

テーマ別パスファインダー



コーパスの使い方



✧ パスファインダーとは？

Pathfinder（パスファインダー）とは、探検者／草分け／開拓者の意。レポート作成や論文作成で、何をすればいいのか、どこへ行けばいいのかわからない！そんな人のための助けになるように作成した、学問の「道しるべ」です。

作成日：2018年11月14日
大阪大学 外国学図書館 | 箕面キャンパス |
ラーニングコモンズ るくす | LS チーム

I. コーパスとは

ある言語の研究のために、その言語で実際に用いられた用例を大量に偏りなく収集して電子化し、検索用情報を付加したものです。（前川 2013：2）

◀ 前川喜久雄編・辻井潤一〔ほか〕著（2013）『コーパス入門』朝倉書店
【外国図-4 階開架 810.8||46||1】

II. なぜコーパスが必要になってきたか

コーパスができる前の言語研究

1. 手作業で少数のデータしか手に入れられません（大変な労力がかかります）。
2. 研究者の内省だけで言語の特徴を判断していました。
3. 特定の作家の文体を個別に分析していました。

コーパスができた後の言語研究

1. 大量のデータを短時間で集めることができるようになりました。
2. 研究者の内省だけでは知り得なかった、新たな言語の特徴を見出すことができます。
3. 単語の周辺的な意味・用法のより精度の高い分析が可能になります。
4. さまざまなジャンルから横断的にデータを収集し、それらを比較して、それぞれのジャンルの特質を明らかにすることができます。
5. 歴史的な言語の変化を概観することができます。

◀ 砂川有里子（2011）「日本語教育へのコーパスの活用に向けて（特集 関連領域の動向と日本語教育）」
『日本語教育』 150 号, pp. 4-18

III. コーパスの種類

書き言葉	話し言葉
- 現代日本語書き言葉均衡コーパス - 青空文庫 - 新潮文庫の 100 冊 - 日英新聞記事対応付けデータ - 日英対訳文対応付けデータ - 京都大学テキストコーパス (Version 4.0) - 新聞記事データベース - 国会会議録検索システム	- 日本語話し言葉コーパス - 名大会話コーパス - BTS による多言語話し言葉コーパス - 女性のことば・職場編 - 男性のことば・職場編 - 戦時中の話しことば

<http://jhlee.sakura.ne.jp/page1.html> (最終アクセス 2018 年 10 月 3 日)

歴史的言語の変化が分析できるコーパス

コーパス	産出年代	生年情報	字数 (万字)	形態素分析
太陽	1895～1925	あり	約 1455	なし
近代女性雑誌	1895～1925	あり	約 214	なし
日本話し言葉コーパス	1999～2001	あり	約 1410	あり
現代日本語書き言葉均衡コーパス	1976～2005	あり	約 19642	あり
青空文庫	主に戦前	あり	約 10027	なし
国会会議録	1947～	なし	約 347179	なし

(服部 (2014:22) を加筆修正)

- ◀ 服部匡 (2014) 「現代日本語の通時変化」 田野村忠温編 (2014) 『コーパスと日本語学』 朝倉書店
【総合図-A 棟 4 階 学習用図書 810.8||K0Z||6】

IV. 現代日本語書き言葉均衡コーパス(BCCWJ)の使い方

現代日本語書き言葉均衡コーパス(BCCWJ)とは？

国語国立研究所が構築したコーパスであり、現在、日本語について入手可能な唯一の均衡コーパスです。書籍全般、雑誌全般、新聞、白書、ブログ、ネット掲示板、教科書、法律などのジャンルにまたがって1億430万語のデータを格納しており、各ジャンルについて無作為にサンプルを抽出しています。

http://pj.ninjal.ac.jp/corpus_center/bccwj/ (最終アクセス 2018年10月3日)

また、BCCWJでは、すべてのテキストは形態素解析がなされ、詳細な形態論情報が付与されているほか、出版社、出版年、ジャンル、などの書誌情報や、著者の氏名、性別、出生年などの情報が加えられています。
(前川 2007)

◀ 前川喜久雄(2007)「コーパス日本語学の可能性——大規模均衡コーパスがもたらすもの——」『日本語科学』22号 pp. 13-28

BCCWJの基本的利用方法

- ① 少納言：文字列検索のみ対応。データのダウンロード不可(登録不要)
<http://www.kotonoha.gr.jp/shonagon/>
- ② 中納言：文字列検索。品詞による検索、長単位検索。データのダウンロード可(登録が必要)
<http://chunagon.ninjal.ac.jp/>

BCCWJにおけるコロケーションの検索方法

- ③ NINJAL-LWP for BCCWJ：確率的言語現象/コロケーション研究用。データのダウンロード不可
<http://nlb.ninjal.ac.jp/> (国語研とLago言語研究所が共同開発したオンライン検索システム、登録不要)

まず、中納言を使ってみましょう！

「キー」の検索用の部分を更に拡大してみると ↓

上の図に出てきている様々な検索キーとは何か？

- 書字形：同じ語で異なる表記がある時に。「図る」「測る」「はかる」など
- 語彙素：標準的なキー。異なる形態であるが同じ語であると考えられるものからなる語の集合をいう。
例えば、泣き、泣いた、泣かせたの語彙素は全て「泣く」
- 語彙素読み：語彙素が分からない時に。カタカナで入力。「ヤバイ」「サビシイ」など
- 語形：同じ語で異なる語形がある時に。「やばい」と「やべえ」など
- 品詞：動詞、助詞、形容詞などの指定ができる
- 活用形：未然、連用、終止など。古典研究
- 活用型：五段、一段、変格など。古典研究

練習

「～てしまう」を中納言で検索してみましょう。

しかし、「てしまう」をそのまま検索キーとして入力すると、何も出てきませんよ。

なぜなら、中納言は「語彙素」で検索することが基本となっているためです。

語彙素の調べ方は、以下の「web 茶まめで試す」という形態素解析システムで試してみてください。



<http://unidic.ninjal.ac.jp/chamame/>

例えば、「花瓶を割ってしまった」という文を「web 茶まめで試す」で形態素解析してみると、次のようになります。

辞書	文境界	書字形 (=表層形)	語彙素	語彙素読み	品詞	大分類	中分類
現代語Unidic (Ver.1603)	B	花瓶	花瓶	カビン	名詞-普通名詞-一般	名詞	普通名詞
現代語Unidic (Ver.1603)	I	を	を	ヲ	助詞-格助詞	助詞	格助詞
現代語Unidic (Ver.1603)	I	割っ	割る	フル	動詞-一般	動詞	一般
現代語Unidic (Ver.1603)	I	て	て	テ	助詞-接続助詞	助詞	接続助詞
現代語Unidic (Ver.1603)	I	しまっ	仕舞う	シマウ	動詞-非自立可能	動詞	非自立可能
現代語Unidic (Ver.1603)	I	た	た	タ	助動詞	助動詞	

上の四角い赤線で示しているように、「てしまう」を検索したい時は、語彙素の「て」と「仕舞う」で検索しなければならない、ということが分かります。

正しい検索の仕方は↓



短単位検索

前方共起条件の追加

i
前方共起1
キーから
1
語
☐ キーと結合して表示
i

語彙素
が
て

短単位の条件の追加

i
キー
--
1
語
☐ キーの条件を指定しない
i

語彙素
が
仕舞う

短単位の条件の追加

後方共起条件の追加

中納言を利用して現代書き言葉均衡コーパス（BCCWJ）での「てしまう」の検索結果は以下の通りです。

縦線で隔ててあるのは各語彙素を分けるためです。

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
LBj9_00038	140	100	捻挫 体験# 右 足首を 捻挫 て	しまっ	た 。# 捻挫 くらい と 思い 、 その まま タクシー に 乗って 仕事 に 出かけた の が	シマウ	仕舞う
OB6X_00099	6890	4410	が 日本 人 に あっ た の で は ない で しょう か 。# 「 に の まま アメリカを 超え て	しまっ	たら 、 その あと 、 どう する の だ と いう こと です 。# 「 追い 抜いた あと 、	シマウ	仕舞う
PB1n_00061	81630	54800	。# 夕方 に 少し 雨 が ふり まし た 。# 今日 で 夏 休み の しゅく だい を 皆 仕上 げ て	しまい	まし た 。# しゅく だ い を 皆 し た の で ま っ と し まし た 。# 九 月 二 日	シマウ	仕舞う
OB4X_00063	67440	43250	カギを かけ て しまっ た んだ 。# その 「 せい で 、 あの 劇場 は 密室 」に なっ て	しまっ	た 。# 犯人 は え も 予想 な か っ た に の 偶然 が 、 不可思議 な 密室 殺人 を 生 み 出 し て しまっ	シマウ	仕舞う
OC05_03256	430	280	閉鎖 、 そして 中国 等 への 海外 移転 が 進み 、 日本 の 製造 業 は どう なっ て	しまう	の が 大 変 不安 に 思っ て い ま す 。# こんな 時 に そ 、 みんな で 、 安い 中国 製	シマウ	仕舞う

次、NINJAL-LWP for BCCWJ を使ってみましょう！

NINJAL-LWP for BCCWJ は「少納言」と「中納言」とは異なるレキシカルプロファイリングという手法を用いたコーパス検索ツールです。名詞や動詞などの内容語の共起関係や文法的振る舞いを網羅的に表示できるのが最大の特長です。

練習

「N を切る」の N にはどのようなものが最も多いのかを NINJAL-LWP for BCCWJ で検索してみましょう。



NINJAL-LWP for BCCWJ

ここに「切る」を入力

切る

絞り込み 元に戻す

すべて 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
いる・非自立	イル	iru	968,064
こと	コト	koto	701,470
の	ノ	no	686,025
する	スル	suru	601,911
なる	ナル	naruru	484,861
ある	アル	aru	479,785
その	ソノ	sono	347,357
よう	ヨウ	you	333,829
言う	イウ	iu	285,579
この	コノ	kono	272,773
もの	モノ	mono	234,829
ん	ン	n	205,301
思う	オモウ	omou	199,915
できる	デキル	dekiru	194,228

1 page / 852 100 ▼ 85,112件中 1 - 100を表示

NINJAL-LWP for BCCWJ

切る 絞り込み 元に戻す

すべて 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

切る キル kiru 13,075

ここをクリック

切る 頻度=13,075

グループ別 パターン頻度順 基本

名詞+助詞

パターン	頻度	比率
...が切る	777	
...は切る	966	
...も切る	179	
...の切る	75	
...を切る	6,387	
...に切る	3,277	
...へ切る	44	
...で切る	1,110	
...と切る	146	
...から切る	228	
...まで切る	80	
...より切る	17	

さらにここをクリックすると →

名詞+複合助詞

パターン	頻度	比率
...として切る	3	
...において切る	1	
...に加えて切る	1	

...を切る 1040種類

コロケーション	頻度	MI	LD
電話を切る	564	8.58	9.19
言葉を切る	171	6.28	6.99
電源を切る	164	10.47	9.33
緑を切る	155	9.25	8.89
首を切る	148	7.98	8.21
髪を切る	134	8.71	8.54
スタートを切る	125	9.42	8.75
水気を切る	124	11.22	9.13
木を切る	123	7.13	7.56
人を切る	120	2.73	3.64
シャッターを切る	113	10.91	8.97
スイッチを切る	110	10.10	8.80
ハンドルを切る	110	10.67	8.90
腹を切る	85	7.93	7.82
手を切る	84	4.85	5.61
口火を切る	77	13.65	8.60
【人名】を切る	76	0.02	0.95
エンジンを切る	72	8.07	7.75
水を切る	68	5.74	5.01

1 page / 11 100

「...を切る」と共起する名詞の頻度が一目瞭然

V. コーパスのデメリット

- 「コーパスのパラドクス」

コーパスが大きくなればなるほど、そこから抽出されるデータも多くなり、分析者にとって手に負えないものになってしまいます。

- 「コーパスの非文」

コーパスには書き誤り（言い誤り）が存在し、ある文がコーパスに存在するからといって、その文が文法的であるとは言えません。

（コーパスに存在するのは文法的な文ではなく、話者が容認可能とみなした文）

VI. 最後に

汎用的な大規模コーパスが公開・提供されることは、コーパス言語学の発展には不可欠のことですが、一方、コーパスは文法研究にとって必ずしも万能な道具ではありません。あらゆる研究に有効であるわけではありませんし、使い方を間違えれば、間違った結果が導き出されてしまいます。

そのため、コーパスの利用者には、そのコーパスが自らの研究課題に合致しているか否かを吟味する姿勢がより強く求められるようになっていきます。

◀ 石井正彦・杉本武（2014）「コーパスを用いた日本語研究の特徴—語彙・文法を中心に—」 田野村忠温編（2014）『コーパスと日本語学』朝倉書店
【総合図-A 棟4階 学習用図書 810.8||K0Z||6】

VII. 分からないときは・・・

赤野一・堀正広編 『コーパスと多様な関連領域』ひつじ書房
【総合図-A 棟4階 学習用図書 830.8||EIG||7】

石川慎一郎・前田忠彦・山崎誠編（2010）『言語研究のための統計入門』くろしお出版
【総合図-A 棟3階 留学生用図書 801.019||ISH】

田野村忠温編（2014）『コーパスと日本語学』朝倉書店
【総合図-A 棟4階 学習用図書 810.8||K0Z||6】

前川喜久雄編・辻井潤一〔ほか〕著（2013）『コーパス入門』朝倉書店
【外国図-4 階開架 810.8||46||1】

砂川有里子編（2016）『コーパスと日本語教育』朝倉書店
【外国図-4 階開架 810.8||46||5】

砂川有里子（2011）「日本語教育へのコーパスの活用に向けて（特集 関連領域の動向と日本語教育）」『日本語教育』 150 号, pp. 4-18

----- MEMO -----

✧ [パスファインダーの凡例]

＜ 図書の情報は以下の順に表記しています。（主に論文の参考文献に使われている書式です。）

著者名（出版年）『本の名前』出版社名，翻訳者名（あれば）

＜ 説明の最後に、【 】で貸し出し可能な図書館と配架場所、請求記号を記しました。

総合図 → 総合図書館（豊中キャンパス）

生命図 → 生命科学図書館（吹田キャンパス）

理工学図 → 理工学図書館（吹田キャンパス）

人図 → 人間科学研究科図書室（吹田キャンパス）

外国図 → 外国学図書館（箕面キャンパス）

外国図-雑誌 → 直近 1～2 年に出版されたものは 3 階雑誌コーナー、バックナンバーは 1 階書庫

電 → 電子ジャーナル、電子ブック

※雑誌、電子ジャーナルは、すべての巻号が利用できるとは限りません。

＜ 検索を容易にするために、ISBN（各図書固有の識別番号）や ISSN（各雑誌固有の識別番号）を記している場合もあります。

＜ 外国学図書館を中心に紹介していますので、記載している場所以外でも貸し出し可能な場合があります。図書館各階にある検索端末で確認するか、カウンター/LS デスクまでお尋ねください。