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

小木曽 智信*‡　岡　照晃‡　小町 守‡　松本 裕治‡

*人間文化研究機構 国立国語研究所 言語資源研究系
‡奈良先端科学技術大学院大学 情報科学研究科

近年、古典文学などの歴史的な日本語テキストの自動形態素解析が可能になった。しかし、一般的な日本語研究者にとって、言語学的な研究のために形態素解析済みのコーパスを利用することは困難であった。これは主として、検索や集計、統計的な分析のための十分なツールが存在しなかったためである。だが、形態素解析済みのコーパスを作成・検索するための汎用コーパス管理ツール「茶器」が開発された。「茶器」は形態素解析や係り受け解析などの多様な言語学的アノテーションを扱うことが可能であり、その機能は歴史的な日本語資料の言語学的研究に十分なものとなってきている。そこで、我々は「茶器」を形態素解析済みの歴史的なテキストに応用した。これにより歴史的な日本語資料を対象にした自由度の高い検索や、n グラムや MI スコアなどを用いた統計的な分析が可能になる。

Application of the corpus management tool ChaKi to morphologically annotated corpora of classical Japanese

Toshinobu Ogiso*‡　Teruaki Oka‡　Mamoru Komachi‡　Yuji Matsumoto‡
* Department of Corpus Studies, National Institute for Japanese Language and Linguisics, National Institutes for the Humanities
‡ Graduate School of Information Science, Nara Institute of Science and Technology

Recently, automatic morphological analysis for historical Japanese texts has become practical. However, it is difficult for general Japanese linguists to use automatically tagged corpora for their linguistic research. This is chiefly because there do not exist effective tools for search, summarization and statistical analysis on linguistic corpus. Nevertheless, a general corpus management tool ChaKi was developed for creating and searching annotated corpora. Since ChaKi can deal with various linguistic annotations including morphological analysis and dependency relation, its function is sufficient for linguistic research of historical Japanese texts. Accordingly, we adopted ChaKi to morphologically analyzed texts of historical Japanese. It allows flexible search and statistical analysis such as n-gram frequencies, mutual information (MI) scores, and frequent sequence mining on historical Japanese.

1. はじめに

近年、日本語学においてコーパスを用いた研究がますます活発になっている。国立国語研究所を中心に開発が行われ、今年完成した「現代日本語書き言葉均衡コーパス」（BCCWJ）は約 1 億語という大規模なコーパスであり、現代語研究の基礎となるデータとして活用が進められている。さらに、現代語のみならず日本語の歴史的な研究においても、コーパスが大きな位置を占めつつある。たとえば、国立国語研究所では「通時コーパス」の構築が計画され研究が始まっている[1]ほか、オックスフォード大学日本学研究センターでは上代日本語を中心とした歴史的な日本語コーパスが開発されている[2]。

こうした歴史的な日本語を対象としたコーパス構築のために、既存の形態素解析技術を応用して古文の形態素解析も実用化されてきた。しかし、その解析結果を研究者が利用するツールが整っていなかったため、これまではその利用が十分に進んでいなかった。

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ている。

現代語と比較してテキスト量が少ない古典語を対象とした研究では、高い精度でタグ付けされたコーパスが期待されるため、人手によるタグ付けが柔軟に行えるプログラムが求

められる。その一方で、多くの文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「茶器」[3]がある。本発表はこの「茶器」に形態素解析済みの古典語のデータを格納し、日本語の歴史研究での活用を試みるものである。「茶器」の高度な検索や統計的処理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を支援する環境を整備する。

2. 古文の形態素解析

現代日本語の自動形態素解析は，JUMAN や ChaSen，MeCab[4]などにより 1990 年代後半以降，高い精度での解析が可能になっている。特に、ChaSen 以降の機械学習に基づく形態素解析技術は、人手でのルール整備を不要にし、辞書と学習用のコーパスをもとにして高精度の解析を行うことを可能にした。

しかし、古典文学作品などの歴史的資料の形態素解析については，様々な先駆的試みがあったものの、本格的な古語の電子化辞書と学習用の古典語コーパスが不足していたため、実用的な精度で実現することは長らくできなかった。

近年， BCCWJ のタグ付けを行うための現代語の形態素解析辞書「UniDic」[5]が新たに開発された。これは言語研究に利用することを念頭に設計されており、短単位という斉一な解析単位、階層化され目的に応じて利用可能な階層化された見出し語を特長としている。

発表者らは、この UniDic をもとに、歴史的な日本語資料の解析に必要な見出し語を追加し、学習用の古文のコーパスを整備することによって、歴史的な日本語資料を対象とした形態素解析辞書の開発を行ってきた。形態素解析器 MeCab とともに用いることで、「近代文語 UniDic」[6]によって明治期の文語論説文の形態素解析が可能となったほか，「中古和文 UniDic」[6][7]によって平安時代の和文系資料についても解析が可能となった。

これにより、古典文学作品を解析して単語情報付きのコーパスを比較的容易に作成することが可能になった。古典文学作品については作品ごとの総索引がすでに整備されているが、紙の索引では検索性・検索結果の利用・統計的処理などの点で難がある上に、作品ごとに異なる単語認定基準によっているため単純に集計ができないといった問題があった。しかし UniDic による単語情報付きのコーパスであればこうした問題を一度に解決することができる。

しかし、これまでは、形態素解析を施してもその解析結果を活用するためのツールがないという問題があった。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、多くの文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ルーツが必要とされる。

また、歴史的資料を対象とした場合、形態素解析の精度はおよそ 95～97%程度である。限られた資料をもとに研究を進める必要がある古典語研究にとっては、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、研究目的にもよるが、この解析精度はそのまま研究に用いるには十分でなく、解析結果を修正して精度を高める必要がある。しかし修正を行うためのツールがないため、十分な活用ができないという側面があった。

3. コーパス管理ツール「茶器」の特長

「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）

多言語対応を目指していることもあり、文字コードは Unicode に対応している（内部文字コードは UTF-16）。そのため、古文には現れるが一般的には使用頻度の低い文字であっても取り扱うことが可能である。

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 [8]による係り受け解析結果である。付属のデータインポート支援ツール「Text Formatter」を用いることで、容易にタグ付きデータをインポートして利用することができる。形態素解析辞書としては IPADIC と UniDic に対応しているため「中古和文 UniDic」で解析された古典語の単語情報付きコーパスも容易に格納することも可能である。

「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。

4. 形態素解析済み古典語コーパスのインポート

「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。図 1 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。

画面上に配置された各種のパネルによって、コーパスの検索・集計・統計情報の取得から、コーパスの修正、新しいデータのインポートまで行うことができる。

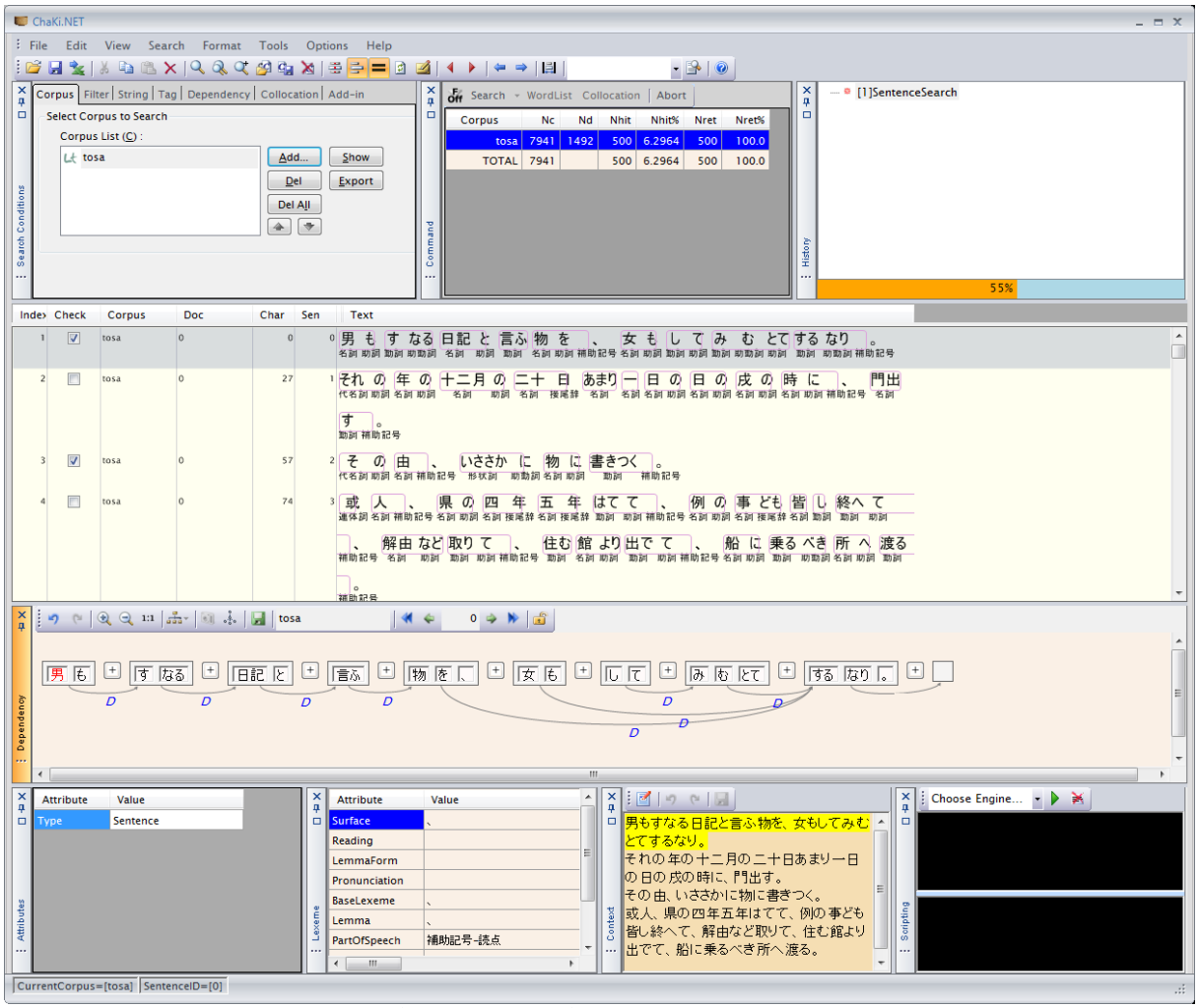

図 1　ChaKi.NET の実行画面（「土佐日記」）

インポートしたデータ

今回「茶器」にインポートした古典語コーパスは表 1 の通りである。合計で約 58.8 万語になる。一部のデータについては、自動解析結果に対して人手による修正を加えている。

表 1　古典語コーパスの作品一覧

作品名	語数	人手修正
伊勢物語	14624	済み
源氏物語（全）	528734	一部のみ
土佐日記	7948	済み
更級日記	16658	済み
紫式部日記	20353	一部のみ

現時点では、係り受けのタグ付けまで行った古典語のデータは存在しないため、今回用意したデータは、人手修正済みのものを含め、すべて形態素解析までしか行われていないものである。したがって、本来であれば「茶器」には形態素解析結果（*.mecab 形式）の取り込みしか行えない。

しかし、今回、これらのデータに対して実際に文節の係り受けをタグ付けを試行するために、次の簡単なルールに基づいて文節相当と考えられる部分をまとめ上げることにした。

1）助詞・助動詞は，直前の自立語または助詞・助動詞とつなげる
2）接尾辞・句読点は直前の形態素とつなげる
3）接頭辞は直後の形態素とつなげる

このルールにより、9 割以上の文節が正しく分割された。そこで、このようにまとめ上げたものを*.cabocha 形式のファイルに変換した後、TextFormatter を用いてインポートした。

形態論情報

「茶器」は、形態論情報として、次の基本 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、その属性数は計 20 以上に上る。これらをすべて同レベルで扱うことはできないため、基本 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格納している。

- custom＝カスタムフィールド

今回の中古和文 UniDic の解析結果では 9 属性以外の情報を次のようにカスタムフィールドに格納している。

custom="goshu 和 pronBase ムカシ kanaBase ムカシ formBase ムカシ"

goshu＝語種（和語・漢語・外来語等）
pronBase＝発音形基本形
kanaBase＝仮名形基本形
formBase＝語形基本形

5. タグ付けツールとしての利用

形態素解析結果の修正

先述したとおり、古典語コーパスを研究に利用する際、一般に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。

文節係り受けのタグ付け

また、先述したとおり、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。しかし、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。「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 2 は、「茶器」の文節係り受けのアノテーション画面（Dependency パネル）である。文節間のリンクのドラッグアンドドロップなどマウス操作だけで係り受けのタグ付けができるほか、文節の切り直しなども可能である。

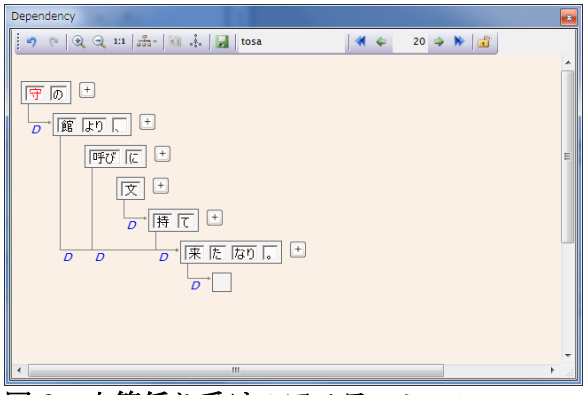

図 2　文節係り受けのアノテーション

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で、一通りのタグ付けを行うことも可能だと思われる。

今回は、検証のためにタグ付けの試用のみを行ったが、今後、係り受け情報付きの古典語コーパスの実現に向けて、一部のデータへのタグ付けを進めていきたいと考えている。

6. 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。

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、「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。

ここでは、このような新しい利用の例として、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、コロケーション統計、そして係り受け検索について紹介する。

こうした処理では、データのサイズと処理速度が絶えず問題になるが、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。今回用意した 58.8 万語のコーパスでは、以下の時間のかかる検索であってもすべて 20 秒以内に取得することができた（Windows7 x64, Core i7 2.53GHz, 8GB RAM の環境で確認）。

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。

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。図 3 の例では、動詞の後 2 語以内に助動詞「つ」が来る例を検索している。

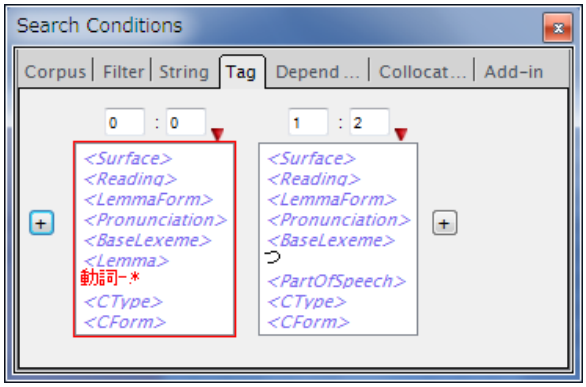

図 3　語検索（共起条件の設定）

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 2 は、この機能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。

表 2　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

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。

コロケーション

検索条件を指定して検索を行った後に、Collocation タブで設定することで、表示されている KWIC を対象にした各種の統計を取ることができる。取得できる情報は、粗頻度、MI スコア（相互情報量）、N-gram 頻度、FSM（Frequent Sequence Mining）である。

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、現時点では人手修正済みのわずかなデータしか利用できないが、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。

6. おわりに

「茶器」を用いることで、課題だった形態素解析済み古典語データの利用環境を整備することができた。今後、単語情報付きコーパスの豊富な情報と「茶器」の高度な検索機能や統計情報を用いて、新しい古典語研究が行われることに期待したい。

システムの面では、今後データ整備を行って古典語の自動係り受け解析を実現し、古典語についても係り受けを用いた検索等を可能にしていきたい。

参考文献

[1] 国立国語研究所基幹型共同研究プロジェクト「通時コーパスの設計」 http://www.ninjal.ac.jp/research/project/a/corpus/
[2] The Oxford Corpus of Old Japanese http://vsarpj.orinst.ox.ac.uk/corpus/
[3] 「茶器」http://sourceforge.jp/projects/chaki/
[4] MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://mecab.sourceforge.net/feature.html
[5] 形態素解析辞書 UniDic http://download.unidic.org
[6] 「近代文語 UniDic」「中古和文 UniDic」 http://www2.ninjal.ac.jp/lrc/index.php?UniDic
[7] 小木曽智信・小椋秀樹・田中牧郎・近藤明日子・伝 康晴（2010）「中古和文を対象とした形態素解析辞書の開発」情報処理学会研究報告 人文科学とコンピュータ Vol.2010-CH-85(No.4) pp.1-8
[8] CaboCha: Yet Another Japanese Dependency Structure Analyzer http://code.google.com/p/cabocha/