Corpus (コーパス)

DOI : https://doi.org/10.60430/digital.e-learning0007

独立行政法人情報処理推進機構
デジタル基盤センター
データ環境勉強会

更新日：2023-05-24
作成日：2023-05-24

目次

1. Corpus (コーパス)とは
2. 例えば、Corpusの考えを基に、文章を品詞分類で分類してみる

1．Corpus (コーパス)とは

- 自然言語による文章などの使用方法を構造化し大規模に集め、説明を記録したもの。構造化された言語情報（動詞、形容詞などの品詞・統語構造）などのタグ付けがされている。

  - 品詞体系の辞書として有名なのものとして、情報処理振興事業協会 (IPAの前身) の IPA 品詞体系 (THiMCO97) に基づいて 作成されたIPA品詞体系日本語辞書（IPADIC）がある。
  - 主な用途として、形態素解析で使われたりする。
  - 入力文から単語の候補を列挙したグラフ構造を生成し、最適な経路を選択して解析する手法を使ったMeCabがある。

2．例えば、Corpusの考えを基に、文章を品詞分類で分類してみる

- 例文１：『あのランチタイム勉強会がパワーアップして帰ってきました。』

あの	ランチタイム	勉強	会	が	パワー	アップ	し	て	帰っ	て	き	まし	た	。
感動詞	名詞	名詞	名詞	**助詞**	名詞	名詞	動詞	助詞	動詞	助詞	動詞	助動詞	助動詞	補助記号

- 例文２：『修正コメントはありませんでしたが、このデザイン最大のポイント』

修正	コメント	は	あり	ませ	ん	でし	た	が	、	この	デザイン	最大	の	ポイント
名詞	名詞	助詞	動詞	助動詞	助動詞	助動詞	助動詞	**助詞**	補助記号	連体詞	名詞	名詞	助詞	名詞

  - このように分解すると、例えば、例文１のように助詞の「が」が、名詞に挟まれている「が」と、例文２のように前に助動詞があるときの「が」では、AI音声でのイントネーションが変わるような仕組みに活用できる。