『万葉集』漢字本文における漢字の使用頻度集計

西山雄大
株式会社 Helpfeel
nishiyama@helpfeel.com

概要

『万葉集』の漢字本文[1)]を漢字の列として処理し、漢字の出現回数を集計して全巻および各巻での使用頻度を求めた。本文の訓読および解釈に伴う形態素解析による解釈の差異や情報の損失を伴わずに、万葉集テキストを定量的に解析するものである。これによって、万葉集テキストにおける漢字の使用状況を表記様式の特徴として記述することができるとともに、全 20 巻中の巻同士の比較から各巻の性格を記述できることを確認した。

1 はじめに

『万葉集』[2)]は現存する日本最古の歌集であり、奈良時代末に成立したとされる [1]。同じく奈良時代に成立した『古事記』や『日本書紀』などの史書が基本的に漢文で書かれているのに対し、万葉集には編纂までの間に詠まれた和歌が当時の日本語のまま収録されている。また、古今和歌集をはじめとする平安時代以降の歌集が漢字ひらがな交じり文であるのに対し、万葉集は全文が漢字で表記されている。特に漢字の音訓を借りて表音文字としたものは万葉仮名と呼ばれるが、一方で漢字を意味通り訓読させる表意文字としての表記法も存在しており、実際の本文の表記様式は仮名と訓字の両方を含んだ多様なものである [2]。

こうした特色から、万葉集は日本語の歴史研究において重要な文献資料と見なされてきた。数十点にわたる写本に見られる校註や注釈書に加え、江戸時代の国学者や明治時代以降の国語学者による上代日本語研究の歴史を有している。その中には本居宣長 [3] および石塚龍麿 [4] によってなされた万葉仮名の書き分けの収集・分類から、上代日本語の音韻論的な規則性を見出した橋本進吉 [5] の研究もあった。他方で、巻三・歌番号 9 番の歌に代表される、今なお訓みが定まらない難読歌も存在する[3)]。

1.1 訓み下し本文の問題点

言語の歴史的研究は残された文献資料に頼らざるを得ないが、情報化時代において言語研究のための用例を集めるには、それらを収集・整備したコーパスを利用するのが最も強力な方法である [6]。単語情報がタグづけされ、読み・品詞などの形態論情報が付与された本格的なコーパスを構築するためには、歴史的資料の形態素解析が必要となる。しかし古典語は形態素解析の自動化がしにくく [7]、このため古文を対象とした形態素解析は 2010 年代まで実現しなかった。小木曽智信ら [8] の報告によると、古文用の辞書データを作成することでコーパス構築に利用可能な精度の解析が可能となるが、歴史的な日本語資料の形態素解析において精度 95%を超えるには約 5 万語の学習用コーパスが必要であった。

2023 年現在、万葉集については国立国語研究所が『日本語歴史コーパス　奈良時代編 I 万葉集』[9] をすでに開発・公開している。しかし小木曽智信ら [10] が認める通り、コーパスとして形態論情報を付与し、かつ現代人に読みやすいものとするためには、漢字本文を校訂して漢字ひらがな交じりの訓み下し本文を用意する必要があるが、その際に原文の持つ情報は少なからず失われてしまう。日本語歴史コーパスにおいては、原文の前後文脈つきで検索結果を表示できる原文 KWIC 表示機能によってこの不足を補っている。とはいえ漢字本文における用字例を直接に収集・分類できない以上、特殊仮名遣に代表されるような用字法に関する洞察をコーパスの使用によって直接に得ることは難しいように思われる。

1） 現行の万葉集の本文は西本願寺本を底本とする校訂本文である。
2） 以下、単に万葉集とする。
3） 莫囂圓隣之大相七兄爪謁氣 吾瀬子之 射立為兼 五可新何本（最初 12 字に定訓なし）

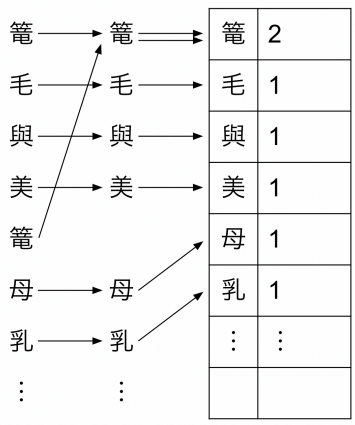

図 1　漢字本文における漢字の使用頻度集計の概念図

2　漢字本文における漢字の使用頻度

本稿では漢字ひらがな交じり文に訓み下した本文を形態素解析する代わりに、漢字本文を漢字の列として文字列処理した。ここでは、文字は前後の文字系列にかかわらず一定の確率で出現するものと仮定してある。漢字ひらがな交じり文を用意しないことで訓みや送り仮名といった解釈による差異はなくなり、漢字本文が持つ情報も失われることがない。さらに、単純な文字列処理なので形態素解析エンジンや学習用コーパスを用いる必要がなく、比較的簡単なスクリプトでプログラマブルに処理が可能である。

文字列処理の結果として、全巻および各巻について漢字ごとの使用頻度を集計した。万葉集は巻によって表記法が異なることが知られており、それらは形態素のレベルではなく用字のレベルで表現されていることから、各字種の使用頻度にも違いが見られることが期待される。なお、巻ごとの頻度には粗頻度ではなく、巻の総文字数で割った調整頻度を用いるべきだろう。また、全巻を通して使用頻度を概観することで、研究上注目すべき字種を抽出することも考えられる。たとえば、用例が極端に少ない漢字を拾い上げ、誤写誤読の可能性を検討したり、書記上の表現意図を問うことが定量的に可能となる。

準備　まず、万葉集に収録されている和歌の歌番号や収録巻、漢字本文などを含むテキストを CSV ファイル形式でデータ化した。校訂資料としては、利用に伴うライセンスの問題や利用しやすさの観点から、Wikisource で CC BY-SA 3.0 ライセンスのもとに提供されているテキスト [11] を本稿では利用した。

辞書型データの作成　つぎに、全 20 巻・総計 4516 首[4] の漢字本文について、漢字の用例とその使用回数を計上し、辞書型データに記録した。具体的には Python の辞書型オブジェクトを作成し、漢字一文字をキー、数値型オブジェクトを値として格納するよう準備した。それから漢字本文を一文字ごとに走査し、キーと値のペアが存在しなかった場合は新たに作成し、そのうえで値を加算した。この処理を全巻と各巻とについて行った。

使用頻度表の作成　つづいて、集計した漢字について使用頻度表を作成した。全巻、各巻について作成した辞書型データを値の降順でソートし、漢字ごとの粗頻度と調整頻度（%）を算出した。ここで粗頻度とは漢字の出現回数であり、調整頻度とは粗頻度を全巻または各巻の総文字数で割った値を百分率で表したものである。

3　結果

3.1　全巻を通しての漢字の使用頻度表

まず、全巻を通して使用の最も多い漢字上位 10 種と、それらの粗頻度・調整頻度を表 1 に示す。

使用回数は「之」が最も多かった。これは写本による校異を考慮しても疑いない結果と思われる。訓みは「し」（音仮名）と「の」（訓仮名）の少なくとも 2 通りあるが、使用頻度表からは用例を判別できない。

つづいて「尓」の使用も際立っている。万葉仮名としての訓みは「に」（音仮名）である。格助詞としての用法が多いものと推測される。

表 1　使用頻度表（全巻、上位 10 種）

漢字	粗頻度	調整頻度
之	4937	3.82
尓	4087	3.16
者	2414	1.87
乃	2391	1.85
乎	2131	1.65
毛	1778	1.38
可	1701	1.32
波	1696	1.31
奈	1620	1.25
能	1557	1.21

4）『国歌大観』の歌番号に従う。数の異同は底本の違いや、長歌（巻六・1010 および 1011）および「或本歌」「一云歌」の取り扱いの違いから生じる。

表 2　使用頻度表（各巻、上位 10 種）

巻一		巻二		巻三		巻四		巻五		巻六		巻七		巻八		巻九		巻十	
乃	3.73	之	5.25	之	3.76	之	3.78	能	3.29	之	3.90	之	5.25	尓	4.06	之	4.00	之	4.73
之	3.37	尓	3.12	尓	3.37	者	3.20	波	3.09	者	3.51	尓	2.59	之	3.94	尓	3.03	者	3.12
尓	2.87	者	2.81	乃	2.95	尓	2.72	尓	3.07	乃	3.45	者	2.58	者	2.76	者	2.73	尓	2.66
者	1.78	乃	2.43	者	2.81	吾	2.22	奈	2.59	尓	2.91	見	1.60	乃	2.71	乃	2.25	不	1.49
見	1.62	乎	1.70	乎	1.64	乎	2.20	良	2.54	見	1.64	不	1.51	毛	1.54	而	1.81	乎	1.30
山	1.62	而	1.58	見	1.57	不	2.16	等	2.49	乎	1.21	乎	1.47	花	1.46	不	1.45	乃	1.29
乎	1.59	見	1.47	不	1.54	有	1.78	可	2.43	山	1.19	吾	1.38	乎	1.41	乎	1.37	而	1.28
良	1.16	不	1.35	而	1.53	乃	1.76	多	2.35	而	1.19	山	1.24	有	1.41	見	1.28	吾	1.28
毛	1.12	有	1.27	有	1.51	毛	1.62	麻	2.31	有	1.17	而	1.23	而	1.35	毛	1.16	来	1.25
有	1.09	吾	1.24	山	1.21	而	1.55	久	2.21	不	1.15	将	1.14	吾	1.34	吾	1.09	有	1.16

巻十一		巻十二		巻十三		巻十四		巻十五		巻十六		巻十七		巻十八		巻十九		巻二十	
之	3.83	之	4.59	之	4.90	可	3.84	尓	3.71	之	3.87	尓	3.86	之	3.64	尓	4.78	波	4.12
者	2.45	者	3.32	者	3.03	尓	3.66	能	3.58	尓	3.69	之	3.58	尓	3.47	之	4.66	之	3.57
不	2.40	尓	2.63	尓	2.21	奈	3.62	安	3.17	乃	2.37	奈	3.12	能	3.46	乎	1.85	尓	3.57
吾	2.00	不	2.45	而	1.75	波	3.52	毛	3.09	者	1.94	多	2.87	奈	2.71	可	1.77	奈	3.18
尓	1.97	乎	1.88	不	1.70	能	3.03	之	3.05	乎	1.41	能	2.50	波	2.67	等	1.65	美	2.71
戀	1.82	毛	1.84	乃	1.64	麻	2.79	可	3.01	吾	1.35	麻	2.36	多	2.53	波	1.63	麻	2.68
乎	1.49	而	1.82	乎	1.51	乎	2.41	多	2.68	為	1.29	波	2.36	可	2.51	都	1.59	伎	2.55
見	1.48	戀	1.80	吾	1.38	安	2.40	麻	2.67	而	1.29	可	2.28	等	2.44	能	1.54	久	2.52
妹	1.28	吾	1.74	来	1.09	乃	2.37	伎	2.64	毛	1.23	久	2.17	安	2.26	奈	1.49	乃	2.48
乃	1.23	見	1.50	有	1.09	久	2.27	奈	2.64	来	1.20	安	2.11	久	2.11	毛	1.47	等	2.48

やや頻度を少なくして「者」「乃」「乎」「毛」が後に並ぶ。これらも典型的には「は」「の」「を」「も」などの格助詞として用いられる万葉仮名である。「者」は訓仮名、その他は音仮名である。

「可」は「か」と音仮名で用いられる一方で、助動詞「べし」としての用例もある。これも使用頻度表からは一概に判別できない。

3.2　各巻での漢字の使用頻度表

つづいて、巻ごとに使用の最も多い漢字 10 種と、それらの調整頻度を表 2 に示す。全巻の結果で上位 10 種にない漢字はセルを桃色でハイライトした。

全巻を通して得られた結果で見た傾向がここでも多くの巻で観察される一方で、巻によってそれらの比率が違っていたり、一部の巻については特徴的な漢字の使用を見て取ることができた。

例として「奈」という漢字を取り上げる。「な」（音仮名）を表すこの万葉仮名は、巻十四以降の巻数が古い巻において多く使用されていることがわかる。これらは万葉集の成立に近い比較的後期に編纂されたものと推定されている。それに対して巻数だけで見れば比較的若い巻五で第 4 位の使用頻度であることは注目に値する。この巻は前後の巻とは異なる文体の特徴が指摘されており、所収歌の年代や書き手の問題がこれまでにも論じられてきた[5)]。

5)　本稿と近い分類をしたものとして山田浩貴 [12] を挙げる。

4　まとめと今後の展望

本稿では万葉集の漢字本文を漢字の列として処理し、漢字の出現回数を集計して全巻および各巻での使用頻度を求めた。得られた漢字の使用頻度表は巻ごとに異なる表記体を反映していると考えられる一方で、実際の用例を確かめるには本文や訓みを参照しないといけなかった。

今回は各歌での集計にまで踏み込まなかったため、和歌の総文字数によって漢字の使用頻度を調整することはしなかった。しかし短歌・長歌・旋頭歌といった歌体によって和歌を分類し、巻間で同じ歌体同士の調整頻度を比較すれば、より精確に巻ごとの文体を論じることが可能であると考えられる。

さらに、文字は前後の文字系列にかかわらず一定の確率で出現するものと仮定したが、実際には特定の組み合わせで一定の単語を表すために使われるということがありうる。品詞という単位を考慮に入れるためには形態論情報を付加する必要があり、漢字本文を漢字の列として扱う手法は少なくとも部分的に修正しなくてはならない。

将来の展望として、統計的・確率論的手法を本文の処理に適用し、漢字同士の結びつきをモデル化する方法を探究するとともに、文字列としての扱いやすさを維持したまま訓みや品詞に関する情報を漢字本文に付加するアノテーション手法も検討したい。

参考文献

[1] 伊藤博．1974．万葉集の構造と成立．

[2] 佐野宏．2015．萬葉集における表記体と用字法について．国語国文:84(4)，161-181．

[3] 本居宣長．仮字の事．古事記伝，1．

[4] 石塚龍麿．1798．仮名遣奥山路．

[5] 橋本進吉．1917．国語仮名遣研究史の一発見：石塚達麿の仮名遣奥山路について．帝国文学．

[6] 講義「コーパスを使って日本語の歴史を探る」（小木曽智信）／言語学レクチャーシリーズ（試験版）Vol.19 - YouTube. https://www.youtube.com/watch?v=dEi0kKdyrWk （最終閲覧日：2023 年 1 月 10 日）

[7] 近藤泰弘．2009．古典語・古典文学研究における言語処理，pp.472-473．共立出版．

[8] 小木曽智信ら. 小木曽智信，小町守，松本裕治．2013．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:20(5)，727-748．

[9] 国立国語研究所．2017．日本語歴史コーパス　奈良時代編 I 万葉集． https://clrd.ninjal.ac.jp/chj/nara.html#manyo （最終閲覧日：2023 年 1 月 10 日）

[10] 小木曽智信，岡照晃，中村壮範，八木豊．2017．『日本語歴史コーパス』における原文 KWIC 表示機能の実装．言語資源活用ワークショップ発表論文集:2，252-257．

[11] 万葉集 - Wikisource． https://ja.wikisource.org/wiki/万葉集 （最終閲覧日：2023 年 1 月 4 日）

[12] 山田浩貴．2001．万葉集の付録的巻々：巻五と末四巻．北海道大学大学院文学研究科研究論集:1，21-41．