

コーパスに基づく 語彙指導を目指して

第15回モンゴル日本語教育シンポジウム

2023/11/05(日)

森篤嗣(武庫川女子大学)

コーパスとは

The body of written or spoken material upon which a linguistic analysis is based. (言語分析に使われる書き言葉または話し言葉の資料の集積)

The Oxford English Dictionary 1956 (second edition, 1989)

言語分析のために, 分析対象となる言語またはさまざまな言語変種を代表するように収集され, コンピュータ処理可能な状態にされた実際に話されたり書かれたりしたテキストの集合体

齊藤他(2005)

日本語教科書の歴史

- 2012年 『みんなの日本語 第二版』
- 1998年 『みんなの日本語』
- 1990年 『新・日本語の基礎』
- 1974年 『日本語の基礎』

→約50年前に日本語教育関係者の経験と勘で作られたのが原型

『現代日本語書き言葉均衡コーパス

(BCCWJ; Balanced Corpus of Contemporary Written Japanese)』

- 対象は出版物として刊行された現代日本語の書き言葉です。
- 従来から語彙調査の対象となってきた新聞、雑誌に加えて、書籍全般を対象にします。白書や教科書も対象とします。
- 非公開の日記や私信の類は対象としません。
- WEB上の文書についてはQ&A 掲示板のテキストなど一部を対象とします。
- 収録対象の刊行年代は、最大30年間(1976～2005)です。メインとなる書籍の場合は、1986～2005年になります。これはサンプリングのために必要とされるISBN(国際標準図書番号)が普及するようになった時期を対象としているためであって、言語学的な基準によるものではありません。

KOTONOHA 計画のウェブページ(http://www.ninjal.ac.jp/kotonoha/ex_8.html)

『BCCWJ』と「NLB」「中納言」の関係



日本語教育・授業にコーパスを活かす

- **コロケーション**に注目しよう！
- コロケーション＝「よく使われる語の組み合わせや、自然な語のつながり」李ほか(2018:8)

- 風____()
- 風邪____()

- 教えるときの例文やテスト作成に役立つ！



【問題】「飛ぶ」の何でしょうか？

- NLB (NINJAL-LWP for BCCWJ) <http://nlb.ninjal.ac.jp/>

The screenshot displays the NLB interface for the verb '飛ぶ' (to fly). The main search results table is as follows:

コロケーション	頻度	MI	LD
...	59	7.13	5.85
...	41	10.04	8.25
...	38	2.06	0.83
...	31	7.88	6.47
...	30	0.84	-0.39
...	22	6.95	5.58
...	21	10.32	8.00
...	19	9.91	7.73
...	19	7.80	6.29
...	18	7.90	6.36
...	18	4.97	3.70
...	15	7.48	5.97
...	14	11.86	8.06
...	13	9.40	7.20
...	12	8.67	6.74
...	12	8.36	6.54
...	11	10.77	7.56
...	11	4.73	3.45
...	11	8.00	6.26
...	11	9.34	7.05
...	10	8.56	6.57
...	10	6.10	4.70

The left sidebar shows the following filters:

- 名詞+助詞: ...が飛ぶ (1,429), ...は飛ぶ (446), ...も飛ぶ (151), ...の飛ぶ (63), ...を飛ぶ (649), ...に飛ぶ (1,036), ...へ飛ぶ (355), ...で飛ぶ (395), ...と飛ぶ (44), ...から飛ぶ (284), ...まで飛ぶ (101), ...より飛ぶ (7)
- 名詞+複合助詞: ...として飛ぶ (2), ...に応じて飛ぶ (1), ...に沿って飛ぶ (5), ...について飛ぶ (1), ...に向かって飛ぶ (7)

The right sidebar shows additional statistics and a list of example sentences:

- 声が飛ぶ 全59件
- 図・書誌: 1.02
- 出・書誌: 0.72
- 雑誌: 0.43
- ベスト: 0.73
- ブログ: 0.06

At the bottom, the footer reads: NINJAL-LWP for BCCWJ Copyright © 2012-2022 National Institute for Japanese Language and Linguistics, Lingo Institute of Language. All rights reserved.

「～が飛ぶ」ランキング

コロケーション	コーパス全体			コロケーション	コーパス全体		
	頻度	MI	LD		頻度	MI	LD
声が飛ぶ	59	7.13	5.85	記憶が飛ぶ	15	7.48	5.97
飛行機が飛ぶ	41	10.04	8.25	打球が飛ぶ	14	11.86	8.06
【一般】が飛ぶ	38	2.06	0.83	鳥が飛ぶ	13	9.40	7.20
首が飛ぶ	31	7.88	6.47	水分が飛ぶ	12	8.67	6.74
【人名】が飛ぶ	30	0.84	-0.39	球が飛ぶ	12	8.36	6.54
機が飛ぶ	22	6.95	5.58	カモメが飛ぶ	11	10.77	7.56
花粉が飛ぶ	21	10.32	8.00	物が飛ぶ	11	4.73	3.45
矢が飛ぶ	19	9.91	7.73	虫が飛ぶ	11	8.00	6.26
鳥が飛ぶ	19	7.80	6.29	蝶が飛ぶ	11	9.34	7.05
ボールが飛ぶ	18	7.90	6.36	弾が飛ぶ	10	8.56	6.57
話が飛ぶ	18	4.97	3.70	意識が飛ぶ	10	6.10	4.70
記憶が飛ぶ	15	7.48	5.97				

名詞「心」のコロケーション

心
頻度=34,724

グループ別

パターン頻度順

基本

>> 助詞+動詞

パターン	頻度	比率
○ 心が...	2,539	
○ 心は...	1,013	
○ 心も...	488	
○ 心の...	431	
○ 心を...	7,165	
○ 心に...	4,524	
○ 心へ...	17	
○ 心で...	509	
○ 心と...	191	
○ 心から...	193	
○ 心まで...	71	
○ 心より...	170	

動詞 <<

パターン	頻度	比率
○ 動詞基本形+心	1,126	
○ 動詞連用形+心	63	

心を... 1,040種類

コロケーション	コーパス全体		
	頻度	MI	LD
心をいる	413	10.75	10.29
心をする	310	22.17	10.47
心を持つ	304	5.87	6.85
心を開く	207	7.47	8.09
心を動かす	179	9.10	8.92
心を奪う	145	9.39	8.83
心を表す	124	9.50	8.70
心を与える	122	9.49	8.68
心を打つ	116	7.02	7.51
心をつかむ	101	8.44	8.16
心を痛める	96	11.14	8.65
心を決する	95	11.17	8.64
心を砕く	87	11.32	8.54
心を癒す	72	9.32	8.05
心をとらえる	71	8.19	7.74
心をいく	70	5.73	6.40
心を許す	69	7.26	7.33
心をしまう	64	9.08	7.87
心を決める	63	6.19	6.67

心をいる 全413件

■ 韻文: 12.90

■ 出・書籍: 4.91

■ 雑誌: 1.73

■ 広報: 0.49

■ ベスト: 9.68

■ ブログ: 2.50

■ 知恵袋: 1.06

■ 白書: 0.42

■ 図・書籍: 5.24

■ 教科書: 1.92

■ 国会: 0.98

私は心を決めている。
(清川妙著『出会いのときめき』, 2002, 914)

人の心を持っていない。
(Yahoo!知恵袋, 2005, 政治、社会問題)

つまり心を病んでいる。
(徳永進著『ホスピス通りの四季』, 2002, 490)

「ひとの心を持っている」
(赤城毅著『帝都探偵物語』, 2001, 913)

千田さんが心を痛めていた
(久美沙織著『修道女マリコ』, 1994, 913)

子供のような心を持っている
(Yahoo!ブログ, 2008, Yahoo!ブログ)

彼女はすでに心を決めていた。

1 ページ / 11
1 ページ / 5
413件中 1-100を表示

「心をいる」はおかしくない？

心をいる 全413件

韻文: 12.90	ベスト: 9.68	図・書籍: 5.24
出・書籍: 4.91	ブログ: 2.50	教科書: 1.92
雑誌: 1.73	知恵袋: 1.06	国会: 0.98
広報: 0.49	白書: 0.42	

私は心を決めている。
(清川妙著 『出会いのときめき』, 2002, 914)

人の心を持っていない。
(Yahoo!知恵袋, 2005, 政治、社会問題)

つまり心を病んでいる。
(徳永進著 『ホスピス通りの四季』, 2002, 490)

「ひとの心を持っている」
(赤城毅著 『帝都探偵物語』, 2001, 913)

千田さんが心を痛めていた
(久美沙織著 『修道女マリコ』, 1994, 913)

子供のような心を持っている
(Yahoo!ブログ, 2008, Yahoo!ブログ)

彼女はすでに心を決めていた。

ここをクリックするとダウンロード可能

1 ページ / 5 100 413件中 1-100を表示

2語比較「とても」と「めっちゃ」

NINJAL-LWP for BCCWJ → 見出し語検索

めっちゃ とても 絞り込み 元に戻す

2語比較

名詞 動詞 形容詞・連体詞 **副詞** オノマトペ

	見出し	読み	ローマ字表記	頻度
<input checked="" type="checkbox"/>	とても	トテモ	totemo	17,964
<input checked="" type="checkbox"/>	めっちゃ	メツチャ	meccha	1,122

1 ページ / 1 100 2件中 1 - 2を表示

NINJAL-LWP for BCCWJ Copyright © 2012-2022 National Institute for Japanese Language and Linguistics, Lago Institute of Language. All rights reserved.

とても > めっちゃ

とても		めっちゃ	
頻度=17,964		頻度=1,122	
グループ別			
パターン頻度順			
基本			
??			
» 動詞			
とても		めっちゃ	
パターン	頻度	パターン	頻度
とても+動詞	2831	めっちゃ+動詞	249
とてもと+動詞	5		
» 形容詞			
とても		めっちゃ	
パターン	頻度	パターン	頻度
とても+形容詞	5042	めっちゃ+形容詞	379

サブコーパス(ジャンル)頻度

サブコーパス			
	とても	めっちゃ	
出版サブコーパス			
書籍	3840 130.92	17 0.58	
雑誌	811 175.31	27 5.84	
図書館サブコーパス			
書籍	4800 152.42	16 0.51	
特定目的サブコーパス			
ベストセラー	815 197.3	10 2.42	
知恵袋	3283 289.76	165 14.56	

法律	0	0	0	0
国会会議録	270	53.08	0	0
広報紙	296	72.05	3	0.73
教科書	60	57.59	0	0
韻文	22	94.62	0	0
白書	10	2.1	0	0
ブログ	3757	312.6	884	73.55

MIスコアやLD (LogDice係数) に注目

とても+動詞 618		めっちゃ+動詞 141						
とても				めっちゃ				LD差
コロケーション	頻度	MI	LD	コロケーション	頻度	MI	LD	
とてもできる	162	5.01	4.75					4.75
とても言う	120	4.02	3.77	めっちゃ言う	2	1.62	-2.12	5.89
とても思える	110	8.92	8.28					8.28
とても喜ぶ	100	9.34	8.54	めっちゃ喜ぶ	3	7.79	3.99	4.55
とても信じる	76	8.27	7.66					7.66
とても言える	69	6.82	6.42					6.42
とても考える	69	4.98	4.69	めっちゃ考える	1	2.37	-1.37	6.06
とても分かる	56	4.73	4.44	めっちゃ分かる	2	3.43	-0.32	4.76
とても困る	48	7.67	7.04	めっちゃ困る	2	6.59	2.81	4.23
とても楽しむ	48	7.77	7.12	めっちゃ楽しむ	8	8.69	4.91	2.21
とても疲れる	42	8.44	7.52	めっちゃ疲れる	11	10.02	6.20	1.32
とても食べる	41	5.67	5.31	めっちゃ食べる	4	5.82	2.07	3.24
とても感謝する	32	9.32	7.78					7.78
とても心配する	32	8.35	7.31					7.31
とても驚く	32	7.14	6.50					6.50
とても感動する	31	9.78	7.92					7.92
とても悩む	29	7.96	7.02	めっちゃ悩む	3	8.20	4.38	2.64
とても楽する	27	10.79	8.02					8.02
とても助かる	27	9.30	7.62	めっちゃ助かる	5	10.37	6.43	1.19
とてもかなう	26	9.20	7.56					7.56
とても役立つ	26	8.67	7.33					7.33
とても役に立つ	25	8.90	7.40					7.40

「NLB」と「中納言」

- 動詞や名詞などの実質語のコロケーションであれば「NLB」で調べられる(お手軽！)
- ただ、日本語教育で「文型」と呼ばれる機能語については、「NLB」で調べられないので「中納言」+エクセルを使う必要がある(ちょっと面倒)
- 機能語についても、中俣(2014)や中俣(2017)を参照すると、文法項目の「よく使われるパターン」がまとめられている(自分で検索や集計をしなくても「その語はどのように使われるのか」がわかる)

日本語の形態素解析

- 「NLB」にしても「中納言」にしても、コーパスデータを扱うには、まず**日本語を単語に分割**する必要がある
- 見出し語から調べたい単語を選ぶ「NLB」の場合はあまり問題にならないが、自分で条件を指定して検索をする「中納言」の場合は、形態素解析ツールの「くせ」を知っておく必要がある。
- [Web茶まめ](#)では、形態素解析辞書UniDic(国立国語研究所)による形態素解析が手軽に体験できる。

形態素解析ツール

- [Web茶まめ](#): 形態素解析器 (MeCab) と形態素解析辞書 (UniDic) によって稼動するGUI (Graphical User Interface) ツール → **任意のテキストに使える**
- 中納言: 既に形態素解析を終えたデータを検索するツール → **BCCWJのみしか使えない**

	茶まめ	中納言
対象テキスト	何でも	BCCWJのみ
長単位	×	○
短単位	○	○

文・文節・長単位・短単位

- 文：固有名詞に関する論文を執筆した。
- 文節：固有名詞に／関する／論文を／執筆した／。
- 長単位：固有名詞／に関する／論文／を／執筆し／た／。
- 短単位：固有／名詞／に／関する／論文／を／執筆／し／た／。

小澤俊介ほか(2014)より

UniDic (短単位解析) の設計方針

①設計方針

【階層的な見出し構造】

異表記・異形態に対して同一の見出しを与える構造

辞書見出し*

語彙素読み：
オオキイ
語彙素：
大きい
語彙素類：
相

語形
(基本形)

オオキイ

オッキイ

書字形・発音形
(基本形)

書	大きい
書	おおきい
音	オーキー
書	おっきい
音	オッキー

* 語彙素読み, 語彙素, 語彙素類を合わせてはじめて当該の階層構造が一意に定まる。

【斉一な言語単位】

単位認定が、ある場合は長く、ある場合は短いという不揃いをなくす

斉一でない言語単位で分割した例：

「梅雨\空」が2単位に分割されるのに対し、「青空」は1単位として扱われる。

梅雨\空\の\間\の\青空\だ\。

斉一な言語単位で分割した例：

梅雨空|の|間|の|青空|だ|。

実装

②UniDicデータベース



エクスポート

③解析用UniDic (本サイトにて配布)

書字形 (基本形)	発音形 (基本形)	語形 (基本形)	語彙素読み	語彙素	...
大きい	オーキー	オオキイ	オオキイ	大きい	...
おおきい	オーキー	オオキイ	オオキイ	大きい	...
おっきい	オッキー	オッキイ	オオキイ	大きい	...

https://clrd.ninjal.ac.jp/unidic/about_unidic.html

【問題】どのように形態素解析される でしょうか？

- 自動車
- 三角形
- 公民館
- 法案
- 平仮名
- スーパーマーケット
- ティッシュペーパー

Web茶まめ
各種UniDicを使用した形態素解析ツール

更新履歴 Web茶まめについて 国立国語研究所 サイトポリシー

使い方

- テキストボックスに入力した文章か、ファイル選択ボタンからアップロードしたテキストデータに対してMeCabによる形態素解析を行います。
- テキストを入力後、解析前処理の有無/形態素解析に使用する辞書/出力項目/出力形式を選択し一番下にある「解析する」ボタンを押すと形態素解析が実行されます。

テキストを入力

クリア

自動車
三角形
公民館
法案
平仮名
スーパーマーケット
ティッシュペーパー

テキストファイルから解析

ファイルを選択

- ✓複数のファイルを選択してアップロードできます。
- ▲CSV形式でダウンロードする場合のテキストデータの容量は、1ファイルにつき5MBまでです。
- ▲辞書、出力項目を増やすと出力ファイルのデータサイズが大幅に増加しますので、ご注意ください。

こうなります

- 自動／車
- 三角／形
- 公民／館
- 法／案
- 平仮名
- スーパーマーケット
- ティッシュ／ペーパー

出力結果

辞書	文境界	書字形 (=表層形)	語彙素	語彙素読み	品詞	活用型	活用形	発音形出現形	仮名形出現
現代語B		自動	自動	ジドウ	名詞-普通名詞-一般			ジドー	ジドウ
現代語I		車	車	シャ	接尾辞-名詞的-一般			シャ	シャ
現代語B		三角	三角	サンカク	名詞-普通名詞-形状詞可能			サンカク	サンカク
現代語I		形	形	ケイ	接尾辞-名詞的-一般			ケー	ケイ
現代語B		公民	公民	コウミン	名詞-普通名詞-一般			コーミン	コウミン
現代語I		館	館	カン	接尾辞-名詞的-一般			カン	カン
現代語B		法	法	ホウ	名詞-普通名詞-一般			ホー	ホウ
現代語I		案	案	アン	名詞-普通名詞-一般			アン	アン
現代語B		平仮名	平仮名	ヒラガナ	名詞-普通名詞-一般			ヒラガナ	ヒラガナ
現代語B		スーパーマーケット	スーパーマーケット-supermarket	スーパーマーケット	名詞-普通名詞-一般			スーパーマーケット	スーパーマーケ
現代語B		ティッシュ	ティッシュ-tissue	ティッシュ	名詞-普通名詞-一般			ティッシュ	ティッシュ
現代語I		ペーパー	ペーパー-paper	ペーパー	名詞-普通名詞-一般			ペーパー	ペーパー

旧JLPT語彙表の形態素解析結果(森2016)

	語数	1語	2語	3語	4語	5語	6語	計	割合
1級	2,933	2,811	108	10	4	0	0	122	4.16%
2級	3,736	3,457	206	49	17	4	3	279	7.47%
3級	725	603	115	6	1	0	0	122	16.83%
4級	735	641	84	10	0	0	0	94	12.79%
	8,129	7,512	513	75	22	4	3	617	7.59%

- 下位級において分割される語の割合が多い
- サ変動詞や「～さん」など, 特定の法則を除いた実数の調査が必要
- 外来語も別立てする必要があるか?
- 【問題】5語や6語に分割された語の正体は?

5語や6語に分割された語の例

「いつ／の／ま／に／か」

「お／また／せ／し／まし／た」

「お／せわ／に／なり／まし／た」

「お／まち／ど／お／さま」

「ご／ちそう／さま／でし／た」

「それ／は／いけ／ませ／ん／ね」

「よく／,／いらっしやい／まし／た」

いわゆる定型表現が多く、単語という概念を逸脱している。しかし、「日本母語話者の直観に沿い、日本語教育に役立つ」という視点に立てば、「お待たせしました」を動詞「待つ」としてカウントするよりも、「お待たせしました」として抽出できる方がメリットがあるのではないか。このあたりはポリシーの問題である。

文の解析における品詞付け処理

選択辞書:

- 現代語

入力文字列:

すもももももものうち。

出力結果

辞書	文境界	書字形 (=表層形)	語彙素	語彙素読み	品詞	活用型	活用形	発音形出現形	仮名形出現形	語種	書字形(基本形)	語形(基本形)
現代語	B	すもも	李	スモモ	名詞-普通名詞-一般			スモモ	スモモ	和	すもも	スモモ
現代語	I	も	も	モ	助詞-係助詞			モ	モ	和	も	モ
現代語	I	もも	桃	モモ	名詞-普通名詞-一般			モモ	モモ	和	もも	モモ
現代語	I	も	も	モ	助詞-係助詞			モ	モ	和	も	モ
現代語	I	もも	桃	モモ	名詞-普通名詞-一般			モモ	モモ	和	もも	モモ
現代語	I	の	の	ノ	助詞-格助詞			ノ	ノ	和	の	ノ
現代語	I	うち	内	ウチ	名詞-普通名詞-副詞可能			ウチ	ウチ	和	うち	ウチ
現代語	I	。	。		補助記号-句点					記号	。	

「中納言」を使ってみよう！

まとめて検索

KOTONOHA

書字形出現形で検索

書字形出現形で検索 語彙素で検索

この検索窓では、『まとめて検索 KOTONOHA』を使い、あなたが利用申請した（【検索対象】が灰色でない）コーパスをまとめて一括検索できます。検索は短単位の書字形出現形と語彙素の2種類が選択できます。検索ボタンをクリックすると、KOTONOHAに移動し、そこに検索結果が表示されます。

【検索対象】 [現代日本語書き言葉均衡コーパス](#) [国語研日本語ウェブコーパス](#) [日本語話し言葉コーパス](#) [日本語日常会話コーパス](#) [昭和話し言葉コーパス](#) [名大会話コーパス](#) [現日研・職場談話コーパス](#)
[日本語歴史コーパス](#) [日本語方言コーパス](#) [中国語・韓国語母語の日本語学習者継断発話コーパス](#) [多言語母語の日本語学習者横断コーパス](#)

【個別検索】 ご利用になりたいコーパス名をクリックしてください。

コーパス名	略称	個別検索	包括的検索	備考
書き言葉 現代日本語書き言葉均衡コーパス 中納言版	BCCWJ	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	従来より利用いただいている BCCWJ のデータです（コーパスの紹介ページ）。こちらのページから BCCWJ アノテーションデータをダウンロードできます。
書き言葉 国語研日本語ウェブコーパス 中納言版	NWJC	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
話し言葉 日本語話し言葉コーパス	CSJ	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	コーパスの紹介ページ
話し言葉 日本語日常会話コーパス	CEJC	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	コーパスの紹介ページ 有償版契約者は関連データを「データ配布」からダウンロードできます。

文字列検索

現代日本語書き言葉均衡コーパス BCCWJ ↑コーパス選択画面 ログアウト

中納言 コーパス検索アプリケーション 中納言 2.7.2 データバージョン 2021.03 マニュアル 語数について

現在のサーバ負荷状況: 現在 41 人ログイン中

ここで切り替える

短単位検索 | 長単位検索 | **文字列検索** | 位置検索

文字列検索 検索フォームで検索 | 検索条件式で検索 | 履歴で検索

検索文字列 | 対象文字列 NumTrans 処理を施したテキスト 原文文字列 | **結果表示単位 短単位 長単位**

キー表示形式 語 文字列全体

検索文字列では、特別な意味を持つメタ文字として以下に挙げる文字を使用できます。

※	任意の文字列	例: 国立 ※ 所 →「国立国語研究所」「国立社会保障・人口問題研究所」など
_	任意の1文字	例: 日 _ 友好 →「日米友好」「日中友好」「日韓友好」など
[]	文字クラス	例: 日本 [人国語] →「日本人」「日本国」「日本語」
[^]	文字クラスの否定	例: 日本 [^ 人国語] →「日本海」「日本史」「日本は」「日本の」など

ここに注目!

文字列検索の罫

13,107 件の検索結果が見つかりました。そのうち 500 件を表示しています。

検索対象語数: 124,100,964

記号・補助記号・空白を除いた検索対象語数: 104,911,460

検索対象サンプル数: 172,675

表示している(新)NDCの出典: 国立国会図書館書誌データ(<https://iss.ndl.go.jp/information/api/> 2019年4月に取得)に基づき、『現代日本語書き言葉均衡コーパス』書籍サンプルの NDC を再確認した。

サンプル ID	開始位置	連番	前文脈	キ	後文脈	語彙素読み	語彙素	語彙素細分類	品詞	活用型	活用形	レジスター	執筆者	書名/出典	編著者等	出版者	出版年	話者名
PN1b_00030	250	160	ママの 社会 進出 応援 #北千住 駅ビル に認 証 保育 所 # 東京	都	が 独自 に 制度 化 した 認証 保育 所 の 第1 号 園 が 1日 、 足立区	ト	都		名詞- 普通名 詞一般			出版・ 新聞	毎日新聞社 (著)	毎日新聞		毎日新聞 社	2001	
OP69_00004	78300	49180	いただいた おかげで す 。 # 今後 は、 この 問題 を 踏まえ て、 地 元 の 反対 運動 が	京都	府 まで 広ま った こと など 、 地域 活動 の 参考 に し て いただけ たら と 思い ます 。	キョウ ト	キョウ ト		名詞- 固有名 詞-地 名一般			特定の 目的・広 報紙		広報京丹波		京都府船 井郡京丹 波町	2008	
OM41_00004	160160	101740	がたい 事実 であるとい うふう に 思 っておりま す 。 # 五十嵐 (は い) 委員 # 東京	都	が その 三百 億 を 出さ ない という こと に な った 場 合 、 いろい ろ 支 障 が 出 て	ト	都		名詞- 普通名 詞一般			特定の 目的・国 会会議 録		国会会議録			1995	

文字列検索で形態素区切りを見る

346 件の検索結果が見つかりました。

検索対象語数: 124,100,964

記号・補助記号・空白を除いた検索対象語数: 104,911,460

検索対象サンプル数: 172,675

表示している(新)NDCの出典:国立国会図書館書誌データ(<https://iss.ndl.go.jp/information/api/> 2019年4月に取得)に基づき、『現代日本語書き言葉均衡コーパス』書籍サンプルの NDC を再確認した。

サンプルID	開始位置	連番	前文脈	キー	後文脈	語彙素読み	語彙素	語彙素細分類	品詞	活用型	活用形	レジスター	執筆者	書名/出典	編著者等	出版者	出版年	話者名
OY03_09598	1370	780	面倒な ので これが 夕飯 代わり でし た。#今日 は 何か 作ら ない といけ ませ	ん	ね。#パスタ か ラーメン か? #それ とも ・ ・ #つ いで に、 どなた でも かまい	ズ	ず		助動 詞	助動 詞-又	終止 形-撥 音便	特定 目的・ ブログ		Yahoo! ブログ		Yahoo!	2008	
LBr7_00008	20950	13200	責める わけ に はいきま せん。#もっ とみんな が 早く 三浦 君 を 捕まえ な い といけま せ	ん	でした。#6 月 二十 五 日 横浜 戦 く 東京 ド ム) #【勝】 前田	ズ	ず		助動 詞	助動 詞-又	終止 形-撥 音便	図書 館・書 籍	原 辰徳 (著)	ジャイア ンツ愛	原辰徳 著	幻冬舎	2003	

文字列検索で形態素区切りを見る

OY03_09598	1280	730	作ら	ツクル	作る		動詞一般	五段-ラ行	未然形一般	ツクラ	和	作ら	特定目的・ブログ	0	0	1
OY03_09598	1300	740	ない	ナイ	ない		助動詞	助動詞-ナイ	終止形一般	ナイ	和	ない	特定目的・ブログ	0	0	1
OY03_09598	1320	750	と	ト	と		助詞-接続助詞			ト	和	と	特定目的・ブログ	0	0	1
OY03_09598	1330	760	いけ	イク	行く		動詞-非自立可能	下一段-カ行	連用形一般	イケ	和	いけ	特定目的・ブログ	0	0	1
OY03_09598	1350	770	ませ	マス	ます		助動詞	助動詞-マス	未然形一般	マセ	和	ませ	特定目的・ブログ	0	0	1
OY03_09598	1370	780	ん	ズ	ず		助動詞	助動詞-又	終止形-撥音便	ン	和	ん	特定目的・ブログ	0	0	1
OY03_09598	1380	790	ね	ネ	ね		助詞-終助詞			ネ	和	ね	特定目的・ブログ	0	0	1
OY03_09598	1390	800	。		。		補助記号-句点				記号	。	特定目的・ブログ	0	0	1

短単位検索方法「～ないといけません」

短単位検索

検索フォームで検索 検索条件式で検索 履歴で検索

▼ 前方共起条件の追加

キー 1 語 キーの条件を指定しない ⓘ

品詞 の 大分類 が 動詞 短単位の条件の追加

後方共起1 キーから 1 語 キーと結合して表示 ⓘ

語彙素 が ない 短単位の条件の追加

後方共起2 キーから 2 語 キーと結合して表示 ⓘ

語彙素 が と 短単位の条件の追加

後方共起3 キーから 3 語 キーと結合して表示 ⓘ

語彙素 が 行く 短単位の条件の追加

後方共起4 キーから 4 語 キーと結合して表示 ⓘ

語彙素 が ます 短単位の条件の追加

後方共起5 キーから 5 語 キーと結合して表示 ⓘ

語彙素 が ず 短単位の条件の追加

▲ 後方共起条件の追加

短単位検索結果「～ないといけません」

316 件の検索結果が見つかりました。

検索対象語数: 124,100,964

記号・補助記号・空白を除いた検索対象語数: 104,911,460

検索対象サンプル数: 172,675

検索ヒット数が500件を超える場合、検索結果からランダムで選ばれた500件が表示されます。その500件の選択は完全ランダムであり、再現性はありません。

↓

ダウンロードすれば10万件まで取得可能

サンプルID	開始位置	連番	前文脈	キー	後文脈	類	動詞-一般	五段-ラ行	未然形-一般	特定目的-ブログ	図書館-書籍	著者	書名	出版社	発行年	
OY15_12214	510	350	ところ 肝心の お 魚さん に ま 遊ん でもらっ て ませ ん が 、 これ で 少し ま	頑張ら	(ない) (と) (い) (け) (ま) (せ) (ん) (ね) (ん) (ご ;) # ガンディー に の プロト ル ア ー です 。 # えっ ? # 関 こ 市販 さ	ガンバル	頑張る	動詞-一般	五段-ラ行	未然形-一般	特定目的-ブログ	Yahoo!ブログ	Yahoo!	2008		
LB08_00014	7970	5170	はい け ない と 、 母 親 は 深 く 反 省 し た と い い ま す 。 # これ は 気 を	つけ	(ない) (と) (い) (け) (ま) (せ) (ん) 。 # 子ども は 、 母 親 の 言 葉 を そ の ま ま 受 け と り ま す 。 # 自 分 は	ツケル	付ける	動詞-非自立可能	下一段-力行	未然形-一般	図書館-書籍	市堀 艶子 (著)	続・心にしみる言葉のごちそう	主婦の友社	2000	
LBn5_00010	5480	3250	が き か な い こ と が 充 分 に 予 想 さ れ ま す 。 # 本 業 以 外 に 自 分 の フ ィ ールド を	持た	(ない) (と) (い) (け) (ま) (せ) (ん) 。 # あ の 人 は 「 わ け の 分 か ら な い 人 」 だ と 思 わ れる	モツ	持つ	動詞-一般	五段-タ行	未然形-一般	図書館-書籍	牧野 昇(著)	創造のちから	牧野昇 著	エイチアンドアイ	1999

短単位検索結果のダウンロード

検索対象 ⓘ 設定を隠す

検索対象を選択 検索対象をクリア

全て

検索動作 ⓘ 設定を隠す

文脈中の区切り記号 ⓘ | 文脈中の文区切り記号 ⓘ #

前後文脈の語数 ⓘ 20 検索対象(固定長・可変長) ⓘ 両方

共起条件の範囲 ⓘ 文境界をまたがない

ダウンロードオプション ⓘ 設定を表示する

検索 **検索結果をダウンロード** 条件クリア キャンセル

CSV(カンマ付きテキスト)でダウンロード

短単位検索結果をCSVで表示

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	サンプル	開始位置	連番	前文脈	キー	後文脈	語彙素 ID	語彙素読み	語彙素	語彙素細分	語形	品詞	活用型	活用形	書字形	発音形出現
2	LBdn_000	39390	24300	メートル	へこまさ	(ない) (&)	177485	ヘコマス	凹ます		ヘコマス	動詞-一般	五段-サ行	未然形-一	へこます	ヘコマサ
3	LBf0_0001	44560	27460	つまむ。	つかわ	(ない) (&)	24457	ツカウ	使う		ツカウ	動詞-一般	五段-ワア	未然形-一	つかう	ツカワ
4	LBf9_0012	17560	11670	てしまう	頑張ら	(ない) (&)	8220	ガンバル	頑張る		ガンバル	動詞-一般	五段-ラ行	未然形-一	頑張る	ガンバラ
5	LBg5_000	11570	7640	何 と いう	い	(ない) (&)	2585	イル	居る		イル	動詞-非自	上一段-ア	未然形-一	いる	イ
6	LBg9_001	65640	40750	接近 し 、	進ま	(ない) (&)	19186	ススム	進む		ススム	動詞-一般	五段-マ行	未然形-一	進む	ススマ
7	LBh5_000	760	460	おやつ に	し	(ない) (&)	19537	スル	為る		スル	動詞-非自	サ行変格	未然形-一	する	シ
8	LBj5_0003	90830	57610	心配 です	もらわ	(ない) (&)	37949	モラウ	貰う		モラウ	動詞-非自	五段-ワア	未然形-一	もらう	モラワ
9	LBk4_000	49350	30970	結石 が	とら	(ない) (&)	26854	トル	取る		トル	動詞-一般	五段-ラ行	未然形-一	とる	トラ
10	LBk9_001	46160	31100	。# 「	おか	(ない) (&)	4811	オク	置く		オク	動詞-非自	五段-カ行	未然形-一	おく	オカ
11	LBln_0001	9750	6460	と いう	砲とりあげ	(ない) (&)	26756	トリアゲル	取り上げる		トリアゲル	動詞-一般	下一段-ガ	未然形-一	とりあげる	トリアゲ
12	LBm1_000	10870	6710	が いい	とやら	(ない) (&)	38541	ヤル	遣る		ヤル	動詞-非自	五段-ラ行	未然形-一	やる	ヤラ
13	LBm2_000	49900	32370	どこ が	遠し	(ない) (&)	19537	スル	為る		スル	動詞-非自	サ行変格	未然形-一	する	シ
14	LBn2_000	19930	12320	「 絶対	二言わ	(ない) (&)	1571	イウ	言う		イウ	動詞-一般	五段-ワア	未然形-一	言う	イワ
15	LBn3_000	54190	33560	対策 を	や変え	(ない) (&)	5977	カエル	変える		カエル	動詞-一般	下一段-ア	未然形-一	変える	カエ
16	LBn4_000	117280	73520	交配 する	離さ	(ない) (&)	29990	ハナス	離す		ハナス	動詞-一般	五段-サ行	未然形-一	離す	ハナサ
17	LBn5_000	5480	3250	が さ	か な 持た	(ない) (&)	37804	モツ	持つ		モツ	動詞-一般	五段-タ行	未然形-一	持つ	モタ
18	LBo3_000	13610	8610	を 求める	おか	(ない) (&)	4811	オク	置く		オク	動詞-非自	五段-カ行	未然形-一	おく	オカ
19	LBo4_000	42650	27620	ない 、	出任せ	(ない) (&)	35500	マカセル	任せる		マカセル	動詞-一般	下一段-サ	未然形-一	任せる	マカセ
20	LBo8_000	7970	5170	は い	け な つけ	(ない) (&)	24585	ツケル	付ける		ツケル	動詞-非自	下一段-カ	未然形-一	つける	ツケ
21	LBo9_001	10680	7020	に 「	# 出	(ない) (&)	25728	デル	出る		デル	動詞-一般	下一段-ダ	未然形-一	出る	デ
22	LBo9_001	56720	38490	から 、	宗つけ	(ない) (&)	24585	ツケル	付ける		ツケル	動詞-非自	下一段-カ	未然形-一	つける	ツケ

【問題】「～ないといけません」と共起する動詞ランキング上位は？

1. 切る
2. 書く
3. いる(居る)
4. 使う
5. 行く
6. 付ける
7. もらう
8. 受ける
9. 見る
10. 来る
11. 払う
12. 待つ
13. 勝つ
14. 出る

短単位検索結果をピボットテーブルで

3	行ラベル ↓	個数 / 語彙素
4	為る	85
5	行く	21
6	付ける	13
7	居る	10
8	貰う	9
9	見る	9
10	払う	8
11	置く	7
12	考える	7
13	書く	6
14	上げる	6
15	遣る	5
16	成る	4
17	待つ	4
18	使う	4
19	読む	3
20	頑張る	3
21	聞く	3
22	勝つ	3
23	変える	3

「～ないといけません」と共起する動詞 ランキング上位

- | | |
|-------------------|----------------|
| 1. 切る(2, 12位) | 8. 受ける(1, 13位) |
| 2. 書く(6, 8位) | 9. 見る(9, 5位) |
| 3. いる(居る)(10, 4位) | 10. 来る(3, 11位) |
| 4. 使う(4, 10位) | 11. 払う(8, 6位) |
| 5. 行く(21, 2位) | 12. 待つ(4, 10位) |
| 6. 付ける(13, 3位) | 13. 勝つ(3, 11位) |
| 7. もらう(9, 5位) | 14. 出る(2, 12位) |

コーパスを日本語教育に活かす方法



日本語教育は「何を」教えるべきなのか。主観をサポートするコーパスの使い方と応用を丁寧に解説する。

「まとめ」に代えて
この部分からお話を...

[内容] **総説** / 日本語教材の分析 / 文型とコロケーション / 学習者話し言葉コーパス / 学習者書き言葉コーパス / 対照言語学的分析 / 付録

日本語教育で何を教えるのか

- 外国人に日本語を教えるとき、「何を」「どうやって」教えるのかということが問題になる。方法と内容である。
- 当然、方法も重要なのであるが、ここで問題にしたいのは「何を」の部分、すなわち内容である。
- 日本語教育が内容として扱うのは、現代日本語である。しかも、実際に外国人が話す・聞く・書く・読むという言語活動で「使える」日本語である。
- さて、このように目標を設定したとき、教えるべき日本語とはどのように決めればよいだろうか。

「基本的な日本語」を取捨選択するには

- 「基本的な日本語」を決めることはそう簡単なことではない。例えば「リンゴ」は基本的な日本語と言えるのか。
- しかも、日本語教育の初級は学習者の負担の観点から、他のレベルに比べて語数が少なくなければならない。実用度という観点から、取捨選択の判断を迫られる。
- そして、この取捨選択の判断が実に難しい。「リンゴ」と「みかん」であれば、比べようがあるが、「リンゴ」と「どういたしまして」であると、比較の観点がまったく異なる。しかし、いずれも「使える」日本語という観点から検討する必要がある語である。

「主観」による日本語教科書作成

- 日本語教育に限らず，教育とは人間と人間のコミュニケーションを扱うものである。したがって，「使える」日本語とは何かを判断するのは人間であるのは自然なことであり，その人間の判断の基本はまず主観である。
- 事実，日本語教育に限らず多くの教材は，主観によって作成される。判断は，教材作成において尊重されるべき。
- また，教材作成においては日本語母語話者だけでなく，日本語学習者である**日本語非母語話者の判断も重要**になってくる。

「あなた」はどの場面で使う？

- しかし、人間の判断は偏りがあることがあることも否めない事実である。例えば、日本語教育においては、二人称の「あなた」を初級の早い段階で必ず学習するが、果たして日本語による会話で「あなた」が使える場面はどれほどあるだろうか。
- もちろん、日本語教育において「あなた」を教えなくてよいということにはならない。ポイントはそこではない。
- 冒頭に述べたことを繰り返せば、「あなた」は基本的な日本語であることは間違いない。しかし、「実際に外国人が話す・聞く・書く・読むという言語活動で「使える」日本語」から考えると、果たしてこの条件に合致するだろうか。

「コーパス」で主観をサポートする

- こうした「よく考えてみれば...」ということは、日本語という言語においては日本語母語話者ほど陥りやすい。むしろ、日本語非母語話者こそ、こうした点を指摘してくれる可能性がある。
- そして、もう一つこのような点を気付かせてくれるのが、「コーパス」である。検索の方法さえ間違えなければ誰でも使える。
- ただし、コーパスから得られるデータは、単なるデータでしかない。ただ、コーパスは実際に書かれたり話されたりした言語データである。コーパスの活用は、現実場面の日本語を扱えるという点で日本語教育に密接な関係がある。

人間の「主観」と コーパスから得られる「傾向」1

- 教育という分野の特質を考えたとき、本当に「何を」の部分が「傾向」でよいのかという問題がつきまとう。なぜなら教育とは、カリキュラムに基づいて教える内容を固定化する必要があるからである。
- それに対して、コーパスから明らかになる傾向は、どちらかといえば不安定であり、時代によってもコーパス化された言語データのジャンルによっても異なる。

人間の「主観」と コーパスから得られる「傾向」2

- コーパスから得られる知見（傾向）は、日本語教育の改善に大きな示唆を与えてくれるが、そこにすべての正解を求めることはできない。あくまでコーパスから得られる知見は、主観のサポートに過ぎない。
- コーパスから得られる知見を参考にしつつも、教育において「何を」教えるかということについては、最終的には教材作成者や教師が決定しなければならない。
- 現時点においては、コーパスによって、「何を」教えるのかということを決めるのではなく、主観のサポートに活用するというのが、コーパスの教育利用の最適解であるだろう。

午後のワークショップ予告

1. 「NLB(NINJAL-LWP for BCCWJ) ユーザマニュアル 1.40」を使ったNLBの使い方解説と体験
2. 「中納言オンラインマニュアル」を使った中納言によるBCCWJ検索の使い方解説と体験
3. 中納言を用いた日本語教育のための文型検索の方法と実践

参考文献

- 李在鎬・石川慎一郎・砂川有里子(2018)『新・日本語教育のためのコーパス調査入門』(くろしお出版)
- 齊藤俊雄・中村純作・赤野一郎(編)(2005)『改訂新版英語コーパス言語学基礎と実践』(研究社)
- 国立国語研究所言語資源開発センター(2023)「中納言オンラインマニュアル」
<https://clrd.ninjal.ac.jp/chu-00.html>
- 国立国語研究所(2016)「NLB(NINJAL-LWP for BCCWJ) ユーザマニュアル 1.40」
<https://nlb.ninjal.ac.jp/static/pdf/NLB.manual.v.1.40.pdf>
- 中俣尚己(2014)『日本語教育のための文法コロケーションハンドブック』(くろしお出版)
- 中俣尚己(2017)『コーパスから始まる例文作り』(くろしお出版)
- 森篤嗣(2016)「旧JLPT語彙表に基づく形態素解析単位の考察」庵功雄・佐藤琢三・中俣尚己(編)『日本語文法研究のフロンティア』くろしお出版, pp.43-63
- 森篤嗣・田中祐輔・中俣尚己・奥野由紀子・建石始・岩田一成(2018)『コーパスで学ぶ日本語学日本語教育への応用』(朝倉書店)
- Simpson, J. A., and E. S. C. Weiner, eds. (1989) The Oxford English Dictionary. 2nd ed. Oxford: Clarendon P.