いろいろな古文の自動品詞分解

― 歴史的日本語資料の形態素解析と「Web茶まめ」―

小木曽智信

- 日本語の歴史を精密に研究するために、大量の古文のテキストを品詞分解してどこにどんな単語が使われているかデータベース化する（＝単語情報付きのコーパスを作る）必要がある。
- この品詞分解のような処理は形態素解析という技術によってコンピューターに自動で行うことができるが、古文は対応した辞書がないため解析ができなかった。
- そこで、いろいろな古文を解析するための電子辞書（古文用のUniDic）を作成して解析を可能にした。
- 作った辞書は「Web茶まめ」というオンラインのツールで誰でも利用できるように公開している。

形態素解析

入力された文章を、単語に分割して読みや品詞、活用形などの情報を付加して出力する技術。MeCabという解析器が有名。これを古文でもできるようにした。

入力

いづれの御時にか女御更衣あまたさぶらひたまひける中に

↓ 形態素解析

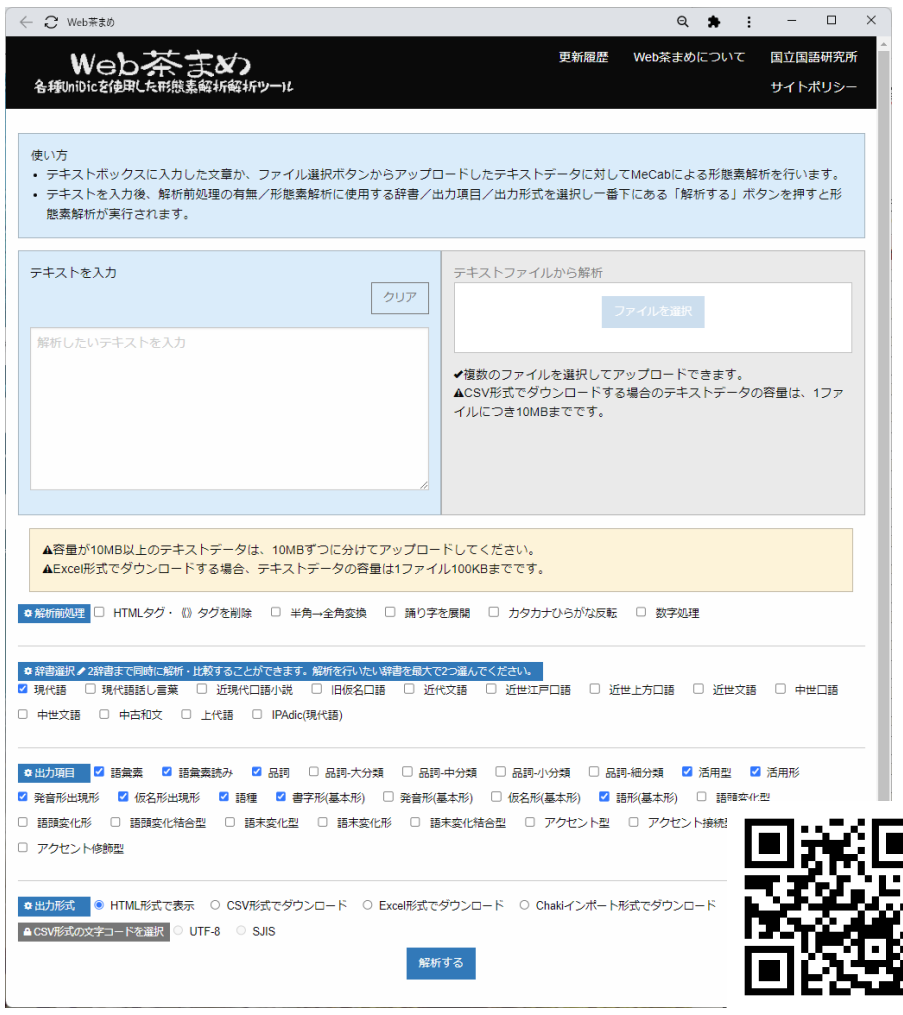
出力

書字形	語彙素読み	語彙素	語形	品詞	活用型	活用形	発音形	語種
いづれ	イズレ	何れ	イズレ	代名詞			イズレ	和
の	ノ	の	ノ	助詞-格助詞			ノ	和
御	オン	御	オオン	接頭辞			オオン	和
時	トキ	時	トキ	名詞-普通名詞-副詞可能			トキ	和
に	ニ	に	ニ	助詞-格助詞			ニ	和
か	カ	か	カ	助詞-係助詞			カ	和
女御	ニョウゴ	女御	ニョウゴ	名詞-普通名詞-一般			ニョーゴ	漢
更衣	コウイ	更衣	コウイ	名詞-普通名詞-サ変可能			コーイ	漢
あまた	アマタ	数多	アマタ	副詞			アマタ	和
さぶらひ	サブラウ	侍ふ	サブラウ	動詞-一般	文語四段-ハ行	連用形-一般	サブライ	和
たまひ	タマウ	給う	タマウ	動詞-非自立可能	文語四段-ハ行	連用形-一般	タマイ	和
ける	ケリ	けり	ケリ	助動詞	文語助動詞-ケリ	連体形-一般	ケル	和
中	ナカ	中	ナカ	名詞-普通名詞-副詞可能			ナカ	和
に	ニ	に	ニ	助詞-格助詞			ニ	和

Web茶まめ

実際に解析が行えるサイト「Web茶まめ」
https://chamame.ninjal.ac.jp/
スマホでも利用できる。

Web茶まめ



いろいろな古文の形態素解析用辞書

万葉集や源氏物語から江戸時代の戯作、明治の論説文など、ひとロに古文といっても全く違うので、専用の辞書が必要となる。時代・文体・地域別に10種類を作成してオープンライセンスで公開中。

古文用UniDicS

1. 近代口語小説UniDic
2. 旧仮名口語UniDic
3. 近代文語UniDic
4. 近世江戸口語UniDic
5. 近世上方口語UniDic
6. 近世文語UniDic
7. 中世口語UniDic
8. 中世文語UniDic
9. 中古和文UniDic
10. 上代語UniDic

近現代口語小説UniDic

主として明治から現代までの小説を短単位自動解析するための解析用辞書です。 近現代の語彙を追加し『日本語歴史コーパス』明治・大正編の小説と国語教科書、『現代日本語書き言葉均衡コーパス』の文学作品（PB_9）のコアデータを学習に利用しています。

営利を目的として利用する場合には、下記問合せ先まで事前にご相談ください。

この解析用辞書を利用して行なった研究等の成果を公表する場合は、その旨を明記してください。 必要に応じて参考文献に挙げた文献を参照してください。

ライセンス

ライセンスに同意して最新版をダウンロード

旧バージョンはこちら

参考文献

- 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).

「日本語歴史コーパス」の構築に利用

日本語の歴史を研究するために広く使われている。

奈良時代	■万葉集(9.8万語) ■宣命 (1.8万語) ■祝詞 (0.9万語)	
平安時代	■仮名文学作品 (85.7万語) ■訓点資料 (0.9万語)	和歌集
鎌倉時代	■説話・随筆 (71.3万語) ■日記・紀行 (11.0万語) ■軍記 (28.1万語)	■和歌集 (26.2万語)
室町時代	■狂言 (23. 5万語) ■キリシタン資料 (12.3万語)	
江戸時代	■洒落本 (20.4万語) ■人情本 (37.3万語) ■近松浄瑠璃 (23.1万語) ■随筆・紀行 (1.4万語)	
明治・大正	■雑誌 (1274.8万語) ■教科書 (70.9万語) ■明治初期口語資料 (20.1万語) ■近代小説 (69.7万語) ■新聞 (38.6万語) ■落語SP盤 (9.3万語)	

こちらもオンラインで公開中！
https://clrd.ninjal.ac.jp/chj/
（利用無料、要登録）