2007.05.28

大規模書き言葉コーパスのオンライン試験公開

～KOTONOHA「現代日本語書き言葉均衡コーパス」～

前川 喜久雄
独立行政法人国立国語研究所 研究開発部門 言語資源グループ長
文科省科学研究費特定領域研究「日本語コーパス」領域代表者

1

要　旨

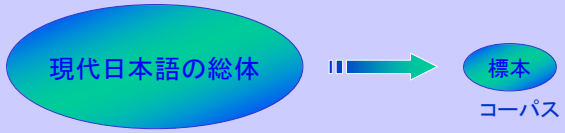
独立行政法人国立国語研究所は、現在構築中の『現代日本語書き言葉均衡コーパス』のデータの一部、約1000万語分をインターネット上で試験公開します。（http://www.kotonoha.gr.jp/demo/）

国語研究所は、明治から現代にいたる日本語の電子化資料をコンピュータ上で公開しようとするKOTONOHA計画を推進しています。『現代日本語書き言葉均衡コーパス』は、その一環として昨2006年度から構築を開始したものであり、2011年の完成時には1億語を超える量の現代日本語の書き言葉データとして公開する予定です。

2

コーパスとは何か

言語分析を行うための基礎資料として、書き言葉や話し言葉の資料を組織的に収集し、研究用の情報を付与したうえで電子的に保存したもの。



偏りのない形で対象の全体像を反映したデータとなっていることが望まれる（均衡コーパス）。

3

どのような領域でどのように利用されるか？

日本語学	実例に基づく分析、定量的な分析
辞書編纂	用例収集、共起関係
日本語教育	基本語彙、基本構文、共起関係
国語教育	基本語彙、漢字の客観的選定
心理学・認知科学	実験における言語刺激の統制
自然言語処理	統計的学習データ、アルゴリズム評価用データ
国語政策	常用漢字の見直し、正書法の提案

4

類義語「風景」と「光景」の分析

毎日新聞記事(2003年)を検索すると

	N	～風景 ～光景	風景～ 光景～
風景	954	246	68
光景	514	3	0

原風景、心象風景
田園風景、日常風景
銀座風景、結婚式風景
Etc. 94種類

日常的光景
歴史的光景
神話的光景
3種類だけ

5

「起きる」と「生じる」

主語となる名詞が「問題」の場合と「事件」の場合とを比較すると、

コーパス	～が	起きる		生じる	
毎日新聞記事 2003年	問題	84	(57.1%)	63	(42.9%)
	事件	301	(99.7%)	1	(0.3%)
国会会議録	問題	85	(31.8%)	182	(68.2%)
	事件	100	(93.5%)	7	(6.5%)

6

「優れた」と「良い」

共起する名詞が「容姿」「性格」「頭」「頭脳」の場合

コーパス	～が（の）	優れた人		良い人	
Yahoo!による インターネット検索	性格	3	(0.0%)	3,687	(99.9%)
	頭	21	(0.0%)	159,675	(99.9%)
	能力	408	(96.5%)	15	(3.5%)
	容姿	83	(23.5%)	270	(76.5%)
	頭脳	12	(50.0%)	12	(50.0%)

Cf. 能力が高い人 22,700
頭脳明晰な人 740

7

コーパスの利点

■実際の用例を、信頼のおける出典情報とともに提供できる
■語や句の特徴を定量的に把握できる

↓

- 日本語学習者（外国人）用日本語辞書の開発に必須
- 母語話者（日本人）の疑問に答えるためにも重要
- 学校教育での利用も考えられる
- 言語研究を定量的な科学に (*e-science*)

8

日本語コーパスの現状と問題点

- 新聞記事テキストデータベース（各社合計で10億語程度、有償）
- ボランティアによる「青空文庫」（著作権の消滅した文芸・学術作品、無償）
- 国会図書館の国会会議録（数億語、無償）
- インターネットを仮想的コーパスとして利用

↓

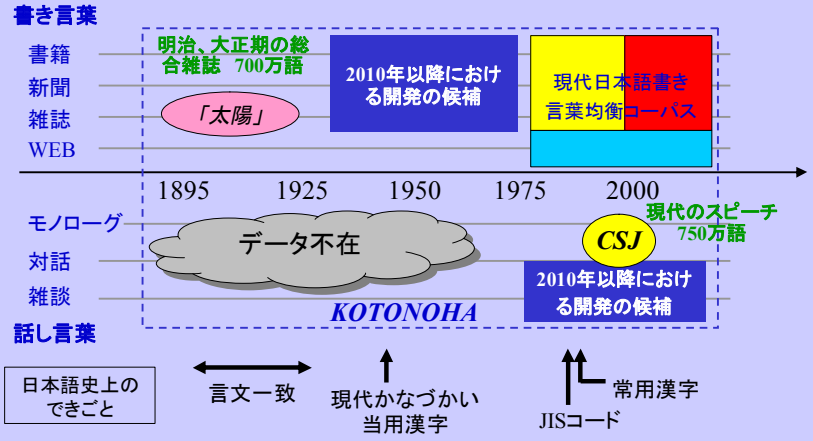
日本語全体をバランスよく反映した均衡コーパスが存在しない

⇒ 均衡コーパスが構築されている言語

英語、米語、スペイン語、ギリシャ語、チェコ語、
韓国語、中国語（台湾）、など

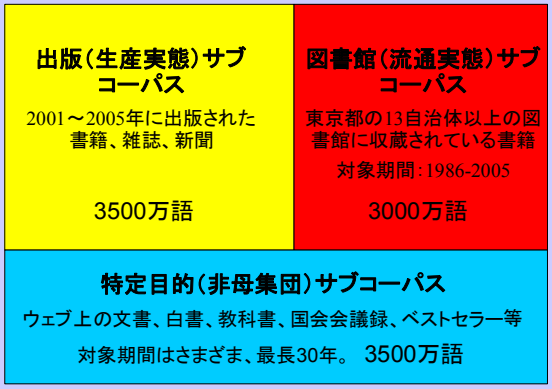
9

国立国語研究所のコーパス開発計画:KOTONOHA



10

『現代日本語書き言葉均衡コーパス』



- 3種のサブコーパスから構成
- 全体で1億語以上
- 著作権処理を施して公開する

11

『現代日本語書き言葉均衡コーパス』の開発

- 開発期間

  2006—2010年の5年間

- 開発費

  国語研運営費交付金

  文科省科学研究費補助金特定領域研究「日本語コーパス」

- 初年度（2006年度）の成果

  データ約1500万語を作成

  そのうち1000万語の著作権処理を終了

  ⇒ 今回デモンストレーション用に公開

12

今回試験公開するデータ

- 白書データ　　あらたまった書き言葉

2001～2005年に発行された白書と過去30年間に継続して発行されつづけた白書の全体を母集団として約500万語を無作為抽出。白書の対象にしたがって9カテゴリー(「安全」「科学技術」「外交」「環境」「教育」「経済」「国土交通」「農林水産」「福祉」)に分類。

- 「Yahoo!知恵袋」データ　くだけた話し言葉の特徴もある

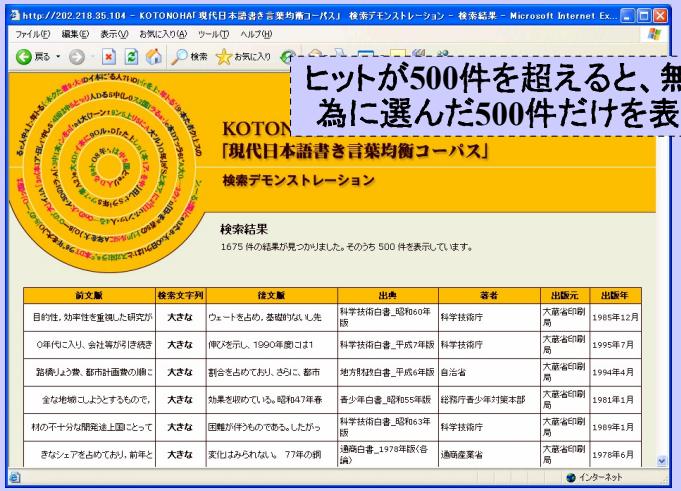
参加者同士で知識を教えあうことを目的とした、Q&A形式のナレッジコミュニティーサービス。総量は1億語以上。今回は500万語分を無作為抽出。話題に沿って14カテゴリー(「Yahoo!Japan」「インターネット」「エンターテインメント」「スポーツ」「ニュース」「ビジネス」「マナー」「教養と学問」etc.)に分類。

⇒ いずれも特定目的サブコーパスの一部

検索デモンストレーションサイト

http://www.kotonoha.gr.jp/demo/

検索結果の表示例



試験公開の目的

- 国語研のKOTONOHA計画を周知
- 著作権処理(3万件)の円滑化

著作権保護、個人情報保護意識の高まりを反映して、著作権者との連絡にかかる手間が著しく増大。フェアユースの通念もない

⇒ 著作権処理団体に加入していない著作権者の場合、連絡先が見つかるのは2割。そのうち回答が返ってくるのが半分。回答はほぼ「承諾」。

⇒ 著作権処理の成否がプロジェクト全体に影響

今後の試験公開予定

- 国会会議録（500万語）
- 文芸作品（500万語）
- 新聞記事（100万語）

等を今後1年回に追加する予定。

17

コーパス完成時の公開方法

- オンライン公開1 （無償）

  簡単な語の検索のみ、出力件数に制約(500件まで)

- オンライン公開2 （有償：年間3000円程度）

  高度な検索インターフェースを提供、全出力をダウンロード可

- データ全体の公開 （アカデミック利用20万円程度）

  利用契約を締結した後、DVD等で配布

18

ご清聴ありがとうございました

KOTONOHAコーパスの説明
http://www.kokken.go.jp/kotonoha/

ご意見、ご質問は下記へお寄せください。

kotonoha@kokken.go.jp

19