「手」の慣用句を指標とした文章ジャンルの判別：現代日本語書き言葉均衡コーパスを用いて

「手」の慣用句を指標とした文章ジャンルの判別

—現代日本語書き言葉均衡コーパスを用いて—

村田　年（慶應義塾大学日本語・日本文化教育センター）
山崎　誠（国立国語研究所）

1．は じ め に

専門分野における学習・研究を目標とする日本語学習者が中・上級レベルに到達し，その後どのように学習を進めるかということについて考える時，文章のジャンル[1)]は一つの重要な視点であると言えよう。日本語学習者が文章ジャンルの違いをより一層意識化し，各ジャンルにおいて特徴的な表現を学んでいくことは中・上級レベル以降の学習の効率化につながると考えられる。

村田 (2007) では論述的な文章に特徴的な接続語句と助詞相当句を選択し，ジャンルによって異なる文章の特徴がそれらの語句の使用傾向の違いに反映されることを実証した。また村田 (2008) では文章ジャンルを判別するための指標として複合動詞の後項動詞を取り上げ，その使用傾向が文章ジャンルによって異なることを明らかにした。本稿では，文章のジャンルを判別するための新たな指標の可能性として慣用句を取り上げたい。従来は個人レベルで収集した文章資料の範囲で分析を行ってきたが，今回は，現代日本語書き言葉均衡コーパスの使用が可能となったので，それを対象資料とする。分析のための指標としてすべての慣用句を取り上げることは数の多さから難しいので，まず「手」を含む動詞慣用句と形容詞慣用句に絞って調査を行った。

2.　分析に用いた文章資料

2.1　『現代日本語書き言葉均衡コーパス』

調査に用いた資料は,『『現代日本語書き言葉均衡コーパス』モニター公開データ（2009 年度版）』である。このデータは，国立国語研究所を中心に構築している大規模コーパス『現代日本語書き言葉均衡コーパス』の一部で，2009 年 7 月現在，著作権者から許諾が得られたデータ約 4,300 万語を収録している。データの内訳は表 1 のとおりである。この調査では，そのうち書籍に該当するデータ（表 1 の (1)〜(3)）を使用した。書籍のデータは暫定的なものであり，最終的にはデータ量は約 2 倍以上になる。『現代日本語書き言葉均衡コーパス』については, 前川 (2008) 同 (2009) を，この調査で使用した書籍の部分については山崎 (2009) を参照されたい。

表 1　モニター公開データ（2009 年度版）の内訳

データ名	語数
(1) 出版サブコーパス・書籍	1,300 万語
(2) 図書館サブコーパス・書籍	1,500 万語
(3) 特定目的サブコーパス・ベストセラー	230 万語
(4) 特定目的サブコーパス・白書	480 万語
(5) 特定目的サブコーパス・Yahoo!知恵袋	520 万語
(6) 特定目的サブコーパス・国会会議録	490 万語

語数については注の 2）参照.

2.2　文章資料データの母集団

この調査で使用した書籍のデータは，3 種類のデータから成っている。同じ書籍であるが，それぞれ母集団が異なる。1 つ目は，「出版サブコーパス・書籍」で，2001 年～2005 年に日本国内で出版された全書籍からコーパス収録条件[3)]で絞り込んだ約 32 万冊が母集団である。2 つ目は，「図書館サブコーパス・書籍」で，東京都内の自治体ごとに ISBN により管理されている蔵書データを利用して，13 自治体以上に共通して所蔵されている書籍からコーパス収録条件で絞り込んだ約 38 万冊が母集団である。ここでは，ISBN の付与が普及した 1986 年～2005 年刊行の書籍を対象とした。3 つ目は，1976 年～2005 年版の『出版年鑑』及び『出版指標年報』に掲載されたベストセラーリスト上位 20 位に挙げられた書籍約 950 冊を対象としている。

2.3　文章資料サンプルの抽出方法

書籍の 1 つのサンプルは 1 万字を超えない範囲のひとまとまりの文章（1 章，1 節など外的な構成を基準として判断したまとまり）である。母集団の全ページをサンプリングフレームと考え，そこからランダムに抽出したページの中の 1 点をランダムに指定し，その点を基準にテキストを抽出している。

2.4　文章資料のジャンル

書籍については以下の 11 に分類され，語数は以下の通りである。

表 2　書籍内の分類

分類	語数
0 総記	642,351
1 哲学	1,799,241
2 歴史	2,674,253
3 社会科学	6,854,656

4 自然科学	1,631,154
5 技術・工学	1,363,963
6 産業	878,954
7 芸術・美術	1,311,554
8 言語	492,255
9 文学	11,267,012
分類なし	478,536

本稿で分析対象とする文章資料は，書籍の中ではジャンルが明示的な 1 哲学から 9 文学までとし，0 総記と分類なしは除くことにする。また，1 から 9 までを大きく 3 つのジャンルに分け，それぞれ「人文科学系」「社会科学系」「自然科学系」として以下のようにまとめた。

人文科学系：哲学，歴史，芸術・美術，言語，文学

社会科学系：社会科学，産業

自然科学系：自然科学，技術・工学

「歴史」を社会科学系に入れる考え方もあるが，専門日本語教育の観点から，大学での学科編成も考慮に入れて本稿では人文科学系に分類した。また，「芸術・美術」についても書籍の内容に芸術理論，美学，美術史，音楽学などが含まれているため，人文科学系に分類した。

３．分析

3.1　指標としての「手」の慣用句の抽出方法

この調査で対象とした慣用句は，『基本慣用句五種対照表』（佐藤理史編 2007：以下，「対照表」と略す）に基づいている。この対照表は 5 種類の資料に現れる慣用句を一覧したものである。5 種類の資料は以下のとおりである。

①金田一春彦，金田一秀穂監修『新レインボー小学国語辞典改訂第 3 版』学研，2005

②金田一京助編『小学館学習国語新辞典全訂第二版』小学館，2006

③宮地裕編『慣用句の意味と用法』明治書院，1982

④米川明彦，大谷伊都子編『日本語慣用句辞典』東京堂出版，2005

⑤金田一秀穂監修『小学生のまんが慣用句辞典』学研，2005

対照表に現れる慣用句は 3,628 句で，この中から「手が～」「手を～」「手に～」を中心に，「手」が使われている動詞慣用句と形容詞慣用句を抽出し，その使用頻度を調査した。検索項目は合計 91 項目となった。検索結果については，活用形の違い，漢字の異同などを同定する作業を行った。例えば，「手がつけられない」「手のつけようがない」は「手をつける」の項目にまとめた。「手を

合わせる」は「手を合わす」と同じ項目として扱った。また,「手の裏を返す」が 2 例あったが「手のひらを返す」の項目にまとめた。このような方法で検索結果を整理したところ，分析対象の手の慣用句は全部で 74 項目となった。

3.2　分析指標としての「手」の慣用句

以下に指標 74 項目のリストを挙げる。

手が上がる，手が空く，手が後ろに回る，手がかかる，手が切れる，手が込む，手が足りる，手がつく，手が出る，手が届く，手がない，手が伸びる，手が入る，手が離れる，手が早い，手がふさがる，手が回る，手が焼ける，手に汗を握る，手に余る，手に入れる，手に負えない，手に落ちる，手に掛かる，手に掛ける，手にする，手につく，手に手を取る，手に取る，手になる，手にのる，手に入る，手に渡る，手のひら／裏を返す，手も足も出ない，手をあける，手をあげる，手を合わせる／合わす，手を入れる，手を打つ，手をかえる，手をかける，手を貸す，手を借りる，手を切る，手を下す，手を組む，手を加える，手をこまねく／こまぬく，手を差し伸べる，手を染める，手を出す，手をつかねる，手を尽くす，手をつける，手を取り合う，手を取る，手を握る，手を抜く，手を濡らす，手を引く，手を広げる，手を施す，手を回す，手を結ぶ，手を焼く，手を休める，赤子の手をひねる（ような），飼い犬に手をかまれる，猫の手も借りたい，胸に手をあてる，手をのばす，手を離す，手をわずらわせる／わざらわす

3.3　調査結果

2. で述べた文章資料を対象に，3.2 で抽出した「手」の慣用句 74 項目の出現回数を調査した。その調査結果を表 3 に示す。表中の代表項目は 3.2 で挙げた第 1 番目の項目によって代表する。

表 3　各文章資料における「手」の慣用句の出現数

	代表項目	1 哲学	2 歴史	3 社会科学	4 自然科学	5 技術・工学	6 産業	7 芸術・美術	8 言語	9 文学
1	手が上がる	0	0	4	0	0	2	0	0	9
2	手が空く	1	0	2	0	0	1	0	0	9
3	手が後ろに回る	0	0	0	0	0	0	0	0	3
4	手がかかる	1	2	8	2	8	2	0	0	14
5	手が切れる	0	1	0	0	0	0	0	0	2
6	手が込む	1	5	3	0	2	1	4	0	23
7	手が足りる	0	0	3	0	0	0	0	0	8
8	手がつく	1	0	2	1	0	0	0	0	5
9	手が出る	1	1	5	1	4	1	1	0	22

10	手が届く	1	4	13	3	2	0	2	2	37
11	手がない	2	3	5	3	1	1	2	2	16
12	手が伸びる	0	3	3	0	0	1	2	0	35
13	手が入る	0	0	1	0	1	2	0	1	9
14	手が離れる	2	0	9	1	0	0	0	0	7
15	手が早い	0	0	2	0	0	0	0	0	4
16	手がふさがる	0	0	0	0	0	0	0	0	1
17	手が回る	1	1	4	0	3	2	0	0	16
18	手が焼ける	0	0	0	0	0	0	0	0	1
19	手に汗を握る	0	1	0	0	0	0	0	0	8
20	手に余る	5	1	6	0	0	0	1	2	16
21	手に入れる	56	74	158	30	48	19	31	7	371
22	手に負えない	0	2	3	0	0	0	0	1	10
23	手に落ちる	1	7	4	0	1	0	0	0	16
24	手に掛かる	1	8	2	0	0	0	5	0	32
25	手に掛ける	0	1	1	0	0	0	0	0	17
26	手にする	55	43	103	18	24	10	47	6	550
27	手につく	3	3	6	1	1	0	0	0	33
28	手に手を取る	0	0	1	0	0	0	0	0	8
29	手に取る	13	21	37	13	14	4	14	9	307
30	手になる	4	11	7	1	2	1	16	0	16
31	手にのる	1	0	0	0	0	0	0	0	9
32	手に入る	18	18	78	24	23	20	10	5	154
33	手に渡る	1	8	10	0	3	3	2	1	27
34	手のひらを返す	1	0	5	1	1	1	1	0	14
35	手も足も出ない	1	2	3	0	2	2	1	0	16
36	手をあける	0	0	1	0	0	0	0	0	0
37	手をあげる	7	5	45	4	1	2	6	3	164
38	手を合わせる	10	8	15	3	0	0	1	0	76

39	手を入れる	7	4	10	3	8	3	12	2	95
40	手を打つ	5	26	35	2	5	0	18	1	102
41	手をかえる	1	1	4	0	3	1	2	1	8
42	手をかける	1	11	14	1	14	2	3	0	204
43	手を貸す	3	8	15	2	1	2	2	1	82
44	手を借りる	1	1	6	0	0	0	1	0	25
45	手を切る	1	4	5	3	1	0	1	0	28
46	手を下す	2	3	4	1	1	0	0	0	34
47	手を組む	3	7	15	1	2	0	2	1	37
48	手を加える	1	5	14	4	8	6	10	1	31
49	手をこまねく	3	3	13	2	1	3	1	1	18
50	手を差し伸べる	12	7	17	7	2	0	2	2	65
51	手を染める	6	2	11	3	3	1	4	3	21
52	手を出す	13	17	73	9	12	3	11	1	197
53	手をつかねる	0	0	0	0	0	0	0	0	4
54	手を尽くす	1	2	4	1	1	1	2	0	26
55	手をつける	12	8	43	7	10	2	7	9	110
56	手を取り合う	1	1	3	0	1	0	3	1	17
57	手を取る	5	7	12	3	2	1	8	0	172
58	手を握る	4	12	10	6	3	0	7	0	140
59	手を抜く	12	1	11	2	3	5	5	1	18
60	手を濡らす	0	0	0	0	0	0	0	0	4
61	手を引く	5	10	25	3	4	2	4	2	119
62	手を広げる	1	4	12	0	3	3	2	0	26
63	手を施す	1	0	1	1	1	0	0	0	8
64	手を回す	0	7	5	0	3	0	0	0	56
65	手を結ぶ	3	9	10	1	1	1	0	0	23
66	手を焼く	1	3	9	1	6	0	0	0	21
67	手を休める	0	0	1	0	0	0	1	0	4

68	赤子の手をひねる	0	0	0	0	0	1	1	0	5
69	飼い犬に手をかまれる	1	0	0	0	0	0	0	0	2
70	猫の手も借りたい	1	0	1	1	0	0	0	0	3
71	胸に手をあてる	1	2	2	0	0	0	2	0	11
72	手をのばす	11	11	18	4	7	1	8	7	347
73	手を離す	1	2	19	3	3	1	3	1	121
74	手をわずらわせる	3	0	2	1	1	0	1	0	9
	合計	311	411	973	178	251	114	269	74	4258

3.4　分析方法

「手」の慣用句 74 項目の中から文章のジャンル判別に特に有効な項目を選択するために，多変量解析の一手法である正準判別分析のステップワイズ法を用いて分析を行い，判別に寄与する項目を選択した。なお，分析の際には，2.1 で述べたように文章資料の語数がそれぞれ異なるので，「手」の慣用句の出現数を 100 万語当たりの出現頻度に換算し直した出現率を用いた。

４．分析結果と考察

3.4 の分析方法により，74 項目の「手」の慣用句を説明変数とし，文章資料グループ（以下，ジャンル）を基準変数としてステップワイズ法を用いて判別分析を行った。その結果，逐次的に 5 個の説明変数が予測式に組み込まれ，その手続き内で削除された変数もなく，2.4 で述べた 3 つのジャンルの判別に有効な，以下の 5 つの慣用句が選択された。

①手に余る　②手を打つ　③手をこまねく／こまぬく
④手をあける　⑤手を取り合う

対象とした文章資料グループは 3 つのため，判別関数は 2 つ算出された。記述的指標としてウィルクスの Λ を用い，Λ に基づく χ^2 値については正規性の仮定が満たされないため，目安としてのみ用いる。これらに関する指標を示すと表 4 のようになる。Λ の値は判別関数 1 と 2 の判別に大きく寄与する情報が含まれていることを示している。関数 1 のみで寄与率は 97.8%である。

次に判別空間におけるジャンル間の関係について検討する。選択された 5 慣用句による判別関数平面での各文章資料の判別得点とジャンルの重心をプロットしたものを図 1 に示す。

表 4　判別関数の固有値等

判別関数	固有値	寄与率	p 値	Λ	χ^2
関数 1	1,115.956	97.8	0.000	0.000	41.144
関数 2	25.248	2.2	0.011	0.038	13.070

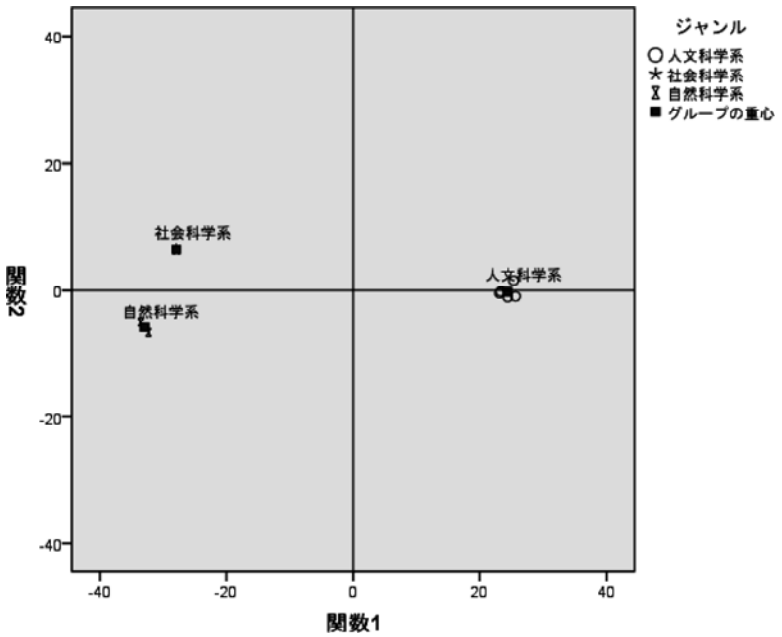

図 1.　判別分析による 9 資料の重心および個体のプロット

図 1 を見るとわかるように，上記 5 慣用句によって 3 つの文章グループがはっきりと分離されているのが分かる。判別関数平面における各文章資料グループの重心の値は表 5 の通りである。

表 5　判別空間における各文章資料グループの重心

ジャンル	関数 1	関数 2
人文科学系	24.351	−0.224
社会科学系	−27.937	6.423
自然科学系	−32.939	−5.862

関数 1 ではまず人文科学系が自然科学系と社会科学系から分離され，関数 2 では社会科学系が自然科学系から分離されている。

次に構造係数と判別空間における各ジャンルの重心の関係から，選択された 5 慣用句がどの文章資料グループを分離するのに有効かを考察する。表 6 に 5 慣用句の構造係数を示す。

表 6　選択された 5 慣用句の構造係数

代表項目	関数 1	関数 2
手に余る	*0.024	0.019
手を打つ	*0.021	−0.005
手をこまねく	−0.010	*0.233
手をあける	−0.014	*0.147
手を取り合う	0.026	*−0.026

* 有意な係数

表 6 から関数 1 では「手に余る」「手を打つ」が人文科学系と自然科学系・社会科学系を分離し，関数 2 では「手をこまねく／こまぬく」「手をあける」「手を取り合う」が社会科学系と自然科学系を分離する特徴的な慣用句となっていることがわかる。ここで判別の可否を評価するための「見かけの的中率」をクロス集計で求めたところ，その結果は 100%となった。さらに交差妥当性を検討したところ，その結果も同様に 100%となった。

以上の結果から，選択された 5 つの「手」の慣用句を指標として，三つの文章資料のジャンルが判別できることが実証されたと考えられる。

最後に日本語教育の観点から，個々の「手」の慣用句の出現率の結果が教育現場の資料としては有意義だと考えられるので，表 7 にその結果を示す。

表 7　各慣用句のジャンルごとの 100 万語当たりの出現頻度

	代表項目	人文科学系	社会科学系	自然科学系
1	手が上がる.	0.81	2.88	0.00
2	手が空く	1.37	1.44	0.00
3	手が後ろに回る	0.27	0.00	0.00
4	手がかかる	2.56	3.48	7.06
5	手が切れる	0.55	0.00	0.00
6	手が込む	7.52	1.59	1.46
7	手が足りる	0.72	0.45	0.00
8	手がつく	1.01	0.30	0.61
9	手が出る	3.67	1.89	3.53
10	手が届く	10.95	1.95	3.29

11	手がない	9.25	1.89	2.56
12	手が伸びる	5.78	1.59	0.00
13	手が入る	2.84	2.43	0.73
14	手が離れる	1.75	1.35	0.61
15	手が早い	0.36	0.30	0.00
16	手がふさがる	0.09	0.00	0.00
17	手が回る	2.37	2.88	2.19
18	手が焼ける	0.09	0.00	0.00
19	手に汗を握る	1.09	0.00	0.00
20	手に余る	9.43	0.90	0.00
21	手に入れる	129.90	45.36	53.34
22	手に負えない	3.67	0.45	0.00
23	手に落ちる	4.59	0.60	0.73
24	手に掛かる	10.20	0.30	0.00
25	手に掛ける	1.90	0.15	0.00
26	手にする	144.11	26.85	28.50
27	手につく	5.76	0.90	1.34
28	手に手を取る	0.72	0.15	0.00
29	手に取る	71.59	10.11	18.15
30	手になる	19.91	2.19	2.07
31	手にのる	1.37	0.00	0.00
32	手に入る	48.35	34.50	31.43
33	手に渡る	9.50	4.92	2.19
34	手のひらを返す	2.58	1.89	1.34
35	手も足も出ない	3.50	2.73	1.46
36	手をあける	0.00	0.15	0.00
37	手をあげる	31.18	9.03	3.17
38	手を合わせる	16.16	2.25	1.83
39	手を入れる	27.13	4.92	7.67

40	手を打つ	37.31	5.25	4.87
41	手をかえる	5.20	1.74	2.19
42	手をかける	25.27	4.38	10.83
43	手を貸す	15.57	4.53	1.95
44	手を借りる	3.94	0.90	0.00
45	手を切る	5.32	0.75	2.56
46	手を下す	5.29	0.60	1.34
47	手を組む	11.15	2.25	2.07
48	手を加える	14.83	8.94	8.28
49	手をこまねく	7.20	5.37	1.95
50	手を差し伸べる	20.74	2.55	5.73
51	手を染める	15.12	2.79	4.02
52	手を出す	41.69	14.37	14.25
53	手をつかねる	0.36	0.00	0.00
54	手を尽くす	5.16	1.74	1.34
55	手をつける	43.17	8.73	11.57
56	手を取り合う	6.77	0.45	0.73
57	手を取る	26.95	2.94	3.29
58	手を握る	24.60	1.50	5.85
59	手を抜く	14.54	7.35	3.41
60	手を濡らす	0.36	0.00	0.00
61	手を引く	24.31	6.03	4.75
62	手を広げる	5.90	5.22	2.19
63	手を施す	1.28	0.15	1.34
64	手を回す	7.63	0.75	2.19
65	手を結ぶ	7.08	2.64	1.34
66	手を焼く	3.56	1.35	4.99
67	手を休める	1.12	0.15	0.00
68	赤子の手をひねる	1.21	1.14	0.00

69	飼い犬に手をかまれる	0.74	0.00	0.00
70	猫の手も借りたい	0.83	0.15	0.61
71	胸に手をあてる	3.81	0.30	0.00
72	手をのばす	61.75	3.84	7.55
73	手を離す	16.50	3.99	4.02
74	手をわずらわせる	3.25	0.30	1.34
	合計	1064.11	275.91	291.81

表 7 の結果は，日本語学習者が人文科学，社会科学，自然科学の各分野の文章において，どのような「手」の慣用句に触れる可能性が高いのかという一つの可能性を示していると言えよう。

５．お わ り に

本研究では，多義性を持つ慣用句について特に意味上の分類は行わず一つの指標として扱って分析を行った。今後の課題としては，多義性を持つ「手」の慣用句について意味分類を行い，その意味機能情報を含む指標を用いて文章ジャンルを判別することの可能性を探っていきたいと思う。

本稿は文部科学省科学研究費基盤研究 C（課題番号 20520429 研究代表者山崎誠）の補助を受けて行った研究成果の一部である。

注

1) ここではジャンルという語を個々人の持つ文体的特徴を超えたところに存在するある特徴パターンを持った文章グループと定義する。
2) 「語数」は，コーパスの構築に際して使われている解析の言語単位である「短単位」で数えたもの。空白，補助記号（句読点など），記号（A, B, C，ア，イ，ウなど）を含まない数である。
3) 書き言葉コーパスの設計と目的に照らして，以下のような書籍は対象外とした。40 ページ以下の書籍，ページ数情報のない書籍（ランダムサンプリングの際にページ情報を利用するため），官公庁刊行物のうち流通していないもの，学習試験図書，電子資料，地図資料，写真集，漫画などである。

謝　　辞

本稿の慣用句データの整理については本塾大学院文学研究科国文学科日本語教育学分野修士 2 年生顧翌清さんの協力を得た。ここに記して感謝の意を表したい。

参 考 文 献

宮地　裕 (1982)『慣用句の意味と用法』　明治書院.
前川喜久雄 (2008),「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」，日本語の研究，4(1), pp. 82-95.
前川喜久雄 (2009)「代表性を有する大規模日本語書き言葉コーパスの構築」人工知能学会誌，24(5), pp. 616-622.

山崎　誠 (2009)「代表性を有する現代日本語書籍コーパスの構築」人工知能学会誌, 24(5), pp. 623-631.
村田　年 (2007)「専門日本語教育における論述文指導のための接続語句・助詞相当句の研究」『統計数理（特集「文化を科学する」）』統計数理研究所 Vol. 55, No. 2, pp. 269-284.
村田　年 (2008)「文章と複合動詞―論述文ジャンルを特徴づける新たな指標を探して―」『日本語と日本語教育』慶應義塾大学日本語・日本文化教育センター36 号 pp. 1-33.

関連 URL

「KOTONOHA 国立国語研究所言語コーパス整備計画」
http://www.ninjal.ac.jp/kotonoha/
「ことば不思議箱」
http://kotoba.nuee.nagoya-u.ac.jp/