様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 １９ 日現在

機関番号：６２６１８
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ００６５５

研究課題名（和文）コーパス分析による書き言葉的「硬・軟」度と話し言葉的「硬・軟」度の語への付与

研究課題名（英文）Annotation of stylistic information of words based on frequency appearing in corpora

研究代表者
柏野　和佳子（KASHINO, Wakako）

大学共同利用機関法人人間文化研究機構国立国語研究所・研究系・准教授

研究者番号：５０３１１１４７

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：「書き言葉的」な語で記述すべき学術的文章（レポート、論文等）に「話し言葉的」な語が混じるという問題を解決するために、「書き言葉的」「話し言葉的」といった注釈のある語を作文技術に関する文献等から2,791語抽出した。それらに、『日本語日常会話コーパス』（CEJC）、『日本語話し言葉コーパス』（CSJ）、『語の文体値データ』の情報を付与し、語の文体差を計量するためのデータベースを作成した。

研究成果の学術的意義や社会的意義
学術的文章（レポート、論文等）の作成時に、「話し言葉的」な語や、軟らかすぎる語を用いるのは避けた方がよい。そのためには、語レベルでの文体情報の把握が必要になる。本研究では、「書き言葉的」な語、「話し言葉的」な語を2,791語集め、それらの話し言葉や書き言葉のコーパス頻度情報に基づいた使用差を数値化することによって、語に文体情報を付したデータベースを構築した。学術文章作成時に、より適切な語句選択が可能になる資料として役立つものと期待される。

研究成果の概要（英文）：To solve the problem of "spontaneous" words being mixed in academic texts (reports, articles, etc.) that should be described in writing-style words, the words with annotations such as "writing style" or "spontaneous" were extracted from literature on composition techniques. For each of those 2,791 words, we newly annotated word frequency information appearing in “Corpus of Everyday Japanese Conversation : CEJC” and “Corpus of Spontaneous Japanese : CSJ”. In addition, we annotated “Word Stylistics Data”. The resulting database enables us to quantitatively measure stylistic differences between words.

研究分野：日本語学

キーワード：文体　書き言葉　話し言葉　コーパス　位相　学術的文章　文章作成　日本語教育

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

学術論文作成時に話し言葉的な語を混在させないためには、「話し言葉的」と「書き言葉的」の区別が必要である。しかし、その文体差には段階的なものがある。また、それらの区別よりも、公的な場面で使用されやすい「硬い」語か、そうではない「軟らかい」語かなどの文体差が使用差に効いている場合がある。

これまで、科研費 基盤研究 (C) 26370554 「「書き言葉的」と「話し言葉的」という文体差のある語の分析」の研究において、日本語の作文技術に関する文献及び、書き言葉と話し言葉の相互関係に関する文献を広く収集し、そこから文体差のある語や表現 2,791 語を抽出していた。また、それらの語に、『現代日本語書き言葉均衡コーパス（BCCWJ）』のレジスター別頻度および、科学技術論文データを用いた頻度情報を付与したデータベースを作成していた。ただし、当時、話し言葉コーパスのデータ収集が十分ではなかったため、それらの頻度情報の付与が大きな課題として残っていた。

２．研究の目的

「書き言葉的な語」と「話し言葉的な語」とされる語を対象に、各種コーパスや、作文、論文の使用例を分析し、文体差を計量的に示す仕組みを実現する。分析し，学術的文書作成のための語の使用指針を策定することが目的である。「書き言葉的」「話し言葉的」という分類に加え、「硬い」「軟らかい」といった位相情報を語に付与する。

３．研究の方法

（１）コーパスに基づく頻度情報の付与

次の話し言葉コーパスの頻度情報を新たに付与する。

①短単位と一致する見出し語は、公開されている頻度表から頻度情報を取得し、見出し語と頻度情報のマトリックスを作成する。

②短単位と一致しない見出し語については、CEJC 中納言の検索 API を使用して頻度情報を取得し、同様に、見出し語と頻度情報のマトリックスを作成する。

『日本語日常会話コーパス（CEJC）』

頻度表：https://www2.ninjal.ac.jp/conversation/cejc/cejc-wc.html

『日本語話し言葉コーパス（CSJ）』

頻度表：https://clrd.ninjal.ac.jp/csj/chunagon.html

（２）文体情報の付与

次の公開されている『語の文体値データ』を用い、短単位と一致する見出し語のみを対象に、語の文体値データの頻度表から頻度情報を取得し、見出し語と頻度情報のマトリックスを作成する。

『語の文体値データ』(2022 年 2 月公開第 1 版)　http://doi.org/10.15084/00003532

４．研究成果

『現代日本語書き言葉均衡コーパス（BCCWJ）』から得られる情報を書き言葉の頻度情報とし、『日本語日常会話コーパス（CEJC）』と『日本語話し言葉コーパス（CSJ）』から得られる情報を話し言葉の頻度情報とし、それらの調整頻度を用いて、書き言葉的、話し言葉的の指標となる数値を算出した。指標は、(書き言葉の調整頻度 - 話し言葉の調整頻度) / (書き言葉の調整頻度 + 話し言葉の調整頻度) を計算した値である。1 に近いほど書き言葉的、-1 に近いほど話し言葉的であることを表す。

図 1 に、書き言葉的　上位語を、図 2 に、話し言葉的　上位語を示す。

番号	対象	書き言葉		話し言葉		指標
	表記	出現頻度	調整頻度	出現頻度	調整頻度	
685	けど。	468	4.46	24,520	2,524.60	-1.00
73	…だったんですけどれども	15	0.14	500	51.48	-0.99
687	けども	762	7.26	11,228	1,156.05	-0.99
642	～くて。	110	1.05	1,639	168.75	-0.99
406	おっきい	34	0.32	472	48.60	-0.98
555	が。	3,020	28.79	28,152	2,898.56	-0.98
748	こっから	44	0.42	271	27.90	-0.96
88	…ですけれども、	4,458	42.49	14,390	1,481.61	-0.94
438	おんなじ	296	2.82	920	94.72	-0.94
47	…から（文末用法）	2,687	25.61	6,870	707.34	-0.93
517	～から。	2,687	25.61	6,870	707.34	-0.93
521	から。	2,687	25.61	6,870	707.34	-0.93
408	おとbetter	205	1.95	537	55.29	-0.93
334	言われまして	54	0.51	137	14.11	-0.92
192	あの	34,506	328.91	53,977	5,557.52	-0.89
324	入れといて	15	0.14	38	3.91	-0.89
212	あるんで	243	2.32	351	36.14	-0.87
736	こうやって	802	7.64	949	97.71	-0.85
152	上がってる	46	0.44	63	6.49	-0.85
11	N でもって	4	0.04	14	1.44	-0.84
276	一番目	184	1.75	194	19.97	-0.83
218	あんまり	2,801	26.70	2,808	289.11	-0.83
409	おととし	71	0.68	70	7.21	-0.81
327	色々な	1,127	10.74	955	98.33	-0.80
2791	～んだって	1,782	16.99	1,455	149.81	-0.80
164	あした	552	5.26	452	46.54	-0.79
661	～ぐらい	14,788	140.96	11,778	1,212.67	-0.79
177	あったかい	172	1.64	143	14.72	-0.79
711	こういうふうな	311	2.96	252	25.95	-0.79
2775	～みたいな	9,751	92.95	7,615	784.05	-0.79

図 1　書き言葉的　上位語

番号	対象	書き言葉		話し言葉		指標
	表記	出現頻度	調整頻度	出現頻度	調整頻度	
114	…のために	20,512	195.52	136	14.00	0.87
134	…ようだ	9,098	86.72	69	7.10	0.85
2759	～によれば	4,807	45.82	67	6.90	0.74
488	かつて	7,703	73.42	112	11.53	0.73
286	一層	3,635	34.65	59	6.07	0.70
2725	～なければならない	22,408	213.59	385	39.64	0.69
457	価格	15,238	145.25	268	27.59	0.68
531	彼	87,535	834.37	1,551	159.69	0.68
242	いかにも	2,903	27.67	52	5.35	0.67
576	企業	36,663	349.47	666	68.57	0.67
5	A や B を	19,348	184.42	353	36.35	0.67
629	金額	8,967	85.47	167	17.19	0.66
593	君	15,823	150.82	307	31.61	0.65
199	あらゆる	5,741	54.72	117	12.05	0.64
2758	～によると	6,364	60.66	130	13.38	0.64
532	彼ら	18,282	174.26	381	39.23	0.63
723	こうした	13,902	132.51	292	30.06	0.63
290	いったん	2,513	23.95	53	5.46	0.62
20	V-ようが	8,556	81.55	197	20.28	0.60
2754	～に基づき	3,785	36.08	90	9.27	0.59
380	大幅に	2,314	22.06	55	5.66	0.59
733	高度	5,423	51.69	132	13.59	0.58
12	N なら	36,927	351.98	905	93.18	0.58
118	…べき N	17,250	164.42	435	44.79	0.57
2698	～でなければ	3,954	37.69	100	10.30	0.57
19	N を…みる	12,822	122.22	326	33.57	0.57
70	…だけでなく	7,164	68.29	184	18.94	0.56
308	いな	2,507	23.90	64	6.59	0.56
117	…ばかりの N	3,231	30.80	83	8.55	0.56
236	いかが	4,605	43.89	119	12.25	0.56

図 2　話し言葉的　上位語

また、論述の場面を想定し、ある表現が使えるのか使えないのかをわかりやすい表にまとめ直した。図 3 から図 5 に、一部を例示する。

なぜかと言うと,なぜかというと,なぜかといえば,なぜって		[×使用しない]
	なぜなら,なぜならば	[〇使用可能]
	これは~ためである	[〇使用可能]
なので		[×使用しない]
	そういうわけで	[〇使用可能]
なのに		[×使用しない]
	それなのに	[〇使用可能]
にしても		[×使用しない]
	それにしても	[〇使用可能]
もしかしたら		[×使用しない]
	場合によっては	[〇使用可能]

図 3　論述の場合 論と論のつなぎ 例

いい		[×使用しない]
	よい,良い	[〇使用可能]
	適切	[〇使用可能]
いい点		[×使用しない]
	よい点	[〇使用可能]
いっぱいの		[×使用しない]
	多くの,数多くの,多数の	[〇使用可能]
いろんな		[×使用しない]
	多様な,多用な,多岐にわたる	[〇使用可能]
	さまざまな,様々な	[〇使用可能]
	種々の,種種の	[〇使用可能]
エグイ		[×使用しない] 俗語
	はげしい	[〇使用可能]

図 4　論述の場合 形容的 例

いいつけ		[×使用しない]
	命令	[〇使用可能]
いいわけ		[×使用しない]
	弁解,弁明	[〇使用可能]
いい加減		[×使用しない]
	おろそか	[〇使用可能]
いっぱい,一杯		[×使用しない]
	大勢	[〇使用可能]
いま,今		[×使用しない]
	現在,ただいま	[〇使用可能]

図 5　論述の場合 名詞 例

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 柏野 和佳子	４．巻 234
２．論文標題 言語研究の成果を発信するさまざまな国語辞典	５．発行年 2022年
３．雑誌名 文学・語学	６．最初と最後の頁 153～161
掲載論文のDOI（デジタルオブジェクト識別子） 10.34492/bungakugogaku.234.0_153	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計9件（うち招待講演　2件／うち国際学会　1件）

１．発表者名 柏野和佳子
２．発表標題 辞書学への応用：『現代日本語書き言葉均衡コーパス』BCCWJ
３．学会等名 Design, construction, and application of Japanese language corpora（招待講演）（国際学会）
４．発表年 2022年

１．発表者名 柏野和佳子
２．発表標題 省略やぼかしに用いられる語の使用実態調査
３．学会等名 シンポジウム「日常会話コーパス」VII
４．発表年 2022年

１．発表者名 柏野和佳子
２．発表標題 コーパスによる語彙教育の精緻化
３．学会等名 専門日本語教育学シンポジウム（招待講演）
４．発表年 2021年

１．発表者名
星野和子

２．発表標題
日本語教育の教科書における「考える」と「思う」の分析

３．学会等名
「学習者辞書用語彙資源の構築」共同研究発表

４．発表年
2023年

１．発表者名
丸山直子

２．発表標題
複合格助詞の位相

３．学会等名
「学習者辞書用語彙資源の構築」共同研究発表

４．発表年
2023年

１．発表者名
馬場俊臣

２．発表標題
『語の文体値データ』について

３．学会等名
「学習者辞書用語彙資源の構築」共同研究発表

４．発表年
2023年

１．発表者名
阿辺川武

２．発表標題
日本語接続表現の計量的分析

３．学会等名
「学習者辞書用語彙資源の構築」共同研究発表

４．発表年
2023年

１．発表者名 柏野和佳子
２．発表標題 語への位相情報の付与の検討
３．学会等名 「学習者辞書用語彙資源の構築」共同研究発表
４．発表年 2023年

１．発表者名 柏野和佳子
２．発表標題 書き言葉的「硬・軟」度と話し言葉的「硬・軟」度の検討
３．学会等名 シンポジウム「日常会話コーパス」VIII
４．発表年 2023年

〔図書〕　計1件

１．著者名 【編著者】石黒圭　【執筆者】青木優子・安部達雄・新城直樹・井伊菜穂子・市江愛・井上雄太・岩崎拓也・王慧雋・赫楊・柏野和佳子・金井勇人・高恩淑・佐野彩子・鈴木英子・田中啓行・董芸・本多由美子	４．発行年 2021年
２．出版社 東京堂出版	５．総ページ数 384
３．書名 日本語文章チェック事典	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	丸山　直子 (MARUYAMA Naoko) (00199936)	東京女子大学・現代教養学部・教授 (32652)	

6．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	佐渡島　紗織 (SADOSHIMA Saori) (20350423)	早稲田大学・国際学術院・教授 (32689)	

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	星野　和子 (HOSHINO Kazuko)		
研究協力者	仁科　喜久子 (NISHINA Kikuko)		
研究協力者	ハー　グエン・ティ・ビック (HA Ngueyen Thi Bich)		
研究協力者	木田　真理 (KIDA Mari)		
研究協力者	前坊　香菜子 (MAEBO Kanako)		

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関