

(2008.04.24)

国立国語研究所のコーパス開発計画 KOTONOHAIにおけるブログデータの 位置と価値

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

1

コーパスとは？

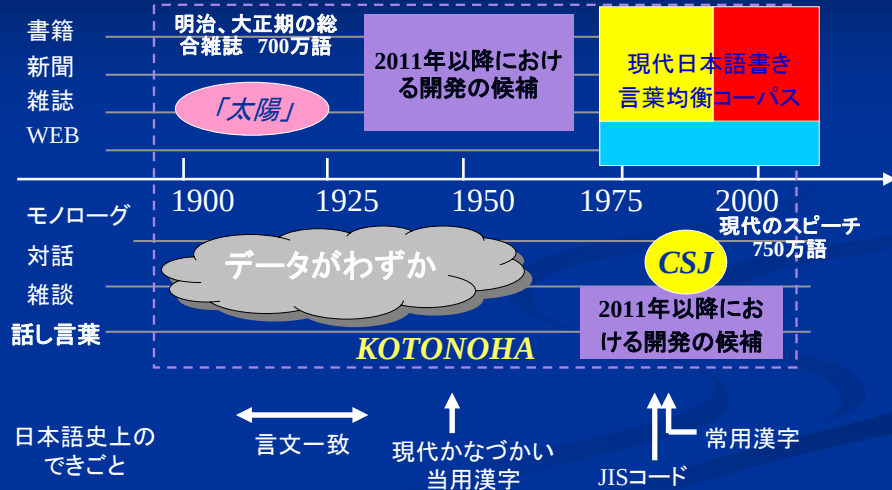
言語研究のための対象とする言語（日本語、英語、中国語、etc.）
の全体をゆがみなく反映したサンプル。その要件は、

- 実際の用例であること
- 規模が大きいこと
- 電子化されていること（コンピュータで検索できること）
- 公開されていること
- 内容が固定されていること
- 研究用の情報が付加していること
- 出典が明らかであること
- 対象言語をうまく代表していること

2

国立国語研究所のコーパス開発計画: KOTONOA

書き言葉



3

『現代日本語書き言葉均衡コーパス』

Balanced Corpus of Contemporary Written Japanese (BCCWJ)



- 全体で1億語以上、著作権処理を施して公開する
- 3種のサブコーパスから構成(そのうち二つは無作為抽出)
- 無作為抽出はできないが欠かせないジャンルがある

4

BCCWJ(2007年版)(約4000万語)にみる 口語表現の文頭での分布

表現	N	白書	知恵袋	文芸	その他	%知恵袋
とすると	46	0	4	25	17	8.7
とはいうものの	21	0	2	5	14	9.5
となると	62	0	6	27	29	9.7
といっても	193	0	46	76	71	23.8
というわけで	38	0	13	16	9	34.2
というか	152	0	115	29	8	75.7
てか	23	0	21	1	1	91.3
つか	9	0	9	0	0	100.0
つか	6	0	6	0	0	100.0

7

「ト」でおわる文

一応第二外国語は中国語だったんですけど中国人のネイティブの先生でもう全然先生言ってること分かんないから授業にならないというような感じで取り敢えず先生が言ってることを繰り返すことしかやってなかったのが何が何の勉強になったんだかちょっとよく分かんないような状態で中国語もできないと

『日本語話し言葉コーパス』の用例 

8

知恵袋の用例

質問: 巨人の西山が引退を表明しましたが、何でわざわざ巨人に移籍したんでしょう？広島に残ってれば良かったのに・・・

回答: 阿部の教育係だったんじゃない？でも阿部のファーストコンバートで教育する必要がなくなったと。

質問: デューク東郷さんは炊事、洗濯などは自分で出来るのですか？全て外食、コインランドリー・オタクですか？

回答: そうですね、「出来る」とは思いますよ。やってる姿は想像したくないけど・・・でも、あんまり外食とか、クリーニングとかばかり利用すると、毒盛られたり、爆発物やら盗聴器やら仕掛けられかねませんしねえ・・・。(そんなコトしても、たちまち見破りそうなのがあのヒトのコワイところですが。)まあ、普段は外食やクリーニング利用でも、必要に迫られれば自分でやる、と。

9

ブログデータの価値

■ 日記的性格

- ・ 自分自身と不特定多数が読み手
Cf. 知恵袋; 顔は見えないが特定の相手が対象
- ・ あまり推敲されておらず、あまり規範にしたがっていない
写真などに依存したテキスト

■ 話し言葉と書き言葉の中間的存在

- ・ 話し言葉と書き言葉の相違を研究するための好材料

10

ブログデータの問題

■ 著作権処理

- ・ 書き手の意向を尊重したテキスト収集手法
- ・ 公開後に問題が生じた場合への対策

■ 個人情報が含まれる可能性

- ・ 情報学研究所との共同研究でガイドラインを設定
- ・ 削除すべきものは削除して公開