

「日本語：数億語のコーパスを作って調べるとみえてくる頻出語,頻出表現」

1. コーパスとは？ (おさらい)

私たちが普段使っている生の言葉を大量に集めた言語データ

- 新聞記事,雑誌記事,小説,教科書,Web ページ... 書き言葉コーパス (written language corpus)
- 会話,講演,TV 字幕... 話し言葉コーパス (spoken language corpus)
- blog, twitter, SNS, 製品レビュー... 中間的なコーパス

→大規模に収集した言語データ (コーパス) を調査すると,言語の特徴が見えてくる (はず)

機械可読(コンピュータで扱える形)の形式であることが最大の利点

- 検索が楽... コーパスの中である「語」や「表現」などが出現する場所をすばやく突き止めることができる.
- 語の統計データなどを取りやすい
 - ある語が何回出現しているか (出現頻度)
 - ある語とよく一緒に使われる語 (共起語)
 - どのくらいよく一緒に使われるか (共起頻度)

→集計したり一覧表にしたり,確率などの計算が可能になりさまざまに応用可能

- 大量のデータを相手にしても比較的容易に処理ができる.

例： 今日の記事の中で「車」という文字が何回でてくるか？

手作業で調べようと思ったら気が遠くなるが,コンピュータの力を借りれば作業は楽.では,一年分の記事に「車」が何回出てくるか？

手作業ではちょっと無理を感じるが,コンピュータの力を借りれば,十分可能な作業.ちなみに,毎日新聞 1995 年には「車」は 7085 回,2000 年は 7690 回.

2. コーパスの入手

- 既存のコーパスを利用
 - 研究目的利用での無償提供されたコーパス
国立国語研究所提供のコーパス (http://pj.ninjal.ac.jp/corpus_center/)
Wikipedia の公開データ
 - レビューなど研究用に提供されているデータ
楽天データ (https://rit.rakuten.co.jp/data_release_ja/)

ライブドアデータ (<https://github.com/livedoor>)

リクルートオープンデータ (<http://atl.recruit-tech.co.jp/opendata/>)

など.

- ▶ 市販されているコーパスを購入
新聞記事データ (毎日新聞, 朝日新聞, 読売新聞, 日本経済新聞...) 非常に高価
- ▶ パブリックドメインのテキスト集を利用
青空文庫 (<http://www.aozora.gr.jp/>) を利用

既存のコーパスを利用すると

- まとまったデータが比較的容易に利用可能
- 著作権処理がされているので使用しやすい
- ×有料だと高価なことも
- ×自分の目的に合う内容だとは限らない

- 自分でコーパスを作成して利用
例: WWW ページ, TV 字幕

3. WWW のホームページをコーパスにする

www の利点は, 大量に存在する, ただで利用できる (研究目的)

- ▶ 作成手順
 1. 元になるデータの収集 (web から収集することを web スクレイピングや web クローリングという)
 2. 収集したデータの加工
 3. 単語や品詞などの言語的情報の付加, 整備

WWW (World Wide Web)の仕組みについて

- インターネット上に分散している情報やサービスをハイパーリンクで関連付けたもの
- アンカー (リンクの付いた語句) やアイコンなどをクリックするだけで情報にアクセス
- WWW に関連するアプリケーション → Web ブラウザ (インターネットエクスプローラ, Google Chrome, FireFox, Opera など)

