

『日本語歴史コーパス』の新しい語彙表とその応用例

おぎぞとしのぶ 智信 (国立国語研究所)

1. はじめに

『日本語歴史コーパス』を利用する上で、各々の見出し語が、各々の作品等においていくつ出現しているかは重要な情報である。調査対象とした語について単にその用例を見るだけでなく、その語が各作品やコーパス全体の中で占める位置や、共起する語との関係を測ることで初めて明らかになることが多いからである。各見出し語の用例数はコーパスを利用した研究にとって基礎的な情報であり、たとえばコロケーション強度などの指標を算出するのに重要である。

一方、『日本語歴史コーパス』の利用のために提供されている検索アプリケーション「中納言」は文脈付き用例を出力するコンコーダンスである。単に語の用例数を計測するために用いるには無駄が多く、システムに大きな負荷をかけるため、別途、集計済みの語彙表を利用することが望ましい。そのために従来、語彙表が一般に公開されてきた。しかし、この語彙表は時代別・作品別にファイルが分割されているため、時代や作品を通じて集計することが行いにくかった。

本発表では、できるだけ時代・作品ごとに分割をしない形で新たに作成した語彙表について報告する。この語彙表をもとに表計算ソフト等を用いることで従来形式のデータも出力することが可能であり、コロケーション強度など、統計的な指標の計算にも適した形式となっている。このような応用の方法についてもあわせて説明する。

2. 『日本語歴史コーパス』の語彙表

従来の語彙表の形式

ここでいうコーパスの語彙表とは、コーパスを構成する全ての語を、見出し語ごとにまとめ、簡易な出典情報を付した頻度付きリストである。『現代日本語書き言葉均衡コーパス』では、短単位語彙表・長単位語彙表・品詞構成表・語種構成表がそれぞれ公開されている¹。また『日本語歴史コーパス』では、従来、短単位語彙表・長単位語彙表の二つが公開されており、それぞれに品詞構成表・語種構成表も含まれている²。

このほかに、いずれのコーパスでも出典（サンプル・作品）ごとの総語数をまとめた語数表が公開されている。語数表は見出し語の情報は含まず、たんに語数のみを提供するもので、語彙表とは別のものである。

従来の『日本語歴史コーパス』語彙表は、図1に示すようにサブコーパス（時代別・ジャ

¹ <https://ccd.ninjal.ac.jp/bccwj/bcc-chu.html>

² <https://ccd.ninjal.ac.jp/chj/chj-wc.html>

ンル別の資料群) ごとにフォルダにまとめられ、作品ごとに 1 つのファイルとなっていた。

各語彙表が持つ列は、時代名、語彙素 ID、語彙素読み、語彙素、品詞、語彙素細分類、語種、頻度、の基本的な情報に加え、『万葉集』『源氏物語』等では巻ごとの頻度情報を加えている。

フォルダ	ファイル
└─SUW	└─SUW
├─和歌集	├─和歌集
├─奈良・万葉集	CHJ_古今和歌集_frequencylist_suw.tsv
├─奈良・宣命	CHJ_古今和歌集_frequencylist_suw_goshu.tsv
├─室町・キリシタン	CHJ_古今和歌集_frequencylist_suw_pos.tsv
├─室町・狂言	CHJ_後撰和歌集_frequencylist_suw.tsv
├─平安・仮名文学	CHJ_後撰和歌集_frequencylist_suw_goshu.tsv
├─明治・大正・初期口語	CHJ_後撰和歌集_frequencylist_suw_pos.tsv
├─明治・大正・教科書	CHJ_拾遺和歌集_frequencylist_suw.tsv
├─明治・大正・雑誌	CHJ_拾遺和歌集_frequencylist_suw_goshu.tsv
├─江戸・人情本	CHJ_拾遺和歌集_frequencylist_suw_pos.tsv
├─江戸・洒落本	:
├─江戸・近松	├─奈良・万葉集
├─鎌倉・日記・紀行	CHJ_万葉集_frequencylist_suw.tsv
└─鎌倉・説話・随筆	CHJ_万葉集_frequencylist_suw_goshu.tsv
	CHJ_万葉集_frequencylist_suw_pos.tsv
	:
	├─奈良・宣命
	CHJ_続日本紀_frequencylist_suw.tsv

図 1 従来形式の語彙表のファイル構成

このため、短単位語彙表だけでファイル数が 128、アーカイブ中のファイル数は品詞構成表・語種構成表とあわせて 385 とたいへん多い。このような形式であるため、個別の作品についてのみ見るのであれば十分であっても、コーパス全体の中での位置づけを見渡すには適していなかった。

新しい語彙表の形式

そこで、新たに、できるだけ時代・作品ごとに分割をしない形で短単位の語彙表を作成した。新形式では、従来フォルダ分けの基準であったサブコーパス名や、ファイル分けの基準であった作品名を語彙表の列に回し、一枚の表となるようにまとめ上げた。さらに、研究上で必要と考えられる情報を極力盛り込むこととした。

データサイズなどを考慮して表に取り込んだ情報は、下記の通りである。下線を付したものは従来形式にはなかった情報、波線を付したものはフォルダ名やファイル名で与えられていた情報である。

語彙素読み、語彙素、語種、語彙素 ID、品詞、語形、書字形、時代名、サブコーパス名、作品名、部、成立年、コアフラグ、本文種別、文体、頻度

こうしてできる新しい形式の語彙表は、全体で一般の PC で Excel 等の表計算ソフトを用いて扱うことのできる 100 万行を優に超えることになる。そのため、100 万行以内に収めるため、近代の雑誌とそれ以外の資料に 2 分割した。

上代から近代まで（雑誌以外）：約 915700 行 約 115 MB
近代の雑誌：約 884900 行 約 115 MB

2 ファイルとも同一の列フォーマットとなっているため、データベースやプログラミング言語などで扱う場合には、単純に結合して利用することが可能である。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	語彙素読み	語彙素	語種	語彙素ID	品詞	語形	書字形	時代名	サブコー	作品名	部	成立年	コアフラグ	本文種別	文体	頻度
300641	コクゲン	制限	漢	53767	名詞-普通	コクゲン	制限	6明治	明治・大正	文明田舎問答		1878	0	会話	文語	1
300642	コクゲン	制限	漢	53767	名詞-普通	コクゲン	制限	6明治	明治・大正	明治の光		1875	0	会話	口語	1
300643	コクゲン	制限	漢	53767	名詞-普通	コクゲン	制限	5江戸	江戸-人情	明治後の正夢		1823	0	会話		1
300644	コクゲン	制限	漢	53767	名詞-普通	コクゲン	制限	5江戸	江戸-人情	春色江戸祭		1864	1	会話		1
300645	コクゲン	制限	漢	53767	名詞-普通	コクゲン	制限	5江戸	江戸-人情	春色連理の梅		1851	0	会話		1
300646	コクゲン	制限	漢	53767	名詞-普通	コクゲン	制限	5江戸	江戸-洒落	花街舞々々江戸		1826	1	会話		2
300647	コクゴ	こくごく	和	172939	副詞	コクゴ	こくごく	6明治	明治・大正	安愚楽編		1872	0	会話	口語	1
300648	コクゴ	国語	漢	12521	名詞-普通	コクゴ	こくごく	8昭和	明治・大正	小学校国語小学校2年		1947	1		口語	1
300649	コクゴ	国語	漢	12521	名詞-普通	コクゴ	こくごく	8昭和	明治・大正	小学校国語小学校3年		1947	1		口語	2
300650	コクゴ	国語	漢	12521	名詞-普通	コクゴ	こくごく	8昭和	明治・大正	小学校国語小学校5年		1947	1		口語	1
300651	コクゴ	国語	漢	12521	名詞-普通	コクゴ	こくごく	8昭和	明治・大正	小学校国語小学校6年		1947	1	引用	口語	1
300652	コクゴ	国語	漢	12521	名詞-普通	コクゴ	国語	6明治	明治・大正	よりあひばなし		1874	0	会話	口語	2
300653	コクゴ	国語	漢	12521	名詞-普通	コクゴ	国語	6明治	明治・大正	吾輩は猫である		1907	0		口語	2
300654	コクゴ	国語	漢	12521	名詞-普通	コクゴ	国語	7大正	明治・大正	小学校国語小学校6年		1918	1		口語	4
300655	コクゴ	国語	漢	12521	名詞-普通	コクゴ	国語	8昭和	明治・大正	小学校国語小学校5年		1933	1		口語	22
300656	コクゴ	国語	漢	12521	名詞-普通	コクゴ	国語	8昭和	明治・大正	小学校国語小学校6年		1933	1		口語	2

図 2 表計算ソフトで開いた新形式の語彙表

3. ピボットテーブルによる従来形式の表の作成

従来のような作品別の語彙表や品詞構成表・語種構成表は、新しい形式のデータから表計算ソフトの機能によって生成することができる。図 3 は、フィルタ機能によって『伊勢物語』の語彙表として表示したものである。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	語彙素	語彙素	語彙素	品詞	品詞	書字形	時代名	サブコーパス	作品名	部	成立年	コアラ	本文種別	文体	頻度	
5674	アイ	相	和	79	接頭辞	アイ	あひ	2平安	平安-仮名	伊勢物語		920	1			12
5735	アイ	相	和	79	接頭辞	アイ	あひ	2平安	平安-仮名	伊勢物語		920	1	歌		1
6312	アイゴト	逢い言	和	231683	名詞-普通	アイゴト	あひごと	2平安	平安-仮名	伊勢物語		920	1			1
6940	アイダ	間	和	168	名詞-普通	アイダ	あひだ	2平安	平安-仮名	伊勢物語		920	1			2
7028	アイダ	間	和	168	名詞-普通	アイダ	間	2平安	平安-仮名	伊勢物語		920	1			1
8074	アウ	会う	和	241	動詞-一般	アウ	あふ	2平安	平安-仮名	伊勢物語		920	1			24
8158	アウ	会う	和	241	動詞-一般	アウ	あふ	2平安	平安-仮名	伊勢物語		920	1	会話		3
8279	アウ	会う	和	241	動詞-一般	アウ	あふ	2平安	平安-仮名	伊勢物語		920	1	歌		19
8819	アウ	合う	和	242	動詞-非自	アウ	あふ	2平安	平安-仮名	伊勢物語		920	1			3
9324	アエル	敗える	和	140696	動詞-非自	アウ	あふ	2平安	平安-仮名	伊勢物語		920	1			1
9747	アオイ	青い	和	269	形容詞-一	アオシ	青し	2平安	平安-仮名	伊勢物語		920	1			1
10883	アカイ	赤い	和	308	形容詞-一	アカシ	赤し	2平安	平安-仮名	伊勢物語		920	1			1
11241	アカス	明かす	和	328	動詞-一般	アカス	明かす	2平安	平安-仮名	伊勢物語		920	1			1
11338	アカス	明かす	和	328	動詞-一般	アカス	明かす	2平安	平安-仮名	伊勢物語		920	1	歌		1
11622	アカツキ	暁方	和	149847	名詞-普通	アカツキ	あかつき	2平安	平安-仮名	伊勢物語		920	1	歌		1
12235	アガタ	県	和	46475	名詞-普通	アガタ	あがた	2平安	平安-仮名	伊勢物語		920	1			1
12405	アル	上がる	和	375	動詞-一般	アル	あがる	2平安	平安-仮名	伊勢物語		920	1			1
12945	アキ	秋	和	382	名詞-普通	アキ	秋	2平安	平安-仮名	伊勢物語		920	1			3
13201	アキ	秋	和	382	名詞-普通	アキ	秋	2平安	平安-仮名	伊勢物語		920	1	歌		13
13322	アキカゼ	秋風	和	385	名詞-普通	アキカゼ	秋風	2平安	平安-仮名	伊勢物語		920	1	会話		1

図 3 作品別の語彙表（伊勢物語）

図 4 は、ピボットテーブルを用いて、「平安-仮名文学」のサブコーパスに絞り、作品別の語種構成表を作成したものである。

	A	B	C	D	E	F	G	H	I
	サブコーパス名	平安-仮名文学							
1									
2	サブコーパス名	平安-仮名文学							
3									
4	合計 / 頻度	列ラベル							
5	行ラベル	伊勢物語	源氏物語	古今和歌集	更級日記	讃岐典侍日記	紫式部日記	大鏡	大和物語
6	外		104	2	1	10	5	37	3
7	漢		171	11783	398	371	703	1008	6272
8	記号		2237	74679	1263	2416	3272	3617	13177
9	国		200	1200	1515	194	95	153	1661
10	混		17	1498	15	44	75	89	344
11	不明					1		1	
12	和		13437	431130	29330	14049	14661	16186	62953
13	空白			1	33		11		85
14	総計		16062	520395	32556	17076	18827	21059	84529
15									26953

図 4 作品別の語種構成表（平安-仮名文学）

4. 新しい語彙表を使ったコロケーション強度の計算

新しい形式の語彙表を用いることにより、コロケーション強度の計算のために必要な、個別作品を超えた語別の頻度情報の抽出を行うことが比較的容易に行えるようになった。『日

本語歴史コーパス』は、「中納言」の検索結果にも語彙素 ID が含まれているため、語彙表をもとにピボットテーブルを用いて語彙素 ID の頻度表を用意すれば、調査対象語がコーパス中でどれだけ用いられているか簡単に調査できる。これを用いることでコロケーション強度などの指標の計算も容易になる。

たとえば、コロケーション強度のなかでも比較的単純なダイス係数は、語 A と語 B の結びつきの強さを、コーパス中の A の用例数と B の用例数、そして A と B がともに用いられた用例数から計算する。

$$\text{ダイス係数} = 2 \times (\text{A} \cap \text{B の用例数} / (\text{A の用例数} + \text{B の用例数}))$$

このとき必要な $\text{A} \cap \text{B}$ の用例数は「中納言」で A と B がともに現れる例を検索した結果を集計して取得し、A の用例数、B の用例数は語彙表をもとに作った表を参照して取得すればよい。

具体例として、平安・仮名文学の助動詞「つ」と上接動詞の組み合わせのすべてについてダイス係数を計算する場合には、次のような手順で行うことになる。

- ① 「中納言」を用いて、検索対象をサブコーパス「平安・仮名文学」に限定して助動詞「つ」の直前に来る動詞を検索する（次の検索条件式）。

キー：品詞 LIKE "動詞%"

AND 後方共起：(語彙素="つ" AND 品詞 LIKE "助動詞%") ON 1 WORDS FROM キー
IN subcorpusName="平安-仮名文学" AND core="true"

- ② ①の検索結果をピボットテーブルで語彙素 ID（動詞）別に集計する。
（例えば動詞「見る」（語彙素 ID：36920）の場合、109 例）
- ③ 語彙表をピボットテーブルで「平安・仮名文学」に限定して語彙素 ID で集計する。
（フィルタで「平安・仮名文学」に限定、列に「語彙素 ID」、値に「頻度」を指定）
集計結果を②のファイルの別シートにコピーしておく。
- ④ ③のシート等で助動詞「つ」（語彙素 ID：24321）の頻度を確認する（3170 例）。
- ⑤ ②のシートから VLOOKUP 関数で語彙素 ID を使って③のシートの各動詞の頻度を参照する。（例えば動詞「見る」（語彙素 ID：36920）の場合、5298 例）
- ⑥ 以上から取得できる数字を用いて、ダイス係数を計算する。

（例えば「見る」と「つ」のダイス係数は $2 \times (109 / (5298 + 3170)) \approx 0.02574$ ）

すべての動詞について上記の方法で参照するように計算式をコピーすれば、一度に全ての動詞についてダイス係数を計算することができる。

このように、中納言の検索結果と語彙表の集計結果を組み合わせることで、個別の動詞ごとに検索して頻度を調べたり、旧語彙表の作品別の頻度を合計したりすることなく、一度に多数の組み合わせのダイス係数が計算できるわけである。T スコアや MI スコアなどコーパスサイズを必要とする場合には別途、語数表も参照する必要があるが、ほとんど変わらない手間で計算が可能になる。

5. おわりに

『日本語歴史コーパス』は、多くのユーザーに利用されるようになったが、その多くが用例検索にとどまり、コーパス言語学で用いられる各種指標が用いられることが少ないように見受けられる。その理由の一端は、語彙表の形式が作品レベルに留まっているために、コーパス全体や、各時代の中での調査対象語の位置づけが見通しづらかったことが挙げられるだろう。この新しい形式の語彙表により『日本語歴史コーパス』の一層の活用の一助となることを願うものである。

参考文献

- 国立国語研究所編（2021）『日本語歴史コーパス』 <https://ccd.ninjal.ac.jp/chj/> （2021 年 9 月 20 日確認）
- 国立国語研究所編（2020）『日本語歴史コーパス』短単位語彙表バージョン 2020.03, <http://doi.org/10.15084/00003258>

謝辞

本研究は、国立国語研究所共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」の成果の一部であり、また JSPS 科研費 20K20411 の助成を受けたものです。