Web 資料 1

「Web 茶まめ」による形態素解析と語彙表作成

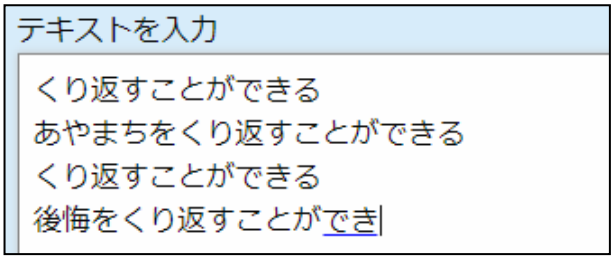
この資料では，「Web 茶まめ」を使った形態素解析と，その解析結果ファイルを使った語彙表作成の手順を説明する。

なお，「Web 茶まめ」で使用される形態素解析用辞書 UniDic の開発は現在も進められており，バージョンアップが行われることがある。本書執筆時点の UniDic と現在公開されている UniDic とは，バージョンが異なるため，本書に示した解析結果と現在公開されている UniDic で解析した結果とが異なる場合がある。この点についてあらかじめ御留意いただきたい。

１．「Web 茶まめ」を使った形態素解析

（１）テキストを入力して形態素解析を行う

① ［テキストを入力］ボックスに，形態素解析対象のテキストを入力する。



② 各項目を以下のとおり設定した上で，［解析してみる］ボタンをクリックする。

辞書選択：「現代語」

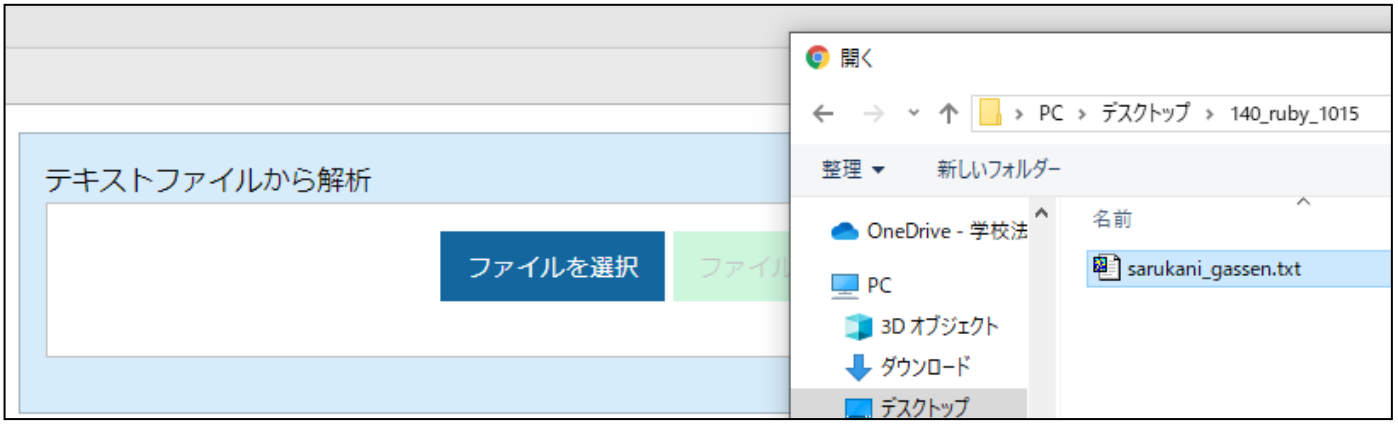
出力項目：解析結果ファイルに出力したい項目

※ 本書で特に指定がない場合は，初期設定のまま

出力形式：「Excel 形式でダウンロード」

（２）ファイルを使って形態素解析を行う

① ［ファイルを選択］をクリックし，解析したいファイルを選択する。



② 上記（１）の②と同様に各項目を設定した上で，［ファイルから解析］ボタンをクリックする。

２．形態素解析前の処理

青空文庫の「テキストファイル（ルビあり）」では，以下のように，ルビが《 》（二重山括弧）でくくられて入力されている。また，「ルビの付く文字列の始まりを特定する記号」として「｜」が入力されている。

蟹《かに》の握り飯を奪った猿《さる》はとうとう蟹に仇《かたき》を取られた。
現に商業会議所会頭某｜男爵《だんしゃく》のごときは

（芥川龍之介「猿蟹合戦」）

また，入力者注（主に外字の説明や，傍点の位置の指定）が以下のような形式で入力されている。

［＃地から１字上げ］（大正十二年二月）

これらのルビや注は，「Web 茶まめ」で形態素解析をする際に，邪魔になるため，テキストからあらかじめ削除する必要がある。削除には，エディターで，正規表現を使った一括置換等を利用するとよい。

以下には，秀丸エディタを使った削除の手順を示す。

（１）青空文庫からダウンロードしたファイルを秀丸エディタで開く。

（２）「ルビの付く文字列の始まりを特定する記号」の削除

メニューから［検索］→［置換］を選択する。［正規表現］をチェックした上で，［検索］［置換］に次のように入力する。

検索：｜　　※ 全角文字
置換：　　※ 何も入力しない。

（３）ルビの削除

メニューから［検索］→［置換］を選択する。［正規表現］をチェックした上で，［検索］［置換］に次のように入力する。

検索：《.*?》　　※ 全て半角文字
置換：　　※ 何も入力しない。

（４）入力者注の削除

メニューから［検索］→［置換］を選択する。［正規表現］をチェックした上で，［検索］［置換］に次のように入力する。

検索：［.*?］　　※ 全て半角文字
置換：　　※ 何も入力しない。

テキストには，冒頭部に凡例（「テキスト中に現れる記号について」），末尾に底本，入力者等に関する情報も入力されている。これらはテキストの電子化に関する情報であり，作品に用いられた語彙等の研究を行う上で不要な情報である。これらも形態素解析前に削除する必要がある。

３．語彙表の作成

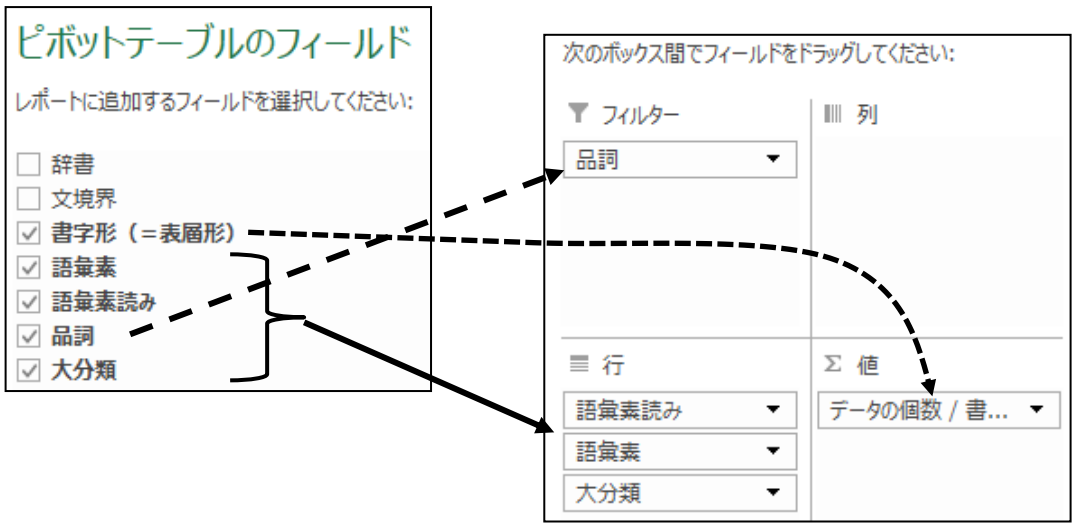
（１）形態素解析結果ファイル（Excel ファイル）を開き，メニューから［挿入］→［ピボットテーブル］を選択する。

（２）以下のとおりフィルターと行，値を指定する。

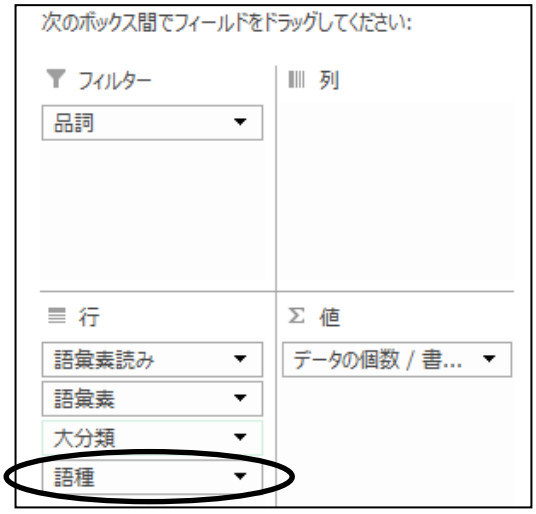
フィルター：「品詞」をドラッグ＆ドロップ。

行：「語彙素読み」「語彙素」「大分類」の順にドラッグ＆ドロップ。

値：「書字形=表層形」をドラッグ＆ドロップ。



※ 語彙表に語種の情報を追加するときは，「語彙素読み」「語彙素」「大分類」に加えて「語種」をドラッグ＆ドロップする

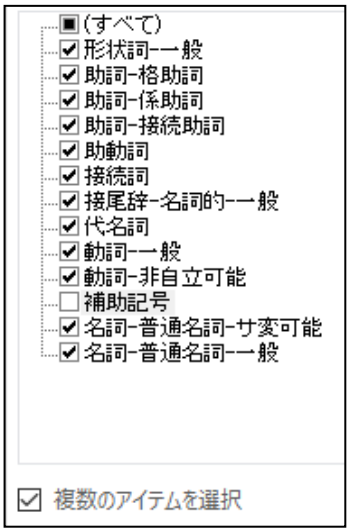


（３）［ピボットテーブルツール］から［デザイン］→［小計］→［小計を表示しない］を選択する。

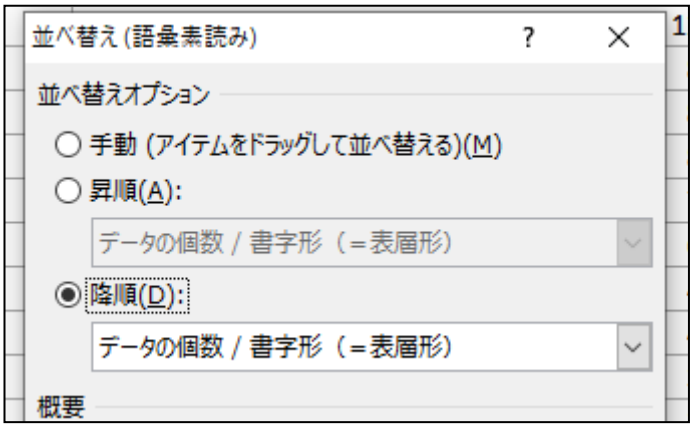
（４）［ピボットテーブルツール］から［デザイン］→［レポートのレイアウト］→［表形式で表示］を選択する。

（５）［ピボットテーブルツール］から［デザイン］→［レポートのレイアウト］→［アイテムのラベルを全て繰り返す］を選択する。

（6）「品詞」のフィルターをクリックし，表示されたメニューで以下のとおり設定する。
①「複数のアイテムを選択」をチェックする。
②「補助記号」のチェックを外す。



（7）「語彙素読み」のいずれかのセルを右クリックし，「並べ替え」→「その他の並べ替えオプション」を選択する。表示されたボックスで「並べ替えオプション」を図のように設定する。



以上の操作で本書 p.25 の表 2.1 のような語彙表を作成することができる。