WebChamame: An Online Tool for Morphological Analysis of Various Historical Japanese Texts using UniDic Dictionaries

Ogiso, Toshinobu
togiso@ninjal.ac.jp
National Institute for Japanese Language and Linguistics

Tsutsumi, Tomoaki
tsutsumi.tomoaki.gn@u.tsukuba.ac.jp
University of Tsukuba

Background

The Japanese text is not divided into separate words and has a large variety of notations. Therefore, morphological analysis is an indispensable tool for analyzing Japanese text data. This is true not only for modern Japanese but also for historical texts, and the older the text, the more important morphological analysis is for research (NINJAL 2022). Hence, we have developed and published the UniDic dictionaries for analyzing various historical Japanese texts (Ogiso et al. 2013). Currently, there are twelve dictionaries for texts of various eras and genres (Ogiso 2022). On the other hand, it is not easy for researchers in the humanities to set up a morphological analyzer and an appropriate dictionary and perform morphological analysis of texts on the command line. To resolve this, we have developed WebChamame, a online tool that can be used without complicated environment construction and can perform morphological analysis with UniDic dictionaries for multiple eras, with the aim of promoting Japanese language research and education using morphological analysis.

Features of WebChamame

WebChamame was developed as a cloud-based service that can be used from a web page. Input from the user is text data written in Japanese, and the system returns the analysis results to the user. The user interface runs on a Web browser (figure 1). Morphological analysis is performed using MeCab (Kudo et al. 2004, http://taku910.github.io/mecab/) and UniDic dictionaries installed on the server. MeCab is a morphological analyzer widely used to analyze Japanese with the UniDic dictionary. This server-based environment allows users to run multiple morphological analyses using UniDic without having to build a complicated environment.

The services provided by WebChamame can be divided into two stages: text preprocessing before analysis and morphological analysis processing by the selected UniDic.

In preprocessing, the following processes can be performed by checking the checkboxes as needed.

(1) Identification of character encoding of text and unification to UTF-8
(2) Deletion of HTML tags, etc.
(3) Conversion of half-width to full-width characters
(4) Unfolding odoriji characters
(5) Katakana-hiragana inversion
(6) Conversion of numeric characters

These are pre-processing steps that are often required for Japanese text processing. (4) and (5) are mainly required for pre-modern texts. Although these can be done in advance by the user, these can be easily analyzed by processing them on the WebChamame.

In the morphological analysis process, one or two dictionaries are selected from the 12 UniDic dictionaries in Table 1 for morphological analysis. If two dictionaries are selected, the user can see the differences in the analysis results by dictionary to see which dictionary is suitable for the analysis of the target text. The default output item for morphological analysis results is the information in Table 2, which is output in tabular format.

The output format can be selected from HTML format, CSV or Excel format, or a format that can be imported by the external tool ChaKi (https://ja.osdn.net/projects/chaki/).

In HTML format, the results can be viewed immediately in the browser, while in Excel format, the downloaded analysis results can be opened with a double-click to perform word counts, etc. using a pivot table.

Figure 1: Screenshot of WebChamame

Use in Research and Education

Users have commented favorably on the ease of use of WebChamame. Although we have not been able to conduct an exhaustive survey, WebChamame is being used in many universities in Japan in Japanese linguistics, corpus linguistics, and text analysis classes. For example, at Ritsumeikan University, all students (two classes, 40 students in total) are using WebChamame to conduct

morphological analysis and quantitative research on vocabulary and sentence structure after digitizing short stories, lyrics, etc. At least 28 studies using WebChamame were also identified in papers and reports. Some of these studies used morphological analysis when it was necessary for research purposes, while others used to guarantee reproducibility when experiments are performed.

Future Works

Since this system has been developed on the assumption that it will be used by researchers and students who understand Japanese, the user interface is only in Japanese. However, we plan to prepare an English version of the user interface (figure 2) in the near future (before the conference presentation), considering that foreign students who are not familiar with Japanese can use the system to analyze Japanese texts for research purposes. We are also considering outputting analysis results in English for fixed items such as parts of speech and conjugations. Furthermore, since word forms and pronunciations can be replaced mechanically, we are considering converting Japanese kana to romaji for output.

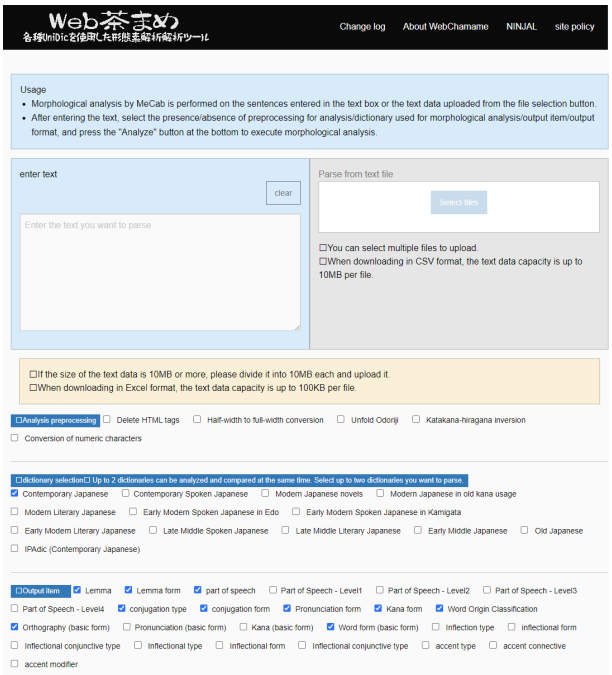

Figure 2: Screenshot of WebChamame English interface version

Table 1: UniDic dictionaries available on WebChamame

Types of UniDic	Description
Contemporary Japanese	Contemporary common written text.
Contemporary Spoken Japanese	Transcribed texts of contemporary spoken language
Modern Japanese novels	Mainly novels from the Meiji period to the present.
Modern Japanese in old kana usage	Mainly editorial texts written in the old kana syllabary.
Modern Literary Japanese	Editorial texts in modern written language.
Early Modern Spoken Japanese in Edo	Colloquial texts in Edo in the 17-18th century.
Early Modern Spoken Japanese in Kamigata	Colloquial texts in Osaka and Kyoto in the 17-18th century.
Early Modern Literary Japanese	Literary texts of the early modern period.
Late Middle Spoken Japanese	Colloquial texts of the Muromachi period (Kyogen, etc.).
Late Middle Literary Japanese	Literary texts of the Kamakura period.
Early Middle Japanese	Kana literature works and waka poetries of the Heian period.
Old Japanese	Man'yoshu, Norito, etc.

Table 2: Output items by morphological analysis (default items only)

Item name	**Example: からく/ karaku (hot)**
lemma	辛い / karai (hot)
lemma form	カライ / karai
Part of speech	形容詞−一般 / Adjective - common
Conjugation Type	形容詞 / adjective type
Conjugation Form	連用形−一般 / consecutive form – common
Pronunciation (appearance form)	カラク / karaku
Kana (appearance form)	カラク / karaku
Word Origin Classification	和 / native Japanese word
Orthography (basic form)	からい
Word Form (basic forms)	カライ/ karai

Acknowledgement: This research was supported by NINJAL “Diachronic Corpus” project, “Lexical Resources” project, and JPSP KAKENHI Grant Number 23H00007.

Bibliography

Kudo ,Taku / **Yamamoto, Kaoru** / **Matsumoro, Yuji** (2004) “Applying Conditional Random Fields to Japanese Morphological Analysis“, in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp.230-237. https://aclanthology.org/W04-3230

National Institute for Japanese Language and Linguistics (2022): "Corpus of Historical Japanese", https://clrd.ninjal.ac.jp/chj/overview-en.html (Accessed 2022-11-02).

Ogiso, Toshinobu / **Komachi, Mamoru** / **Matsumoto, Yuji** (2013). “Morphological analysis of historical Japanese text“, in *Journal of Natural Language Processing* 20(5): 727–748 (in Japanese). https://doi.org/10.5715/jnlp.20.727

Ogiso Toshinobu (2022) UniDics for Historical Texts, National Institite for Japanese Language and Linguistics (Online), https://clrd.ninjal.ac.jp/unidic/download_all.html#unidic_chj (Accessed 2022-11-02).