# 歴史的資料を対象とした複数の UniDic 辞書による形態素解析支援ツール『Web 茶まめ』

堤 智昭（東京農工大学 工学府）
小木曽 智信（国立国語研究所）

近代文語 UniDic や中古和文 UniDic の登場により，近代以前の歴史的な日本語資料に対しても形態素解析が可能となった．しかし，近代以前の現存する日本語資料は時代幅があり，ジャンルも多岐にわたる．そのため，資料ごとに文法・単語が適した辞書を用いて形態素解析を行う必要がある．日本語研究者が形態素解析技術を用いた研究に取り掛かるには，煩雑な形態素解析実行環境の用意と辞書を切り替えた解析作業が必要となり，その難易度が研究推進の妨げとなっている．そこで本研究では，形態素解析を用いた言語研究の支援を目的とし，煩雑な計算機における形態素解析実行環境の用意を必要とせず，容易に複数の辞書を切り替えて形態素解析が可能な形態素解析サポートソフトウェア，Web 茶まめの開発を行った．

# Web ChaMame: A Support Tool for Morphological Analysis of Historical Japanese Texts Using UniDic Dictionaries

Tomoaki TSUTSUMI (Tokyo University of Agriculture and Technology)
Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)

UniDic for modern literary language and classical Japanese could make a morphological analysis of classical Japanese text. However, preparation of the cumbersome morphological analysis execution environment and the analytic work to which a UniDic dictionary was changed are needed for a Japanese researcher to begin to study using morphological analysis technology, and the degree of difficulty is obstruction of study promotion. Therefore we have developed morphological analysis support software "Web Chamame". The feature of “Web Chamame” is that execution is possible by web browser, and the morphological analysis execution environment isn't needed.

## 1.はじめに

近年，国語学・日本語学の分野において，自然言語の文章を電子化・構造化し大規模なデータベースとした，いわゆるコーパスを用いた研究が行われている．コーパス言語学的な研究では，形態素解析を行い語のレベルで調査を行う．そのため，これまでは一般的な形態素解析辞書が存在する現代語を中心に研究が行われてきた．しかし近代文語 UniDic[1] [2]や中古和文 UniDic[1][3]の登場により，近代以前の歴史的な資料についても形態素解析が可能となり，近代語コーパスや日本語歴史コーパスのように歴史的資料に対してもコーパス言語学的な研究が行われている．

しかし，一般的な日本語研究者にとって，様々な形態素解析辞書を用いて形態素解析を行うことは，形態素解析実行環境を用意する難易度と，実際の解析作業の煩雑さから容易ではない．そのためにローカルで動作する形態素解析補助ツール「茶まめ」[4]等が公開され，利用されてきた．しかし，計算機を用いた形態素解析を行うには，MeCab[5]に代表される形態素解析エンジンと UniDic のような形態素解析辞書をそれぞれ計算機にセットアップする必要があるうえ，近代以前の資料を対象とした形態素解析辞書は，種類が多くそれら全てをセットアップすることは煩雑である．また，形態素解析辞書は，解析対象の特徴に応じて適切なものを使用する必要がある．MeCab のようにコマンドラインで動作するソフトウェアを用いて，形態素解析を実行するたびに，適切な辞書を切り替えて使用することは煩雑な作業となる．そこで本稿では，形態素解析を用いた言語研究の支援を目的とし，煩雑な環境構築をすることなく利用でき，辞書選択を容易に行えるようサポートするソフトウェア，Web 茶まめの開発を行った．

## 2.UniDic 辞書の特徴

形態素解析では，対象言語の文法や単語リストから，対象の文を形態素に分割する．そのため，文法，単語が異なる言語間では同一の辞書を用い

ることができない．このことは，同じ日本語である現代語と古文との間でも問題となる．また古文の中でも，現存する資料は 8 世紀ごろから 19 世紀まで時代幅があり時代によって文法・単語が移り変わっていく．ジャンルとしても，和歌・軍記物・狂言・洒落本等，様々なものが存在し，多彩な文法・単語が存在するため，それぞれに適した辞書を用意する必要がある．

UniDic ではこうした問題を解決し，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い，古代から現代に至る通時的な観察を可能とする通時コーパスを実現するために，現代語用 UniDic，中古和文 UniDic，近代文語 UniDic が開発された．中古和文 UniDic，及び近代文語 UniDic は，現代語用 UniDic をベースとして実装されており，UniDic の一貫した原理にもとづいて実装されている．

UniDic は現代日本語書き言葉均衡コーパス（BCCWJ）でも利用されている，代表的な電子化辞書である．MeCab の辞書として利用でき，日本語テキストを単語に分割し形態論情報を付与することができる．近代文語 UniDic と中古和文 UniDic には，現在合わせて 9 個の辞書が存在する．

## 3.Web 茶まめの概要

本システムは，サーバにインストールされた複数の辞書を用いて，Web ブラウザ上で MeCab エンジンによる形態素解析を行う Web アプリケーションである．図 1 に示すように，ユーザは Web ページを介して解析対象文字列を入力する．Web 茶まめサーバは，解析対象文字列に形態素解析を行う前処理を施し Web 茶まめサーバにインストールされた MeCab に入力し，形態素解析を行う．形態素解析に用いる辞書は，サーバにインストールされたものの中から，ユーザが指定した辞書を使用する．その後 MeCab の出力結果をユーザへ返す．このように形態素解析処理は，全て茶まめサーバで行われるため，利用者の計算機に特別なソフトウェアをインストールすることなく，ウェブブラウザがインストールされている計算機なら OS を問わず利用可能である．

## 4. 設計と実装

### 4.1 システム設計

Web 茶まめの目的は，煩雑な環境構築をすることなく利用でき，辞書選択を容易に行えるようサポートすることである．そのため Web ページの作成には，Flash プレイヤー等に代表されるユーザ側においてインストールが必要なソフトウェアを用いず，HTML と JavaScript のみを用いた．Web ページと形態素解析エンジンの間のやり取りは，茶まめサーバにインストールされた PHP を用いて行う．

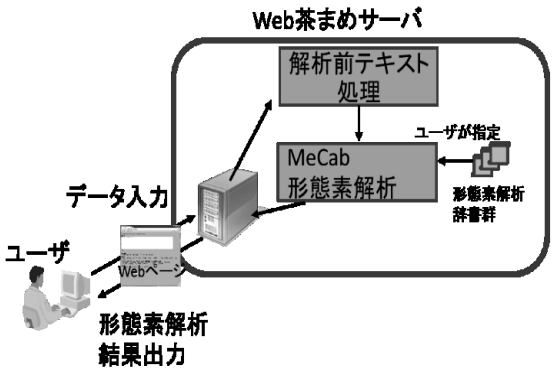


図 1　Web 茶まめの概要図

Web 茶まめサーバの OS には，Linux ディストリビューションの一つである CentOS7 を用いた．形態素解析エンジンには，Linux 版 MeCab を用いた．利用できる形態素解析辞書は，近代文語 UniDic・中古和文 UniDic を中心とした 10 辞書(現代語／旧仮名口語／近代文語／洒落本／近世口語／中世口語／和漢混淆文／中古和文／万葉集／話し言葉)とした．HTTP サーバには Apache2 を用いた．

### 4.2GUI 設計

Web 茶まめの GUI は，ユーザからの入力を受け付けるメイン画面（検索画面）と，形態素解析結果を出力する結果出力画面の 2 つから構成される．

#### 4.2.1 メイン画面

図 2 にシステムのメイン画面を示す．メイン画面の主な機能は以下の 4 つである．

（1） 解析対象文字列の入力

Web 茶まめでは，テキストフィールドに解析対象文字列を直接入力する方式と，ユーザのローカルディスクにあるテキストファイルをアップロードする方式の 2 つを実装した．テキストフィールドは，数行分の短い文字列に対して解析を行う場合を想定している．テキストファイルのアップロードは，何らかの方法で別途作成済みの大量のテキストデータに対して解析を行う場合を想定している．そのため，図 3 示すように，複数のテキストファイルを選択して同時にアップロード可能とした．

入力文字列のサイズが大きすぎると，サーバの計算リソースを大量に消費してしまう．そのため，テキストフィールドへの入力は文字列のサイズを 100KB まで，ファイルのアップロードではファイルサイズを 10MB までに制限した．

（2） 実行する解析前処理の選択

Web 茶まめが行う，形態素解析前処理を項目ごとに，実行するべきか否かをユーザが選択可能


図 2　Web 茶まめメイン画面

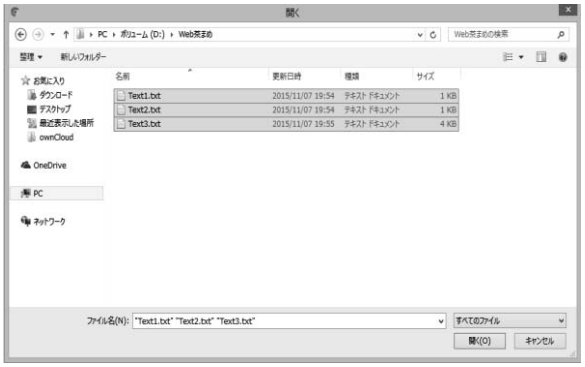


図 3　アップロード画面

とする. 項目の表示にはチェックボックスを用いた. 並び順は, システムが実行する順番に従って左から順に並べた.

（3） 解析に用いる辞書の選択

形態素解析に用いる辞書をユーザが選べるよう, ラジオボックスを用いて一つの辞書を選択する方式とした.

（4） 解析結果の出力方式の調整

解析結果の出力は, 結果のみをすぐに見たい場合と, 結果を用いてユーザが何らかの処理を別途行いたい場合を想定して,「HTML 形式で表示」「csv 形式でダウンロード」「Excel 形式でダウンロード」の 3 つの中から選択可能とした. 選択項目はラジオボックスを用いて表示し, デフォルトでは「HTML 形式で表示」にチェックをつけることとした.

表 1　出力項目一覧

| 項目名 | デフォルトでの出力 |
|---|---|
| 語彙素 | する |
| 語彙素読み | する |
| 品詞 | する |
| 活用型 | する |
| 活用形 | する |
| 発音形出現形 | する |
| 仮名形出現形 | する |
| 語種 | する |
| 書字形(基本形) | する |
| 発音形(基本形) | する |
| 仮名形(基本形) | する |
| 語形(基本形) | する |
| 語頭変化型 | しない |
| 語頭変化形 | しない |
| 語頭変化結合型 | しない |
| 語末変化型 | しない |
| 語末変化形 | しない |
| 語末変化結合型 | しない |
| アクセント型 | しない |
| アクセント接続型 | しない |
| アクセント修飾型 | しない |

また, MeCab と UniDic 辞書を用いて形態素解析を行った場合, 21 項目が出力可能である. ユーザがこれらの項目を常に全て必要としているとは限らないため, チェックボックスを用いて

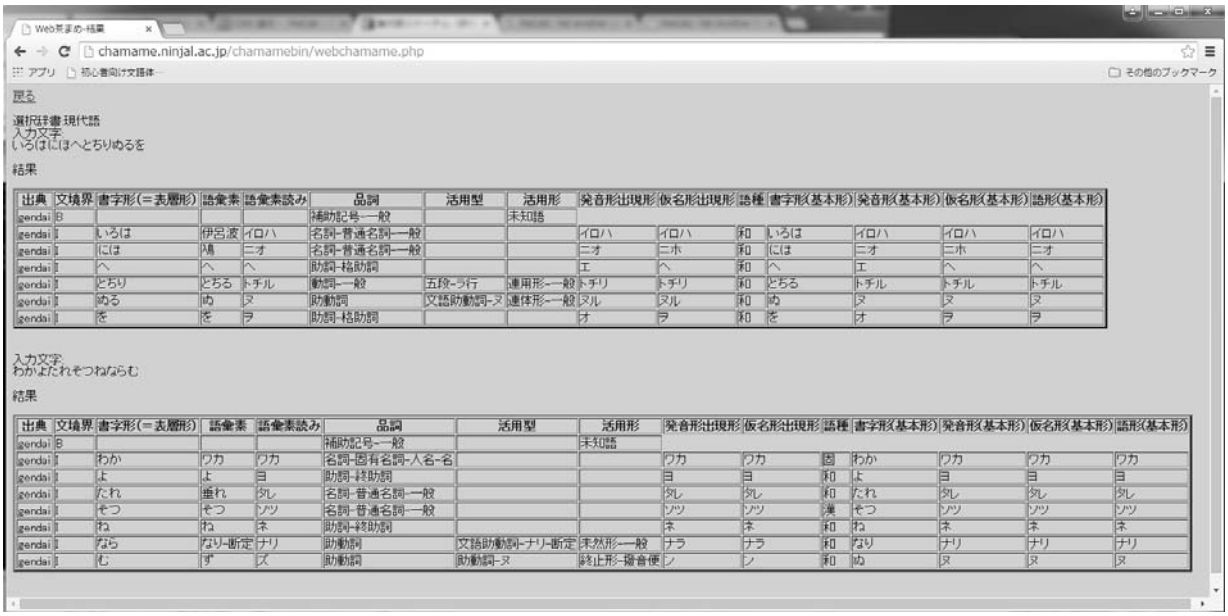


図 4　結果出力画面

出力項目を選択可能とした．また，一般的に使用率が高いと考えられる項目はデフォルトでチェックが付くようにした．表 1 に項目名の一覧と，その項目をデフォルトで出力するかどうかを示す．

#### 4.2.2 出力画面

図 3 に結果出力画面の例を示す．結果出力画面は，出力方式に HTML を指定した場合に表示される．表示項目は解析に使用した「辞書名」と「入力文字列」，「解析結果」の 3 つである．

「入力文字列」には，解析前処理を施した後の文字列が出力する．これは，入力した文字列に解析前処理を施すとどのようになるかを確認できるようにするためである．

「解析結果」では，メイン画面において指定した出力項目を列の項目とする表を出力する．複数のテキストファイルを同時に入力した場合，図 3 に示すように 1 ファイルにつき 1 つの表にまとめ，1 ページ中に全ての結果を表示する．

#### 4.2.3 形態素解析実行手順

Web 茶まめを用いた形態素解析は，メイン画面において次の手順で実行する．

（1）テキストフィールドを用いた場合
①テキストフィールドに解析対象文字列を入力する
②解析前処理の有無，辞書，出力項目，出力形式を選択する
③「実行する」ボタンをクリックする
（2）ファイルアップロードする場合
①解析前処理の有無，辞書，出力項目，出力形式を選択する
②「ファイル選択」ボタンをクリックし，アップロードするファイルを選択する
③アップロードボタンをクリックする

### 4.3 形態素解析処理の設計と実装

Web 茶まめサーバにおける形態素解析処理は次の（1）～（4）の手順で実行する．以下にそれぞれの詳細について論ずる．

（1）解析前処理
（2）辞書の選択と出力項目の指定
（3）MeCab を用いた形態素解析
（4）形態素解析結果の出力

#### 4.3.1 解析前処理

本システムは入力された解析対象文字列に対して簡単な 6 つの文字列整形機能を提供する．これらの機能は，MeCab を用いた形態素解析を行う前に実行される．

① 入力ファイルの確認
② 文字コードを UTF-8 へ変換
③ 半角文字を全角に変換
④ 踊り字を展開
⑤ カタカナ平仮名反転
⑥ 全角数字を漢数字に変換

処理の実行順番は，①②③④⑤⑥の順で行われる．このうち，①②は全ての解析対象文字列に対して実行される．③～⑥はユーザがメイン画面において実行を許可した場合のみ実行される．

①では，ユーザから入力されたデータが形態素解析可能なものか確認し，必要に応じて整形を施す．具体的には，アップロードされたファイルがテキストファイルかどうかを確認し，テキストファイルでなかった場合は図 5 に示すようなエラ

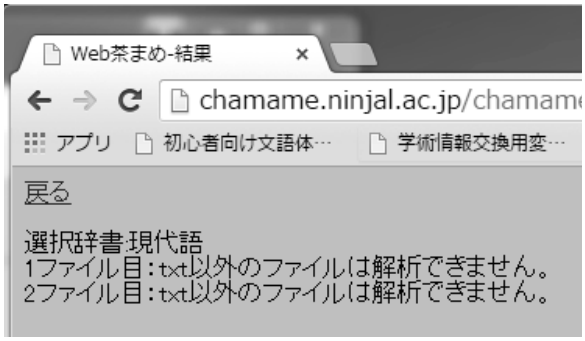

図 5　入力ファイル確認時のエラー画面

一文を表示し，形態素解析を行わない．また，入力ファイルにHTMLやJavaScript等のソースコードが含まれていた場合，該当部分を削除する処理も行っている．

②では，ユーザから入力された文字コードを認識して，UTF-8 へ自動的に変換する．これは，ユーザが用いる文字環境と，Web 茶まめサーバが用いる文字環境との差異を吸収するためである．入力テキストはメモ帳等で開ける平文テキストのみとした．変換には PHP の mb_convert_encoding 関数を用い，WindowsOS や Mac の OSX で使用できる一般的なエンコード（Shift-JIS, Windows-31J, EUC-JP, ASCII）に対応した．

③では，英数字を含む半角文字を全角に変換する．変換には PHP の mb_convert_kana 関数を用いる．変換の一例として「芸者 2 人ヤソトメ」は「芸者２人ヤソトメ」のように変換される．

④では，日本語における繰り返し記号である踊り字を，繰り返し対象の文字に変換する．仮名一文字の繰り返しを表す「ゝ」「ヽ」は，一文字前の仮名に置き換える．「ゞ」「ヾ」は，一文字前の仮名の濁点付きに置き換える．古典資料では多く見られる，「〜〜」「〜〵」「／＼」「／〵＼」のような繰り返し表記は，Unicode の「／＼」「／〵＼」に変換する．変換の一例として，「アハヽヽヽ」は「アハハハハ」に，「いすゞ」は「いすず」に，「やい／＼」は「やい／＼」に，「さま／〵＼」は「さま／〵＼」のように変換される．

⑤では，入力テキスト中の全てのカタカナをひらがなに，ひらがなをカタカナへ変換する．変換には PHP の mb_convert_kana 関数を用いる．本処理では，初めに全てのカタカナを半角カタカナに変換する．その後，平仮名を全て全角カタカナに変換し，最後に半角カタカナを全角平仮名に変換する．変換の一例として，「左様サ，犬の世界とハありがてへ」は「左様さ，犬ノ世界トはアリガテへ」のように変換される．

⑥では，入力テキスト中の全角文字列を漢数字へ変換する．変換には Numtrans [6]を用いた．Numtrans では，半角数字は変換されないため，テキスト内に半角数字が存在する場合は，②の処理を先に実行し全角数字に変換してから，漢数字へ変換することで対応可能とした．また，Numtrans は XML 形式のデータに対して処理を行うため，入力テキストを Numtrans で解析可能な形式に変換する．Numtrans から出力された結果は，オリジナルテキストが保存された XML 形式で出力される．MeCab を用いて形態素解析を行う場合には，プレーンテキストを入力する必要があるため，出力結果には XML タグとオリジナルテキストを取り除く処理を施す．変換の一例として，「３３間堂」は「三十三間堂」に変換される．また，「1512」等半角の数字は，⑥の処理を有効にしただけでは「1512」のまま出力されるが，③の処理と同時に有効にした場合，「１５１２」に変換された後「千五百十二」に変換される．

#### 4.3.2 辞書の選択と出力項目の指定

形態素解析に用いる辞書と，出力フォーマットは，それぞれ MeCab 実行時に，引数として入力する．ここでは，ユーザの入力に応じた引数を生成する．

#### 4.3.3MeCab を用いた形態素解析

解析前処理の終わった文字列に対して MeCab を用いて形態素解析を行う．形態素解析結果は一時ファイルに出力する．このファイルは，ユーザに出力した後削除する．

#### 4.3.4 形態素解析結果の出力

ユーザの選択した出力形式に従って MeCab から受け取った形態素解析結果を出力する．結果は表形式とし，1 行目には各列の項目を出力する．解析データを用いた調査の利便性を向上するため，ユーザが指定した出力項目に加えて，文頭の語であるか文中の語であるかを示す「文境界」項目を付与する．境界面であった場合には「B」を，そうでない場合には「I」を出力する．また一つの入力文字列に対して，複数の辞書で形態素解析した結果を比較・利用することを想定し，解析に使用した「辞書名」項目も付与する．

HTML 形式で出力する場合は，形態素解析結果にHTMLタグを付与してWebページとして出力する．csv 形式の場合は，形態素解析結果をカンマ（，）区切りのテキストデータに整形し，ファイルとして出力する．Excel 出力の場合は，PHPExcel ライブラリを用いて Excel2007 形式のファイルに変換後，出力する．

ユーザから複数のファイルがアップロードされた場合，出力ファイルは 1 入力ファイルにつき 1 つとする．しかし，ブラウザを介して同時に複数のファイルをダウンロードするのは煩雑な作

業となるため，全ての出力ファイルを 1 つの zip ファイルに圧縮し，それをユーザに出力する方式とした．ファイルの圧縮には，Linux の zip コマンドを用いる．

## 5.まとめ

今回，形態素解析を用いた言語研究の支援を目的とし，煩雑な環境構築をすることなく，複数の辞書を切り替えて容易に形態素解析が行えるサポートソフトウェア Web 茶まめの開発を行った．開発したツールは HP 上で公開しており（http://chamame.ninjal.ac.jp/），誰でも自由に利用することができる．

今後の課題には，一つのテキスト内でも複数の辞書切り替えが必要な場合へ対応することが挙げられる．例えば，基本的には中古和文辞書を用いて形態素解析を行いたいが，文中に和歌が入っていた場合のみ辞書を切り替えたい，といった使い方が考えられる．また，利用可能な辞書の追加や入力した文字列に対して適切な辞書をシステムが推定する機能の実装を考えている．

### 謝辞

本研究は国立国語研究所共同研究プロジェクト「通時コーパスによる日本語史研究の新展開」による研究成果の一部である．

## 参考文献

[1] 小木曽 智信，小町 守，松本 裕治 (2013)「歴史的日本語資料を対象とした形態素解析」『自然言語処理』20(5)，pp. 727-748.

[2]国立国語研究所：近代文 UniDic,(オンライン)，〈http://www2.ninjal.ac.jp/lrc/index.php?UniDic%2F%B6%E1%C2%E5%CA%B8%B8%ECUniDic>(参照 2015-09-15).

[3] 国立国語研究所：中古和文 UniDic,(オンライン)，〈http://www2.ninjal.ac.jp/lrc/index.php?UniDic%2F%C3%E6%B8%C5%CF%C2%CA%B8UniDic〉.

[4] 小木曽智信 (2014)「形態素解析ツール」『講座 日本語コーパス 書き言葉コーパス 設計と構築』朝倉書店.

[5] 工藤拓 :MeCab : Yet Another Part-of-Speech and Morphological Analyzer, (online), 入手先〈http:// mecab.sourceforge.net/〉(参照 2015-09-15)

[6] 山田篤,小磯花絵 (2008)『NumTrans マニュアル』, The UniDic Consortium.