Statistical Information of the NINJAL Corpora : BCCWJ Frequency Table, Vocabulary Table, etc.

計量国語学 32 巻 1 号　（2019 年 6 月）　pp.33-40.

研究資料

国立国語研究所コーパスの統計情報

―BCCWJ 語数表・語彙表ほか―

山崎　誠（国立国語研究所）

要旨

本稿は，国立国語研究所のコーパス開発センターで公開しているコーパスのうち，Web 検索ツール「中納言」に搭載されているコーパスの統計情報について紹介するものである．具体的に取り上げるコーパスは，『現代日本語書き言葉均衡コーパス』『日本語話し言葉コーパス』『日本語歴史コーパス』の 3 つである．

キーワード：語数表，語彙表，現代日本語書き言葉均衡コーパス，日本語話し言葉コーパス，日本語歴史コーパス

1．はじめに

2000 年代に入って，『日本語話し言葉コーパス』（以下，CSJ），『現代日本語書き言葉均衡コーパス』（以下，BCCWJ），『日本語歴史コーパス』（以下，CHJ）など，本格的な研究に利用できるコーパスが着々とリリースされ，コーパスの利用は人文系の日本語研究において一般的なものとなっている．コーパスの利用を大いに促進したものの一つが Web 検索ツール「中納言」[1]である．それまでは，多くのデータを扱う研究はプログラミングの技術が必須であったが，用意されたツールを使うことで一気にコーパスを利用するハードルが低くなったからである．中納言はコンコーダンサーの 1 つで，検索結果として，キーとなる検索語を中心に前後文脈が表示されるというものである．

2．コーパスの語数表

2.1　語数表の利用

コーパスを利用する目的の一つに，目的の語や用法がどれくらい用いられているか，あるいは，各ジャンル[2]における分布はどうなっているかなどの，量的な傾向の把握がある．ここで注意したいのは，中納言では粗頻度（raw frequency）しか示されないという点である．表 1 は，BCCWJ で語彙素「矢張り」を短単位で検索した結果をレジスターと語形[3]

[1] 無償だが，利用には登録が必要．

[2] BCCWJ ではレジスターと呼んでいる．ただし，ジャンルとレジスターを厳密に使い分けているわけではない．

[3] 語形とは，BCCWJ の形態素解析に用いた電子化辞書 UniDic における，語彙素の下位区分で，例えば，語彙素「矢張り」には，語形「ヤハリ」「ヤッパリ」「ヤッパシ」「ヤッパ」などが含まれている．

により頻度分布をまとめたものである．作成には Excel のピボットテーブルを利用した．

表 1　BCCWJ における語彙素「矢張り」の語形の頻度分布（粗頻度）

レジスター	ヤハリ	ヤッパリ	ヤッパシ	ヤッパ	ヤパ	ヤパリ
出版・雑誌	565	511	2	52	0	0
出版・書籍	3,510	1,637	4	131	0	0
出版・新聞	72	27	0	1	0	0
図書館・書籍	4,833	2,375	31	96	0	1
特定目的・ブログ	2,129	3,161	38	853	35	0
特定目的・ベストセラー	759	364	7	13	0	0
特定目的・韻文	8	10	0	0	0	0
特定目的・教科書	42	42	0	0	0	0
特定目的・広報誌	34	19	0	0	0	0
特定目的・国会会議録	5,032	1,041	1	0	0	0
特定目的・知恵袋	2,855	2,186	13	360	0	0
特定目的・白書	37	0	0	0	0	0
特定目的・法律[4]	0	0	0	0	0	0
計	19,876	11,373	96	1,506	35	1

表 1 の各レジスターにおける頻度を単純に比較することはできない．それぞれのレジスターの延べ語数が異なるからである．正しい比較のためには，各レジスターの総延べ語数を知り，相対頻度を出さなければならない．そこで，コーパス開発センターでは BCCWJ を構成する各サンプルの語数を示した語数表を作成し公開している．この語数表を使うことにより，検索結果の相対頻度を知ることができる．表 2 は相対頻度によるものである．

表 2　BCCWJ における語彙素「矢張り」の語形の頻度分布（相対頻度（PMW））

レジスター	ヤハリ	ヤッパリ	ヤッパシ	ヤッパ	ヤパ	ヤパリ
出版・雑誌	127.1	115.0	0.4	11.7	0.0	0.0
出版・書籍	122.9	57.3	0.1	4.6	0.0	0.0
出版・新聞	52.5	19.7	0.0	0.7	0.0	0.0
図書館・書籍	159.1	78.2	1.0	3.2	0.0	0.0
特定目的・ブログ	208.8	310.1	3.7	83.7	3.4	0.0
特定目的・ベストセラー	202.8	97.3	1.9	3.5	0.0	0.0
特定目的・韻文	35.5	44.4	0.0	0.0	0.0	0.0
特定目的・教科書	45.2	45.2	0.0	0.0	0.0	0.0
特定目的・広報誌	9.1	5.1	0.0	0.0	0.0	0.0
特定目的・国会会議録	986.2	204.0	0.2	0.0	0.0	0.0
特定目的・知恵袋	278.3	213.1	1.3	35.1	0.0	0.0
特定目的・白書	7.6	0.0	0.0	0.0	0.0	0.0
特定目的・法律	0.0	0.0	0.0	0.0	0.0	0.0
計	2235.3	1189.4	8.7	142.4	3.4	0.0

[4] 特定目的・法律には「矢張り」が 1 例もなかったため，Excel のピボットテーブルでは作成されないが，あとから追加した．

なお，表 2 の数値は単純に度数を延べ語数で割ると，少数点以下の桁数が多くなるため，100 万語当たりの調整頻度（PMW）で表示している．

2.2　語数表の構成

BCCWJ 語数表は, Excel ファイルで, 短単位および長単位の 2 種類がある. 公開の URL は本稿の末尾に示した[5]．フィールド構成は，短単位，長単位とも表 3 のようになっている．いずれのファイルも先頭行は見出し行で，総行数は 172,676 である．これらの情報を使えば，例えば，厚生・医療というジャンルに属する雑誌のサンプル数が 35 であり，その延べ語数が 65,991 語（短単位，記号等除外，全て）である，というようなことが分かる．

表 3　BCCWJ 語数表の構成

フィールド名	内容
サンプル ID	サンプルを一意に特定する ID.
レジスター	サンプルの属するレジスター.
レジスター（略）	レジスターをアルファベット 2 文字で表した略称.
コア	当該サンプルがコアデータ[6]の場合 1．それ以外は 0.
生年代	執筆者[7]の生年代．10 年刻み.
性別	執筆者の性別.
ジャンル	図書分類（NDC），C コード[8]などを示す.
出版年	サンプルの出版年.
語数（固定長）	固定長部分の語数で，除外なし.
語数（可変長）	可変長部分の語数で，除外なし.
語数（全て）	固定長と可変長を足し，重複を除いた語数．除外なし.
語数（記号等除外・固定長）	品詞欄が補助記号，記号，空白となっているものを除外した場合の語数.
語数（記号等除外・可変長）	
語数（記号等除外・全て）	

CSJ の語数表は，現在短単位のみであり，そのフィールド構成は表 4 のようになっている．先頭行は見出し行で，総行数は 3,361 である．この語数表を使うと，例えば学会講演で，20-24 歳の男性によるサンプルが 171 個あり，その延べ語数が 449,364 語（記号等除外の場合）であることが分かる．

[5] 公開ページには，BCCWJ の旧バージョン（1.0）に対応する語数表もあるが，現在中納言で検索できるのは 1.1 なので，1.1 のほうを選択するとよい.
[6] コアデータとは，人手修正を加えて形態素解析の精度を高くしたデータである．人手修正なしの場合，約 98%の精度であるが，コアデータの精度は約 99%となっている.
[7] 書籍全体の著者と該当サンプルの著者が異なる場合があるため，サンプルに採られた部分の著者を執筆者として区別している.
[8] C コードとは，日本図書コードのことで，書籍の流通などに用いられている.

表 4　CSJ 語数表の構成

フィールド名	内容
講演 ID	各講演を一意に特定する ID.
音声のタイプ	独話・対話，学会講演・模擬講演等の区別.
コア	コア・非コアの区別．人手修正・自動解析の区別.
講演者 ID	講演者を一意に特定する ID.
性別	講演者の性別.
収録時の年齢	講演者の収録時の年齢．5 歳刻み.
生年代	講演者の生年代．5 歳刻み.
語数（全て）	短単位の語数．除外なし.
語数（記号等除外・全て）	品詞欄が補助記号，記号，空白となっているものを除外した場合の語数.

CHJ の語数表は，短単位と長単位が公開されている．そのフィールド構成は表 5 のようになっている．先頭行は見出し行で，総行数は 23,250 である．CHJ は現在も追加され続けているため，この数字は今後増えることが予定されている．

表 5　CHJ 語数表の構成

フィールド名	内容
サブコーパス名	奈良，平安，鎌倉，室町，江戸，明治・大正の別.
サンプル ID	サンプルを一意に特定する ID.
コア	当該サンプルがコアデータの場合は 1．それ以外は 0.
本文種別	引用，歌，ト書きなどの別.
文体	韻文，文語，漢文，散文，口語，外国語などの別.
ジャンル	歌集，歌物語，紀行，随筆などの別.
作品名	作品の名称.
成立年	作品の成立年.
巻名等	作品が巻別になっている場合の巻.
部	作品の部の名称.
作者	作者の名前.
生年	作者の生年.
性別	作者の性別.
語数（全て）	短単位（あるいは長単位）の語数．除外なし.
語数（記号等除外・全て）	品詞欄が補助記号，記号，空白となっているものを除外した場合の語数.

３．コーパスの語彙表

3.1　BCCWJ 語彙表

BCCWJ が公開された際に期待されたことの一つに現在の日本語の語彙についての統計的な情報があった．例えば，語彙頻度表や語種の構成（外来語は何パーセントくらいを占めるのかなど）である．教育等への応用で利用価値が高いものに語彙頻度表（以下，語彙表）がある．語彙表はかつての国語研究所の語彙調査でも成果としてたびたび作成されてきた．それらは，単一のレジスターを対象としたものであったが，BCCWJ のような複数のレジスターから構成されるデータの語彙表はこれが初めてであろう．BCCWJ 語彙表は関係するファイルを含めると以下の 5 種類がある．長単位語彙表が 2 つあるのは，全体で 240 万行以上と大きなファイルなので，Excel に読み込むことができないことを考慮して頻度 2 以上のファイル（約 84 万行）を別に用意したためである．

1. BCCWJ 語彙表解説
2. BCCWJ 短単位語彙表（Version 1.1）
3. BCCWJ 長単位語彙表（Version 1.1）
4. BCCWJ 長単位語彙表（頻度 2 以上）（Version 1.1）
5. BCCWJ 品詞構成表（Version 1.1）
6. BCCWJ 語種構成表（Version 1.1）

BCCWJ 語彙表は最初 2013 年に公開され，2017 年にバージョン 1.1 に対応したデータに更新された．BCCWJ 語彙表は以下の集計方法に拠っている．以下，「『現代日本語書き言葉均衡コーパス』語彙表 ver1.1 解説」より引用する．

> (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）。

上記の（1）で示された集計方法は，UniDic の語彙素 ID を使って集計した場合と異なることに注意が必要である．例えば，語彙素「余り」は，品詞が副詞の場合と形状詞の場合とがある．上記（1）の集計ではこれらは別語となる．しかし，両者は同じ語彙素 ID を持っているため，語彙素 ID で集計した場合は，これらは同じ語となる．

また，BCCJW 短単位語彙表に収められた語の頻度の総計は 104,612,418 語（補助記号等除外）であり，BCCWJ 語数表で得られる総語数 104,911,460 語と比べると，約 30 万語ほど少なくなっている．これは，上記（2）の 2）で語彙素が空（null）のもの[9]を差し引いているためである．

[9] その多くは品詞欄が，未知語，カタカナ文，URL，英単語，漢文，言いよどみ，方言などとなっているものである．

3.2　語彙表の構成

BCCWJ 語彙表の構成は以下の表 6[10]のとおりである．BCCWJ 全体での頻度，順位，PMW のほかに，それぞれのレジスターにおいても，頻度，順位，PMW を示しているので，レジスターごとの傾向を見ることができる．

表 6　BCCWJ 語彙表の構成

番号	見出し	備考
1	rank	BCCWJ 全体の順位[11]
2	lForm	語彙素読み
3	lemma	語彙素[12]
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（PMW）
…		以下，各レジスターにおける，順位，頻度，100 万語当たりの頻度（PMW）
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

3.3　品詞構成表および語種構成表

BCCWJ および CSJ では，コーパスにおける品詞および語種の構成を示した集計表を公開している．BCCWJ の場合，品詞は，「名詞，代名詞，動詞，形容詞，形状詞，連体詞，副詞，接続詞，感動詞，助詞，助動詞，接頭辞，接尾辞」について，延べ語数での粗頻度と割合，異なり語数でも粗頻度と割合を示している．これらの集計結果はコーパス全体だけでなく，レジスターごとにも用意されている．語種についても同様に，「和語，漢語，外来語，混種語，固有名」などについて集計している．語種に関しては過去の語彙調査と比べて外来語の割合が増えているかという問いがあるが，BCCWJ における解析単位と過去の語彙調査の解析単位は語の認定基準に違いがあり，かつ，サンプリングの方法も異なる

[10] 表 6 は，「BCCWJ 語彙表解説」の表 3 に変更を加えたものである．

[11] 語彙表の順位は同じ頻度の場合，同じ順位を付けている．例えば，「確実」「特殊」「同士」はそれぞれ BCCWJ における頻度が 5,030 であるため，1,881 という順位になる．同じ頻度の語が 3 つあるため，1,882，1,883 は欠番となり，次の語の順位は 1,884 になる．

[12] 語彙素および語彙素読みに■を含む語が全体で 155 レコードあるが，その多くは人名，住所などの伏せ字処理により生じたものである．

ことから厳密な比較は残念ながらできない.

４．データの利用について

本稿で紹介した語数表，語彙表などのデータは基本的に研究，教育目的であれば無償で自由に利用でき，申し込みの必要はない．中納言ユーザでなくても使うことができる．利用に当たっては注意事項などがあるので,念のため HP やドキュメント等を参照されたい.

参考 URL

中納言

https://chunagon.ninjal.ac.jp/

現代日本語書き言葉均衡コーパス（BCCWJ）

短単位語数表

https://pj.ninjal.ac.jp/corpus_center/bccwj/bcc-chu-suw.html

長単位語数表

https://pj.ninjal.ac.jp/corpus_center/bccwj/bcc-chu-luw.html

語彙表，品詞構成表，語種構成表

https://pj.ninjal.ac.jp/corpus_center/bccwj/bcc-chu.html

日本語話し言葉コーパス（CSJ）

語彙表，語数表，品詞構成表，語種構成表（いずれも短単位）

https://pj.ninjal.ac.jp/corpus_center/csj/chunagon.html#data

日本語歴史コーパス（CHJ）

語彙表，語数表

https://pj.ninjal.ac.jp/corpus_center/chj/chj-wc.html

（2019 年 4 月 24 日受付）

Mathematical Linguistics, Vol.32 No.1 (June 2019) pp.33-40.

Resource

Statistical Information of the NINJAL Corpora: BCCWJ Frequency Table, Vocabulary Table, etc.

YAMAZAKI Makoto (National Institute for Japanese Language and Linguistics)

Abstract:
This paper reports the statistical information of the corpora developed by the Center for Corpus Development and installed in *Chunagon*. The corpora introduced in this article are Balanced Corpus of Contemporary Written Japanese (BCCSJ), Corpus of Spontaneous Japanese (CSJ) and Corpus of Historical Japanese (CHJ).

Keywords: frequency table, vocabulary table, Balanced Corpus of Contemporary Written Japanese, Corpus of Spontaneous Japanese, Corpus of Historical Japanese