堤 智昭，小木曽 智信
複数の UniDic 辞書による形態素解析支援ツール『Web 茶まめ』の実装と運用
『情報処理学会誌』64(3) pp. 749-757.
2023 年 3 月刊行

http://doi.org/10.20729/00225271

情報処理学会論文誌 Vol.64 No.3 1–9 (Mar. 2023)

複数のUniDic辞書による形態素解析支援ツール『Web茶まめ』の実装と運用

堤 智昭[1,a)] 小木曽 智信[2]

受付日 2022年5月22日, 採録日 2022年12月2日

概要：本稿では，複数のUniDic辞書による形態素解析を行えるWebアプリケーション，Web茶まめの開発実装と運用について述べる．文章の形態素解析を行う場合に必要となる辞書は，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．そのため日本語の形態素解析を行う場合でも，現代語と古文では同一の辞書を用いることは難しい．また，古文の中にも和歌・軍記物・狂言・洒落本など，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．こうした問題を解決するためにUniDicでは時代やジャンルに応じた複数の形態素解析辞書を提供している．Web茶まめはこれらの辞書を容易に切り替え，形態素解析の実行環境を提供し，コーパス言語学的な研究の推進やオンライン授業など教育活動を推進することを目的に開発を行った．

キーワード：コーパス，形態素解析，アプリケーション開発

Implementation and Operation of "WebChamame", A Morphological Analysis Support Tool Using Multiple UniDic

Tomoaki Tsutsumi[1,a)] Toshinobu Ogiso[2]

Received: May 22, 2022, Accepted: December 2, 2022

Abstract: In this paper, describes the development, implementation, and operation of WebChamame, a web application that can perform morphological analysis using multiple UniDic. When performing morphological analysis of sentences, it is not possible to use the same dictionaries for languages with different grammars and vocabulary. For this reason, it is difficult to use the same dictionary for both modern and archaic Japanese texts when performing morphological analysis of Japanese. In addition, there are a variety of ancient texts, such as *waka* poems, war chronicle, *kyogen*, and *sharebon* books (gay-quarter novelettes), with a wide range of grammars and words, so it is necessary to prepare appropriate dictionaries for each. To solve these problems, UniDic provides multiple morphological analysis dictionaries for different periods and genres, and WebChamame was developed to easily switch between these dictionaries and provide an environment for performing morphological analysis to promote corpus linguistics research and educational activities such as online classes. WebChamame was developed for the purpose of promoting corpus linguistics research and educational activities such as online classes.

Keywords: corpus, morphological analysis, application development

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コーパスを用いた研究が広く行われている．コーパス言語学的な

[1] 筑波大学
University of Tsukuba, ●●●●●, ●●●●● 000–0000, Japan
[2] 国立国語研究所
National Institute for Japanese Language and Linguistics, ●●●●●, ●●●●● 000–0000, Japan
[a)] tsutsumi.tomoaki.gn@u.tsukuba.ac.jp

研究では通常，形態素解析を行って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 や中古和文 UniDic [1], [2] 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困難さと，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」[3] などが公開され，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 [4], [5] に代表される形態素解析エンジンと UniDic [6] 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，それらすべてをセットアップすることは簡単ではない．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．さらに，大学など各種学校では授業などで利用できるコンピュータに導入可能なソフトウェアに制限があるといった理由で，これらの環境を用意することが難しい場合も少なくない．加えて，近年ではコロナ禍におけるオンライン授業の実施にともない，コンピュータの扱いに慣れていない学生の場合，各自に形態素解析器を用いた演習環境を用意することは難しいこともある．そこで，形態素解析を用いた言語研究，教育の推進を目的とし，煩雑な環境構築をすることなく利用でき，複数の UniDic 辞書による形態素解析を行えるソフトウェア，Web 茶まめの開発を行った．

2. 関連研究

2.1 UniDic 辞書

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 20 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・洒落本など，様々なものが存在し，同時代であっても伝統的な文語が用いられるものから，口語表現が用いられるものまで幅広い．また，資料ごとに各種の文法的な変異があり異なる語彙が用いられるため，それぞれに適した辞書を用意する必要がある．

日本語研究に適した形態素解析用の辞書 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 のほかに，古文用の各種の UniDic が開発された [1], [2]．古文用 UniDic は，現代語用 UniDic をベースとして実装されており，UniDic の設計原理のもと均質な見出し語の単位に基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」をはじめとする多数のコーパスでも利用されている代表的な電子化辞書である．形態素解析器 MeCab の辞書として利用でき，多様な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．UniDic には現在合わせて 12 種類の辞書が存在する．そのうち現代語用 UniDic には 2 種類，古文用 UniDic には 10 種類が存在している．

2.2 関連ツール

ここでは，Web 上でテキスト解析が可能なツールと Web 茶まめとの比較を行う．

（1）形態素解析 API

Web 上で使える形態素解析ツールとしては，Goo ラボが提供する形態素解析 API [7] がある．このツールは，WebAPI として提供されており，リクエスト先の URL に解析を行いたい解析対象テキストなどの情報を HTTP の POST メソッドを用いて送信することで，形態素解析結果を JSON フォーマットで受け取ることができる．Web ページ上から API を試すことも可能だが，主にはこの API を使ったシステムやアプリの開発を利用目的としている．また，形態素解析に用いる辞書や形態素解析器は明記されていない．

（2）Voyant Tools

Web ページ上から操作でき，テキストを入力することで，簡単なテキスト解析と視覚化を行うツールとして Voyant Tools [8] がある．このツールはカナダ・マギル大学の Sinclair と，カナダ・アルバータ大学の Rockwell によって作成された，テキスト解析ツールである．文章の特徴を可視化し大まかな特徴を掴むことや，複数文章を読み込んで文章間の関係を図示できることが特徴である．このツールは，入力したテキストから，単語の個数をカウントしたり，単語がどのようなコンテクストで使われているかを前後の文章とともに表示したりするといった機能を備えている．一方で，検出された単語の品詞の情報や活用などに関する情報は表示されない．そのため，複数辞書を切り替え，品詞の種類などの形態素解析結果を容易に取得可能とすることを目的とした Web 茶まめとは利用方法が異なるテキスト解析ツールであるといえる．

Voyant Tools は Web 上で実行する環境を提供されている．また，Java で作成されたプログラムをダウンロードして，ローカル環境で利用することも可能である．

3. システム設計

本システムは，図 1 に示すように Web ページから利用するクラウド型のサービスとして開発した．ユーザからの入力は日本語で書かれたテキストデータ（以下解析対象テキスト）とし，ユーザへの出力は解析結果のデータとした．ユーザとやりとりをするインタフェースは Web ブラウザとする．形態素解析はサーバのリソースを用いて，サーバにインストールされた MeCab，UniDic 辞書を利用して行う．これにより，ユーザは煩雑な環境構築をすることなく，システムを利用する環境に依存せず複数の UniDic を用いた形態素解析を実行できる．

Web 茶まめが提供するサービスは，大きく 2 つに分類できる．1 つ目は形態素解析実行前に，解析対象テキストに施す解析前処理である．2 つ目は MeCab を利用した形態素解析である．

3.1 UniDic を用いた形態素解析処理の設計

3.1.1 解析前処理の設計

解析前処理は，解析対象テキストの中にある解析にノイズとなるデータを除去する作業の中から，汎用的に可能な，不要なデータの削除，表記揺れの削除を行うこととした．具体的には，以下の 6 項目である．それらのうち，ユーザが実行の有無を選べるのは ②〜⑥ の 5 項目とした．

① 解析対象テキストの文字コードの識別，UTF-8 への統一
② HTML タグ・《》タグを削除
③ 半角 → 全角変換
④ 踊り字を展開
⑤ カタカナひらがな反転
⑥ 数字処理

上記項目のうち ① では，入力された解析対象テキストの文字コードを判別し，UTF-8 に変換を行う．ユーザから送られてくるテキストデータは，様々な文字コードが使われていることが想定される．これらを，一般的に利用されることの多い日本語漢字文字コードのうち，包括される文字の多い UTF-8 に変換し，システム内では統一の文字コードのデータとして扱うこととした．

② では＜＞で囲まれたタグ，および《》で囲まれたタグを削除する．主に Web 上のテキストや，XML 化されたテキストの一部を解析する場合を想定しており，タグの内容にかかわらず削除を行う．

③ では半角文字を全角へ変換する．これは，UniDic 辞書内に搭載されている内容はすべて全角で記載されているため，意図しない表記揺れによる未知語と判定されてしまうことを防ぐことを主な目的としている．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，日本語における繰返し記号である踊り字を，繰返し対象の文字に変換する．仮名 1 文字の繰返しを表す「ゝ」「ヽ」は，1 文字前の仮名に置き換える．「ゞ」「ヾ」は，1 文字前の仮名の濁点付きに置き換える．古典資料で多く見られる「くの字点」と呼ばれる繰返し記号は，しばしばプレーンテキストで「〜〜」「〜〵」「／＼」「／〃＼」のように表記されてきたが，これらは Unicode の「／＼」「𠮟＼」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／〃＼」は「さま𠮟＼」のように変換される．

⑤ では，解析対象テキストの中にあるひらがなを全角カタカナに，半角および全角カタカナをひらがなに変換する．これは古文の一部に見られる漢字カタカナ混じり文に対応するためのオプションである．変換の一例として，「天ハ人ノ上ニ人ヲ造ラズ」は「天は人の上に人を造らず」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．変換の一例として，「２０２２年」は「二千二十二年」に変換される．また，「1512」など半角の数字は，「1512」のまま形態素解析にかけられる．

3.1.2 形態素解析処理の設計

（1）利用可能な辞書とその選択

本システムでは，MeCab と UniDic 辞書を使った形態素解析サービスを提供する．形態素解析に利用できる UniDic 辞書は，国立国語研究所コーパス開発センターが提供する最新のものを用いる．そのため，時期によって用いる辞書が変化することがあるが，2022 年 9 月現在では，**表 1** に示す 12 種類の辞書を使用できる．辞書の変更があった場合は，更新履歴に記載をすることとした．Web 茶まめではこれら 12 種類の辞書の中から，1 つ，または 2 つの辞書を選択し，形態素解析を行う．

2 つの辞書を選択した場合，2 つの辞書による解析の違いを確認し，時代が近い辞書どうしで求める結果に近いのはどちらの辞書かを比較するような利用方法を想定している．また，デフォルトで出力設定になっている項目が出力

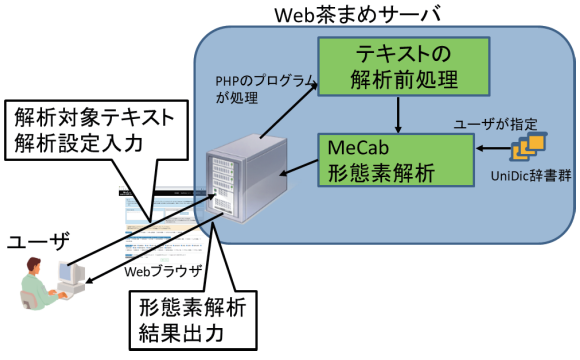

図 1 Web 茶まめ概要図
Fig. 1 Overview of WebChamame.

されることを想定しているため，3 辞書，4 辞書と多くの辞書を同時に指定すると出力結果がいたずらに複雑化することになる．そのため，2 つの辞書まで選択可能とした．2 つの解析結果は比較しやすいように横並びで表示する設計とした．また，2 つの解析結果で行数がずれる場合，行数の少ない解析結果に次の一致する行まで空行を挟み，解析結果の表示を整えることとした．たとえば「三月ばかりになる程に」という一文を解析した結果，冒頭の「三月」が辞書 ① では「三」「月」辞書 ② では「三月（ヤヨイ）」となるような解析結果が出力され，出力結果の行数が 1 行と 2 行と別れる場合，**図 2** に示すように行数の少ない辞書 ② の解析結果に空行を 1 行追加し「三月」以降の解析結果の行を揃える．

表 1　利用可能な UniDic 辞書一覧
Table 1　List of UniDic dictionaries.

辞書名	概要
現代語	現代の一般的な書き言葉テキスト．UniDic unidic-cwj-3.1.1
現代語話し言葉	現代の話し言葉の転記テキスト．UniDic unidic-cwj-3.1.1
近現代口語小説	主として明治から現代までの小説．UniDic ver. 2022.03
旧仮名口語	主として旧仮名遣いの口語で書かれた論説文．UniDic ver. 2022.03
近代文語	近代の文語論説文．UniDic ver. 2022.03
近世江戸口語	近世後期の江戸の口語資料（人情本など）．UniDic ver. 2022.03
近世上方口語	近世の大坂・京都の口語資料（世話物浄瑠璃・洒落本など）．UniDic ver. 2022.03
近世文語	近世の文語資料．UniDic ver. 2022.03
中世口語	室町時代の口語資料（狂言など）．UniDic ver. 2022.03
中世文語	鎌倉時代の文語文（説話・軍記物など）．UniDic ver. 2022.03
中古和文	平安時代の仮名文学作品・和歌集など．UniDic ver. 2022.03
上代語	万葉集・宣命などの上代語のテキスト．UniDic ver. 2022.03

（2）出力項目

形態素解析結果の出力項目は UniDic 辞書に登録されている項目の中から任意のものを選択して出力できるようにした．出力項目の一覧を**表 2** に示す．よく利用される項目はデフォルトで出力し，あまり使われない項目や他の項目で代用可能なものはデフォルトでは出力しないように設計した．このほかに本文に現れている表層形（書字形（出現形））は必ず出力される．なお，語頭変化型以下は，語形変化やアクセント変化を記述する UniDic 特有の項目である（たとえば表中の「カ濁」は語頭のカが濁音化しうることを示す）．各項目の詳細については伝ほか（2007）[6] を参照されたい．

（3）出力形式

出力形式は，以下の 4 方式とした．

1）　HTML 形式
2）　CSV 形式
3）　Excel 形式
4）　Chaki インポート形式

1）はブラウザ上で表示するための HTML テキストを出力する．2 種類の辞書で解析を行った場合，解析結果に違いがある行は図 2 に示すように赤字で強調表示する．授業での利用や，大きなファイルを解析する前の確認を主な利用用途として想定している．

2）はカンマ（,）区切りの CSV ファイルとし，文字コード

辞書	文境界	書字形（＝表層形）	語彙	語彙素読み	品詞	←→	辞書	文境界	書字形（＝表層形）	語彙素	語彙素読み	品詞
現代語	B	三	三	サン	名詞-数詞	←→	中古和文	B	三月	弥生	ヤヨイ	名詞-普通名詞-一般
現代語	I	月	月	ガツ	名詞-普通名詞-助数詞可能	←→	中古和文	I				
現代語	I	ばかり	ばかり	バカリ	助詞-副助詞	←→	中古和文	I	ばかり	ばかり	バカリ	助詞-副助詞
現代語	I	に	に	ニ	助詞-格助詞	←→	中古和文	I	に	に	ニ	助詞-格助詞
現代語	I	なる	成る	ナル	動詞-非自立可能	←→	中古和文	I	なる	成る	ナル	動詞-非自立可能
現代語	I	程	ほど	ホド	助詞-副助詞	←→	中古和文	I	程	程	ホド	名詞-普通名詞-副詞可能
現代語	I	に	に	ニ	助詞-格助詞	←→	中古和文	I	に	に	ニ	助詞-格助詞

図 2　Web 茶まめを用いた形態素解析結果の表示例（HTML 形式の場合）
Fig. 2　Example of analysis results using Web Chamame (html version).

表 2　出力項目一覧
Table 2　List of output items.

項目名	デフォルトでの出力	例：表層形「からく」
語彙素	する	辛い
語彙素読み	する	カライ
品詞	する	形容詞–一般
品詞–大分類	しない	形容詞
品詞–中分類	しない	一般
品詞–小分類	しない	
品詞–細分類	しない	
活用型	する	形容詞
活用形	する	連用形–一般
発音形出現形	する	カラク
仮名形出現形	する	カラク
語種	する	和
書字形(基本形)	する	からい
発音形(基本形)	しない	カライ
仮名形(基本形)	しない	カライ
語形(基本形)	する	カライ
語頭変化型	しない	カ濁
語頭変化形	しない	基本形
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	2
アクセント接続型	しない	C1
アクセント修飾型	しない	

を UTF-8 と SJIS の 2 種類から選べる設計とした．UTF-8 は BOM 付きとしている．出力データは，利用者が用意したプログラムなどへ読み込み，テキスト処理をすることを想定している．

3）は Excel 2010 および Excel 2007 の既定の XML ベースファイル形式（XLSX）で出力することとした．Excel を用いる場合は，Excel の機能による解析などを行うことを想定している．そのため，解析結果に違いがある行を赤字で強調表示はしないこととした．

4）はコーパスの構築・検索ツールの ChaKi[*1] にインポートするための形式で出力することとした．

3.2　GUI 設計

Web 茶まめの GUI では，解析対象テキストの入力と形態素解析を実行するときの各種設定を行う．**図 3** に，システムを Web ブラウザから開いた画面を示す．

解析対象テキストの入力は，テキストを直接入力する方式と，テキストファイルをアップロードする方式の 2 種類を用意することとした．テキストファイルを用いる場合，第一話，第二話のように複数のファイルに分割されていることを想定し，同時に複数ファイルをアップロード可能とした．

ユーザが GUI を経由して設定できる項目は，解析前処理の有無，選択辞書，出力項目，出力形式とした．解析前

図 3　Web 茶まめの入力画面
Fig. 3　Input page of WebChamame.

[*1] https://ja.osdn.net/projects/chaki/

処理の有無，選択辞書，出力項目は複数選択可能とし，出力形式は 1 つのみを選択することとした．

4. 実装

4.1 実装環境

本システムは，Web 茶まめサーバの OS には Linux ディストリビューションの 1 つである CentOS7 を用い，形態素解析エンジンには，Linux 版 MeCab（version 0.996）を用いた．Web サーバには Apache2 を用い，ユーザとサーバ間のデータのやりとりには PHP で作成したプログラムを用いた．

4.2 形態素解析処理の実装

Web 茶まめは以下のフローで形態素解析を実行する．

（1） ユーザからの解析対象テキスト，解析設定の入力
（2） 解析対象テキストの解析前処理
（3） MeCab と UniDic を用いた形態素解析
（4） 出力データの整形

4.2.1 ユーザからの解析対象テキスト，解析設定の入力

解析対象テキストは，HTML の <input> タグを用いて入力する実装とした．Textbox 型と file 型の 2 種類の入力を受け付ける．file 型の受付では，txt 形式の 1 つ，ないし複数のファイルを受け付ける実装とした．

解析設定の入力は，HTML の <input> タグの checkbox 型，radio 型を用いる実装とした．解析前処理，辞書選択，出力項目は複数選択をするため checkbox 型を用い，出力項目は単一の方式を選択するため radio 型を用いた．辞書選択では，1 つまたは 2 つの辞書を選択する設計のため，3 つ以上の辞書を選択した場合は**図 4** のように辞書選択欄の背景色を黄色く変更し，赤字で警告を表示することとした．

4.2.2 解析対象テキストの解析前処理

ユーザから入力された解析対象テキストは，3.1.1 項に記した 6 項目を行う．これらの処理は項目に付した番号の若い順に行われる．

① は PHP の mb_convert_encoding 関数を用い，解析対象テキストの文字コードを識別し UTF-8 への変換を行う．② は preg_replace 関数を用い，< > および《》で囲まれたタグを正規表現を用いて抽出し削除する方法で実装した．③ では PHP の mb_convert_kana 関数を用いて，英数字を含む半角文字を全角に変換する実装とした．④ は該当する踊り字を展開し，直前の文字で置き換える処理を PHP 言語で実装した．⑤ は mb_convert_kana 関数を用いてカタカタとひらがなの変換を行う実装とした．初めにすべてのカタカナを半角カタカナに変換する．その後，平仮名をすべ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こととした．⑥ は，NumTrans [7] を用いて入力テキスト中の全角文字列を漢数字へ変換する実装とした．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，③ の処理にチェックを付け，先に実行して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なる．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こととした．

4.2.3 MeCab と UniDic を用いた形態素解析

Web 茶まめはサーバにインストールした MeCab と UniDic 辞書を用いて形態素解析を行う．解析はユーザから入力された解析対象テキスト 1 つごとに行い，その形態素解析結果は一時ファイルとして，1 解析対象テキストファイル，1 辞書ごとに出力する．この一時ファイルは，ユーザに結果を出力した後に削除する．そのため，サーバ内では利用者がどのようなテキストを解析したかは記録しない．

4.2.4 出力データの整形

出力データは，4.2.3 項で生成した一時データから生成する．HTML 出力の場合はブラウザ上で表示し，そのほかの形式の場合はファイルとしてダウンロードする．解析対象テキストが複数あった場合，HTML 出力の場合は 1 ファイル分の結果を出力したのち，続いて次のファイルの結果を出力する．そのほかの形式の場合は，解析対象テキスト 1 ファイルごとに形態素解析結果のファイルを 1 つ生成し，ZIP 形式に圧縮し 1 つのファイルとしてダウンロードする実装とした．HTML，CSV，ChaKi インポート形式はそれぞれテキストデータとして生成し，Excel 形式では PhpSpreadsheet ライブラリを用いて Excel ファイルを生成することとした．

解析を行う辞書が複数選択されていた場合，形態素解析結果は横並びで表示し，3.1.2 項の（1），および図 2 で示したとおり行数を揃える．そのために，本システムでは Linux の diff コマンドを用いることとした．diff コマンド

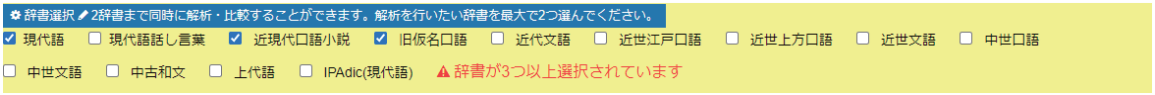

図 4 3 つ以上の辞書選択時の警告表示
Fig. 4 Warning message when more than three dictionaries are selected.

から得られた差異のある行数の情報をもとに，前述のとおり空行を挿入し出力を整形するように実装した．HTML 出力の場合はそれに加えて，横並びになる行の解析結果が異なる場合は文字色を赤色に変更し，差異のある部分を分かりやすく表示する実装とした．

5. Web 茶まめの運用

Web 茶まめは 2015 年より大学共同利用機関法人人間文化研究機構に属する国立国語研究所のドメイン[*2]で運用を行っている．大学共同利用機関法人は学術研究の発展・振興に資することを目的の 1 つとしている機関である．

今回，Web 茶まめの利用方法を理解し，主に大学における授業などで利用機会があると考えられる，国立国語研究所の「通時コーパス」・「語彙資源」プロジェクト共同研究員，および所内の教職員を対象に 2022 年度時点での教育・研究利用に関する活用事例の収集を行った．事例収集には Google Forms を用いて作成した記述式アンケートを用い，メールで対象者に回答用の URL を送信する形で実施した．その結果，14 件の活用事例を得ることができ，その内容から Web 茶まめを実装，運用することで研究・教育において以下のような成果をあげたことが分かった．

5.1 授業など教育での活用

Web 茶まめの利用について，情報提供があった中では，言語学の演習授業において活用されている事例が報告された．以下に事例収集の結果から一部を抜粋する．

「1. 授業などの内容とそこでの Web 茶まめの使い方について簡単に教えてください」という質問に対して以下のような回答が得られた．

1-①：個人またはグループで興味のあるテキストを電子化し，計量言語学的手法により語彙・文体研究を行う．

1-②：日本語の文法についてコーパスを使いながら自分で用例などを調査する．さらに，Web 茶まめに青空文庫などのデータを入れ形態素解析をしたものをエクセルに入力し，コーパスと同じように利用する．

1-③：基礎講読では，受講者全員（2 クラス，計 40 名）が短編小説，歌詞などを電子化したうえで，Web 茶まめで形態素解析を行い，語彙・文体の計量的な研究を行っている．

1-④：専門演習では，毎年，数名が Web 茶まめで形態素解析を行い，文学作品や歌詞の計量的な分析を行い，卒業論文をまとめている．

また，「2. 授業などが Web 茶まめがあるとないとでどのように変わるか教えてください」という質問に対して，以下のような回答が得られた．

2-①：計量語彙論のためには形態素解析が欠かせないが，Web 茶まめはきわめて簡便にこれを実施できる．Web 茶まめが使えない場合，文系のコンピュータ操作に慣れない学修者は手作業で単語分析をすることになり，計量研究に必要となる大規模な調査は時間がかかりすぎて不可能である．

2-②：「形態素解析」という操作を自分で実感できるため，なじみのあるテキストを使って研究ができて，研究への動機付けができる，また「形態素」についてより自覚的に考えることができる．また，（コーパス検索アプリケーションの）中納言からのダウンロードデータと構造的に共通性があるため，それらをプログラミングの対象とするときに教えやすい．

2-③：コンピュータの利用に不慣れな学生が一定数いる．Web 茶まめがなければ，基礎講読の授業で受講生全員に対して形態素解析を行い，そのデータを基に研究するという課題を課することができない．現在の授業が成立しなくなる．

2-④：専門演習でも，簡便に形態素解析ができる Web 茶まめがなければ，形態素解析済みデータを使った研究をしようと思わないと考えられる．形態素解析が簡便にでき，さらに Excel を使って簡単に集計できることから，計量的な分析の面白さを感じることができるのではないか．

回答のあった授業のうち，和洋女子大学，および東京女子大学における語彙調査を行う授業における事例では，小説，新聞記事，歌詞，商品パッケージなど様々な文章を対象に Web 茶まめによる形態素解析機能が利用されている．

5.2 研究活動における活用

研究の一部に Web 茶まめを利用した文献・研究報告は，28 件が確認された．これらのうち，多くのものは土山玄 [10], [11]，中西太郎 [12]，鈴木一史 [13]，小椋秀樹 [14] のように研究の中で形態素解析が必要になった場合のツールとして利用している．それに加えて，中村一夫 [15]，笹島眞実 [16] のように実験を行ったときの再現性を考慮した実行環境の基準としても Web 茶まめを用いている．中村一夫 [15] の例では「語の認定は，形態素解析支援アプリケーション『Web 茶まめ』の生成するデータに従った」と凡例作成のさいに言及している．さらに，相良かおる [17] では Web 茶まめの解析前処理の利用についても言及している．研究成果の中には安岡孝一 [18] のようにシステムの中に API のように組み込んで利用している例も存在した．

5.3 考察

5.1 節で報告のあった利用事例から，実際に Web 茶まめの利用を前提とした授業が実施されており，語彙・文体の計量的な研究を行うための教育などが行われている．コンピュータ技術にあまり明るくない学生に対しても，従来よ

[*2] https://chamame.ninjal.ac.jp/

りも簡単に形態素解析を用いた教育を行えるようになったことが分かる．事例から，Web 上で利用可能としたこと，解析対象テキストをテキストボックスに貼り付けるか，テキストファイルとしてアップロードするだけで簡単に形態素解析結果が得られること，出力形式に表計算ソフトの Excel で読み込める形式を実装したことが効果的であったと考えられる．

これらから，解析対象テキストの入力方法や出力形式は妥当な実装であったと考える．また，これらによって簡単に形態素解析とその結果の集計ができるようになったことで，形態素解析を用いた計量的な言語教育の推進に寄与できたと考えられる．コロナ禍におけるオンライン授業のように受講生によって共通の実行環境を用意するのが難しい場合にも，Web 上で同様のインタフェースを用いて，だれでも同様の結果を得られるという Web 茶まめの利点がスムーズな授業運営に寄与したものと考えられる．さらに，青空文庫などのなじみのあるテキストを簡単に形態素解析できることで，研究への動機づけにもつながっている可能性が示唆された．

5.2 節より Web 茶まめは人文学・国語学分野において形態素解析を行う際のツールとして用いられており，これらの分野における利用に耐えうる設計実装であったと考えられる．また，共通の形態素解析環境を提供できるという特性から，Web 茶まめは形態素解析実行環境の基準として用いられていると考えられる．形態素解析の結果は，用いた形態素解析器や辞書，事前に行う前処理の内容によって結果が異なる．そのため，研究成果に再現性を持たせるためには，形態素解析に使用したこれらのツールについてすべて言及し筆者が管理を行う必要がある．しかし Web 茶まめを利用することで，一元的に辞書や形態素解析器を指定することができ，コンピュータ技術に明るくない研究者であっても容易に実験結果の再現性を確保できるようになる．このことによる，研究の信頼性の向上，研究推進に効果を発揮したと考えられる．

また，安岡 [18] のように Web 茶まめをシステムの一部として利用する例もあることから，それに適したサービス提供方法が求められていると考える．現在の Web 茶まめは Web ブラウザからの利用を前提としてサービス提供を行っているため，出力形式も人が閲覧しやすい形式として HTML，CSV，Excel を提供している．これらの形式は，プログラムなどを用いて機械処理するにあたっては必ずしも扱いやすいとはいえない．これまでに試験的に実装した WebAPI [19] では解析できるテキスト量に大きな制限がかかっており，ユーザの利用には結び付いていないものと考えられる．今回の調査から一定の利用ニーズはあるものと考えられるため，Web インタフェースから実行する場合に近いテキスト量を解析可能にし，解析結果を JSON や XML といった機械処理と親和性の高い形式で提供するなど，より使いやすい形で外部システムから連携可能な仕組みを用意することで，言語研究の推進効果が期待できると考えられる．

しかしそのようなサービスを提供する場合，人手でブラウザからシステムを利用する場合に比べて過度なアクセスが発生し，システムに想定以上の負荷がかかりサービス提供に影響を及ぼす可能性がある．その場合の解決策としては，ユーザ登録を行い，利用者を限定する方法や，一定時間内のアクセス数を制限する方法が考えられる．

6. おわりに

本稿では，Web ブラウザ上で各種 UniDic を用いた形態素解析を実行可能なアプリケーション，Web 茶まめの設計と実装，およびその運用と効果について述べた．Web 茶まめを運用することで，言語研究，教育の場で容易に形態素解析を利用できるようになった．これにより，教育現場では Web 茶まめの利用を前提とした授業が実施されるに至っている．研究における利用では，研究の中で形態素解析が必要になったときのツールとして Web 茶まめを用いる例が確認できた．また，実験を行ったときの再現性を考慮した実行環境の基準としても Web 茶まめは用いられており，研究の促進と信頼性向上に寄与したと考えられる．

今後の課題として，Web 茶まめをシステムの一部として利用するためのモジュールを開発し，研究に利用する例があり当初想定していたものとは異なる利用のされ方も見られた．これに対応するために，WebAPI のような外部システムからの利用を想定したサービス提供方式を拡充することが必要であると考えられる．

謝辞 本研究は国立国語研究所共同研究プロジェクト「開かれた共同構築環境による通時コーパスの拡張」による成果の一部である．

参考文献

[1] 小木曽智信，小町 守，松本裕治：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).
[2] 国立国語研究所：古文用 UniDics，国立国語研究所（オンライン），入手先 ⟨https://clrd.ninjal.ac.jp/unidic/download_all.html#unidic_chj⟩（参照 2022-09-20）.
[3] 小木曽智信：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 (2014).
[4] 工藤 拓：MeCab: Yet Another Part-of-Speech and Morphological Analyzer，（オンライン），入手先 ⟨http://taku910.github.io/mecab/⟩（参照 2022-05-20）.
[5] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. 2004 Conference on Empirical Methods in Natural Language Processing* (July 2004), pp.230–237 (2004).
[6] 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，日本語科学，Vol.22, pp.101–123 (2007).

[7] Goo ラボ：形態素解析 API（オンライン），入手先 ⟨https://labs.goo.ne.jp/api/jp/morphological-analysis/⟩（参照 2022-09-20）．
[8] Sinclair, S. and Rockwell, G.: VoyantTools（オンライン），入手先 ⟨https://voyant-tools.org/⟩（参照 2022-09-20）．
[9] 山田 篤，小磯花絵：NumTrans マニュアル，The UniDic Consortium (2008)．
[10] 土山 玄：絵入源氏物語のテキストデータに対する統計解析 web アプリケーションの設計，研究報告人文科学とコンピュータ，Vol.2016-CH-112, No.1, pp.1–4 (2016)．
[11] 土山 玄：絵入源氏物語の統計的なテキスト解析 web アプリケーションの設計利，じんもんこん 2016 論文集，Vol.2016, pp.75–80 (2016)．
[12] 中西太郎：場面設定会話と自由会話の特徴の比較『生活を伝える方言会話』，『COJADS』の共通語訳テキストを用いて，計量国語学，Vol.32, No.6, pp.346–356 (2020)．
[13] 鈴木一史：作文語彙と学習成績との関連性からわかる語彙力の諸相，茨城大学教育学部紀要，Vol.68, pp.1–12 (2019)．
[14] 小椋秀樹（編）：コーパスで学ぶ日本語学 日本語の語彙・表記，朝倉書店 (2020)．
[15] 中村一夫：承空本『小野篁集』語彙表・総索引—付属語編，国士舘人文学，Vol.10 (2020)．
[16] 笹島眞実：児童作文における使用語彙と文章様式からみたその量的特徴，学芸国語国文学，Vol.51, pp.259–244 (2019)．
[17] 相良かおる：『養生訓』の自動形態素解析における辞書の影響，人文科学とコンピュータシンポジウム論文集，Vol.2018, No.1, pp.153–160 (2018)．
[18] 安岡孝一：形態素解析部の付け替えによる近代日本語（旧字旧仮名）の係り受け解析，研究報告人文科学とコンピュータ，Vol.2020-CH-124, No.3, pp.1–8 (2020)．
[19] 川口寛治，薦田龍輝，堤 智昭：形態素解析ソフトウェア『Web 茶まめ』の改良と Web API の試作，言語資源活用ワークショップ 2016 発表論文集，pp.265–272 (2017)．

堤 智昭 （正会員）

2010 年東京農工大学工学部情報工学科卒業．2012 年同大学大学院工学研究科博士前期課程修了．2015 年同大学院博士後期課程修了．東京電機大学情報環境学部助教を経て，現在，筑波大学人文社会学系助教．モバイルネットワークエミュレータ，時刻情報応用システム，自律分散型インターネットセキュリティ基盤に関する研究に従事する一方，漢字・訓点の情報処理，通時コーパスの構築・応用に関する研究にも従事．日本語学会会員．博士（工学）．

小木曽 智信 （正会員）

国立国語研究所．1995 年東京大学文学部日本語日本文学（国語学）専修課程卒業．1997 年同大学大学院人文社会系研究科日本文化研究専攻修士課程修了．2001 年同博士課程中途退学．2014 年奈良先端科学技術大学院大学情報科学研究科博士課程修了．博士（工学）．2001 年明海大学講師，2006 年独立行政法人国立国語研究所研究員を経て，2009 年人間文化研究機構国立国語研究所准教授，2016 年より教授．専門は日本語学，自然言語処理．言語処理学会，日本語学会各会員．