『日本語話し言葉コーパス』語彙表解説

『日本語話し言葉コーパス』語彙表 ver.2018.03　解説

2018.3.1 作成

１．データの概要

本データは『日本語話し言葉コーパス』（以下、CSJ と略す。）の中納言データ Ver. 2018.3.1 に基づく語彙表である。

CSJ 全体および CSJ を構成する各音声のタイプおよびコアデータについて頻度 1 までの見出し語を収録した。コアデータ、非コア・人手修正、非コア・自動解析の区別があるレジスターについては、コアデータ、非コア・人手修正、非コア・自動解析それぞれの語彙表も作成した。また、品詞構成と語種構成に関する集計表もあわせて公開する。

なお、本語彙表は CSJ および『現代日本語書き言葉均衡コーパス』（BCCWJ）を元にした語彙表として出版した"A Frequency Dictionary of Japanese"(Routledge,2013)とは対象も集計の方法も別のものであるので注意されたい。

音声のタイプと個々の語彙表との関係は以下のとおりである。

表１　音声のタイプと語彙表の種類

音声のタイプ	語彙表の種類
独話・学会	コアデータ、非コア・人手修正、非コア・自動解析
独話・模擬	コアデータ、非コア・人手修正、非コア・自動解析
独話・朗読	非コア・自動解析
独話・再朗読	コアデータ、非コア・自動解析
独話・その他	非コア・人手修正、非コア・自動解析
対話・学会	コアデータ、非コア・自動解析
対話・模擬	コアデータ、非コア・自動解析
対話・課題	非コア・自動解析
対話・自由	非コア・自動解析

２．集計方法

(1)語彙素、語彙素読み、品詞、語彙素細分類、語種の 5 つの組で見出し語を特定した。長単位は、語彙素、語彙素読み、品詞、語種の 4 つの組で見出し語を特定した。

(2)(1)で得られた見出し語の集合から以下の条件に該当するものを除外した。

1) 品詞に「空白」「補助記号」「記号」の文字列を含むもの。

2) 語彙素が空(null)のもの（この場合、語彙素読みも同時に空になっている）。

(3)上記の集計方法は、CSJ-USB のマニュアル及び「中納言」に記載されているものとは異なる方法であるため、レジスターの語数はそれらとは一致しない。

(4)非コア・自動解析の CSJ は誤解析を含む。そのため、本語彙表のデータも同様にエラーを含んでいる。

３．語彙表の見方

３．１　CSJ 短単位語彙表

・ファイル名：CSJ_frequencylist_suw_ver201803.tsv

・42,543 行、UTF8、タブ区切り。

・第 1 行目は見出し。2 行目以降がデータである。各行には以下の表 3 に示す 92 の項目が並んでい

る。

・pmw（100 万語当たりの頻度）は、小数点以下第 7 位まで示した。

・同順位の語があった場合は、語彙素読み、語彙素、品詞の順に文字コード昇順で並べた。

・Excel 等のソフトに読み込んで、目的の列で並べ替えれば、音声のタイプ別の語彙表を得ることができる。

表 2　語彙表の各項目

番号	見出し	備考
1	rank	CSJ 全体の順位
2	lForm	語彙素読み
3	lemma	語彙素
4	pos	品詞
5	subLemma	語彙素細分類
6	wType	語種
7	frequency	CSJ 全体の頻度
8	pmw	CSJ 全体での 100 万語当たりの頻度
9	独話・学会_rank	独話・学会の順位
10	独話・学会_frequency	独話・学会の頻度
11	独話・学会_pmw	独話・学会全体での 100 万語当たりの頻度
12	独話・模擬_rank	独話・模擬の順位
13	独話・模擬_frequency	独話・模擬の頻度
14	独話・模擬_pmw	独話・模擬全体での 100 万語当たりの頻度
15	独話・朗読_rank	独話・朗読の順位
16	独話・朗読_frequency	独話・朗読の頻度
17	独話・朗読_pmw	独話・朗読全体での 100 万語当たりの頻度
18	独話・再朗読_rank	独話・再朗読の順位
19	独話・再朗読_frequency	独話・再朗読の頻度
20	独話・再朗読_pmw	独話・再朗読全体での 100 万語当たりの頻度
21	独話・その他_rank	独話・その他の順位
22	独話・その他_frequency	独話・その他の頻度
23	独話・その他_pmw	独話・その他全体での 100 万語当たりの頻度
24	対話・学会_rank	対話・学会の順位
25	対話・学会_frequency	対話・学会の頻度
26	対話・学会_pmw	対話・学会全体での 100 万語当たりの頻度
27	対話・模擬_rank	対話・模擬の順位
28	対話・模擬_frequency	対話・模擬の頻度
29	対話・模擬_pmw	対話・模擬全体での 100 万語当たりの頻度

30	対話・課題_rank	対話・課題の順位
31	対話・課題_frequency	対話・課題の頻度
32	対話・課題_pmw	対話・課題全体での 100 万語当たりの頻度
33	対話・自由_rank	対話・自由の順位
34	対話・自由_frequency	対話・自由の頻度
35	対話・自由_pmw	対話・自由全体での 100 万語当たりの頻度
36	独話・学会_コア_rank	独話・学会、コアの順位
37	独話・学会_コア_frequency	独話・学会、コアの頻度
38	独話・学会_コア_pmw	独話・学会、コア全体での 100 万語当たりの頻度
39	独話・学会_非コア・人手修正_rank	独話・学会、非コア・人手修正の順位
40	独話・学会_非コア・人手修正_frequency	独話・学会、非コア・人手修正の頻度
41	独話・学会_非コア・人手修正_pmw	独話・学会、非コア・人手修正全体での 100 万語当たりの頻度
42	独話・学会_非コア・自動解析_rank	独話・学会、非コア・自動解析の順位
43	独話・学会_非コア・自動解析_frequency	独話・学会、非コア・自動解析の頻度
44	独話・学会_非コア・自動解析_pmw	独話・学会、非コア・自動解析全体での 100 万語当たりの頻度
45	独話・模擬_コア_rank	独話・模擬、コアの順位
46	独話・模擬_コア_frequency	独話・模擬、コアの頻度
47	独話・模擬_コア_pmw	独話・模擬、コア全体での 100 万語当たりの頻度
48	独話・模擬_非コア・人手修正_rank	独話・模擬、非コア・人手修正の順位
49	独話・模擬_非コア・人手修正_frequency	独話・模擬、非コア・人手修正の頻度
50	独話・模擬_非コア・人手修正_pmw	独話・模擬、非コア・人手修正での 100 万語当たりの頻度
51	独話・模擬_非コア・自動解析_rank	独話・模擬、非コア・自動解析の順位
52	独話・模擬_非コア・自動解析_frequency	独話・模擬、非コア・自動解析の頻度
53	独話・模擬_非コア・自動解析_pmw	独話・模擬、非コア・自動解析全体での 100 万語当たりの頻度
54	独話・朗読_非コア・自動解析_rank	独話・模擬、非コア・自動解析の順位
55	独話・朗読_非コア・自動解析_frequency	独話・模擬、非コア・自動解析の頻度
56	独話・朗読_非コア・自動解析	独話・模擬、非コア・自動解析での 100 万語当たりの頻度

	_pmw	
57	独話・再朗読_コア_rank	独話・再朗読、コアの順位
58	独話・再朗読_コア_frequency	独話・再朗読、コアの頻度
59	独話・再朗読_コア_pmw	独話・再朗読、コア全体での 100 万語当たりの頻度
60	独話・再朗読_非コア・自動解析_rank	独話・再朗読・非コア・自動解析の順位
61	独話・再朗読_非コア・自動解析_frequency	独話・再朗読・非コア・自動解析の頻度
62	独話・再朗読_非コア・自動解析_pmw	独話・再朗読・非コア・自動解析全体での 100 万語当たりの頻度
63	独話・その他_非コア・人手修正_rank	独話・その他、非コア・人手修正の順位
64	独話・その他_非コア・人手修正_frequency	独話・その他、非コア・人手修正の頻度
65	独話・その他_非コア・人手修正_pmw	独話・その他、非コア・人手修正全体での 100 万語当たりの頻度
66	独話・その他_非コア・自動解析_rank	独話・その他、非コア・自動解析の順位
67	独話・その他_非コア・自動解析_frequency	独話・その他、非コア・自動解析の頻度
68	独話・その他_非コア・自動解析_pmw	独話・その他、非コア・自動解析全体での 100 万語当たりの頻度
69	対話・学会_コア_rank	対話・学会、コアの順位
70	対話・学会_コア_frequency	対話・学会、コアの頻度
71	対話・学会_コア_pmw	対話・学会、コア全体での 100 万語当たりの頻度
72	対話・学会_非コア・自動解析_rank	対話・学会、非コア・自動解析の順位
73	対話・学会_非コア・自動解析_frequency	対話・学会、非コア・自動解析の頻度
74	対話・学会_非コア・自動解析_pmw	対話・学会、非コア・自動解析全体での 100 万語当たりの頻度
75	対話・模擬_コア_rank	対話・模擬、コアの順位
76	対話・模擬_コア_frequency	対話・模擬、コアの頻度
77	対話・模擬_コア_pmw	対話・模擬、コア全体での 100 万語当たりの頻度
78	対話・模擬_非コア・自動解析_rank	対話・模擬、非コア・自動解析の順位
79	対話・模擬_非コア・自動解析_frequency	対話・模擬、非コア・自動解析の頻度
80	対話・模擬_非コア・自動解析_pmw	対話・模擬、非コア・自動解析全体での 100 万語当たりの頻度
81	対話・課題_コア_rank	対話・課題、コアの順位

82	対話・課題_コア_frequency	対話・課題、コアの頻度
83	対話・課題_コア_pmw	対話・課題、コア全体での 100 万語当たりの頻度
84	対話・課題_非コア・自動解析_rank	対話・課題、非コア・自動解析の順位
85	対話・課題_非コア・自動解析_frequency	対話・課題、非コア・自動解析の頻度
86	対話・課題_非コア・自動解析_pmw	対話・課題、非コア・自動解析全体での 100 万語当たりの頻度
87	対話・自由_非コア・自動解析_rank	対話・自由、非コア・自動解析の順位
88	対話・自由_非コア・自動解析_frequency	対話・自由、非コア・自動解析の頻度
89	対話・自由_非コア・自動解析_pmw	対話・自由、非コア・自動解析全体での 100 万語当たりの頻度
90	core_rank	コアデータにおける順位
91	core_frequency	コアデータにおける頻度
92	core_pmw	コアデータにおける 100 万語当たりの頻度

・短単位の場合のレジスター等の語数を表 4、表 5 に示す。

表 3　短単位の語数（延べ語数）

音声のタイプ	コア	非コア・人手修正	非コア・自動解析
独話・学会	21,4171	280,077	2,760,920
独話・模擬	222,413	208,096	3,161,572
独話・朗読			156,502
独話・再朗読	18,319		30,447
独話・その他		14,227	265,085
対話・学会	15,265		13,203
対話・模擬	14,439		27,984
対話・課題	10,900		18,711
対話・自由			47,352

CSJ 全体	7,479,773
コアデータ	495,597

表 4　短単位の語数（異なり語数）

音声のタイプ	コア	非コア・人手修正	非コア・自動解析
独話・学会	6,034	7,287	21,152
独話・模擬	8,879	9,111	30,804
独話・朗読			1,058
独話・再朗読	1,255		2,608
独話・その他		1,181	7,724

対話・学会	1,007		1,013
対話・模擬	1,298		2,016
対話・課題	960		1,444
対話・自由			2,590

CSJ 全体	42,542
コアデータ	12,399

4．CSJ 品詞構成表

・ファイル名：CSJ_frequencylist_pos_ver201803.tsv
・72 行、UTF8、タブ区切り。
・以下の 4 個の表を納めた。
(1)短単位における品詞の語数（延べ語数）
(2)短単位における品詞の語数（異なり語数）
(3)短単位における品詞の割合（延べ語数）
(4)短単位における品詞の割合（異なり語数）
・いずれの表も第 1 行目は見出し。2 行目以降がデータである。列は、CSJ 全体、各音声のタイプ、各音声タイプのコアデータ、非コア・人手修正、非コア・自動解析、全体のコアデータの順に並んでいる。
・品詞の割合（百分率）は小数点以下第 3 位まで示した。

5．CSJ 語種構成表

・ファイル名：CSJ_frequencylist_wtype_ver201803.tsv
・40 行、9.40KB、UTF8、タブ区切り。
・CSJ 品詞構成表と同様に 8 個の表を納めた。表の種類は品詞構成表と同じ。
・いずれの表も第 1 行目は見出し。2 行目以降がデータである。列は、CSJ 全体、各音声のタイプ、各音声タイプのコアデータ、非コア・人手修正、非コア・自動解析、全体のコアデータの順に並んでいる。
・語種の割合（百分率）は小数点以下第 3 位まで示した。

6．利用上の注意

(1)研究、教育目的であれば無償で自由に利用できる。申し込みの必要はない。
(2)再配布は不可。商業使用（営利目的での利用）は要相談。
(3)論文等に引用する際は出典とバージョンを明記すること。以下に、出典とバーションの例を示す。
『日本語話し言葉コーパス』短単位語彙表 ver. 2018.3.1
『日本語話し言葉コーパス』品詞構成表 ver. 2018.3.1
『日本語話し言葉コーパス』語種構成表 ver. 2018.3.1
(4)本データの著作権（編集著作権）は国立国語研究所が有する。
(5)データの瑕疵による損害についてはいかなる場合でも補償しない。
(6)内容の改善のため予告なく更新することがある。

本データに関する問い合わせ先：kotonoha@ninjal.ac.jp　（@を半角に変えること）

以上

更新履歴

2017.3.1　CSJ 中納言 データ Vers. 2018.3.1 を公開