国立国語研究所　現代日本語書き言葉均衡コーパス(BCCWJ)の概要

言語コーパスとは

言語コーパス（language corpus)とは、言葉のデータベースのこと。日本語をはじめとするさまざまな言語を分析するための基礎資料として、書き言葉や話し言葉の資料を体系的に収集し、文法に関する情報を付与したもの。資料は全て著作権に関する処理を施し、広く自由な活用が可能。

どのように役立つか

コーパスは1960年頃から最初は言語研究のために構築されるようになったが、近年では狭い意味での言語学の領域を超えて、幅広い研究領域で利用されるようになった。学術目的での利用だけでなく、産業界でも利用されている。

○活用の具体例

言語研究	言語学、日本語学、国語学など個別言語の研究複数言語のコーパスの比較による対照言語学
情報処理	音声自動認識のための言語モデル、音響モデルの構築　自然言語処理のための言語モデルの構築　音声素片の編集に基づく音声合成
言語教育	外国人のための日本語教材開発　日本人のための教材開発
言語政策	常用漢字表などを検討するための基礎資料
辞書編纂	用例の検索　語と語のつながりの把握
心 理 学	言語に関する実験の設計、刺激の統制

平成23年、国立国語研究所が**現代日本語書き言葉均衡コーパス(BCCWJ**:Balanced Corpus of Contemporary Written Japnese)を公開。
現代日本語の書き言葉の全体像を把握するために構築したコーパス。現在、日本語について入手可能な唯一の均衡コーパス（統計学的にバランスのとれたサンプルに基づく）。ただし、2005年のデータが最新であり、それ以降のデータ更新はなされていない。

ある言葉が、どのような種類の文書に、どのような文脈で使用されているかが一目で分かる。有償版では文法的情報も提供

公開・運用

「少納言」「中納言」というアプリケーションを通して、誰もが活用できる形で公開

- 現代日本語の書き言葉として以下を対象
  - 新聞
  - 雑誌
  - 書籍全般
  - 白書
  - 教科書
  - WEB上のQ&A掲示板のテキスト
- 収録対象の刊行年代は、1976～2005年までの最大30年間分。約1億語を収録。

活用例:BCCWJ（無償版アプリ「少納言」）で「なおざり」という語を検索した結果（一部）

検索結果

73 件の結果が見つかりました。そのうち 73 件を表示しています。

表示番号	前文脈	検索文字列	後文脈	執筆者	生年代	性別	メディア／ジャンル	タイトル	副題	巻号	編著者等*	出版者	出版年
1	ながら「義務を意識して、さまざまな破壊を注視し、証言すること」（ＵＴ．１４１）を	なおざり	にすることはなく、逆に作家は最初に見学した被災対象を皮切りにますます詳細な描写を	斎藤　松三郎(著)	1940	男	書籍/9 文学	夢のありかを求めて	ペーター・ハントケ論		斎藤松三郎|著	鳥影社・ロゴス企画部	2001
2	れているようにも見える。だが、ニュルンベルク裁判の目的は理想にすぎなかったとして	なおざり	に扱っては、無知と無理解によって、私たちはいっそう悲劇的な道を歩むことにもなりか	ジョゼフ・Ｅ・パーシコ(著)/白幡　憲之(著)/白幡　憲之(訳)	1930/1950/1950	男/男/男	書籍/3 社会科学	ニュルンベルク軍事裁判		下	ジョゼフ・E.パーシコ|著;白幡憲之|訳	原書房	2003
3	半から二〇世紀初めのアカデミックな中世史学では、集団としての貴族	なおざり	にされていた。それには理由がある。そのころ中世史学の中心的地位	江川　温(著)	1950	女	書籍/2 歴史	西欧中世史		中		ミネルヴァ書房	1995