新聞記事における慣用表現の出現頻度調査

著者	北村 達也, 川村 よし子
雑誌名	甲南大学紀要. 知能情報学編
巻	10
号	1
ページ	25-33
発行年	2017-07-31
URL	http://doi.org/10.14990/00002882

論　文

新聞記事における慣用表現の出現頻度調査

北村達也[a], 川村よし子[b]

[a] 甲南大学 知能情報学部 知能情報学科
神戸市東灘区岡本８－９－１, 658-8501
[b] 東京国際大学 言語コミュニケーション学部 英語コミュニケーション学科
埼玉県川越市的場北１－１３－１, 350-1197

（受理日 2016 年 5 月 9 日）

概 要

日本語学習者を対象とした慣用表現の効率的な教育（学習）を実現するため，朝日新聞，毎日新聞，読売新聞の 2014 年の記事データベースを対象にして出現頻度の高い慣用表現のリストを作成した．まず，宮地（1982）『慣用句の意味と用法』，森田（2010）『日本語の慣用表現』に掲載された慣用表現に異表記を追加し，6,378 個の慣用表現を得た．次に，3 紙の合計で約 490 万文を対象として，慣用表現の出現頻度を計数した．そして，各紙における出現頻度上位 100 位のうち，3 紙すべてに共通して現れる慣用表現を求めた．これによって，新聞による偏りを排除し，使用頻度が高く学習者に有用な慣用表現リストが得られた．

キーワード： 日本語教育，慣用表現，新聞記事データベース，出現頻度，カバー率

1　はじめに

外国語の学習において慣用表現（慣用句，慣用語，イディオム）の習得は難しいと言われている [1]. ミン・佐藤 [2] は，日本語学習における慣用表現学習の困難さの原因として，(1) 単語個々の意味が分かっても慣用表現の意味は分からない，(2) 慣用表現は文化や社会習慣に基づく場合が多い，(3) どの慣用表現から学習したら良いか分からず，教材もない，の 3 点を挙げている．中国にて日本語を学んだ張 [3] は,（執筆当時の）現地の日本語教育の状況として，参考書や辞書の慣用表現に関する記述が不十分であり，誤りも少なくないと指摘している．

以上のように，慣用表現は日本語学習者にとって習得が難しい上にそれを学ぶための教材も十分ではない．しかも，学習時間は限られているため，効率的かつ効果的に慣用表現を教育／学習することが望まれている．そこで，本研究では，日本語教育における慣用表現の教育／学習方法設計のための基礎データを提供することを目的に，新聞記事における慣用表現の出現頻度を調査し，高頻度に使用されている慣用表現を明らかにする．

ミン・佐藤 [2] は，新聞記事および文学作品における慣用表現の使用頻度を調査している．彼らの

表 1: 各紙の記事データベースにおいて分析対象とした文の数

新聞	文の数
朝日新聞	1,803,290
読売新聞（東京版のみ）	1,723,305
毎日新聞	1,302,196

報告によると，2000 年の毎日新聞の記事における上位 3 位は,「軌道に乗る」,「手を出す」,「手を打つ」であり，新潮文庫の 67 作品における上位 3 位は「手を出す」,「恥をかく」,「クビになる」であった．

近年の計算機能力の向上により，上記の先行研究よりも大規模なデータを対象とした調査が容易になった．そこで，本研究では，新聞 3 紙の 1 年分の記事データベースを対象として，慣用表現の出現頻度ランキングを計測し，日本語教育における慣用表現の教育／学習方法設計のための基礎データを作成する．

2 方法

2.1 分析対象

データとしては，朝日新聞，読売新聞，毎日新聞の 2014 年 1 月から 12 月までの記事データベースを用いた．このデータベースには，日付，見出し，記事本文，記事内容分類等の情報が含まれている．この中から，句点または逆三角形（▼）で終了する日本語文字列を 1 文と定義し，分析対象にした．従って，見出し等の句点で終了しないものは対象にしなかった．形態素解析をより正確に行うため，記事において丸括弧を用いて記されているよみがなや年齢は削除した．例えば,「同地方で神馬藻（じんばそう）と呼ばれる縁起物の.....」という文の場合には，よみがなの部分を削除して「同地方で神馬藻と呼ばれる縁起物の.....」という文にした．

以上の処理を経て，各紙の記事データベースから抽出した文の数を表 1 に示す．読売新聞記事データベースには本社版に加えて地方版の記事も含まれていたため，東京版の記事のみを抽出した．本研究ではこれらの計約 490 万文を対象にした．

2.2 慣用表現リスト

慣用表現のリストは宮地 [4] および森田 [5] に基づいて作成した．これら 2 つの資料に掲載されている慣用表現はやや異なる傾向があるため，双方のリストを統合することによって広範囲の慣用表現をカバーできると考えた．

宮地 [4] の巻末（268 から 285 ページ）のリストと森田 [5] の巻末（331 から 362 ページ）の索引に掲載されている慣用表現を電子化し，慣用表現リストを作成した．慣用表現に含まれる感嘆符は削除

した．これらの書籍の本文には索引に掲載されない慣用表現が現れるが，それらは慣用表現リストに含まなかった．

また，記事の表記揺れに対応するため，得られた慣用表現の異表記を手作業で追加した．追加した異表記は以下のような種類である．

1. 常用漢字以外の文字のひらがな表記を追加（例：「遅蒔きながら」→「遅まきながら」を追加）
2. 難読漢字のひらがな表記を追加（例：「恰幅がいい」→「かっぷくがいい」を追加）
3. 漢字表記を追加（例：「あぐらをかく」→「胡坐をかく」を追加）
4. 漢字の異表記を追加（例：「沽券に関わる」→「估券に関わる」を追加）
5. 常用漢字音読表にない読みのひらがな表記を追加（例：「瓜田に履を納れず」→「瓜田に履をいれず」を追加）
6. 動植物名についてひらがな，カタカナ表記を追加（例：「独活の大木」→「うどの大木」,「ウドの大木」を追加）
7. 送り仮名に揺れのある物を追加（例：「有らん限りの力を尽くす」→「有らん限りの力を尽す」を追加）
8. 助詞が余分に入る表現を追加（例：「首の皮一枚つながっている」→「首の皮一枚でつながっている」を追加）
9. 漢字になる可能性のあるものを追加（例：「お題目を並べる」→「御題目を並べる」を追加）
10. 読点の付く可能性のあるものを追加（例：「壁に耳あり障子に目あり」→「壁に耳あり，障子に目あり」を追加）
11. 文語表現には口語表現を追加（例：「瓜の蔓に茄子はならぬ」→「瓜の蔓に茄子はならない」を追加）
12. 話し言葉で出てくる表現を追加（例：「鴨が葱を背負って来る」→「鴨が葱をしょってくる」を追加）
13. 話し言葉で出てくる縮約表現を追加（例：「鴨が葱を背負って来る」→「カモネギ」を追加）

以上のようにして得られた異表記を含む慣用表現の総数は 3,347 である．なお，佐藤 [6] は，複数の辞書間で慣用表現の表記が様々な形で揺れていることを報告しているが，上記の 13 種のルールによってほぼそれらを全て網羅したものと考えられる．

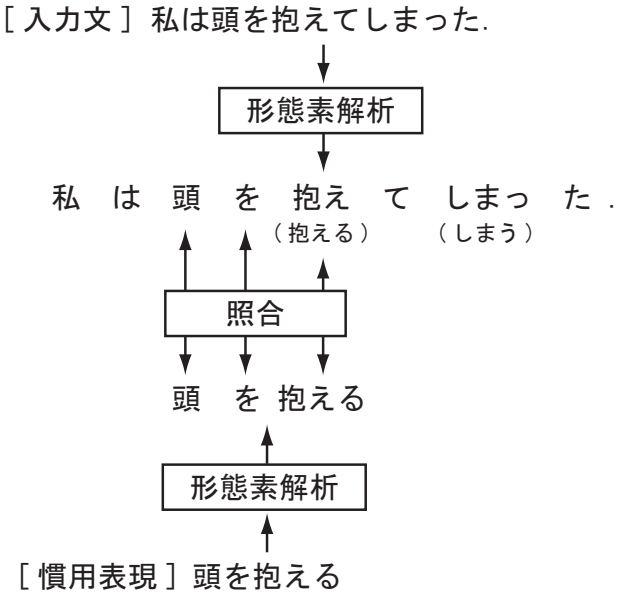

図 1: 入力文と慣用表現との照合

2.3 慣用表現の計数

3 紙の記事データベースから抽出した文を対象に，各慣用表現を照合し，出現頻度を求めた．照合の手続きを図 1 に示す．照合においては，まず，形態素解析システム MeCab バージョン 0.996 [7] により，文と慣用表現を形態素に分割した．次に，文の先頭の形態素（図 1 では「私」）から順に慣用表現の形態素と一致するか否かを調べた．なお，慣用表現の最後の形態素は文中で活用する場合があるため，その形態素に関しては形態素解析システムにより得られる基本形（図 1 の例では「抱える」）との照合を行った．そして，慣用表現の全ての形態素が一致した場合に当該の慣用表現が出現したと判定した．

なお，慣用表現リストに含まれる挨拶表現である「今日は（こんにちは）」は,「今日は（きょうは）」という文字列と区別して検出することが困難であったため，ひらがな表記の「こんにちは」のみを検出対象とした．また,「たとえ～であったにせよ」のように慣用表現の中に「～」が含まれるものについては，その部分に 1 個以上の形態素が入りうるものとした．

以上の方法で 3 紙の記事データベースに対して上述したリストの異表記を含む全ての慣用表現の出現頻度を求め，さらに，各紙それらの上位 100 位のうち 3 紙全てに共通して現れる慣用表現を求めた．これによって，新聞による偏りを排除して使用頻度の高い慣用表現のリストを得ることができると考えた．

3 結果と考察

2014 年の朝日新聞，毎日新聞，読売新聞の記事における慣用表現の出現頻度の上位 100 位に共通して現れたものを表 2 に示す．この表の慣用表現は，文の数が最も多かった朝日新聞（表 1 参照）における出現頻度順に並べてある．また，この表には，各紙における出現頻度およびその順位も示してある．この表には 73 個の慣用表現が含まれる．これは，つまり各紙の上位 100 位のうち，73 %の慣用表現が 3 紙に共通して高頻度に用いられていたことを意味している．このような高頻度の慣用表現を優先的に学ぶことによって，日本語学習者は新聞を読むために必要な語彙を効率的に身につけることができる．

この表において出現頻度が 355 回（365 日から休刊日 10 日分を除いた日数）を超えている慣用表現は，紙面に 1 日 1 回以上現れることになる．朝日新聞では上位 26 位まで，読売新聞では上位 28 位まで，毎日新聞では上位 20 位までがこのような慣用表現に該当した．

3 紙の 1,000 字あたりに出現する慣用表現ののべ数は，朝日新聞で 0.52 個，読売新聞で 0.51 個，毎日新聞で 0.53 個でとほぼ同数であった．これらは全ての慣用表現を対象にして求めた値である．朝日新聞のホームページ [8] によると，新聞 1 部の文字の量は約 179,000 字である．これは 1 部あたり 40 ページで計算された値である．この数字に基づくと，1 部あたりのべ約 90 個の慣用表現が現れ，1 ページあたり 2 個から 3 個の慣用表現が現れる計算となる．また，1,000 文あたりに出現する慣用表現ののべ数は，朝日新聞が 32.4 個，読売新聞で 32.5 個，毎日新聞で 32.4 個とほぼ同数であった．

図 2 に出現頻度の順位とカバー率の関係を示す．カバー率とは，新聞記事に現れた全ての慣用表現の何パーセントを占めるかを表す値である．この結果から，3 紙ともほぼ同じカーブを描いてカバー率が上昇することがわかる．上位 100 位で 60%前後のカバー率であり，上位 100 個の慣用表現を学んだだけでは，新聞記事に現れる約 6 割の慣用表現しか理解できないことがわかる．一方，上述したように，新聞における慣用表現の出現頻度自体は低い．そのため，学習者にとって慣用表現をただ覚えるだけでは，努力が新聞読解力の向上およびその実感に結びつきにくいことが予想される．そこで慣用表現を教える際には，単に出てきた慣用表現の意味を教えるのではなく，出てきた表現が慣用表現であることに気づく方法を教えるとともに，慣用表現の意味の類推方法を教えていく必要があると言えよう．

なお，本研究では宮地 [4]，森田 [5] の巻末の索引に掲載されている慣用表現を対象にしたが，これらの本の本文中には類似表現や派生表現も掲載されている．これらの表現を調査対象に加えることによって，出現頻度は増加すると考えられる．

4 おわりに

本研究では，朝日新聞，読売新聞，毎日新聞の 1 年分の記事における慣用表現の出現頻度を調査し，各紙の出現頻度ランキング上位 100 位までに共通して現れる慣用表現のリストを示した．3 紙の記事データベースに基づいた分析を行うことにより，従来の研究に比べて，より高精度の情報の提示が可能となった．今後は，この知見をより効率的で効果的な日本語教育の設計に役立てるとともに，日本語学習支援技術（例えば，渡邊・川村 [9]，北村 [10] など）に反映させていく計画である．

表 2: 2014 年の朝日新聞，毎日新聞，読売新聞の記事データベースにおける慣用表現出現頻度（回数）ランキング上位 100 位のうち，3 紙に共通して出現する慣用表現

順位	慣用表現	朝日新聞における出現頻度	読売新聞における出現頻度	毎日新聞における出現頻度
1	力を入れる	2,065 (1)	2,214 (1)	758 (6)
2	声をかける	1,940 (2)	1,743 (2)	666 (9)
3	気になる	1,267 (3)	996 (6)	801 (4)
4	気がする	1,194 (4)	1,010 (5)	930 (3)
5	身につける	1,058 (5)	786 (11)	407 (17)
6	声が上がる	919 (6)	1,399 (3)	934 (2)
7	口にする	912 (7)	901 (9)	1,090 (1)
8	求めて＊する	914 (8)	969 (7)	779 (5)
9	手にする	897 (9)	935 (8)	757 (7)
10	にもかかわらず	841 (10)	890 (10)	755 (7)
11	足を運ぶ	788 (11)	1,074 (4)	403 (18)
12	声を上げる	749 (12)	383 (26)	435 (14)
13	間に合う	744 (13)	751 (12)	492 (11)
14	ざるを得ない	644 (14)	398 (25)	376 (20)
15	耳を傾ける	634 (15)	656 (14)	380 (19)
16	気にする	623 (16)	6466 (21)	6422 (16)
17	気を付ける	588 (17)	687 (13)	289 (23)
18	役に立つ	531 (18)	504 (17)	267 (26)
19	役割を果たす	522 (19)	558 (16)	470 (12)
20	だけに＊だ	491 (20)	587 (15)	575 (10)
21	ものになる	464 (21)	216 (46)	243 (28)
22	気に入る	436 (22)	428 (23)	226 (30)
23	目にする	425 (23)	425 (24)	349 (21)
24	ことになっている	416 (24)	343 (31)	278 (24)
25	余儀なくされる	387 (25)	475 (19)	446 (13)
26	手に入れる	366 (26)	258 (39)	191 (34)
27	ではないが	345 (27)	354 (29)	310 (22)
28	顔を出す	335 (28)	290 (34)	124 (62)
29	ものにする	329 (29)	495 (18)	424 (15)
30	頭を下げる	319 (30)	364 (28)	246 (27)
31	口をそろえる	299 (31)	232 (43)	175 (42)
32	手を合わせる	292 (32)	378 (27)	242 (29)
33	やむを得ない	284 (33)	242 (41)	187 (38)
34	合わせて＊も	275 (34)	348 (30)	187 (38)
35	かなわない	271 (35)	238 (42)	181 (40)

順位	慣用表現	朝日新聞における出現頻度	読売新聞における出現頻度	毎日新聞における出現頻度
36	わけにはいかない	257 (36)	226 (45)	199 (33)
37	残念ながら	244 (38)	146 (75)	154 (49)
38	一つとして	244 (38)	288 (35)	174 (43)
39	目に入る	240 (40)	211 (48)	174 (43)
40	公算が大きい	235 (41)	282 (36)	172 (45)
41	耳にする	221 (42)	202 (50)	189 (36)
42	工夫をこらす	218 (43)	263 (38)	125 (61)
43	目を引く	217 (44)	274 (37)	160 (48)
44	あっという間に	216 (45)	205 (49)	164 (47)
45	命を落とす	214 (46)	172 (60)	154 (49)
46	ものはない	213 (47)	180 (55)	137 (55)
47	たまらない	211 (48)	171 (61)	139 (54)
48	間違いなく	210 (49)	186 (54)	176 (41)
49	歯止めをかける	208 (50)	298 (33)	222 (31)
50	仕方がない	197 (51)	153 (70)	136 (56)
51	昔ながらの	192 (53)	227 (44)	93 (80)
52	知恵を絞る	188 (54)	167 (62)	85 (89)
53	せざるを得ない	187 (55)	124 (90)	112 (73)
54	手を出す	180 (57)	164 (63)	120 (68)
55	顔をする	179 (58)	137 (78)	188 (37)
56	とどまらず	178 (59)	174 (59)	142 (53)
57	難色を示す	175 (60)	175 (58)	201 (32)
58	身を寄せる	173 (61)	162 (65)	135 (57)
59	なっていない	161 (64)	131 (82)	124 (62)
60	力が入る	157 (65)	155 (69)	122 (65)
61	幕を閉じる	146 (67)	157 (67)	150 (51)
62	肩を落とす	143 (68)	309 (32)	91 (34)
63	からには	143 (68)	127 (87)	93 (80)
64	が上がらない	136 (73)	132 (81)	78 (100)
65	目にとまる	132 (74)	138 (77)	80 (98)
66	脚光を浴びる	127 (79)	125 (89)	120 (68)
67	笑みを浮かべる	118 (84)	247 (40)	277 (25)
68	足並みをそろえる	118 (84)	145 (76)	150 (51)
69	磨きをかける	117 (87)	193 (53)	90 (83)
70	注目を浴びる	116 (88)	133 (79)	87 (87)
71	実を結ぶ	115 (90)	157 (67)	128 (60)
72	力を尽くす	107 (96)	113 (95)	90 (83)
73	波に乗る	107 (96)	129 (85)	120 (68)

各慣用表現の頻度には異表記も含まれている．順位は朝日新聞の記事データベースにおける結果に基づく．出現頻度欄のカッコ内の数字は当該紙における出現頻度の順位を表している．

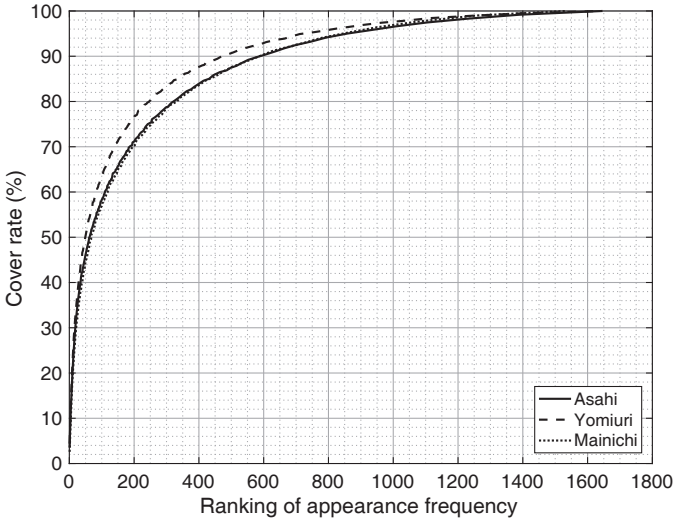

図 2: 2014 年の朝日新聞，読売新聞，毎日新聞の記事における慣用表現の出現順位とカバー率の関係

謝辞

本研究の一部は，平成 28 年度科学研究費（基盤研究（B）15H03219）の支援を得て行われた．

参考文献

[1] C. T. Cooper, “Teaching idioms,” *Foreign Language Annals*, vol. 31, no. 2, pp. 255–266, 1998.

[2] ダニー・ミン，佐藤洋, “日本語学習者のための慣用句データベースの作成：統計処理を用いた一手法の提案,” 情報処理学会研究報告, vol. 2001, no. 122, pp. 55–62, 2010.

[3] 張淑華, “中国人に分かり易い日本語の慣用句の記述について,” 信大国語教育, no. 6, pp. 16–26, 1996.

[4] 宮地裕編, 慣用句の意味と用法. 明治書院, 1982.

[5] 森田良行, 日本語の慣用表現. 東京堂出版, 2010.

[6] 佐藤理史, “基本慣用句五種対照表の作成,” 情報処理学会研究報告, vol. 35, no. 2007-NL-178, pp. 1–6, 2007.

[7] 工藤拓, MeCab 0.996, `http://taku910.github.io/mecab/` (2017 年 3 月 28 日閲覧).

[8] 朝日新聞社, 数字で見る朝日新聞, `http://www.asahi.com/shimbun/honsya/j/number.html` (2017 年 3 月 28 日閲覧).

[9] 渡邊飛雄馬，川村よし子, “やさしい日本語書き換えシステムの基本設計,” 日本語教育方法研究会誌, vol. 21, no. 1, pp. 48–49, 2013.

[10] 北村達也, “単語リストに基づく単語分類機能を持つテキストエディタ,” 日本語学, no. 8, pp. 80–87, 2016.