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

小木曽 智信（国立国語研究所言語資源研究系）†

Application of the Corpus Management Tool ChaKi to the Annotated Corpus of Early Middle Japanese

Toshinobu Ogiso (Dept. Corpus Studies, NINJAL)

１．はじめに

『現代日本語書き言葉均衡コーパス』（BCCWJ）の完成を受けて、新たに歴史的な日本語のコーパスの構築が進められつつある。国立国語研究所では「通時コーパスの設計」プロジェクト（近藤泰弘リーダー）を中心に構築のための研究が始まっている。こうした歴史的なコーパスのテキストを解析するために、古文のための形態素解析も研究開発が進められており、古典文学作品のテキストについては、実用的な精度で解析を行うことが可能になっている。

しかし、これまでは形態素解析済みの古文のデータの利用が十分に進んでいなかった。その理由の一つに、形態素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ことがある。古典語は現代語と比較して利用可能なテキスト量が少ないため、研究のためには貴重なテキストに対して高い精度でタグ付けを行う必要がある。そのためには、自動処理に加えて、人手によるタグ付けが柔軟に行えるツールが必要である。また、ツールは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。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。

このようなニーズを満たすものとして奈良先端科学技術大学院大学で開発された「茶器」がある。本発表はこの「茶器」に形態素解析を施した中古和文のコーパスを格納し、研究への応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視点からの古典語研究が可能になると思われる。

２．「中古和文 UniDic」

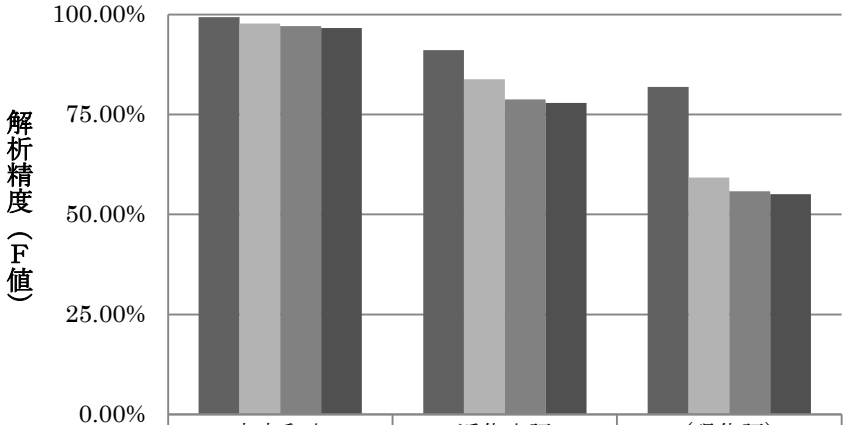
日本語の自動形態素解析は 1990 年代後半から実用化が進んだ。特に ChaSen 以降の機械学習に基づく形態素解析技術は、人手でのルール整備を不要にし、辞書と学習用のコーパスをもとにして高精度の解析を行うことを可能にした。しかし、古典文学作品などの歴史的な資料については，様々な先駆的試みがあったものの、本格的な古語の電子化辞書と機械学習用の古文のデータが不足していたため、実用的な精度で実現することは長らくできなかった。

こうした中、発表者らは歴史的な日本語コーパスの構築に備えるために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進めてきた（小木曽ほか 2010）。BCCWJ のタグ付けを行うために開発された形態素解析辞書「UniDic」をベースとし、さまざまな古文の解析に必要な見出し語を追加し、当該古文の機械学習用コーパスを整備することで古文のための形態素解析辞書を実現したものである。現在、近代の文語論説文を対象にした「近代文語 UniDic」と、中古の和文系資料を対象とした「中古和文 UniDic」を作成、公開している。UniDic はもともと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。

「中古和文UniDic」は、『源氏物語』をはじめとする平安時代の仮名文学作品を主たる対

† togiso@ninjal.ac.jp

象とした辞書である。図 1 は、文単位でランダムサンプリングした平安時代の仮名文学作品のテキストを、現代語用のUniDicと「近代文語UniDic」、「中古和文UniDic」のそれぞれで解析した精度を比較したものである[1]。グラフから分かるように、平安時代の仮名文学作品の形態素解析は、現代語用の辞書による解析では実用にならないものだったが、「中古和文UniDic」を用いることで高い精度で行うことが可能となった。



	中古和文 UniDic0.5	近代文語 UniDic1.1	(現代語) UniDic1.3.12
■単位境界	99.31%	91.09%	81.86%
■品詞認定	97.77%	83.78%	59.25%
■語彙素認定	97.10%	78.73%	55.77%
■発音形認定	96.64%	77.89%	55.05%

図 1　中古和文の解析精度比較

仮名文学作品の文体は、古典文学としてその後の規範となり長い間使われ続けていくため、「中古和文 UniDic」は中世・近世の擬古文をはじめとするさまざまなテキストの解析にも利用できる。明治期の雅文、和歌の解析にも利用可能である。

図 1 に示したとおり、形態素解析の精度はおよそ 95～97%程度である。先述したとおり、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、この解析精度は目的によっては必ずしも十分でなく、解析結果を修正してさらに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問題があった。

３．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，

[1] テストデータは『伊勢物語』『源氏物語』『大和物語』『土佐日記』『紫式部日記』『更級日記』から抽出した約 2.5 万語。ただし未知語なし。

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、茶器は 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、容易に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 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コーパスも格納することが可能である。

「茶器」は、近年「ChaKi.NET」としてシステムが一新され、簡易なデータベースである SQLite 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．「茶器」による中古和文コーパスの利用

４．１．中古和文コーパスのインポート

「中古和文 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。図 2 は「茶器」に形態素解析済みの『更級日記』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。

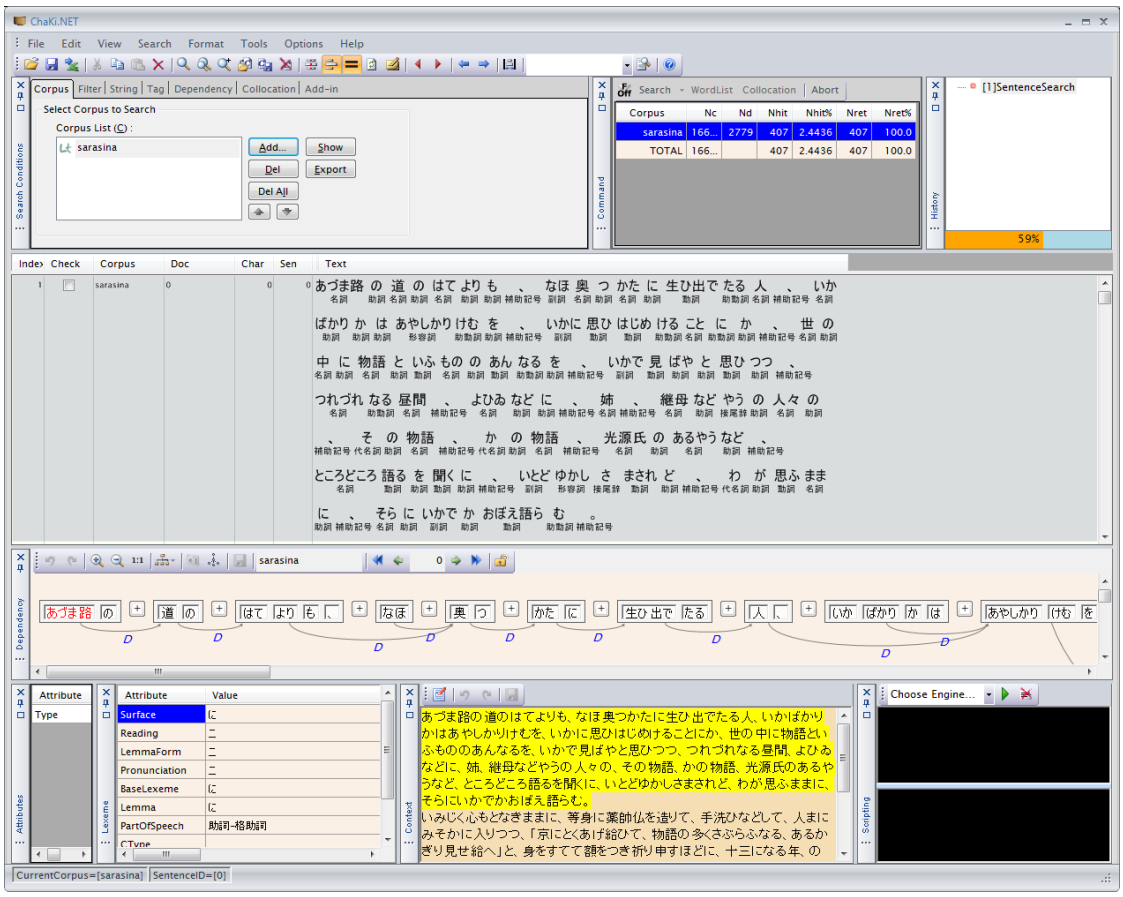


図 2　ChaKi.NET の実行画面（『更級日記』）

今回、表 1 にあるデータを「茶器」にインポートして利用した（以下、このデータを中古和文コーパスと呼ぶ[2]）。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　中古和文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・助動詞と結合する

2）接尾辞・句読点は直前の短単位と結合する

3）数詞の連続と数詞に続く助数詞は結合する

4）接頭辞は直後の短単位と結合する

このルールにより、簡易検査で 9 割以上の文節が正しく分割されることが確認された。このルールでまとめ上げた*.mecab 形式のファイルを、係り受け情報付きのデータ（*.cabocha 形式）に変換した後、TextFormatter を用いてインポートした。

4．2．形態論情報の格納

「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

[2] 「中古和文 UniDic」の作成と、形態素解析済み古典語データ利用の検証のために作成したもので公開予定はない。

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。そのため、基本 9 属性に対応しない情報については、カスタムフィールド（custom）に格納して対処した。中古和文コーパスでは 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た。

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

中古和文コーパスのカスタムフィールド内のデータは次の通りである。

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

５．１．形態素解析結果の修正

先述したとおり、研究に利用する古典語コーパスには、現代語以上に高いタグ付けの精度が求められる。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（図 3）。

図 3　解析結果修正のための辞書見出し語選択画面

５．２．文節係り受けのタグ付け

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、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 4 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。

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、極めて難しい作業となる。特に散文では一文が長く、係り先が曖昧な場合が多いため、完全なアノテーションを行うことは困難である。しかし、古典語のデータは量が限られているため、タグ付けする必要がある係り受けの種類を厳選し、十分な時間をかけることで、作品全体に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を試みたに過ぎないが、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、連体修飾や動詞の項構造などに限定して、一部の作品へのタグ付けを進めていきたいと考えている。

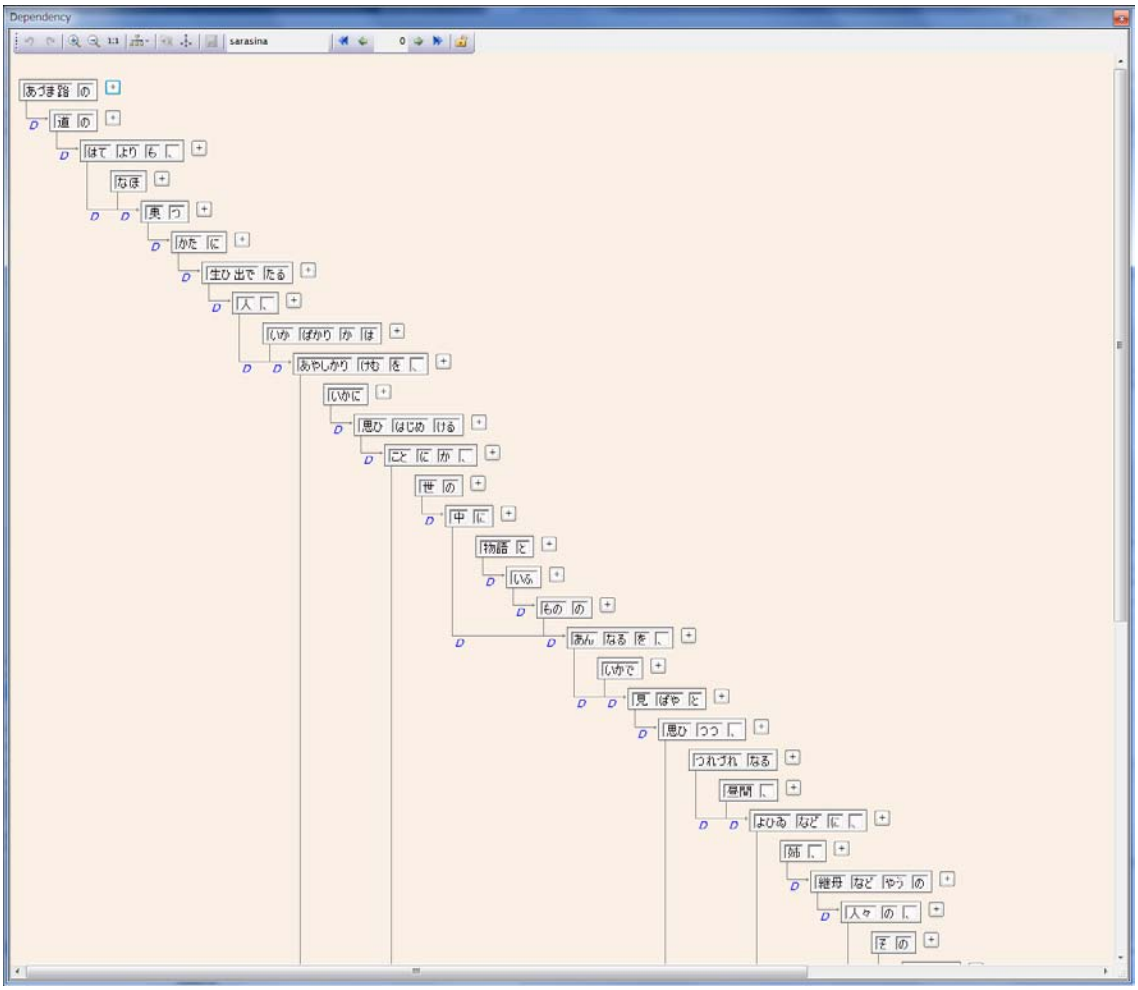


図 4　長い文の文節係り受けアノテーション（『更級日記』冒頭）

６．検索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、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、検索が可能な古典語データ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 10 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。

タグ検索（TagSearch）

タグ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、共起条件を複数設定した検索も可能である。図 5 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。

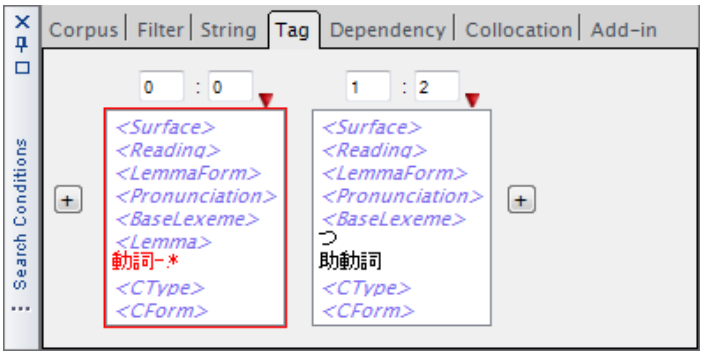


図 5　タグ検索（共起条件の設定）

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。

図 7 は図 5 の検索結果を元に、助動詞「つ」の前 2 語以内に来る動詞のワードリストを作り、頻度順並び替えたものである。

	Su	Re	Le	Pr	Ba	Lemma_0	Pa	C	Cf	Su	Re	Le	Pr	Ba	Le	Pa	C	Cf	genji	ise	murasa	sarasin	tosa	All	Ratio(%)
▸TOTAL																			1843	37	36	31	18	1965	100
1	*	*	*	*	*	諦める	*	*	*	*	*	*	*	*	*	*	*	*	259	5	2	0	0	266	13.536...
2	*	*	*	*	*	会う	*	*	*	*	*	*	*	*	*	*	*	*	105	0	0	0	0	105	5.3435...
3	*	*	*	*	*	有る	*	*	*	*	*	*	*	*	*	*	*	*	82	0	2	1	1	86	4.3765...
4	*	*	*	*	*	表わす	*	*	*	*	*	*	*	*	*	*	*	*	78	2	0	0	2	82	4.1730...
5	*	*	*	*	*	急ぐ	*	*	*	*	*	*	*	*	*	*	*	*	80	0	0	1	0	81	4.1221...
6	*	*	*	*	*	坐す	*	*	*	*	*	*	*	*	*	*	*	*	73	0	0	1	0	74	3.7659...
7	*	*	*	*	*	入り揉む	*	*	*	*	*	*	*	*	*	*	*	*	63	1	0	3	0	67	3.4096...
8	*	*	*	*	*	承る	*	*	*	*	*	*	*	*	*	*	*	*	62	0	1	0	0	63	3.2061...
9	*	*	*	*	*	置く	*	*	*	*	*	*	*	*	*	*	*	*	39	0	1	0	0	40	2.0356...
10	*	*	*	*	*	遅らす	*	*	*	*	*	*	*	*	*	*	*	*	25	2	0	1	0	28	1.4249...
11	*	*	*	*	*	怠る	*	*	*	*	*	*	*	*	*	*	*	*	24	0	0	0	0	24	1.2213...
12	*	*	*	*	*	押し立てる	*	*	*	*	*	*	*	*	*	*	*	*	20	0	0	1	1	22	1.1195...
13	*	*	*	*	*	大人びる	*	*	*	*	*	*	*	*	*	*	*	*	19	1	0	0	0	20	1.0178...
14	*	*	*	*	*	驚く	*	*	*	*	*	*	*	*	*	*	*	*	20	0	0	0	0	20	1.0178...
15	*	*	*	*	*	驚かす	*	*	*	*	*	*	*	*	*	*	*	*	19	0	0	1	0	20	1.0178...
16	*	*	*	*	*	おわす	*	*	*	*	*	*	*	*	*	*	*	*	13	1	0	5	0	19	0.9669...
17	*	*	*	*	*	おわします	*	*	*	*	*	*	*	*	*	*	*	*	17	0	0	0	0	17	0.8651...
18	*	*	*	*	*	覚える	*	*	*	*	*	*	*	*	*	*	*	*	14	1	0	1	0	16	0.8142...

図 6　ワードリスト検索結果

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。

ただし、ここで計算に使われる頻度は、コーパス全体を対象としたものではなく、検索結果の KWIC に表示されているものが利用されるので注意が必要である。

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。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。

7．おわりに

「茶器」を用いることで、懸案だった形態素解析済みの古典語データの利用環境を整備することができた。今後、単語情報付きの古典語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。

将来的には、「茶器」で整備した係り受けデータを元に古典語の係り受け解析を実現していきたい。係り受けまで整備されたコーパスが用意できれば、検索結果のいわゆる「ゴミ取り」作業が軽減できるだけでなく、動詞の項構造や連体修飾関係などを使う高度なコーパス利用が可能になる。これによりコーパスを利用した古典語研究がさらに大きく前進するはずである。

文　献

松本 裕治（2009）「コーパスへの自動アノテーションツールとアノテーション支援環境の構築」人工知能学会誌 24(5), pp.632-639

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

小椋秀樹・須永哲矢・小木曽智信・近藤明日子・田中牧郎（2011）「「中古和文 UniDic」における言語単位の設計」『言語処理学会第 17 回年次大会発表論文集』pp.312-315

小木曽智信・岡照晃・小町守・松本裕治（2011）「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」『人文科学とコンピュータシンポジウム論文集「デジタル・アーカイブ」再考』情報処理学会 pp. 255-260

関連 URL

茶器　http://sourceforge.jp/projects/chaki/

中古和文 UniDic　http://www2.ninjal.ac.jp/lrc/index.php?UniDic

MeCab　http://mecab.sourceforge.net/feature.html

CaboCha　http://code.google.com/p/cabocha/

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/

国立国語研究所萌芽・発掘型共同研究プロジェクト「統計と機械学習による日本語史研究」
http://www.ninjal.ac.jp/histlingstat/