

日本語の用例の集め方

—BCCWJ を中納言で検索—

概要

- 日本語の実例を集めたいけど、どこからデータ集められる？
 - 日本語についてのコーパスってどんなものがある？
 - コーパスってどうやって使う？
- ★日本語専攻の人はもちろん、日本語と他言語の対照研究をしたい人、留学生のみんな、このような問題はありませんか？
- ★用例を収集するときに知っていたら便利なコーパスと、その検索方法をご紹介します。

もっと知りたい人はこちらへどうぞ

図書館 4F@ラボ 学習相談デスク

HP : <http://www.tufs.ac.jp/blog/is/g/sodan/>

時間 : 月～金 12:00～19:00

予約不要

制作 : 東京外国語大学附属図書館 学習相談デスク

制作日 : 2018年11月29日



BCCWJ とは？

・BCCWJ とは、国立国語研究所で開発された「現代日本語書き言葉均衡コーパス」(Balanced Corpus of Contemporary Written Japanese) のことで、現代日本語書き言葉について、現在唯一入手可能な均衡コーパス¹。

BCCWJ を検索するには中納言を使ってみよう！

・その規模はなんと **1 億語**

中納言とは？

- ・コーパス検索アプリケーション (要申し込み)。
- ・中納言で検索できるコーパスは、BCCWJ 以外に、日本語歴史コーパスや、話し言葉コーパス、日本語学習者コーパスなども、ユーザ登録を申請すれば、**無償**で利用可能。



コーパス名	略称	状態	関連サイト	備考
現代日本語書き言葉均衡コーパス 通常版	BCCWJ-NT	利用可能	関連データ配布	従来より利用いただ いている BCCWJ の データです。関連 データの紹介ペー ジへのリンク
現代日本語書き言葉均衡コーパス 非 numTrans 版	BCCWJ-OT	利用可能		非 NumTrans 版 (OT)とは
日本語歴史コーパス	CHJ	利用可能		
日本語話し言葉コーパス	CSJ	利用できません	利用申請	
多言語母語の日本語学習者横断コーパス	I-JAS	利用できません	利用申請	発話の音声ファイル や書き起こしたプレ インテキストは、関連 データ配布サイトから ダウンロードできま す。その他の資料に ついては、I-JAS 関連資料を 参照ください。
名大会話コーパス	名大会話コーパス	利用可能		

¹ 「均衡コーパス」とは、「ある言語の特徴や性質を知るために、その言語の多様性をできるだけ忠実に反映するようにバランスよくサンプルを収集して構築される言語資源」を意味する。

検索結果画面：

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
という国のことがやっとわかりかけてきた気がする。イタリア料理に通じる	美味しさ	リビアで初めてのディナーはトリポリのダウンタウンのレストランへ。外食の習慣が少ない	オイシイ	美味しい		形容詞一般	形容詞	語幹一般	図書館・書籍	井原 三津子(著)
方法そのものと言ってもよいのだが、その中に悲しみを訴えるもの、	美しさ	を訴えるもの、怒りを訴えるものなど、いろいろあって、おかしさを訴えるもの	ウツクシイ	美しい		形容詞一般	形容詞	語幹一般	出版・書籍	阿刀田 高(著)
ハムステッド・ヒース」駅はない。いずれにしても、公園の数とその	広さ	で言えば、東京とロンドンとでは、比較すること自体に無理がある	ヒロイ	広い		形容詞一般	形容詞	語幹一般	図書館・書籍	林 信吾(著)

検索結果を Excel にダウンロードして、必要な項目だけを Excel で編集することができる。

1	前文脈	キー	後文脈	レジスター	執筆者	生年代	性別	ジャンル	書名/出典	副題/分類	巻号	編著者等	出版者	出版年	反転前文脈
2	ほどだ。だが、反面、日常のささいな部分で批判を受ける“甘	さ	”があるのも事実だ。「両陛下や秋篠宮ご一家とご一緒に	出版・雑誌	実著者不明			総合/一般/一般週刊誌	週刊現代				講談社	2004	甘”るけ受を判批で分らないさの常日、面反、がだ。だどほ
3	事の償いとして十分なのか。消費者団体などからは処分の甘	さ	を問う声が上がった。午後6時すぎ、会見場に紺のスーツ姿	出版・新聞	山中純枝(著)/毎日	/	女/	全国紙/毎日新聞	毎日新聞	朝刊	2002/8/21		毎日新聞社	2002	甘の分処はらかどな体団者費消。かのな分十てしとい償の事
	を逆手に取って、チョコレート市場に参戦してきたのがサッポロビール。甘	さ	を抑えたチョコに濃縮したビールを隠し味として練りこんだ「サッポロビールチョコレート	出版・新聞	産業経済新聞社(著)			全国紙/産経新聞	産経新聞	朝刊	2004/2/8		産業経済新聞社	2004	甘。ルービロポッサがのたきてし戦参に場市トーレコヨチ、てっ取に手逆を

★ Tips : もっと効率的に活用しよう！

① 検索結果がたくさん出てきて分析しきれない？

- ▶ 「知恵袋」「ブログ」「国会会議録」を検索対象から外す。(➤ 執筆者が不明のため)
- ▶ 「検索対象」を「コア」に設定する。(➤ コアデータとは、形態論情報の付与が人によって行われたデータのこと。なので、コアデータの解析精度が非コアデータよりも高い)

② Excel を使って効率的に分析しよう！

- ▶ 「後文脈」や「反転前文脈」を Excel の「昇順」(「降順」でも良い) 機能で並べ替えると、「キー」の直前と直後の語が五十音順になるので、もっと分析しやすくなる。

③ ランダムに用例を収集する方法は？そのメリットは？

- ▶ Excel の関数「RANDBETWEEN」を活用する。
 - 例えば 1 万の用例が収集されて、とりあえず 200 例を分析してみたいというように、用例を絞る方法にもなる。
 - いろいろなジャンルから用例を抽出でき、偏りを避けられる。

「RANDBETWEEN」の使い方を更に知りたいなら：

https://kokodane.com/kansu_math_55.htm

BCCWJ へのアクセスはこちらをご覧ください。

https://pj.ninjal.ac.jp/corpus_center/bccwj/

