],%R[1],%R[2],%R?[3],%R?[7],%R?[8]
BIGRAM GOW_GOW05:%L[0],%L?[1],%L?[7],%L?[8]/%R[0],%R?[7],%R?[8]
BIGRAM GOW_GOW06:%L[0],%L?[1],%L?[7],%L?[8]/%R[0],%R?[1],%R?[7],%R?[8]
BIGRAM GOW_GOW07:%L[0],%L?[1],%L?[7],%L?[8]/%R[0],%R[1],%R?[2],%R?[7],%R?[8]
BIGRAM GOW_GOW08:%L[0],%L?[1],%L?[7],%L?[8]/%R[0],%R[1],%R[2],%R?[3],%R?[7],%R
    ?[8]
BIGRAM GOW_GOW09:%L[0],%L[1],%L?[2],%L?[7],%L?[8]/%R[0],%R?[7],%R?[8]
BIGRAM GOW_GOW10:%L[0],%L[1],%L?[2],%L?[7],%L?[8]/%R[0],%R?[1],%R?[7],%R?[8]
BIGRAM GOW_GOW11:%L[0],%L[1],%L?[2],%L?[7],%L?[8]/%R[0],%R[1],%R?[2],%R?[7],%R
    ?[8]
BIGRAM GOW_GOW12:%L[0],%L[1],%L?[2],%L?[7],%L?[8]/%R[0],%R[1],%R[2],%R?[3],%R
    ?[7],%R?[8]
BIGRAM GOW_GOW13:%L[0],%L[1],%L[2],%L?[3],%L?[7],%L?[8]/%R[0],%R?[7],%R?[8]
BIGRAM GOW_GOW14:%L[0],%L[1],%L[2],%L?[3],%L?[7],%L?[8]/%R[0],%R?[1],%R?[7],%R
    ?[8]
BIGRAM GOW_GOW15:%L[0],%L[1],%L[2],%L?[3],%L?[7],%L?[8]/%R[0],%R[1],%R?[2],%R
    ?[7],%R?[8]
BIGRAM GOW_GOW16:%L[0],%L[1],%L[2],%L?[3],%L?[7],%L?[8]/%R[0],%R[1],%R[2],%R
    ?[3],%R?[7],%R?[8]
BIGRAM CO_COW01:%L?[4],%L?[5],%L?[6]/%R?[4],%R?[5],%R?[6],%R?[8]
BIGRAM COW_C01:%L?[4],%L?[5],%L?[6],%L?[8]/%R?[4],%R?[5]
BIGRAM COW_COW01:%L?[4],%L?[5],%L?[6],%L?[8]/%R?[4],%R?[5],%R?[6],%R?[8]
BIGRAM I_I01:%L?[11]/%R?[9],%R?[10]
BIGRAM IO_I01:%L?[11],%L?[7]/%R?[9],%R?[10]
BIGRAM I_IO01:%L?[11]/%R?[9],%R?[10],%R?[7]
BIGRAM IO_IO01:%L?[11],%L?[7]/%R?[9],%R?[10],%R?[7]
BIGRAM F_F01:%L?[9],%L?[10]/%R?[11]
BIGRAM F_FO01:%L?[9],%L?[10]/%R?[11],%R?[7]
BIGRAM FO_F01:%L?[9],%L?[10],%L?[7]/%R?[11]
BIGRAM FO_FO01:%L?[9],%L?[10],%L?[7]/%R?[11],%R?[7]
BIGRAM G_GI01:%L[0]/%R[0],%R?[9],%R?[10]
BIGRAM G_GI02:%L[0]/%R[0],%R?[1],%R?[9],%R?[10]
BIGRAM G_GI03:%L[0]/%R[0],%R[1],%R?[2],%R?[9],%R?[10]
BIGRAM G_GI04:%L[0]/%R[0],%R[1],%R[2],%R?[3],%R?[9],%R?[10]
BIGRAM G_GI05:%L[0],%L?[1]/%R[0],%R?[9],%R?[10]
BIGRAM G_GI06:%L[0],%L?[1]/%R[0],%R?[1],%R?[9],%R?[10]
BIGRAM G_GI07:%L[0],%L?[1]/%R[0],%R[1],%R?[2],%R?[9],%R?[10]
BIGRAM G_GI08:%L[0],%L?[1]/%R[0],%R[1],%R[2],%R?[3],%R?[9],%R?[10]
BIGRAM G_GI09:%L[0],%L[1],%L?[2]/%R[0],%R?[9],%R?[10]
```

```
BIGRAM G_GI10:%L[0],%L[1],%L?[2]/%R[0],%R?[1],%R?[9],%R?[10]
BIGRAM G_GI11:%L[0],%L[1],%L?[2]/%R[0],%R[1],%R?[2],%R?[9],%R?[10]
BIGRAM G_GI12:%L[0],%L[1],%L?[2]/%R[0],%R[1],%R[2],%R?[3],%R?[9],%R?[10]
BIGRAM G_GI13:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R?[9],%R?[10]
BIGRAM G_GI14:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R?[1],%R?[9],%R?[10]
BIGRAM G_GI15:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R[1],%R?[2],%R?[9],%R?[10]
BIGRAM G_GI16:%L[0],%L[1],%L[2],%L?[3]/%R[0],%R[1],%R[2],%R?[3],%R?[9],%R?[10]
BIGRAM GF_G01:%L[0],%L?[9],%L?[10]/%R[0]
BIGRAM GF_G02:%L[0],%L?[9],%L?[10]/%R[0],%R?[1]
BIGRAM GF_G03:%L[0],%L?[9],%L?[10]/%R[0],%R[1],%R?[2]
BIGRAM GF_G04:%L[0],%L?[9],%L?[10]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM GF_G05:%L[0],%L?[1],%L?[9],%L?[10]/%R[0]
BIGRAM GF_G06:%L[0],%L?[1],%L?[9],%L?[10]/%R[0],%R?[1]
BIGRAM GF_G07:%L[0],%L?[1],%L?[9],%L?[10]/%R[0],%R[1],%R?[2]
BIGRAM GF_G08:%L[0],%L?[1],%L?[9],%L?[10]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM GF_G09:%L[0],%L[1],%L?[2],%L?[9],%L?[10]/%R[0]
BIGRAM GF_G10:%L[0],%L[1],%L?[2],%L?[9],%L?[10]/%R[0],%R?[1]
BIGRAM GF_G11:%L[0],%L[1],%L?[2],%L?[9],%L?[10]/%R[0],%R[1],%R?[2]
BIGRAM GF_G12:%L[0],%L[1],%L?[2],%L?[9],%L?[10]/%R[0],%R[1],%R[2],%R?[3]
BIGRAM GF_G13:%L[0],%L[1],%L[2],%L?[3],%L?[9],%L?[10]/%R[0]
BIGRAM GF_G14:%L[0],%L[1],%L[2],%L?[3],%L?[9],%L?[10]/%R[0],%R?[1]
BIGRAM GF_G15:%L[0],%L[1],%L[2],%L?[3],%L?[9],%L?[10]/%R[0],%R[1],%R?[2]
BIGRAM GF_G16:%L[0],%L[1],%L[2],%L?[3],%L?[9],%L?[10]/%R[0],%R[1],%R[2],%R?[3]
```

参考文献

[1] Yasuharu Den, Junichi Nakamura, Toshinobu Ogiso, and Hideki Ogura. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008), Marrakech, Morocco*, pp. 1019–1024, 2008.

[2] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , Barcelona, Spain*, pp. 230–237, 2004.

[3] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning, Williamstown, MA., USA*, pp. 282–289, 2001.

[4] Yuji Matsumoto, Masayuki Asahara, Kou Kawabe, Yurika Takahashi, Yukio Tono, Akira Ohtani, and Toshio Morita. Chaki: An annotated corpora management and search system. In *Proceedings from the Corpus Linguistics Conference Series, Vol.1, no.1*, 2005.

[5] Toshinobu Ogiso, Mamoru Komachi, Yasuharu Den, and Yuji Matsumoto. Unidic for early middle Japanese: a dictionary for morphological analysis of classical Japanese. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey*, pp. 911–915, 2012.

[6] 上田英代. 源氏物語語彙用例総索引. 勉誠社, 1994.

[7] 星野雅英, 国文学研究資料館. 古典テキストデータ用データベースシステムの開発. 国文学研究資料館報告, 第 11 号. 国文学研究資料館, 1983.

[8] 伊藤雅光. 古典語文の自動分割（要旨）. 計量国語学, Vol. 16, No. 3, 1987.

[9] 伊藤雅光. コンピューターによる『今昔物語集』本文の自動分割の問題点, pp. 76–88. 桜楓社, 1988.

[10] 伊藤雅光. 漢文訓読文のための仮名列自動分割法–“長い単位”の場合. 計量国語学, Vol. 20, No. 4, pp. 139–167, 1996.

[11] 丸山岳彦. 第 5 章 日本語コーパスの発展, pp. 105–133. 講座日本語コーパス, No. 1. 朝倉書店, 2013.

[12] 前川喜久雄. 第 1 章 コーパスの存在意義, pp. 1–29. 講座日本語コーパス, No. 1. 朝倉書店, 2013.

[13] 風間喜代三, 荻野綱男, 豊島正之. 「ぎやどぺかどる」の読解に於ける電子計算機の利用の試み : 科学研究費研究報告書. [風間喜代三], 1983.

[14] 吉澤義則, 木之下正雄. 對校源氏物語用語索引. 平凡社, 1952.

[15] 金水敏. 古文献の計算機処理 : 東京大学国語研究室蔵恵果和上之碑文. [金水敏], 1984.

[16] 山元啓史. 和歌のための品詞タグづけシステム. 日本語の研究, Vol. 3, No. 22, pp. 33–39, 2007.

[17] 宮田光, 稲賀敬二. 恋路ゆかしき大将. 山路の露. 中世王朝物語全集, No. 8. 笠間書院, 2004.

[18] 大塚光信. 大蔵虎明能狂言集 翻刻註解. 清文堂出版, 2006.

[19] 竹内孔一, 松本裕治. 隠れマルコフモデルによる日本語形態素解析のパラメータ推定. 情報処理学会論文誌, Vol. 38, No. 3, pp. 500–509, 1997.

[20] 池田幸恵, 須永哲矢. 『五国史』宣命のコーパス化. 第 4 回コーパス日本語学ワークショップ予稿集, pp. 187–194, 2013.

[21] 西端幸雄. マイコンによる索引作り. 樟蔭国文学, Vol. 21, pp. 108–126, 1983.

[22] 西端幸雄, 木村雅則, 志甫由紀恵. 平安日記文学総合語彙索引 : 土佐日記・蜻蛉日記・和泉式部日記・紫式部日記・更級日記. 勉誠社, 1996.

[23] 西端幸雄, 藤田久, 成田徹. パーソナル・コンピュータによる語彙索引自動作成の試み. [西端幸雄], 1989.

[24] 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. 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 (特集 コーパス日本語学の射程). 日本語科学, Vol. 22, pp. 101–123, 2007.

[25] 国文学研究資料館. 古典選集本文データベース. http://base1.nijl.ac.jp/~anthologyfulltext/.

[26] 国立国語研究所. 総合雑誌の用語 : 現代語の語彙調査. 国立国語研究所報告, No. 12,13. 国立国語研究所, 1957.

[27] 国立国語研究所. 国定読本用語総覧 : CD-ROM 版. 三省堂, 1997.

[28] 国立国語研究所. 雑誌『太陽』による確立期現代語の研究 : 『太陽コーパス』研究論文集. 国立国語研究所報告, No. 122. 博文館新社, 2005.

[29] 国立国語研究所. 太陽コーパス : 雑誌『太陽』日本語データベース. 国立国語研究所資料集, No. 15. 博文館新社, 2005.

[30] 国立国語研究所. 近代女性雑誌コーパス. 国立国語研究所, 2006.

[31] 国立国語研究所. 明六雑誌コーパス. 国立国語研究所, 2012.

[32] 国立国語研究所, 見坊豪紀, 水谷静夫, 石綿敏雄, 宮島達夫. 現代雑誌九十種の用語用字. 国立国語研究所報告, No. 21-22, 25. 国立国語研究所, 1962.

[33] 国立国語研究所, 情報通信研究機構, 古井貞熙, 前川喜久雄, 井佐原均. 日本語話し言葉コーパス. 国立国語研究所, 情報通信研究機構, 2004.

[34] 国立国語研究所コーパス開発センター. 現代日本語書き言葉均衡コーパス. 国立国語研究所コーパス開発センター, 2011.

[35] 京都大学黒橋・河原研究室. 京都大学テキストコーパス. http://nlp.ist.i.kyoto-u.ac.jp/index.php?%E4%BA%AC%E9%83%BD%E5%A4%A7%E5%AD%A6%E3%83%86%E3%82%AD%E3%82%B9%E3%83%88%E3%82%B3%E3%83%BC%E3%83%91%E3%82%B9.

[36] 中野三敏, 神保五弥, 前田愛. 洒落本・滑稽本・人情本. 新編日本古典文学全集, No. 80. 小学館, 2000.

[37] 紫式部, 阿部秋生, 秋山虔, 今井源衛. 源氏物語. 日本古典文学全集 / 秋山虔 [ほか] 編, No. 12-17. 小学館, 1970.

[38] 小学館. 新編日本古典文学全集. 小学館, 1994–2002.

[39] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内左夜香, 小西光, 原裕. 『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（上・下）. 国立国語研究所内部報告書, No. LR-CCG-10-05. 国立国語研究所, 第 4 版, 2011.

[40] 小椋秀樹, 須永哲矢. 中古和文 UniDic 短単位規程集（科研費基盤研究（C）課題番号 21520492「和文系資料を対象とした形態素解析辞書の開発」研究成果報告書 2）. Technical report, 国立国語研究所, 2012.

[41] 小椋秀樹, 冨士池優美. 第 4 章 形態論情報. 『現代日本語書き言葉均衡コーパス』マニュアル, pp. 39–74. 国立国語研究所コーパス開発センター, 2011.

[42] 小木曽智信, 中村壮範. 『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装. 国立国語研究所内部報告書, No. LR-CCG-10-06. 国立国語研究所, 改訂版, 2011.

[43] 小澤俊介, 内元清貴, 伝康晴. BCCWJ に基づく中・長単位解析ツール. 特定領域「日本語コーパス」平成 22 年度公開ワークショップ予稿集, pp. 331–338, 2011.

[44] 安永尚志. 国文学研究とコンピュータ. 勉誠社, 1998.

[45] 山口昌也. 全文検索システム『ひまわり』を利用した言語資料検索環境の構築手法. 日本語科学, Vol. 21, pp. 111–123, 2007.

[46] 岡島昭浩. 日本文学等テキストファイル. `http://www.let.osaka-u.ac.jp/~okajima/bungaku.htm`.

[47] 松本裕治. 日本語形態素解析システム JUMAN 使用説明書,version 1.0. Technical report, 京都大学工学部長尾研究室, 1992.

[48] 松本裕治. 形態素解析システム『茶筌』. 情報処理, Vol. 41, pp. 1208–1214, 2000.

[49] 松本裕治, 浅原正幸, 橋本喜代太, 投野由紀夫, 大谷朗, 森田敏生. タグ付きコーパス管理/検索システム「茶器」使用説明書 version 2.1. Technical report, 奈良先端科学技術大学院大学, 2007.

[50] 岡照晃, 小町守, 小木曽智信, 松本裕治. 未整備の歴史的文献への濁点の自動付与アプリケーション. じんもんこん 2012 論文集, 第 2012 巻, pp. 191–198, 2012.

[51] 岡照晃, 小町守, 小木曽智信, 松本裕治. 統計的機械学習を用いた歴史的資料への濁点付与の自動化. 情報処理学会論文誌, Vol. 54, No. 4, pp. 1641–1654, 2013.

[52] 岡照晃, 小町守, 小木曽智信, 松本裕治. 表記のバリエーションを考慮した近代日本語の形態素解析. 人工知能学会全国大会 (JSAI2013), pp. 323–332, 2013.

[53] 新情報処理開発機構（RWCP）テキスト・サブ・ワーキンググループ. 研究開発用知的資源: タグ付きテキストコーパス報告書. Technical report, 新情報処理開発機構, 1998.

[54] 大野晋. 日本語と私. 朝日新聞社, 1999.

[55] 菊池真一, 深沢秋男. J-TEXT 日本文学電子図書館. http://www.j-texts.com/.

[56] 長瀬真理. 日本語-英語対照「源氏物語」のテキストデータベースの作成に関する基礎的研究. 情報知識学会誌, Vol. 1, pp. 40–53, 1990.

[57] 村上征勝. シェークスピアは誰ですか？—計量文献学の世界. 文春新書, 2004.

[58] 浅原正幸. NAIST Japanese Dictionary. http://sourceforge.jp/projects/naist-jdic/.

[59] 小林正行, 市村太郎. 『虎明本狂言集』コーパスの構造化 ー仕様と事例の検討ー. 第 3 回コーパス日本語学ワークショップ予稿集, pp. 323–332, 2013.

[60] 豊島正之. キリシタン版ぎやどぺかどる本文・索引. 清文堂, 1987.

[61] 時枝誠記. 徒然草總索引. 至文堂, 1955.

[62] 浅原正幸. IPADIC ユーザーズマニュアル. Technical report, 奈良先端科学技術大学院大学, 2002.

[63] 高山善行, 青木博史. ガイドブック日本語文法史. ひつじ書房, 2010.

[64] 市村太郎, 河瀬彰宏, 小木曽智信. 洒落本コーパスの構造化 —仕様と事例の検討—. 第 3 回コーパス日本語学ワークショップ予稿集, pp. 249–258, 2013.

[65] 近藤泰弘. 文法研究と電子化テキスト (国語研究資料の「電子化」とその利用〈国語学会〔平成 4 年度〕春季大会テーマ発表〉). 国語学, pp. p128–123, 1992.

[66] 近藤泰弘. 古典語・古典文学研究における言語処理, pp. 472–473. 共立出版, 2009.

[67] 近藤泰弘. 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 VSARPJ 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, 2012.

[68] 工藤拓. CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer. `https://code.google.com/p/cabocha/`.

[69] 工藤拓. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. `https://code.google.com/p/mecab/`.

[70] 宮島達夫. 総索引への注文. 国語学, Vol. 76, pp. 110–120, 1969.

[71] 宮島達夫. 古典対照語い表. 笠間索引叢刊, No. 4. 笠間書院, 1971.

[72] 宮島達夫. 古典対照語い表. 笠間書院, フロッピー版, 1989.

[73] 小木曽智信. 旧仮名遣いの口語文を対象とした形態素解析辞書. じんもんこん 2012 論文集, 第 2012 巻, pp. 25–32, 2012.

[74] 小木曽智信. 中古仮名文学作品の形態素解析. 日本語の研究, Vol. 9, 4（通巻 255 号）, pp. 49–62, 2013.

[75] 小木曽智信, 市村太郎, 鴻野知暁. 近世口語資料の形態素解析の試み. 第 4 回コーパス日本語学ワークショップ予稿集, pp. 145–150, 2013.

[76] 小木曽智信, 小町守, 松本裕治. 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, 2013.

[77] 小木曽智信, 小椋秀樹, 小磯花絵, 宮内佐夜香, 渡部涼子, 伝康晴. 形態素解析辞書のベンチマークテスト—ipadic・NAIST-jdic・UniDic のジャンル別精度比較—. 言語処理学会第 16 回年次大会発表論文集, pp. 326–329, 2010.

[78] 小木曽智信, 小椋秀樹, 近藤明日子. 近代文語文を対象とした形態素解析辞書・近代文語 UniDic. 日本語学会 2008 年度春季大会予稿集, pp. 211–218, 2008.

[79] 小木曽智信, 岡照晃, 小町守, 松本裕治. コーパス管理ツール「茶器」による単語情報付き古典語コーパスの活用. じんもんこん 2011 論文集, 第 2011 巻, pp. 255–260, 2011.

[80] 小木曽智信, 中村壮範. 『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用. 自然言語処理, Vol. 21, No. 2, 印刷中.

[81] 小木曽智信, 須永哲矢. 「近代文語 UniDic」「中古和文 UniDic」を利用した総索引作成システムの開発. じんもんこん 2010 論文集, 第 2010 巻, pp. 119–124, 2010.

[82] 小木曽智信, 須永哲矢, 冨士池優美, 中村壮範, 田中牧郎, 近藤泰弘. 「日本語歴史コーパス 平安時代編」先行公開版について. 第 3 回コーパス日本語学ワークショップ予稿集, pp. 269–276, 2013.

[83] 市古貞次, 三角洋一. 鎌倉時代物語集成. 笠間書院, 1988.

[84] 正宗敦夫. 萬葉集總索引. 白水社 : 萬葉閣, 1929.

[85] 内元清貴, 小澤俊介, 伝康晴. 長・中単位解析ツール Comainu ver. 0.6 ユーザーズマニュアル. Technical report, The UniDic consortium, 2011.

[86] 日本国語大辞典第二版編集委員会, 小学館国語辞典編集部. 日本国語大辞典. 小学館, 第 2 版, 2000.

[87] 安武満佐子, 吉村賢治, 首藤公昭. 古文の形態素解析システム. 福岡大学工学集報, Vol. 54, pp. 157–165, 1995.

[88] 山本靖, 松本裕治. 日本語形態素解析システム JUMAN による古文の形態素解析とその応用. 情報処理語学文学研究会 第 19 回研究発表大会要旨, 1996.

[89] 冨士池優美. 中古和文における長単位の概要. 第 2 回コーパス日本語学ワークショップ予稿集, pp. 51–58, 2012.

[90] 冨士池優美, 河瀬彰宏, 野田高広, 岩崎瑠莉恵. 『今昔物語集』のテキスト整形. 第 4 回コーパス日本語学ワークショップ予稿集, pp. 125–134, 2013.

[91] 築島裕. 平安時代語新論. 東大人文科学研究叢書. 東京大学出版会, 1969.

[92] 上田裕一, 上田英代, 村上征勝. 源氏物語の自動単語分割と計量分析. 文献情報のデータベースとその利用に関する研究会, 1992.

[93] 洒落本大成編集委員会. 洒落本大成. 中央公論社, 1978.

外部発表一覧

学術雑誌

- 小木曽 智信，中村壮範「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」，自然言語処理 21(2) ，印刷中，2014 年 4 月
- 小木曽 智信，小町 守，松本 裕治「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 20(5)，pp.727-748　2013 年 12 月
- 小木曽 智信「中古仮名文学作品の形態素解析」日本語の研究　9 巻 4 号（通巻 255 号），pp.49-62　2013 年 10 月
- 岡 照晃，小町 守，小木曽 智信，松本 裕治「統計的機械学習を用いた歴史的資料への濁点付与の自動化」情報処理学会論文誌 54(4)，pp.1641-1654　2013 年 4 月

国際会議

- Toshinobu Ogiso，Mamoru Komachi，Yasuharu Den and Yuji Matsumoto. 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 (LREC2012)*，　pp.911-915　2012 年 5 月 23 日，Istanbul，Turkey.
- Teruaki Oka，Mamoru Komachi，Toshinobu Ogiso and Yuji Matsumoto . Automatic Labeling of Voiced Consonants for Morphological Analysis of Modern Japanese Literature, In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP2011)*，　pp.410-419　2011 年 11 月 9 日，Chiang Mai，Thailand

その他の発表

- 河瀬 彰宏，市村 太郎，小木曽 智信「TEI P5 に基づく近世口語資料の構造化とその問題点」，情報処理学会 人文科学とコンピュータシンポジウム「じんもんこん 2013」（PNC/ECAI 共催），京都大学，2013 年 12 月 11 日
- 河瀬 彰宏，小木曽 智信 “The Current Situation and Role of TEI P5 as an XML Standard for the Corpus of Historical Japanese”，国際シンポジウム デジタル時代の人文学と仏教学の役割について，東京大学，2013 年 11 月 17 日

- 近藤 明日子，高田 智和，小木曽 智信，堤 智昭「原本画像参照機能付き『明六雑誌コーパス』の開発」日本語学会 2013 年度秋季大会，静岡大学，2013 年 10 月 27 日
- 小木曽 智信，市村 太郎，鴻野 知暁「近世口語資料の形態素解析の試み」第 4 回コーパス日本語学ワークショップ，国立国語研究所，2013 年 9 月 5 日
- 岡 照晃，小町 守，小木曽 智信，松本 裕治「表記のバリエーションを考慮した近代日本語の形態素解析」人工知能学会全国大会(JSAI2013)，富山国際会議場，2013 年 6 月
- 小木曽 智信，須永 哲矢，冨士池 優美，中村 壮範，田中 牧郎，近藤 泰弘（2013）「日本語歴史コーパス 平安時代編」の先行公開　第３回コーパス日本語学ワークショップ，国立国語研究所，2013 年 3 月
- 市村 太郎，河瀬彰宏，小木曽 智信「洒落本コーパスの構造化　―仕様と事例の検討―」，第３回コーパス日本語学ワークショップ，国立国語研究所，2013 年 3 月
- 近藤 明日子，小木曽 智信，須永 哲矢，田中 牧郎「形態論情報付き近代語コーパスのアノテーション　−『明六雑誌コーパス』を例として−」言語処理学会第 19 回年次大会，名古屋大学，2013 年 3 月 13 日
- 小木曽 智信，伝 康晴「UniDic2: 拡張性と応用可能性にとんだ電子化辞書」言語処理学会第 19 回年次大会，名古屋大学，2013 年 3 月 15 日
- 市村 太郎，河瀬 彰宏，小木曽 智信「近世口語テキストの構造化とその課題」，情報処理学会 人文科学とコンピュータ研究会（CH96），国文学研究資料館，2012 年 10 月 12 日
- 小木曽 智信，「旧仮名遣いの口語文を対象とした形態素解析辞書」，情報処理学会 人文科学とコンピュータシンポジウム「じんもんこん 2012」，北海道大学，2012 年 11 月 17 日
- 岡 照晃，小町 守，小木曽 智信，松本 裕治「未整備の歴史的資料への濁点の自動付与アプリケーション」，情報処理学会 人文科学とコンピュータシンポジウム「じんもんこん 2012」，北海道大学，2012 年 11 月 17 日
- 小木曽 智信，中村 壮範「通時コーパス用 Web アプリケーション「中納言」のデモンストレーション」，NINJAL「通時コーパス」プロジェクト・Oxford VSARPJ 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，2012 年 7 月 31 日

- 小木曽 智信「コーパス管理ツール「茶器」による中古和文コーパスの利用」第 3 回コーパス日本語学ワークショップ」，2012 年 3 月
- 小木曽 智信，岡 照晃，小町 守，松本 裕治「コーパス管理ツール「茶器」による単語情報付き古典語コーパスの活用」，情報処理学会 人文科学とコンピュータシンポジウム「じんもんこん 2011」　龍谷大学 査読有 2011 年 11 月
- 小木曽 智信「通時コーパスの構築に向けた古文用形態素解析辞書の開発」，情報処理学会人文科学とコンピュータ研究会（CH92），国立国語研究所，2011 年 1 月
- Toshinobu OGISO「歴史的資料を対象とした形態素分析辞書によるテキスト解析」, The 13th international conference of the European association for Japanese Studies (EAJS) Tallinn，Estonia，2011 年 8 月
- 岡 照晃，小町 守，小木曽 智信，松本 裕治「機械学習による近代文語文への濁点の自動付与」情報処理学会第 201 回自然言語処理研究会，東京大学，2011 年 5 月
- 小木曽 智信，須永 哲矢「「近代文語 UniDic」「中古和文 UniDic」を利用した総索引作成システムの開発」，情報処理学会 人文科学とコンピュータシンポジウム「じんもんこん 2010」，東京工業大学，2010 年 12 月
- 近藤 明日子，小木曽 智信，加藤 文明子「『高等小学読本』の形態論情報付きコーパス」, 情報処理学会 人文科学とコンピュータシンポジウム「じんもんこん 2010」, 東京工業大学，2010 年 12 月

受賞

- 小木曽 智信　2010 年度（平成 22 年度）情報処理学会 山下記念研究賞「中古和文を対象とした形態素解析辞書の開発」（2010-CH-85），2011 年 3 月
- Akihiro Kawase，Taro Ichimura，Toshinobu Ogiso“Problems in TEI P5 Encoding on Colloquial Japanese Documents of the Early Modern Period”，PNC/ECAI & Jinmoncom (IPSJ SIG-SH) Joint Meeting 2013 (Kyoto Univerisity)，Best Poster Award Gold Prize, 2013 年 12 月