

国立国語研究所学術情報リポジトリ

『現代日本語書き言葉均衡コーパス』語彙表解説

| | |
|-----|---|
| 著者 | 国立国語研究所 コーパス開発センター |
| URL | http://doi.org/10.15084/00003282 |



1. データの概要

本データは『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と略す。）の中納言データ Vers. 1.1(NumTrans 版)に基づく語彙表である。

BCCWJ 全体およびBCCWJを構成する各レジスター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。固定長データと可変長データの区別があるレジスターについては、固定長、可変長それぞれの語彙表も作成した。以上の語彙表を短単位と長単位の 2 つの場合について作成した。語彙表の種類は合計 50 個である。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 BCCWJ および『日本語話し言葉コーパス』（CSJ）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。

レジスターと個々の語彙表との関係は以下のとおりである。

表 1 レジスターと語彙表の種類

| レジスター(略称) | 語彙表の種類 |
|--------------------|--------------|
| 出版・書籍(PB) | 固定長、可変長、統合形式 |
| 出版・雑誌(PM) | 固定長、可変長、統合形式 |
| 出版・新聞(PN) | 固定長、可変長、統合形式 |
| 図書館・書籍(LB) | 固定長、可変長、統合形式 |
| 特定目的・白書(OW) | 固定長、可変長、統合形式 |
| 特定目的・教科書(OT) | 可変長 |
| 特定目的・広報紙(OP) | 可変長 |
| 特定目的・ベストセラー(OB) | 可変長 |
| 特定目的・Yahoo!知恵袋(OC) | 可変長 |
| 特定目的・Yahoo!ブログ(OY) | 可変長 |
| 特定目的・韻文(OV) | 可変長 |
| 特定目的・法律(OL) | 可変長 |
| 特定目的・国会会議録(OM) | 可変長 |

統合形式とは、重複のないように固定長と可変長をあわせたものである。

表 2 全体およびコアデータについて

| | |
|----------|--|
| BCCWJ 全体 | PM～OW は統合形式、OT～OM は可変長で集計 |
| コアデータ | PB、PM、PN、OW、OC、OY から合計約 100 万語を抽出し、人手のチェックを経て解析精度を高めたデータ |

2. 集計方法

(1)短単位は、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、BCCWJ-DVD 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)BCCWJ は約 2%の誤解析を含む。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 BCCWJ 短単位語彙表

- ・ファイル名：BCCWJ_frequencylist_suw_ver1_1.tsv
- ・185,137 行、73MB、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 80 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、レジスター別の語彙表を得ることができる。Excel のバージョンは 2007 以降でないと全行読み込めないで注意されたい。

表 3 語彙表の各項目

| 番号 | 見出し | 備考 |
|----|--------------|--------------------------|
| 1 | Rank | BCCWJ 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 | BCCWJ 全体の頻度 |
| 8 | Pmw | BCCWJ 全体での 100 万語当たりの頻度 |
| 9 | PB_rank | 出版・書籍における順位 |
| 10 | PB_frequency | 出版・書籍における頻度 |
| 11 | PB_pmw | 出版・書籍における 100 万語当たりの頻度 |
| 12 | PM_rank | 出版・雑誌における順位 |
| 13 | PM_frequency | 出版・雑誌における頻度 |
| 14 | PM_pmw | 出版・雑誌における 100 万語当たりの頻度 |
| 15 | PN_rank | 出版・新聞における順位 |
| 16 | PN_frequency | 出版・新聞における頻度 |
| 17 | PN_pmw | 出版・新聞における 100 万語当たりの頻度 |
| 18 | LB_rank | 図書館・書籍における順位 |
| 19 | LB_frequency | 図書館・書籍における頻度 |
| 20 | LB_pmw | 図書館・書籍における 100 万語当たりの頻度 |
| 21 | OW_rank | 特定目的・白書における順位 |
| 22 | OW_frequency | 特定目的・白書における頻度 |
| 23 | OW_pmw | 特定目的・白書における 100 万語当たりの頻度 |
| 24 | OT_rank | 特定目的・教科書における順位 |

| | | |
|----|-----------------------|----------------------------------|
| 25 | OT_frequency | 特定目的・教科書における頻度 |
| 26 | OT_pmw | 特定目的・教科書における 100 万語当たりの頻度 |
| 27 | OP_rank | 特定目的・広報紙における順位 |
| 28 | OP_frequency | 特定目的・広報紙における頻度 |
| 29 | OP_pmw | 特定目的・広報紙における 100 万語当たりの頻度 |
| 30 | OB_rank | 特定目的・ベストセラーにおける順位 |
| 31 | OB_frequency | 特定目的・ベストセラーにおける頻度 |
| 32 | OB_pmw | 特定目的・ベストセラーにおける 100 万語当たりの頻度 |
| 33 | OC_rank | 特定目的・Yahoo!知恵袋における順位 |
| 34 | OC_frequency | 特定目的・Yahoo!知恵袋における頻度 |
| 35 | OC_pmw | 特定目的・Yahoo!知恵袋における 100 万語当たりの頻度 |
| 36 | OY_rank | 特定目的・Yahoo!ブログにおける順位 |
| 37 | OY_frequency | 特定目的・Yahoo!ブログにおける頻度 |
| 38 | OY_pmw | 特定目的・Yahoo! ブログにおける 100 万語当たりの頻度 |
| 39 | OV_rank | 特定目的・韻文における順位 |
| 40 | OV_frequency | 特定目的・韻文における頻度 |
| 41 | OV_pmw | 特定目的・韻文における 100 万語当たりの頻度 |
| 42 | OL_rank | 特定目的・法律における順位 |
| 43 | OL_frequency | 特定目的・法律における頻度 |
| 44 | OL_pmw | 特定目的・法律における 100 万語当たりの頻度 |
| 45 | OM_rank | 特定目的・国会会議録における順位 |
| 46 | OM_frequency | 特定目的・国会会議録における頻度 |
| 47 | OM_pmw | 特定目的・国会会議録における 100 万語当たりの頻度 |
| 48 | PB_fixed_rank | 出版・書籍・固定長における順位 |
| 49 | PB_fixed_frequency | 出版・書籍・固定長における頻度 |
| 50 | PB_fixed_pmw | 出版・書籍・固定長における 100 万語当たりの頻度 |
| 51 | PB_variable_rank | 出版・書籍・可変長における順位 |
| 52 | PB_variable_frequency | 出版・書籍・可変長における頻度 |
| 53 | PB_variable_pmw | 出版・書籍・可変長における 100 万語当たりの頻度 |
| 54 | PM_fixed_rank | 出版・雑誌・固定長における順位 |
| 55 | PM_fixed_frequency | 出版・雑誌・固定長における頻度 |
| 56 | PM_fixed_pmw | 出版・雑誌・固定長における 100 万語当たりの頻度 |
| 57 | PM_variable_rank | 出版・雑誌・可変長における順位 |
| 58 | PM_variable_frequency | 出版・雑誌・可変長における頻度 |
| 59 | PM_variable_pmw | 出版・雑誌・可変長における 100 万語当たりの頻度 |
| 60 | PN_fixed_rank | 出版・新聞・固定長における順位 |
| 61 | PN_fixed_frequency | 出版・新聞・固定長における頻度 |
| 62 | PN_fixed_pmw | 出版・新聞・固定長における 100 万語当たりの頻度 |
| 63 | PN_variable_rank | 出版・新聞・可変長における順位 |

| | | |
|----|-----------------------|------------------------------|
| 64 | PN_variable_frequency | 出版・新聞・可変長における頻度 |
| 65 | PN_variable_pmw | 出版・新聞・可変長における 100 万語当たりの頻度 |
| 66 | LB_fixed_rank | 図書館・書籍・固定長における順位 |
| 67 | LB_fixed_frequency | 図書館・書籍・固定長における頻度 |
| 68 | LB_fixed_pmw | 図書館・書籍・固定長における 100 万語当たりの頻度 |
| 69 | LB_variable_rank | 図書館・書籍・可変長における順位 |
| 70 | LB_variable_frequency | 図書館・書籍・可変長における頻度 |
| 71 | LB_variable_pmw | 図書館・書籍・可変長における 100 万語当たりの頻度 |
| 72 | OW_fixed_rank | 特定目的・白書・固定長における順位 |
| 73 | OW_fixed_frequency | 特定目的・白書・固定長における頻度 |
| 74 | OW_fixed_pmw | 特定目的・白書・固定長における 100 万語当たりの頻度 |
| 75 | OW_variable_rank | 特定目的・白書・可変長における順位 |
| 76 | OW_variable_frequency | 特定目的・白書・可変長における頻度 |
| 77 | OW_variable_pmw | 特定目的・白書・可変長における 100 万語当たりの頻度 |
| 78 | core_rank | コアデータにおける順位 |
| 79 | core_frequency | コアデータにおける頻度 |
| 80 | 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 4 短単位の語数（延べ語数）

| レジスター(略称) | 固定長 | 可変長 | 統合形式 |
|--------------------|-----------|------------|------------|
| 出版・書籍(PB) | 6,363,435 | 27,039,539 | 28,450,509 |
| 出版・雑誌(PM) | 1,157,252 | 4,196,696 | 4,424,572 |
| 出版・新聞(PN) | 930,600 | 877,202 | 1,369,772 |
| 図書館・書籍(LB) | 6,685,183 | 28,828,228 | 30,307,622 |
| 特定目的・白書(OW) | 1,041,559 | 4,712,324 | 4,880,892 |
| 特定目的・教科書(OT) | | 924,939 | |
| 特定目的・広報紙(OP) | | 3,750,468 | |
| 特定目的・ベストセラー(OB) | | 3,737,668 | |
| 特定目的・Yahoo!知恵袋(OC) | | 10,235,490 | |
| 特定目的・Yahoo!ブログ(OY) | | 10,125,783 | |
| 特定目的・韻文(OV) | | 223,181 | |
| 特定目的・法律(OL) | | 1,079,083 | |
| 特定目的・国会会議録(OM) | | 5,102,439 | |

| | |
|----------|-------------|
| BCCWJ 全体 | 104,612,418 |
| コアデータ | 1,097,933 |

表 5 短単位の語数（異なり語数）

| レジスター(略称) | 固定長 | 可変長 | 統合形式 |
|-----------|-----|-----|------|
|-----------|-----|-----|------|

| | | | |
|--------------------|--------|---------|---------|
| 出版・書籍(PB) | 82,393 | 123,139 | 125,772 |
| 出版・雑誌(PM) | 42,552 | 63,506 | 65,244 |
| 出版・新聞(PN) | 35,308 | 34,390 | 41,581 |
| 図書館・書籍(LB) | 85,618 | 126,923 | 129,309 |
| 特定目的・白書(OW) | 16,005 | 27,357 | 27,763 |
| 特定目的・教科書(OT) | | 27,372 | |
| 特定目的・広報紙(OP) | | 37,391 | |
| 特定目的・ベストセラー(OB) | | 55,574 | |
| 特定目的・Yahoo!知恵袋(OC) | | 61,530 | |
| 特定目的・Yahoo!ブログ(OY) | | 78,537 | |
| 特定目的・韻文(OV) | | 18,419 | |
| 特定目的・法律(OL) | | 5,106 | |
| 特定目的・国会会議録(OM) | | 27,840 | |

| | |
|----------|---------|
| BCCWJ 全体 | 185,136 |
| コアデータ | 36,649 |

3. 2 BCCWJ 長単位語彙表

- ・ファイル名：BCCWJ_frequencylist_luw_ver1_1.tsv
- ・2,434,722行、972MB、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。各項目は上記の表3に同じである。ただし、語彙素細分類は、長単位にその属性がないため、列はあるが、値はすべて空 (null) になっている。
- ・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。
- ・このデータは現行の Excel では全行を読み込むことができない。
- ・長単位の場合の各レジスター等の語数を表6、表7に示した。

表6 長単位の語数 (延べ語数)

| レジスター(略称) | 固定長 | 可変長 | 統合形式 |
|--------------------|-----------|------------|------------|
| 出版・書籍(PB) | 5,080,050 | 21,644,025 | 22,767,278 |
| 出版・雑誌(PM) | 903,144 | 3,286,053 | 3,461,006 |
| 出版・新聞(PN) | 675,469 | 646,596 | 997,074 |
| 図書館・書籍(LB) | 5,510,362 | 23,821,192 | 25,031,768 |
| 特定目的・白書(OW) | 659,831 | 2,991,194 | 3,098,691 |
| 特定目的・教科書(OT) | | 742,686 | |
| 特定目的・広報紙(OP) | | 2,303,793 | |
| 特定目的・ベストセラー(OB) | | 3,182,019 | |
| 特定目的・Yahoo!知恵袋(OC) | | 8,591,787 | |
| 特定目的・Yahoo!ブログ(OY) | | 8,217,362 | |
| 特定目的・韻文(OV) | | 200,866 | |
| 特定目的・法律(OL) | | 706,250 | |
| 特定目的・国会会議録(OM) | | 4,007,806 | |

| | |
|----------|------------|
| BCCWJ 全体 | 83,308,386 |
| コアデータ | 836,849 |

表 7 長単位の語数（異なり語数）

| レジスター(略称) | 固定長 | 可変長 | 統合形式 |
|--------------------|---------|---------|---------|
| 出版・書籍(PB) | 322,656 | 842,108 | 879,475 |
| 出版・雑誌(PM) | 108,370 | 250,972 | 263,925 |
| 出版・新聞(PN) | 97,918 | 90,480 | 128,438 |
| 図書館・書籍(LB) | 303,208 | 786,586 | 821,025 |
| 特定目的・白書(OW) | 77,951 | 217,352 | 222,744 |
| 特定目的・教科書(OT) | | 59,746 | |
| 特定目的・広報紙(OP) | | 213,818 | |
| 特定目的・ベストセラー(OB) | | 139,732 | |
| 特定目的・Yahoo!知恵袋(OC) | | 289,954 | |
| 特定目的・Yahoo!ブログ(OY) | | 439,121 | |
| 特定目的・韻文(OV) | | 27,171 | |
| 特定目的・法律(OL) | | 18,174 | |
| 特定目的・国会会議録(OM) | | 125,808 | |

| | |
|----------|-----------|
| BCCWJ 全体 | 2,434,721 |
| コアデータ | 92,103 |

3. 3 BCCWJ 長単位語彙表（頻度 2 以上）

- ・ファイル名：BCCWJ_frequencylist_luw2_ver1_2.tsv
- ・841,977 行、338MB、UTF8、タブ区切り。
- ・頻度 2 以上の語にしぼった以外は 3. 2 の語彙表と同じ。
- ・Excel 等のソフトに読み込んで、目的の列で並べ替えれば、レジスター別の語彙表を得ることができる。Excel のバージョンは 2007 以降でないと全行読み込めないで注意されたい。

4. BCCWJ 品詞構成表

- ・ファイル名：BCCWJ_frequencylist_pos_ver1_1.tsv
- ・132 行、21KB、UTF8、タブ区切り。
- ・以下の 8 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）
 - (5)長単位における品詞の語数（延べ語数）
 - (6)長単位における品詞の語数（異なり語数）
 - (7)長単位における品詞の割合（延べ語数）
 - (8)長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、BCCWJ 全体、各レジスター、

固定長、可変長、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. BCCWJ 語種構成表

- ・ファイル名：BCCWJ_frequencylist_wtype_ver1_1.tsv
- ・80行、14KB、UTF8、タブ区切り。
- ・BCCWJ 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、BCCWJ 全体、各レジスター、固定長、可変長、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、出典とバージョンの例を示す。
 - 『現代日本語書き言葉均衡コーパス』短単位語彙表 ver.1.1
 - 『現代日本語書き言葉均衡コーパス』長単位語彙表 ver.1.1
 - 『現代日本語書き言葉均衡コーパス』品詞構成表 ver.1.1
 - 『現代日本語書き言葉均衡コーパス』語種構成表 ver.1.1
- (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

2013.02.22 作成

2015.11.18 BCCWJ 短単位語彙表、BCCWJ 長単位語彙表、BCCWJ 長単位語彙表（頻度2以上）のファイルサイズを実際のサイズに変更した。

2017.12.15 BCCWJ 中納言 データ Vers. 1.1 (NumTrans 版) により再作成