

BCCWJモニター公開データの利用実態について

山崎 誠 (国立国語研究所言語資源研究系)

BCCWJモニター公開データ



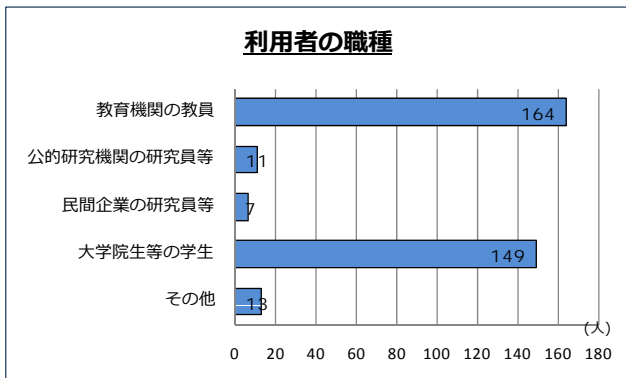
著作権者から利用許諾を得たデータ約4,500万語を収録したDVDディスク。2010年7月26日現在 934名に配布。

利用者アンケートの実施

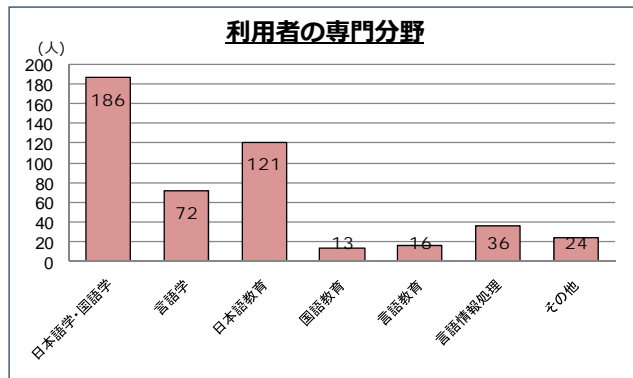
対象者 : 897名 (2008年版255名, 2009年版642名)
 実施時期 : 2010年7月23日～同年8月6日
 実施方法 : 電子メールで送付, 回答メール不達: 47名
 返信数 : 344名 (有効回答340名)
 回答率 : 37.9% (対象者における有効回答の割合)

※2008年版利用者の回答率: 18.4%
 2009年版利用者の回答率: 45.6%

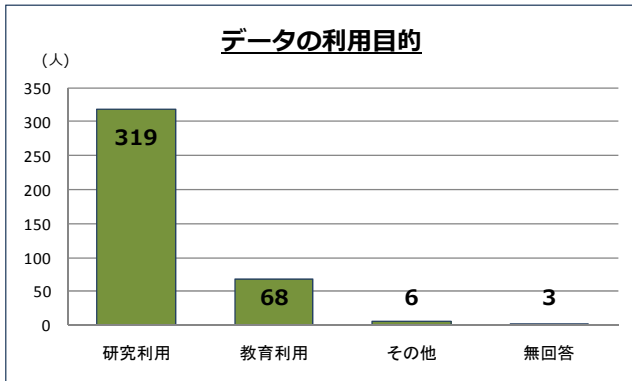
利用者の職種



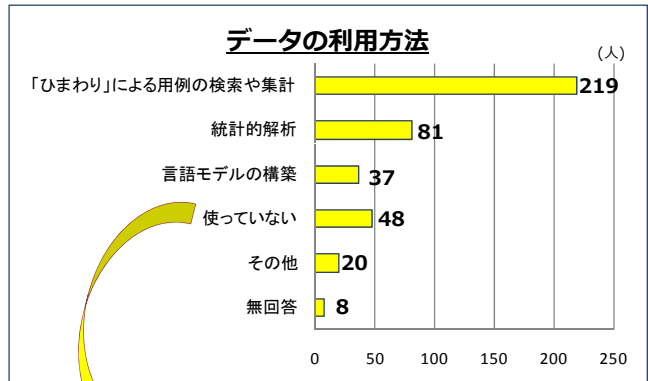
利用者の専門分野



データの利用目的



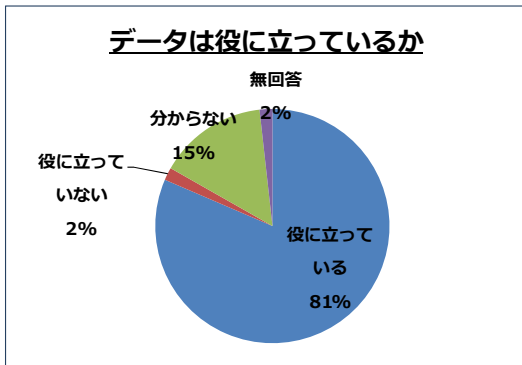
データの使用方法

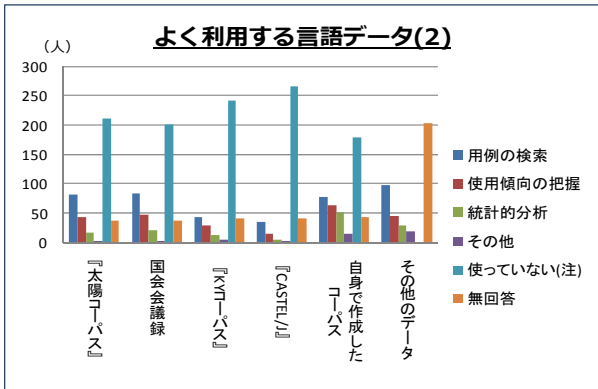
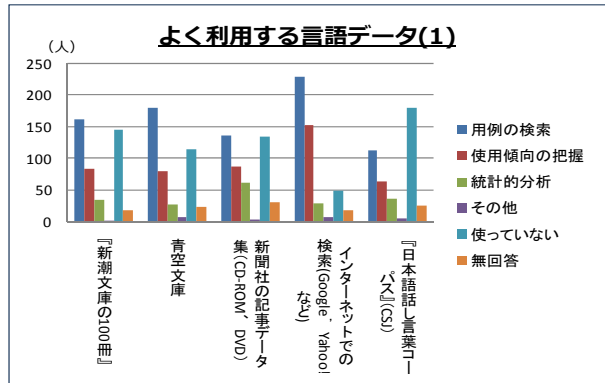
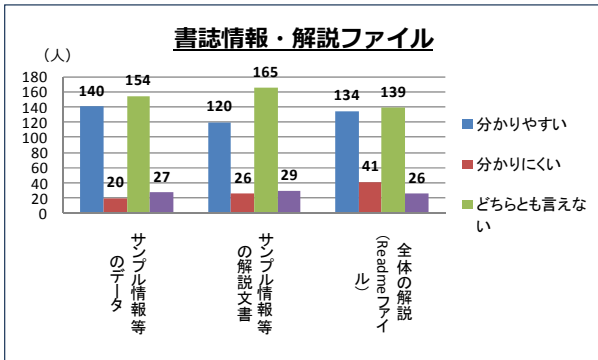
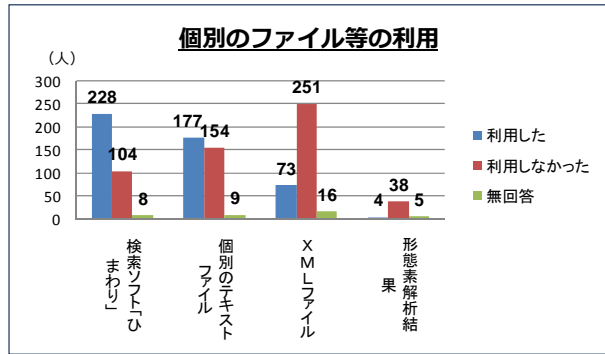
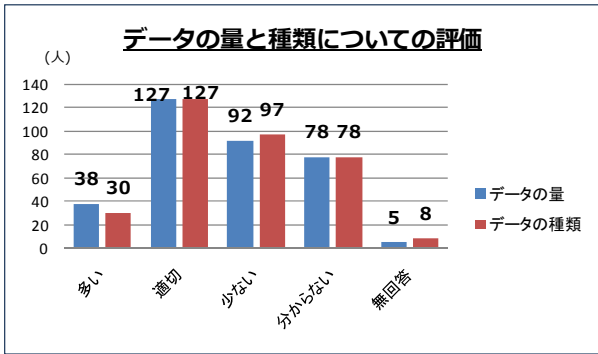


使っていない理由 (主なもの)

- 【使い方が分からない】**
 - ・インストールの仕方がまだ分からない。／・Windowsの使用法しか説明されていない。／・うまく動作しない。／・文字化けが頻繁に起こるため。／・ディスクを入れたときにウイルスが検出された。／・説明を読んで難しく感じた。
- 【忙しくて時間が取れない】**
 - ・授業や仕事がいそがしい。／・(研究の)時間がとれていない。
- 【研究上の都合による】**
 - ・具体的な研究方法が固まっている。／・研究に利用可能かどうかを検討している段階。／・まだ実験準備が整っていないため。
- 【今後使う】**
 - ・帰国後使用予定。／・これから使いたい。
- 【目的と違う】**
 - ・研究目的に沿う書籍(雑誌など)のサンプルが少なかったため。
 - ・欲しいデータがなかった。
- 【必要がない】**
 - ・現時点で必要性がない。
- 【他のデータを使っている】**
 - ・特定領域内公開データ, 国語研究所内部データを使っているため。

データは役に立っているか





(注) 自身で作成したコーパスの設定では、選択肢「使っていない」は「自身で作成したコーパスはない」になっている。

その他利用する言語データ

新潮文庫明治の文豪／新潮文庫大正の文豪／新潮文庫絶版100冊
 閑蔵／近代女性雑誌コーパス／小松左京コーパス
 検定教科書／大学入試問題／日本語教材／日本語教科書データベース
 言語政策班作成の教科書コーパス
 岩波国語辞典／角川類語辞典
 日本古典文学大系(国文学研究資料館)
 PERC CORPUS ONLINE
 Googleブック検索／JPwacL2／ブログ
 名大会話コーパス／BTS話し言葉コーパス／『男性のこぼれ 職業編』／『女性のこぼれ 職業編』／CALL Home Japanese Transcripts／『戦時中の話しことば』
 中日対訳コーパス／日本語学習者会話データベース／作文対訳データベース
 全国方言談話データベース／方言談話資料
 21世紀世宗計画
 京大コーパス／NAISTテキストコーパス／GSK2007-C Web 日本語Ngram第1版
 連想概念辞書／NTICR-4,5,6 PATENT／格フレーム辞書
 日英新聞記事対応付けデータ／EDR日本語コーパス
 NTT『日本語の語彙特性』シリーズ

改善点・要望

〔手続き・利用条件〕

- ・手続き等、もう少し簡便になるとありがたい。
- ・データの使用条件が厳しい。
- ・データの利用期間がもっと長ければ良い。
- ・モニター公開データから抽出した統計データや言語モデルの公開・配布についての権利や条件について明確にしてほしい。
- ・正式公開まで期間が長いので、プロジェクト中間段階での商業利用についても顧慮してほしい。

〔初心者への対応〕

- ・専門家以外を対象にした平易なマニュアルがあるとよい。
- ・データの平易な解説、「やってよい検索」と「やってはいけない検索」の例示。
- ・使用方法などについて、コーパス言語学の初心者にも分かりやすい解説があるとうれしい。
- ・初心者向けのガイドをホームページ上に掲載してほしい。
- ・データの効果的利用方法のワークショップなどを開催していただければありがたい。
- ・年1度行われる大会が別の場で院生や研究者のためのオリエンテーションがあればぜひ聞きにきたい。

〔支援体制〕

- ・オンラインサポート体制
- ・ネット上で質疑応答ができる掲示板などがあればよい。
- ・特定領域研究のウェブページに研究成果やデータを使用した文献目録などを掲載するとモニター公開データの利用者の参考になるのではないかと。

〔データについて〕

- ・研究のタイプによっては「顔」は「顔」と入力されているより、「顔」であった方がいいものもある。いろいろな研究に対応できるようにさまざまな形での提供を望む。
- ・本当に入力ミスなのか、もとの誤植なのか確認できるように、画像データでの提供もあるとよい。
- ・語彙数などの情報や、検索ソフトの使用方法が、いくつかのファイルに分かれて記載されている場合があり、どのファイルを見ればいいのか分かりにくい。

〔ツールについて〕

- ・日本語以外のOS対応についての説明がちょっと足りない。
- ・使用方法についても、ウィンドウズだけでなく、Macの場合の説明もほしい。
- ・統計的な分析ツールの充実。
- ・用例の長さが短いのでもう少し長くしてほしい。
- ・データがでるまでの待ち時間がすこし長い。タグの選択ができるようにしていただきたい。
- ・BNCのようにインターネットでアクセスできるコーパス(内容に制限があったとしても)があれば便利。

〔用語について〕

- ・「短単位」などの概念は、分野外の者にとっては用語としてもなじみにくい。
- ・生産・流通というような呼び方は、BCCWJの知識がないととまどう。