

新聞記事における慣用表現の出現頻度の調査

Appearance frequency of idiomatic phrases in newspaper articles

北村 達也+・川村 よし子++

KITAMURA Tatsuya+・KAWAMURA Yoshiko++

甲南大学知能情報学部+・東京国際大学言語コミュニケーション学部++

Faculty of Intelligence and Informatics, Konan University+

School of Language Communication, Tokyo International University++

〒658-8501 神戸市東灘区岡本 8-9-1+

t-kitamu@konan-u.ac.jp+

Abstract: This study aims to reveal how much time and effort Japanese language learners should devote to learning idiomatic phrases, and which idiomatic phrases should be learned preferentially. We measured the frequency of use of idiomatic phrases in numerous of sentences appearing in articles from the newspapers Asahi, Mainichi, and Yomiuri. Idiomatic phrases were extracted from “Nihongo no Kan'youhyougen Jiten” (2014), written by Yoshiyuki Morita, and notational variants of these phrases were added manually. The results showed that less than one idiomatic phrase occurred in every 1,000 characters of the newspaper articles. The 30 idiomatic phrases with the highest frequency were revealed in this study.

キーワード：慣用表現、新聞記事データベース、出現頻度、使用頻度

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。その理由として、ミン・佐野（2001）は、

（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点を指摘している。限られた学習時間の中で学習者が効率的に慣用表現を習得するためには、やみくもに暗記するのではなく、使用頻度の高い慣用表現を把握した上で学んで行くべきであろう。そこで、本研究では、新聞記事データベースを対象にして、慣用表現の出現頻度を調査する。

ミン・佐野（2001）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べている。それによると、前者における上位3位は「軌道に乗る」、「手を出す」、「手を打つ」であり、後者における上位3位は「手を出す」、「恥をかく」、「首になる」であった。

我々も「現代日本語書き言葉均等コーパス（BCCWJ）」（前川, 2008）の固定長タグセット（約1,000字のサンプルを1サンプルとしたデータセット）を対象にして慣用表現の出現頻度を調査した（Kitamura ら, 2014）。調査の結果、BCCWJ収録の新聞記事では約1,000字あたり0.65個の慣用表現が用いられていることがわかった。また、出現頻度の高い慣用表現は、

「気がする」、「手にする」、「身に付ける」などであった。この調査では、雑誌記事、書籍、白書のデータに関しても出現頻度を求め、傾向の違いを報告している。

本研究では、データの信頼性を向上させるため、より大規模なデータを用いて新聞記事における慣用表現の出現頻度を調査する。

2. 新聞記事データベース

2014年の朝日新聞、毎日新聞、読売新聞の記事データベースから句点で終了するものを一文と見なし、分析対象とした。なお、記事の中でカッコを使って記されているよみがなは分析対象から外している。各紙の文の数は以下の通りである。

表1：各新聞データベースの規模

新聞	文の数
朝日新聞	約181万文
毎日新聞	約130万文
読売新聞	約287万文

3. 慣用表現リスト

慣用表現リストは、森田良行著『日本語の慣用表現辞典』（東京堂出版, 2010）をもとに作成した。巻末の索引に掲載されている慣用句を電子化し、異表記については手作業で追加した。感嘆符は削除した。このようにして得られた慣用表現の総数は3,347である。

4. 慣用表現の検索

文中の慣用表現を検出するプログラムを開発し、出現頻度を調査した。

なお、慣用表現リストにある挨拶の「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できないため、平仮名表記の「こんにちは」のみを検出の対象とした。

また、行為の失敗を嘆く「しまった」ということば（原文では最後に感嘆符が付与されている）は、文字列パターンの照合において「～て（で）しまった」と区別することができない。これは、文脈や意味を考慮していないためである。そこで、本研究では、「しまった」の出現頻度から「～て（で）しまった」の出現頻度を減じた値を上記の意味合いの「しまった」の出現頻度とした。

この他、「たとえ～であったにせよ」のように、表現の間に任意の語が入りうる慣用表現においては、その部分に1語以上何語でも入れることができることとした。

5. 結果

2014年の3紙の記事に出現した慣用表現のうち、出現頻度が上位30位までのものを表1に示す。前報（Kitamuraら, 2014）とは慣用表現リストが異なるため、上位の慣用表現も異なっている。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、学習者は新聞を読むために必要な表現を効率的に身に付けることができる。

例えば、表1において出現数が365回を超えているものは、1日1回以上の確率で新聞紙面に現れることになる。表2では上位6位までがそのような慣用表現に該当する。

各紙の1,000字あたりに出現する慣用表現の数は、朝日新聞で0.23個、毎日新聞で0.25個、読売新聞で0.20個であった。

なお、本研究では森田（2010）の巻末の索引に掲載されている慣用句を対象としたが、この本の本文中で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。

6. おわりに

本研究では、新聞記事にて用いられる頻度の高い慣用表現、および慣用表現の出現頻度を調査した。今後、上記の類似表現や派生表現をリストに加えると同時に、新聞記事以外のテキストに対しても同様の調査を実施する計画である。

表2: 出現頻度の高い慣用表現上位30 (2014年3紙平均)

No.	慣用表現	出現数
1	求めて～する	1,075
2	にもかかわらず	960
3	だけに～だ	646
4	ざるを得ない	529
5	余儀なくされる	497
6	ではないが	390
7	合わせて～も	354
8	やむを得ない	289
9	かなわない	273
10	わけにはいかない	267
11	昔ながらの	245
11	あつという間に	227
13	間違いなく	214
14	ものはない	211
15	たまらない	205
16	残念ながら	201
17	駒を進める	198
18	仕方がない	193
19	とどまらず	192
20	一役買う	182
21	目を楽しませる	170
22	せざるを得ない	160
23	足並みをそろえる	158
24	なっていない	153
25	しょうがない	148
25	頭を抱える	148
27	脚光を浴びる	147
28	とんでもない	146
29	その場で～する	145
30	からには	144

謝辞 本研究の一部は、平成27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

- 前川喜久雄 (2008) 「KOTONOHA『現代日本語書き言葉均等コーパス』の開発」『日本語の研究』 Vol. 4, No. 1, pp. 82-95
- ダニー ミン・佐野洋 (2001) 「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(CE)』 Vol. 2001, No. 122, pp. 55-62
- 森田良行 (2010) 『日本語の慣用表現』東京堂出版
- Kitamura・Tanijiri・Kanazawa・Kawamura (2014) Appearance frequency of idiomatic phrase in fixed-length tag set of BCCWJ corpus, Sydney ICJLE2014