

『現代日本語書き言葉均衡コーパス』に対する MIPに基づく比喩表現情報の付与

加藤祥¹ 菊地礼² 浅原正幸³¹目白大学 ²中央大学 ³国立国語研究所

s.kato@mejiro.ac.jp

概要

日本語比喩表現の実態調査を目指し、『現代日本語書き言葉均衡コーパス』に対し、手作業で比喩表現情報付与を行った。作業対象は均衡性のある新聞・書籍・雑誌の34万語とし、言語対照のためにMIP及びMIP VUに基づく比喩表現情報付与を進めた。中村(1977)の基準を用いて、結合や種別を付与したほか、一般的な読み手による印象評定情報も付与した。本稿は、作業と作業結果の基礎統計を報告する。また、BNCにMIPに基づく情報付与を行ったVU Amsterdam Metaphor Corpusとの対照についても報告する。

1 はじめに

現在までのところ、日本語の比喩表現に関しては、均衡性のあるデータに基づく実態的な調査が大規模に行われているとは言い難い。主に文学作品を用いた日本語比喩表現の収集は、中村(1995)[1]の『比喩表現辞典』の8,201例をはじめ、レトリックに関する事典やデータベースなどに見られるが、均衡性を有した大規模な収集は行われていないためである。

比喩表現の同定手法は、選択選好(Willksら, 2013[2]), 具象度(Turneyら, 2011[3]), 意味空間モデルの使用(Maoら, 2018[4], Liu, 2020[5])などが考案されているものの、大規模な日本語比喩表現を収集、調査した例としては、現在まで、人手による中村(1977)[6]が最も網羅的であると考えられる。同書は、現代日本語との対照や検証の求められるデータであるが、文学作品から収集した用例に基づく日本語比喩表現の収集手順が整備された実態調査資料である。主に近現代の文学作品50編から比喩的な表現を含むと判定した2万の用例を収集し、比喩表現を構成する要素の結合(慣用から意味的なずれのある場合)について5,537類型(B型把握)を整理したほか、

何らかの言語形式として比喩の把握に影響していた359の比喩指標要素(A型把握)、場面や文脈との関わりにおいて比喩となる2,434の文脈比喩類型(C型把握)を示している。本書で設定された手順と類型データは、日本語比喩表現用例の収集に有効と考えられる。

しかし、言語対照の可能性のためには、比喩情報付与の基準として提唱されたMIP(The Pragglejaz Group(2007)[7])に基づくVU Amsterdam Metaphor CorpusのMIP VUを採用し、日本語における比喩情報の付与手順を整備することが有用であろう。MIPによる日本語比喩表現の自動検出を目標とした宮澤ら(2016)[8]の試みでは、新聞記事に含まれる1,100文の動詞と目的語について2名の作業員により試行した際、VU Amsterdam Metaphor Corpusの結果よりも判定揺れが生じたと報告される。よって、日本語比喩表現の実態調査を目指すためには、判定基準を整備することはもちろん、複数の作業員の判定揺れの確認のみならず、一般的な読者に比喩と認識されるのかという判定なども必要と考えられる。

そこで、日本語比喩表現の実態調査を目指し、情報付与の手順を整備するとともに、均衡性のある日本語コーパスに対し網羅的な比喩表現情報の付与を行うことを目指した。具体的には、『現代日本語書き言葉均衡コーパス(以降BCCWJ)』の分類語彙表番号が付与された新聞(以降PN)・雑誌(PM)・書籍(PB)サンプル(347,094語, 以降BCCWJ-WLSP, 加藤ら2019a[9]b[10])を作業対象とし、MIP VUの手順に基づいて、中村(1977)[6]の基準や類型を用いた情報付与を行う。BCCWJ-WLSPは、文脈上の語義が人手で付与されているため、本比喩性の判断作業に利用することができる。さらに、一般的な読み手による判定情報をクラウドソーシングにより大規模に収集する。

日本語比喩表現の実態調査が可能となれば、VU Amsterdam Metaphor Corpusなどにおける英語メタフ

アナーなどの対照や、比喩に関する理論検証が可能となることが期待される。

2 比喩情報付与作業

2.1 比喩情報付与手順

VU Amsterdam Metaphor Corpus は、BNC-Baby コーパスから抽出した4レジスタ約19万語について、Steenら(2010) [11]がThe Pragglejaz Group(2007)の比喩情報付与手順であるMIP (Metaphor Identification Procedure) を拡張したMIP VUによってメタファーの判定 (MRW; Metaphor Related Word) を付与したものである。また、Seminoら(2005) [12]が特定コーパスで示したメタファー類型 (combining と mixing) や、MIPからの拡張として付与されたMFlag (指標)、概念マッピング情報 (擬人化) を用いた比喩の種別分類も試みられている。

よって、本稿の作業は、MIPを用いて比喩性に関わると判定された短単位を抽出し、VU Amsterdam Metaphor Corpusを参照したMRWとしてマークすることにより進めることとした。BCCWJ-WLSPは、作業対象範囲の全ての自立語には分類語彙表番号、助動詞には用法が、人手で文脈上該当する意味において付与されており、選択制限違反の判断時に参照可能である。このほかの参照資料及び付与する日本語の比喩情報としては、中村(1977) [6]の指標、比喩種別、結合等を援用することとしたほか、動詞の選択制限確認には宮島(1972) [13]の格情報、語義及び歴史的な順序の確認に際しては『日本国語大辞典』を用いた。

なお、本手順は、加藤ら(2020) [14]の指標比喩データベースを構築した際にも同様に比喩表現の認定を手作業で行うことにより、実作業による微調整を加えながら整備を進めたものであり、概要は以下の通りであるⁱⁱ。

- ① 当該形態素が文脈上、比喩表現に関わるかどうか判定する。比喩表現に関わると判断された場合は、当該形態素をマークする。

- ② 比喩表現の喩辞と被喩辞を抽出するとともに、把握の根拠として、
 - A型：比喩表現の認識に関わった「何らかの言語形式」に該当する短単位の範囲に「MFlag (指標)」を付与する。
 - B型：異質性を読んだ「結合」 (選択制限違反ⁱⁱⁱ) が確認できれば「結合」を記述する。結合または異質な表現の該当部には、比喩表現の種別として、結合 (何らかの類似性に基づく転換、隠喩) ・換喩 (事物・事象の隣接性という類縁関係に基づく質的転換) ・提喩 (類と種という類的転換) および結合比喩の下位分類 (擬人化・具象化など^{iv}) を付与する。
- ③ 「結合」の構成要素に分類語彙表番号を付与する。(意味上は別の分類番号が付与されているため、選択制限違反となる番号を再付与する)
- ④ 用例 (文脈の把握が可能な2~3文程度まで) および個別の「結合」に一般の日本語母語話者の判定を付与する。

2.2 付与した比喩情報例

BCCWJの対象範囲に付与した情報は以下の通りである。一部に付与漏れや作業中の部分が残るが、概ね情報付与が終了している (2021年12月現在)。

- 比喩を構成する要素の結合類型と分類語彙表番号：手順②のB型および③
- 種別情報 (擬人化・具象化などの隠喩 (類似性の転換) における概念マッピング、換喩、提喩など)：手順②のB型
- 指標となった言語形式、その他比喩に関わると考えられる語句：手順①および②のA型
- 比喩度 (クラウドソーシングによる数十名規模の被験者実験による判定結果 (加藤・浅原別発表[15]) として、0~5の6段階の平均値を示す)：手順④、以下に付与した評定値を示す。
(自) 自然さ

ⁱ ジャパンナレッジ (<https://japanknowledge.com/>) 版。

ⁱⁱ MIP VUを援用した作業手順の詳細は、加藤ら(2020)の3.1節 (図3およびpp.861-862)。

ⁱⁱⁱ 選択制限違反の判断に際しては、参照可能な語義情報を活用する (詳細は加藤ら(2020)の3.1節 (pp.861-862))。

^{iv} 付与した種別の詳細は、加藤ら(2020)の3.5節内「種別」 (p.868)。

- (わ) わかりやすさ
- (古) 古さ
- (新) 新しさ
- (比) 比喩性

情報付与結果例を以下に示す。例文中の【 】内はMRWとしてマークされている部分を示す。

[]内は分類語彙表番号であり、比喩性の判定された要素の結合には、文脈上のBCCWJ-WLSPと異なる場合、再付与した番号(類型化のために固有名詞等は概念化も行った)を記述してある。

- (1) 今年から家の場所や危険個所の【確認に重点を置く】。(PN1a_00002)
[1.3062に1.3070を2.1513] (具象化)
(自) 4.25 (わ) 4.05 (古) 2.30 (新) 1.65 (比) 1.75
- (2) こんな英語のことわざが書かれた【額に出会った】のです。(PB11_00013)
[1.4570に2.1550] (擬人化)
(自) 2.35 (わ) 2.20 (古) 1.75 (新) 1.00 (比) 1.30
- (3) 急速な変化に【戸惑う声もある】。(PN1a_00005)
[2.3067 1.3061/1.3061が2.1200] (換喩)
(自) 4.00 (わ) 3.55 (古) 2.70 (新) 1.50 (比) 2.35
- (4) 【メーカー】は(中略)容器を【作る】ことになるだろう。(PM23_00007)
[1.2450が2.3860] (換喩)
(自) 3.75 (わ) 3.50 (古) 1.80 (新) 2.05 (比) 0.65
- (5) 浦和レッズ—柏レイソルのサテライト(二軍)【戦】(PN2c_00002)(提喩)
(自) 3.60 (わ) 3.15 (古) 1.65 (新) 2.35 (比) 1.25
- (6) 大きな【ギャップが生まれ】ている(PN2c_00006)
[1.1830が2.5710] (擬生)
(自) 3.85 (わ) 3.80 (古) 1.9 (新) 2.20 (比) 1.95
- (7) 【脚が棒になる】まで歩き回ります(PM51_00077)
[1.5603が1.4150に2.1500] (転換・慣用)
(自) 4.35 (わ) 4.50 (古) 3.30 (新) 0.65 (比) 4.45

3 結果

3.1 比喩情報付与結果

比喩情報付与結果(データ画面例は付録の図1を参照)を示す。表1では、レジスタ別に語数および、結合(選択制限違反)数と比喩性に関わるとマークされた語(MRW)数、指標(MFlag)数を示している。なお、表1は結合(選択制限違反)数を示したが、作業者が選択制限違反を認めなかった場合(換喩や提喩などは比喩性の判断が作業者によって揺れる場合もある)や結合の取得しにくい指標比喩などの場合、結合が取得されていないことがある。また、複数の種別が記載されている場合もあるため、比喩表現数は結合数と一致するものではない。

表1 比喩情報付与結果の統計

レジスタ	語数	結合数	MRWs	M-Flags
PN(新聞)	111,754	5,018	9,074	183
PB(書籍)	111,983	1,789	7,326	54
PM(雑誌)	117,568	3,176	10,577	100
計	341,305	9,983	26,977	337

新聞(PN)において結合数が最も多いが、MRWとしてマークされた語数は雑誌(PM)が最も多い。書籍(PB)では結合とMRWのどちらも少なく、比喩表現が少ない傾向が確認される。また、比喩指標として抽出された要素は結合の確認された比喩表現における3%程度と僅かであり、いわゆる直喩の出現頻度は比喩表現中においても低い割合であることがわかる。

表2には、レジスタ別の比喩表現種別を示す。

表2 レジスタ別の比喩表現種別の統計

	PN(新聞)	PB(書籍)	PM(雑誌)	計
類似性(隠喩)	2,705	888	1,501	5,094
換喩	1,581	284	471	2,336
慣用	887	299	263	1,449
文脈	71	359	938	1,368
提喩	248	55	67	370
その他	6	3	2	11
計	5,498	1,888	3,242	10,628

表 2 では、1 つの結合に対し「文脈・慣用」などのように複数種別が記載されていた場合、2 つ目までを取得して集計した。新聞は換喩や提喩が他レジスタよりも多い傾向があり、いわゆる隠喩（類似性に基づく転換）も最も多く取得されており、比喩表現の取得される頻度が高い。但し、文脈比喩は少ない傾向にある。雑誌はいわゆる隠喩の頻度が高いが、慣用表現はやや低く、文脈比喩が多く見られる傾向が確認されている。書籍は、種別の分布としては雑誌に似るものの、雑誌との違いとして、慣用表現の割合がやや高い傾向がある。

また、類似性に基づく転換（いわゆるメタファー）の下位種別を表 3 に示す。新聞から取得されるメタファー表現の 7 割が具象化であるため、一般的に比喩表現とは意識されにくいと考えられる。新聞は、慣用表現の出現頻度も高い（表 2 参照）ことから、いわゆる文章技巧的な比喩表現が出現しにくいことが、新聞^vの特徴と考えられる。反対に、書籍では擬人化のようにわかりやすい比喩表現の頻度が高いともいえる。しかし、一般的な読み手による評定値では、(3)や(7)のような慣用的な用例において自然さやわかりやすさとともに比喩性が高く、(4)のような換喩用例において比喩性が低く評定される傾向が見られる。(2)や(6)のような擬人（生）化の用例は比喩性の高さが期待されるが、(2)のようにわかりやすさの低い例では、比喩性も低く評定される傾向がある。(1)のような具象化や(5)のような提喩の例からも、わかりやすさが比喩性に影響する可能性が考えられる。

表 3 種別「類似性（隠喩）」の下位種別の統計

	PN（新聞）	PB（書籍）	PM（雑誌）	計
具象	1,876	482	586	2,944
その他の転換	441	227	613	1,281
擬人	214	111	189	514
擬物	84	25	63	172
擬生	90	43	50	183
計	2,705	888	1,501	5,094

^v 但し、BCCWJ の PN については、記事情報が付与されており（加藤ら 2021）、新聞においても記事の内容や種類によって比喩表現の分布には差異が見られる。

3.2 英語コーパスとの対照

Steen ら(2010)の BNC-baby に情報付与した VU Amsterdam Metaphor Corpus と対照することが可能である。表 4 に、BNC-baby^{vi}と BCCWJ^{vii}の付与結果をあわせて示す。なお、Steen ら(2010)は、borderline case を‘WIDLII’と分けて示している。ここでは、BCCWJ の付与結果についても、作業員間で合意のなかった比喩表現に関わる短単位（1 名または 2 名のみが MRW としてマークした短単位）について、WIDLII に分類してある。

表 4 英語コーパスと本稿コーパス

	NON-MRWs	WIDLII	MRWs	MFlags	Total
BNC-baby	161,105	1,831	23,611	141	186,688
	86.3%	1.0%	12.6%	0.1%	100.0%
BCCWJ	306,586	7,405	26,977	337	341,305
	89.8%	2.2%	7.9%	0.1%	100.0%

M-Flag（指標）は、英語コーパスと同程度の割合となっていたが、MRW としてマークされた語数には差が見られる。本作業は短単位へのマークであることや、日英語の品詞差により、単純に対照することはできないが、作業員間の揺れや作業漏れの可能性も考えられるため、付与作業漏れの整備とともに、今後確認を進める予定である。

4 おわりに

BCCWJ の約 34 万語に比喩表現情報の付与を行った。MIP に基づいた比喩表現抽出を行うことで、他言語の比喩表現情報付与コーパスとの対照が可能である。作業対象には、文脈情報として NDC の補助分類や新聞記事内容や種別の情報が付与されているため、種類の異なるコーパスとも、該当部分の対照が可能となっている。

また、比喩表現に関する情報として、種別や印象評定等を付与しており、日本語比喩表現の実態調査に有用なデータと期待される。

^{vi} Steen ら(2010)の Table10.1 から該当列を転載した。

^{vii} 表 1 を Steen ら(2010)の Table10.1 に対応させたもの。

謝辞

本研究は JSPS 科研費 JP18K18519 の助成を受けたものです。

参考文献

1. 中村明.『比喩表現辞典』, 1995 年, 角川学芸出版.
2. Willks, Yorick, Galescu, Lucian, Allen, James, and Adam Dalton. "Automatic Metaphor Detection using Large-Scale Lexical Resources and Conventional Metaphor Extraction," Proceedings of the Workshop on Metaphor in NLP, 2013, 36-44.
3. Turney, Peter D., Neuman, Yair, Assaf, Dan, and Yohai Cohen. "Literal and Metaphorical Sense Identification through Concrete and Abstract Context," Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, 680-690.
4. Mao, Rui, Lin, Chenghua, and Guerin, Frank, "Word Embedding and WordNet Based Metaphor Identification and Interpretation", Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, 1222-1231.
5. Liu, Jerry, O'Hara, Nathan, Rubin, Alexander, Draelos, Rachel, and Rudin, Cynthia. "Metaphor Detection Using Contextual Word Embeddings From Transformers", Proceedings of the Second Workshop on Figurative Language Processing, 2020, 250-255.
6. 中村明.『比喩表現の理論と分類』国立国語研究所報告 57, 1977 年, 秀英出版.
7. Pragglejaz Group. "MIP: A Method for Identifying Metaphorically Used Words in Discourse," Metaphor and Symbol 22, 1, 2007, 1-39.
8. 宮澤彬, 吉田奈央, 宮尾祐介. 「日本語メタファーコーパス作成のためのガイドライン」『言語処理学会第 22 回年次大会発表論文集』, 2016 年, 150-153.
9. 加藤祥, 浅原正幸, 山崎誠. 「『現代日本語書き言葉均衡コーパス』新聞・書籍・雑誌データの助動詞に対する用法情報付与」『日本語学会 2019 年度春季大会予稿集』, 2019 年 a, 161-166.
10. 加藤祥, 浅原正幸, 山崎誠. 「分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の新聞・書籍・雑誌データ」『日本語の研究』2019 年 b, 15(2): 134-141.
11. Steen, Gerard J., Dorst, Aletta G., Herrmann, Berenike, Kaal, Anna, Krennmayr, Tina, and Trijntje Pasma. A Method for Linguistic Metaphor Identification; From MIP to MIPVU, John Benjamins Publishing Company, 2010.
12. Semino, Elena. The Metaphorical Construction of Complex Domains: The Case of Speech Activity in English, Metaphor and Symbol 20, 1, 2005, 35-69.
13. 宮島達夫.『動詞の意味・用法の記述的研究』国立国語研究所報告 43, 1972 年, 秀英出版.
14. 加藤祥, 菊地礼, 浅原正幸. 「『現代日本語書き言葉均衡コーパス』に基づく指標比喩データベース」『自然言語処理』2020 年, 27(4):853-887.
15. 加藤祥, 浅原正幸. 「『現代日本語書き言葉均衡コーパス』に対する印象評定情報付与」『言語処理学会第 28 回年次大会発表論文集』, 2022 年, 発表予定.

付録

サンプル	開始位置	短単位語彙素	短単位書字形	MRW	品詞	分類語彙表番号	結合種別	結合	分類語彙表番号(結合)
PN5b_00007	9440	今季	今季	1	名詞-普通名詞-副詞可能	1.1641			
PN5b_00007	9460	に	に	1	助詞-格助詞				
PN5b_00007	9470	入る	入っ	1	動詞-一般	2.1532	具象	「今季に入る」	1.1641 に 2.1535
PN5b_00007	9490	て	て		助詞-接続助詞				
PN5b_00007	9500	漸く	ようやく		副詞	3.1670			
PN5b_00007	9540	プロ	プロ		名詞-普通名詞-一般	1.2340			
PN5b_00007	9560	初	初		名詞-普通名詞-一般	1.1650			
PN5b_00007	9570	勝利	勝利	1	名詞-普通名詞-サ変可能	1.3570			
PN5b_00007	9590	を	を	1	助詞-格助詞				
PN5b_00007	9600	上げる	挙げ	1	動詞-非自立可能	2.3700	具象	「勝利を挙げた」	1.3570 を 2.1540
PN5b_00007	9620	た	た		助動詞				
PN5b_00007	9630	二	2	2	名詞-数詞	1.1960			
PN5b_00007	9640	七	7	7	名詞-数詞	1.1960			
PN5b_00007	9650	歳	歳		接尾辞-名詞的-助数詞	1.1962			
PN5b_00007	9660	は	は		助詞-係助詞				
PN5b_00007	9670	「	「		補助記号-括弧開				
PN5b_00007	9680	彼処	あそこ	1	代名詞	1.1700			
PN5b_00007	9710	を	を	1	助詞-格助詞				
PN5b_00007	9720	乗り切る	乗り切れる	1	動詞-一般	2.1524	具象	「あそこを乗り切れる」	1.1700 を 2.1524
PN5b_00007	9770	様	よう		形状詞-助動詞語幹	3.1300			
PN5b_00007	9790	だ	に		助動詞				
PN5b_00007	9800	成る	なれ		動詞-非自立可能	2.1112			
PN5b_00007	9820	ば	ば		助詞-接続助詞				
PN5b_00007	9830	…	…		補助記号-一般				
PN5b_00007	9840	」	」		補助記号-括弧閉				
PN5b_00007	9850	と	と		助詞-格助詞				
PN5b_00007	9860	課題	課題	1	名詞-普通名詞-一般	1.3070			
PN5b_00007	9880	を	を	1	助詞-格助詞				
PN5b_00007	9890	見据える	見据える	1	動詞-一般	2.3091	具象	「課題を見据える」	1.3070 を 2.3091
PN5b_00007	9930	。	。		補助記号-句点				
PN5b_00007	9940				空白				
PN5b_00007	9950	イトウ	伊東		名詞-固有名詞-人名-姓	1.2390			
PN5b_00007	9970	監督	監督		名詞-普通名詞-サ変可能	1.2450			
PN5b_00007	9990	は	は		助詞-係助詞				
PN5b_00007	10000	「	「		補助記号-括弧開				
PN5b_00007	10010	ゲーム	ゲーム	1	名詞-普通名詞-一般	1.3374	換喩	「ゲームを作れる」	1.3374 を 2.3801
PN5b_00007	10040	を	を	1	助詞-格助詞				
PN5b_00007	10050	作る	作れる	1	動詞-一般	2.1220			
PN5b_00007	10080	投手	投手		名詞-普通名詞-一般	1.2450			
PN5b_00007	10100	。	。		補助記号-句点				
PN5b_00007	10110	良い	いい		形容詞-非自立可能	3.1332			
PN5b_00007	10130	仕事	仕事	1	名詞-普通名詞-サ変可能	1.3320	提喩		1.3320 を 2.3430
PN5b_00007	10150	を	を	1	助詞-格助詞				
PN5b_00007	10160	為る	し	1	動詞-非自立可能	2.3430			
PN5b_00007	10170	て	て		助詞-接続助詞				
PN5b_00007	10180	呉れる	くれる		動詞-非自立可能	2.3770			
PN5b_00007	10210	」	」		補助記号-括弧閉				
PN5b_00007	10220	と	と		助詞-格助詞				
PN5b_00007	10230	褒める	ほめ		動詞-一般	2.3682			
PN5b_00007	10250	、	、		補助記号-読点				

図1 情報付与結果の例(作業画面)