第11章
コーパスの作成

7513105
言語文化学部フィリピン語3年
佐藤一輝

目次

1．はじめに

- 自前のコーパスを作る意義

　COCAやBNCのような汎用コーパスでも個人の使用目的に適しているとは限らない

　⇒使用目的に合わせたコーパスが必要

　　e.g.) 教科書の語彙を分析する

　　センター入試の語法問題の傾向を調べる

　　生徒の習得過程を調べる

2. 学習者コーパスの構築

２.１　構築時の留意点

(1) 作文のテーマを統一するかどうか

　　統一する場合：使われる表現の傾向を把握しやすい

　　統一しない場合：多様な表現を得られる

(2) 辞書の使用を認めるかどうか

　　認める場合：辞書を使っても書けない表現がわかる

　　認めない場合：生徒自身の知識を計ることができる

(3) 時間制限を設けるかどうか

　　設ける場合：生徒の実態を把握しやすい

　　設けない場合：調べても書けない表現がわかる

2.2　データの収集と保存（1）

- ワープロソフトを利用してデータを収集すると効率的

- 手作業で電子化する場合はデータの書式に気を付ける
  e.g.) 分析しやすいように1行につき1文を入力する

- スペルチェック機能を使用するかを検討する
  　スペリングミスの傾向を把握するか各単語の頻度を維持するか

2.2　データの収集と保存（2）

- データを保存するときはWord文書として保存するだけでなく、テキストエディタを使用してテキストファイルとしても保存するとよい
  e.g.) サクラエディタの使用

- テキストファイルの保存時に文字コードと改行コードに注意する

- 学年ごと、作文ごとなどトピックごとで表現を比較する場合は、生徒ごとに作成したデータファイルを1つのファイルにまとめる必要がある

２.３　形式のチェックとデータ修正

・不要な空行が含まれていないか、１行に複数の文が含まれていないかを確認する

・できるだけ目視で各行を確認する

３．教科書コーパスの構築

３.１　ファイル名の付け方

- 出版社から教科書のテキストを入手した場合、調査目的に応じてファイルをチャプター別などに分割し、ファイル名も変更する必要がある

  e.g.) 中学１年生用の*New Crown*のLesson5であれば

  「j_ncrown_0105.txt」

- lessonの冒頭にタイトルを記載し、ファイル名と該当するlessonの出典やジャンル情報を表計算ソフトに保存しておけばファイルの詳細がわかるようになる

3.2　サンプリング

- 教科書のテキストファイルが出版社から入手できない場合、「メモ帳」や「サクラエディタ」を使って手入力で作成する

- サンプリングの留意点
  （1）どの学校課程の教科書からコーパスを構築するのか
  （2）どの教科書からコーパスを構築するのか
  （3）教科書のどの部分からコーパスを構築するのか

3.3　テキストの入力

・テキストの入力とは「紙面上の文字データをデジタル化する一連の作業」のこと

・教科書のテキストを電子化する方法
（1）キーボードを用いた手入力
（2）OCRソフトを用いた半自動的な文字の読み取り（修正作業を伴う）
※脚注に掲載された単語やフレーズは対象外にする

3.4　テキストの整形

- テキストの整形とは、テキストに含まれる特定の文字を異なる文字に変えたり、余分な文字列を削除するといった作業のこと
  e.g.) テキストエディタの置換機能を使用した、改行や一行一文の整形

- 正規表現はp138を参照

３.５　タグ付け

- 取り込んだテキストの各単語に品詞タグをつけることで、句レベルの検索やコロケーションの検索、構文レベルの検索が可能になる

- ランカスター大学が提供しているFree CLAWS WWW taggerの使用により、テキストファイルを張り付けるだけでBNCで使われているものと同じ品詞タグが各単語に振られる

４．おわりに

・ 自作のコーパスは小規模であっても有用な資料

・自分用のコーパスを作成し、日ごろの学習指導や英語表現の蓄積・活用に役立ててみよう！