E-004

近代文語に対応した形態素解析辞書の作成についての検討
Generating a Dictionary for Morphological Analyzer for Modern Literary Japanese Text

山本 理紗子[1)]　来住 伸子[1)]
Risako Yamamoto　Nobuko Kishi

1　はじめに

現在多くの近代文語で書かれた小説が青空文庫 ([1]) などの形で、インターネット上で公開されている。形態素解析器 MeCab に対応した、近代文語のための形態素解析辞書には近代文語 UniDic([3]) がある。しかし、現在の近代文語 UniDic では青空文庫にある旧字旧仮名の作品の形態素解析には十分に対応していない。本研究では、青空文庫に旧字旧仮名づかいのテキストと新字新仮名づかいのテキストが両方存在する作品は 186 作品あることを利用して辞書を作成することを目指している。

本報告では 186 作品の中から、芥川龍之介「蜜柑」のテキスト一作品を対象に分析を行なった結果を報告する。

2　本研究の手法

本研究では、文語体とは、近代から明治・大正時代に使われた書き言葉とする。昭和 24 年の内閣告示で交付された当用漢字字体表により定められた漢字を新字、それ以前に用いられてきた漢字を旧字とする。([5]) 昭和 21 年の内閣訓令で交付された仮名遣いを新仮名、それ以前に用いられてきた仮名遣いを旧仮名とする。([6])

本研究は以下のような流れで行う予定である。

- 青空文庫に収録されている作品のうち、同著者同タイトルの仮名遣い種別が旧字旧仮名と新字新仮名のファイルがあるものを選ぶ。
- 選んだファイルのテキストを加工したファイルを作成する。
- 仮名遣い種別ごとに加工したファイルを比較する。

このような手法用いて近代文語 UniDic と ipadic を用いた形態素解析で対応できなかった文字列を利用して、近代文語 UniDic に追加登録する形態素を生成することを目指している。

2.1　使用したデータ

本研究では次の二つの辞書を使用した。

- ipadic
  形態素解析のための辞書であり、情報処理振興事業協会により作成された ipa 品詞体系を元にしている。([4])
- 近代文語 UniDic
  小木曽他によって作成された近代文語文を対象とした形態素解析のための辞書である。近代文語 UniDic の作成には公文書、青空文庫の 9 編の評論を含む評論文等の文語論説文を中心としたデータが用いられている。([3])

本研究では次のテキストデータを使用した。

- 青空文庫に収録されている作品
  青空文庫には現在 18058 ファイルが収録されている。本研究では下記の条件をもとに対象とするデータを選んだ。同じ著者名と同じタイトルの仮名遣い種別が旧字旧仮名と新字新仮名の両方の版が存在する作品である。この条件を満たしかつ、旧字旧仮名版、新字新仮名版が複数存在するタイトルについては登録された時期が一番古いものを選んだ。対象となる作品は 186 作品である。全データは青空文庫の全データを保管している aozorahack([2]) より入手した。

2.2　データの加工

本報告では先述したように一作品のみを対象に以下のような処理を行なった。一作品のみとした理由は解析結果の詳細をひとつひとつ確認し、186 作品を処理するとどの程度の大きさの辞書が生成できるか推定するためである。

2.2.1　記号の除去

青空文庫からダウンロードした作品テキストファイルに正規表現を利用して以下のような処理を行なった。

- 文字コードを shift-jis から Utf-8 に変換する。
- 文中の脚注、作品の発表された年月日、底本の情報を削除した。
- 文中のふりがなを削除する。
- 次のような注に関しては、ふりがなを残し、注を削除した。
  ※［＃「釣のつくり」、第 3 水準 1-14-75］《にほひ》

2.2.2　形態素解析

テキストファイルの形式を以下のように変更した。

- テキストを一行一文に分割した。
- 新字新仮名版テキストの句読点の入力ミスによる文の区切りを修正した。
- 一文を形態素解析を用いて形態素のリストにした。
- この際、旧字旧仮名版テキストには近代文語 UniDic を、新字新仮名版テキストには ipadic を MeCab の辞書として利用した。

2.2.3　旧字を新字に置換

形態素のリストの、旧字を新字に置換した。次の手順で置換した。

- 旧字・新字表 ([8]) に「ゐ い」と「ゑ え」を追加した。([9])
- 旧字を key、新字を value とする json ファイルに変換した。
- 旧字旧仮名の形態素リスト中の旧字を新字に、一部の旧字を新字に置換して新字旧仮名の形態素リストを作成した。

2.2.4　形態素リストの比較

旧字旧仮名版と新字旧仮名版のテキストをそれぞれ新字新仮名版のテキストと比較した。比較は両者の一文一文の形態素ごとに python のモジュールである

1)　津田塾大学

difflib の SequenceMatcher を利用することで行なった。SequenceMatcher はゲシュタルトパターンマッチングを利用して二つの文字列の類似度を計算する。ゲシュタルトパターンマッチングアルゴリズムは最小の編集距離を生成することは保証していないが、レーベンシュタイン距離アルゴリズムよりは処理が高速なことが知られている。([10])

3　現在の解析結果

3.1　現在の追加形態素の候補

まず、それぞれのリストの形態素数は以下のようになる。

	旧字旧仮名版	新字旧仮名版	新字新仮名
形態素数	2033	2033	2009

表 1　旧字旧仮名版・新字旧仮名・新字新仮名の形態素数

旧字旧仮名の形態素リストと新字旧仮名の形態素リストをそれぞれ新字新仮名の形態素リストを比較した結果は以下のようになる。

	旧字旧仮名版	新字旧仮名版
新字新仮名と一致しない形態素数	472	378
文中の形態素数	2033	2033

表 2　旧字旧仮名版・新字旧仮名の新字新仮名版と一致しない形態素数

表 2 の結果から、旧字旧仮名版の 94 の形態素の差異は旧字または一部の旧仮名を用いて表記されていることが原因で生じた差異であるということがわかった。以降は新字旧仮名版の 378 の差異について説明する。新字旧仮名の差異は以下の 3 種に分類できた。

1 種類目は旧仮名づかいによって生じた差異である。

新字旧仮名版形態素	新字新仮名版形態素
曇つ	曇っ

表 3　旧仮名づかいによる差異

2 種類目は、新字新仮名版作成時の表記の違いである。

新字旧仮名版形態素	新字新仮名版形態素
ポケット	ポッケット
頓著	頓着
著物	着物

表 4　表記の違いによる差異

3 種類目は、近代文語 UniDic を用いて形態素解析を行った際に形態素解析をできなかったものである。

	元のテキスト	分かち書き
新字旧仮名	銀杏返し	銀杏 返し
新字新仮名	銀杏返し	銀杏返し

表 5　形態素解析時に生じる差異

4　考察

- 旧仮名遣いと新仮名遣いの違いが原因の差異については正しく形態素解析できていないものは新たな追加項目となる。
- 新字新仮名版入力時の表記が異なっているものの差異のうち、「頓著」のように近代文語 UniDic に登録されている項目は追加する必要がない。しかし「著物」のように近代文語 UniDic に登録されていない場合は追加する必要がある。
- 形態素解析時に生じる差異について、「銀杏返し」のように正しく形態素解析できていない場合は新たに登録する必要がある。

186 作品全てを調査の対象とした場合、検討が必要になる項目は全形態素の 2 割程度になることを現在推定する。ただし、作品ごとに使われている語彙の違いから、この推定割合は、辞書登録が必要な項目数と異なる可能性が高い。検討が必要な項目の分布にはばらつきが見られることが大いに予想できる。今後、青空文庫の他の作品についても形態素解析し、辞書登録が必要な項目数を正確に調査する予定である。

参考文献

[1] 「青空文庫」
https://www.aozora.gr.jp/index.html

[2] aozorahack
https://github.com/aozorahack

[3] 小木曽智信　国立国語研究所　研究開発部門「近代文語文を対象とした形態素解析のための電子化辞書の作成とその活用」

[4] 工藤拓『実践・自然言語処理シリーズ 第 2 巻 形態素解析の理論と実装』近代化学社、2018 年。

[5] 文化庁「3. 当用漢字字体表（昭和 24 年 4 月）【国語審議会】」『国語施策沿革資料１２（平成 9 年 1 月 17 日）漢字字体資料集（諸案集成２・研究資料）』
https://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/sisaku/enkaku/enkaku12.html

[6] 文化庁「8. 昭和 21 年 11 月 内閣訓令第８号、内閣告示 33 号現代かなづかい」『国語施策沿革資料１（昭和 55 年 3 月 31 日）仮名遣い資料集（諸案集成）』
https://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/sisaku/enkaku/enkaku1.html

[7] 青空文庫編【テキスト中に現れる記号について】
https://www.aozora.gr.jp/KOSAKU/txt_chu_kigo.html

[8] 旧字体新字体相互変換表
https://www.gaoshukai.com/lab/0039/

[9] 三省堂編集所編「新旧かなづかい辞典」三省堂、2017 年。

[10] Python 3.10.4 Documentation 「difflib ― 差分の計算を助ける」
https://docs.python.org/ja/3/library/difflib.html