テキストアナリシスによる明治期日本語教科書『日語指南』の検証：特集多様化する日本語研究の現在

テキストアナリシスによる明治期日本語教科書『日語指南』の検証

伊藤孝行

1．はじめに

テキストアナリシスによる近代語資料の調査および分析はどれほど可能であろうか。本稿では『日語指南』を資料とし，テキストアナリシス（text analysis）[(1)]による調査を報告する。近代に刊行された日本語を母語としない人を対象につくられた日本語教科書のことばについてテキストアナリシスによる基礎データとなる調査をし，分析のための第１次調査として報告する。

近代に刊行された日本語教科書の多くはこの数年来国立国会図書館デジタルコレクション等により，インターネットに接続されている環境であれば時間や場所を問わず閲覧およびダウンロードが可能になっている。とはいえ閲覧できるのは画像ファイルであるため，言語資料として用例収集を行う際は従来の方法，すなわちすべて人力で確認し，収集したい用例を見つける作業にかなりの時間を使わざるを得ないのが現状である。

そこで本稿では『日語指南』の日本語による記述箇所全文をテキストデータ化したテキストファイルを使用し，テキストアナリシスによる調査を試みる。検索性を高め，テキストアナリシスによる調査により，近代に刊行された日本語教科書のことばの実態を従来とは異なる視点から探る。小林（2017）に「全ての単語を網羅的に調べていく過程で何か有用な情報を発見しようとする仮説発見型（hypothesis finding）のアプローチであり，後者は，より明確な分析目的について検討する仮説検証型（hypothesis testing）のアプローチであるといわれています。」とある。本稿は仮説発見型（hypothesis finding）のアプローチにあたる。

2．『日語指南』のこれまで

『日語指南』のテキストアナリシスに入る前に，本稿でとりあげる『日語指南』の著者，書誌等，先行研究について簡単にふれる。

著者について，『日語指南』の著者は金井保三（かないやすぞう，1872-1917）。金井についてもっとも詳細に記されてある諸星（2000）によると，金井は1872（明治４）年，長野県北佐久郡中佐津村（現在の佐久市根々井）生まれ。中国で10年ほど過ごし，帰国後は教育者として国語や中国語を担当し，獨協中学，東京高等商業学校（一橋大学の前身）付属外国語学校，台湾協会専門学校（拓殖大学の前身），帝大（東京大学の前身），早稲田大学で教鞭をとった。諸星（2000）によれば，教師としての金井保三は厳格な中にもユーモアをまじえた和やかな授業を行っていたことがわかる。享年47歳。

書誌等について，『日語指南』は1904（明治37）年に『日語指南壹』が早稲田大学出版部から刊行され，1905（明治38）年に『日語指南貳』が同じく早稲田大学出版部から刊行された。金井が32歳から33歳にかけての著作である。『日語指南壹』は全168ページ，『日語指南貳』は全194ページある。中国語を母語とする学習者向けの日本語教科書である。すべて縦書きで，各課(2)の題名はすべて中国語，本文も中国語による説明があり，日本語の単語や例文が並べられている。1905（明治38）年から1910（明治43）年まで開設されていた早稲田大学清国留学生部の開設と『日語指南』２冊の刊行年が重なり，金井も早稲田大学清国留学生部で日本語を担当していたことが明らかであることから，『日語指南』は新設される早稲田大学清国留学生部の日本語のクラスを念頭に置いてつくられたものであることはまちがいあるまい。

所蔵状況等について，CiNii Booksによると『日語指南』所蔵図書館は２館，拓殖大学図書館と東京大学文学部言語学研究室に所蔵が確認できる。2010年にクロスカルチャー出版から刊行された『近代日本語教科書選集 第３巻』にも所収されている。さらに，国立国会図書館デジタルコレクションにも所収されたため，インターネット環境があれば容易に『日語指南』の閲覧等が可能になった。単著によるこれだけの分量のある日本語教科書は，現代に於いてはまれである。伊藤（2017）でも指摘したが，ひらがなで記されている課とカタカナで記されている課がバランスよくほぼ交互に記されてあるところ等，配慮の行きとどいた日本語教科書と言える。金井自らの中国語学習者としての経験，そして帰国後さまざまな教育機関においての中国語および日本語の教歴がこのようなアイディアの源となっていたことは想像に難くない。

３．テキストアナリシスによる『日語指南』

３．１．テキストアナリシスのための環境設定

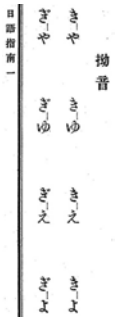
本稿ではテキストアナリシスのためにR（Mac版，R version 3.5.1）を使用し，形態素解析にはMeCabを使用した。MeCabをRで使用できるようにするために，RMeCabパッケージ（石田基広氏作成）をインストールした。形態素解析用の辞

書は，国立国語研究所が公開・配布している現代書き言葉UniDicをインストールし，現代書き言葉UniDicをRMeCabで使用できるように開発されたRMeCabUniパッケージ（石田基広氏作成）をインストールした。

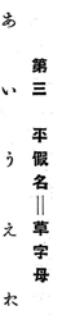
参考１　『日語指南』テキストデータ

テキストアナリシスのために，まず『日語指南』の日本語で記されてあるところすべてを対象に，テキストエディタを使用し，人力で全文テキストデータ化した。使用したテキストエディタは，MicrosoftのVisual Studio Code（Mac版）を使用した。テキストデータ化にあたっては，テキストの文字コードはUTF-8を使用した。形態素解析の結果を出力する段階で誤解析を減らすため，旧漢字・表音式仮名遣い・歴史的仮名遣いを現代仮名遣いに整形した版（ファイル名：1_shinan.txt）を作成した（参考１）。さらに，『日語指南』の場合は拗音の表記が近代に刊行された他の日本語教科書には見られない独自の表記がある（参考２）ため，独自の表記も修正した。ひらがなの「お」のみ「於」を崩した変体仮名が用いられているので「お」に変換した（参考３）。

参考２　『日語指南』の拗音表記



参考３　『日語指南』の「お」



３．２．テキストアナリシスによる『日語指南』の基礎データ

テキストアナリシスによる『日語指南』の基礎データとなりうる項目を調査する。『日語指南』の総語数，異語数，異語率，頻度表が『日語指南』全文を収めたテキストファイルを読みこむ。下記を入力すると読みこむファイルを選択するウィンドウが開くので，該当するファイルを選択する。今回はデスクトップにテ

キストファイルを保存している場合の記述である。

Rを開き，Rで今回の調査に必要となるパッケージを使用できるよう設定する（この設定を「呼び出す」といわれることが多い）。前述したとおり，インストールしたRMeCabとRMeCabUniを呼び出す。

```
> library(RMeCab)
> library(RMeCabUni)
```

3．2．1．『日語指南』の総語数，異語数，異語率

```
> RMeCabFreq.res <- RMeCabFreq(file.choose())
file = /Users/xxx/Desktop/1_shinan.txt
length = 2136
```

読みこんだファイルの場所とファイル名がfileに，読みこんだテキストファイルの異語数（types：重複を除いた単語数）が出力される。『日語指南』の異語数は2136。

総語数を計算する。『日語指南』の総語数（tokens：テキストにある単語数）は15665。

```
> sum(RMeCabFreq.res$Freq)
[1] 15665
```

総語数と異語数が明らかになったので，異語率（TTR あるいは type-token ratio）が計算できる。異語率は0.1363549。異語率は０から１までの値をとり，１に近いほどそのテキスト中の単語の種類があるので，『日語指南』は単語の種類が豊富とは言えないことがわかる。異語率については諸説あり，たんに異語数を総語数で割った値を異語数とすることに対してさまざま検討されていることも付しておく。本稿では『日語指南』のような近代日本語教科書についてテキストアナリシスの可能性をさぐることを目的としているので，たんに異語数を総語数で割った異語率の算出にとどめる。

```
> nrow(RMeCabFreq.res) / sum(RMeCabFreq.res$Freq)
[1] 0.1363549
```

3．2．2．『日語指南』の全文頻度表

『日語指南』の頻度表を作成する。頻度表には形態素と品詞頻度表の作成のため，

write.tableという関数を使用する。今回はshinan.csvというファイル名をつける。OSがMacの場合，文字化けしたファイルが出力されることがあるので，最後に`fileEncoding = "UTF-8"`と付けた。

```
> write.table(RMeCabFreq.res, file = "shinan.csv", sep = ",",
+ row.names = TRUE, col.names = NA, fileEncoding = "UTF-8")
```

３．２．３．『日語指南』の品詞別頻度表

『日語指南』全文の頻度表が出力されたので，次は品詞別にみていく。

本稿ではUniDicを形態素解析用の辞書としているので，形態素解析結果にはUniDicの品詞体系が出力される。UniDicの品詞体系は，学校文法とおおむね同じであるが，異なるところがある。UniDicの品詞分類は4層構造になっており，大分類・中分類・小分類・細分類となっている。大分類にある形状詞について，伝・山田・小椋・小磯・小木曽（2008）の定義は下記に引用する。中分類・小分類・再分類については，詳しくは伝・山田・小椋・小磯・小木曽（2008）を参照されたい。

> ■形状詞－一般　「静か」「健やか」など、いわゆる形容動詞の語幹部分。ただし、名詞としての用法があるものは、「名詞－普通名詞－形状詞可能」に分類する。
> ■形状詞－タリ　「釈然」「錚々」など、いわゆるタリ活用の形容動詞の語幹部分。
> ■形状詞－助動詞語幹　一般に助動詞とされる「そうだ（様態）」「ようだ」「みたいだ」の語幹部分。

品詞大分類の項目であるInfo1の内訳を確認する。なお，一番左にある番号は自動的に付けられた番号で，Rのウィンドウの大きさ等により変動する。今回は１行に５つの品詞名が列挙されたため，２行目の行頭に[６]，３行目の行頭に[11]と自動的に付けられた。本稿では動詞，形容詞，名詞，副詞を扱う。

```
> unique(RMeCabFreq.res$Info1)
 [1] "補助記号"   "助詞"       "助動詞"     "動詞"       "名詞"
 [6] "感動詞"     "代名詞"     "連体詞"     "形容詞"     "接頭辞"
[11] "接尾辞"     "形状詞"     "副詞"       "記号"       "接続詞"
```

同様に品詞細分類の項目であるInfo2の内訳も確認しておく。

```
> unique(RMeCabFreq.res$Info2)
 [1] "読点"       "係助詞"     "*"          "格助詞"
 [5] "接続助詞"   "非自立可能" "普通名詞"   "フィラー"
 [9] "終助詞"     "一般"       "数詞"       "準体助詞"
[13] "名詞的"     "助動詞語幹" "副助詞"     "形容詞的"
[17] "固有名詞"   "動詞的"     "文字"       "タリ"
[21] "形状詞的"   "括弧閉"     "括弧開"
```

品詞大分類および品詞細分類を確認したところで，品詞別の数を計算する。動詞479，形容詞122，名詞1116，副詞86。

```
> nrow(RMeCabFreq.res.v <- RMeCabFreq.res[RMeCabFreq.res$Info1==
+ "動詞",])
[1] 479
> nrow(RMeCabFreq.res.a1 <-
+ RMeCabFreq.res[RMeCabFreq.res$Info1=="形容詞",])
[1] 122
> nrow(RMeCabFreq.res.n <- RMeCabFreq.res[RMeCabFreq.res$Info1==
+ "名詞",])
[1] 1116
> nrow(RMeCabFreq.res.ad <-
+ RMeCabFreq.res[RMeCabFreq.res$Info1=="副詞",])
[1] 86
```

品詞別の頻度表を作成する。Info1が動詞となっているものにしぼりこみ，そのFreqの降順に並べる。形容詞・名詞・副詞についても同様である。

```
> RMeCabFreq.res.v <- RMeCabFreq.res[RMeCabFreq.res$Info1==
+ "動詞", ]
> RMeCabFreq.res.v[rev(order(RMeCabFreq.res.v$Freq)), ]
```

表１ 『日語指南』動詞頻度表（頻度10以上，上位31）

Term	Freq
為る	201
有る	189
行く	106
言う	87
居る	70
成る	54
為さる	50
来る	36
出来る	31
見る	26
買う	22

Term	Freq
書く	20
出す	20
遣る	20
過ぎる	19
下さる	18
乗る	17
出る	16
飲む	15
仕舞う	15
降る	14
落ちる	13

Term	Freq
思う	12
拵える	12
知れる	12
聞く	12
入る	11
持つ	11
取る	10
変える	10
返る	10

表２ 『日語指南』形容詞頻度表（頻度10以上，上位8）

Term	Freq
無い	64
良い	41
酷い	18
早い	15

Term	Freq
悪い	13
多い	10
寒い	10
旨い	10

表３ 『日語指南』名詞頻度表（頻度13以上，上位31）

Term	Freq
人	184
事	91
時	45
日	34
月	31
物	29
度	28
方	24
所	23
者	23

Term	Freq
店	22
皆	21
品物	20
年	19
今日	18
家	18
御覧	18
今	17
船	17
風	17

Term	Freq
雨	16
手紙	15
車	15
話	14
金	14
馬	14
上	13
今年	13
本	13
着物	13

表 4『日語指南』副詞頻度表(頻度10以上，上位14)

Term	Freq
未だ	30
余り	27
急度	23
もう	21
若し	19
大変	13
そう	12

Term	Freq
どうぞ	12
是非	12
直ぐ	12
一寸	11
もっと	10
余程	10

３．２．４．『日語指南』のn-gram

『日語指南』全文の頻度表，品詞別の頻度表の次はn-gramについて調査する。n-gramとは文字あるいは形態素の連なりのことである。よって文字のn-gram，単語のn-gram，品詞のn-gram等のn-gramがある。nの値は２や３が用いられることが多い。2-gramはバイグラム（bigram），3-gramはトライグラムあるいはトリグラム（trigram）と呼ばれる。どのような文字等の連なりが多いのかを知ることができる。本稿では文字の2-gram，単語の2-gram，品詞の2-gramを調査し，それぞれ頻度順上位５つに並べかえた結果を報告する。

まず，文字の2-gramを調査する。

```
> ngram.res.1 <- Ngram(file.choose(),type = 0)
file = /Users/xxx/Desktop/1_shinan.txt Ngram = 2
length = 7470

> head(ngram.res.1,5)
    Ngram Freq
1 […-…]   28
2 […-て]   3
3 […-と]   2
4 […-な]   2
5 […-ば]   2
```

文字の2-gramは出たが，頻度順ではないので頻度順降順の結果をだす。

```
> ngram.res.1 <- ngram.res.1[order(ngram.res.1$Freq,
+ decreasing = TRUE), ]
```

『日語指南』の文字の2-gram上位５つは下記となった。「…は，…」という連なり，助動詞「です」「ます」とその否定形が多い連なりであることがうかがえる。

```
> head(ngram.res.1,5)
      Ngram          Freq
2811  [は-、]        457
3117  [ま-す]        324
3465  [り-ま]        287
2083  [で-す]        280
3119  [ま-せ]        225
```

同様に単語の2-gram，品詞の2-gramを調査する。

```
      Ngram          Freq
489   [人-人]        18
1069  [品物-皆]      7
3389  [風-風]        7
142   [一-度]        6
149   [一-月]        6

      Ngram          Freq
73    [名詞-助詞]    1853
54    [助詞-補助記号] 1230
47    [助詞-動詞]    1054
59    [動詞-助動詞]  985
75    [名詞-名詞]    897
```

４．おわりに

以上,『日語指南』のテキストデータよりテキストアナリシスの面から調査した。「多様化する日本語研究の現在」というお題は，まさに時宜を得たお題であった。テキストデータ化の際，”tidy data”という考えがとある分野に於いてさかんに議論されていたことを知った。近代語資料をテキストデータ化する際，どのようなデータが”tidy data”なのか，今後さまざまな分野との連携が必要であることは言うまでもない。このような方法についてはさまざまな立場があり，賛否両論あろうかと思う。しかし，近代語資料についてもテキストアナリシスによる可能性は，

「ある」のではないだろうか。検索性やさまざまな障害をクリアしたテキストファイルを作成する段階を越えれば，これまで明らかにされてこなかったデータが得られる可能性は高い。今後もテキストアナリシスによる調査のさらなる可能性について別稿を記す。

註

（1）このような調査方法の呼称について書籍や論文では「計量テキスト分析」「テキストマイニング」「テキストアナリティクス」「テキストアナリシス」といったことばが使われている。本稿では「テキストアナリシス」を使用する。

（2）『日語指南』に「課」という記述はない。便宜上「課」とした。

参考文献等

MeCab
http://taku910.github.io/mecab/
R
https://www.r-project.org
RMeCab
http://rmecab.jp/wiki/index.php?RMeCab
https://sites.google.com/site/rmecab/home
RMeCabUni
https://sites.google.com/site/rmecab/home/unidic
https://github.com/IshidaMotohiro/RMeCabUni
Wickham, H. (2014) Tidy data. Journal of Statistical Software, 59 (10)
*Wickham（2014）は下記にて閲覧可能。
http://dx.doi.org/10.18637/jss.v059.i10
伊藤孝行（2017）『近代日本語史に見る教育・人・ことばの交流　日本語を母語としない学習者向け教材と資料を通して』，大空社出版
金井保三（1904）『日語指南壹』
*金井（1904）は下記にて閲覧可能。
http://dl.ndl.go.jp/info:ndljp/pid/862070
金井保三（1905）『日語指南貳』
*金井（1905）は下記にて閲覧可能。
http://dl.ndl.go.jp/info:ndljp/pid/864024
小林雄一郎（2017）『Rによるやさしいテキストマイニング』，オーム社
伝康晴・山田篤・小椋秀樹・小磯花絵・小木曽智信（2008）「UniDic version 1.3.9 ユーザーズマニュアル」
https://www.gavo.t.u-tokyo.ac.jp/~mine/japanese/nlp+slp/UNIDIC_manual.pdf
諸星美智直（2000）「解説〔金井保三著：日本俗語文典〕」北原保雄・古田東朔編『日本語文法研究書大成 7』，勉誠出版
吉岡英幸（1994）「早稲田大学清国留学生部―そのカリキュラムと日本語教師―」，「講座日本語教育」29，pp.83-104
*吉岡（1994）は下記にて閲覧可能。
http://hdl.handle.net/2065/3279

吉岡英幸（2001）「金井保三著『日語指南』の文法学習項目」,「講座日本語教育」37, pp.14-26
*吉岡（2001）は下記にて閲覧可能。
http://hdl.handle.net/2065/3372