認知言語学とコーパス分析

李在鎬

背景

理論言語学の盲点
- 作例による一般化は本質的に循環論→実証性の乏しい考察→応用性の欠如
- 認知言語学が例外ではなく、問題視する声が出始めている現在

コーパス分析が切り開く可能性
- 自らの分析モデルを評価する良い手段→意味研究への貢献
- 言語単位の認定に関する新たな方法論→構文文法、語彙意味論への貢献

どんなデータベースが使えるのか

ウェブコーパス（キネット検索）の位置づけ
- ウェブ上のデータをコーパス化したもの
- 自然言語処理の分野で利用

格フレーム辞書、Sketch Engine、ウィキペディアなど
- 長所：大規模でかつ、著作権フリー、簡単、無料
- 短所：書き手が不明、メンテなし、ノイズ多し

長所を生かし、コロケーション調査などでは有効利用可能

書き手によるコーパスの現状と位置づけ
- 新聞、小説など複数のジャンルのテキストあり
- 代表的な問題をクリアするのが難しい
- 利用には制限が多いため、要確認

日英対応記述対応付けデータ、日本語均衡コーパス（国研）、青空文庫
- 長所：書き手・出発が明確、表現の規範性高い
- 短所：手間、コスト高、データのパワーアップ

複数のデータを比較評価しながら利用するので望ましい

どうやって用例を集めるのが

形態素解析
- 対象言語の文法の知識（文法のルールの集まり）や辞書（単語毎の情報付与の単語リスト）を用いて、自然言語で書かれた文章を形態素（Morpheme：語彙で意味を持つ最小単位）の列に分割し、それぞれの分類を理解する作業を指す

Chase, J, M, Mace
- 英語
- 韓国日本語コーパス電子辞典

KICコード
- 棱形表（名著大学校）作成
- 計量社会学ソート
- 分析機能（本）抽出機能が充実

E-KWIC
- 韓国日本語コーパス作成
- 論文（名著大学校）作成
- 英語による形態素解析におけるデータを提供
- エクセル上で動作するインストール不要

どう分析するのか

Excelでできるもの
- エクセルでデータ処理する際の関数を覚える
- 例えば、left, right, mid, findなど
- Excel（エクセル）基本関数：Excel（エクセル）関数一覧表
  (http://www.nrms.dit.ac.jp/~yosenma/Excel/kensei.html)

専属の頻度を利用した分析
- 分布を調べる→ピボットテーブルを使う
- グループ間の類似度を調べる→関係係数を求める
- 語と語の共起性を調べる→共起指数、MIスコアを求める
- 分布の差を計算する→差異係数&カイ二乗検定を行う

Excelでできるもの
- エクセルでデータの下処理をしておく
- 主成分分析で特等を誘導する
- クラスター分析でグループ化する
- 分散分析でグループを比較する

関連研究を見る
- コーパス日本語学文献リスト：http://www30.atwiki.jp/corpus-ling/