日本語教師のためのコーパスシステム Co-Chu の使い方

山本裕子・小森早江子・本間妙（中部大学）
ラニガン・マシュー（中部大学大学院修了生）

近年のコンピューターの発達に伴って，研究・教育の現場でもさまざまな利用法が提案されるようになった（庵・山内編 2015 他）。しかし，言語の研究者や教師が言語データを分析可能な形式に整えて解析し結果を分析することは，容易であるとは言えない。そこで，本発表ではコンピューターに苦手意識がある人も平易に使えるツールとして開発している，コーパスシステム Co-Chu (Chubu University Corpus System の略)の使い方を紹介する。

テキスト分析には，データ，分析可能なフォーマットと分析するソフトの３つが統一したシステムとして機能しなければならない。しかし，日本語教育や研究に携わる文系研究者ではこうしたシステムを自力で開発することは難しい。そこで，日本語教師や研究者が集めたオリジナルデータを解析可能な形式に変換し，形態素解析にかけ，分析できるようにするコーパスシステム Co-Chu の開発を試みた。Co-Chu はコーパス日本語学のためのウェブアプリケーションであり，【Build】【Import】【Edit】【Analyze】の４つの機能を一つのインターフェイスで使えるようにデザインされている。今後 Co-Chu をフリーソフトウェアとして公開し，様々な研究で利用できるように，現在準備を進めている。

Co-Chu では，語（表現）の検索，頻度，コロケーションの分析が可能である。検索，頻度分析では，話者の属性による表現の違いや，対話の相手による表現の使い分けの有無等を分析することができる。コロケーション分析については，Wei and Li (2013)で提案された N グラム解析と MI スコアの計算方法に基づいて日本語用に改良したものを利用している。コーパスごとに高頻度表現の MI スコアを比較し，語の使われ方にどのような違いがあるのかを分析することができる。

本発表では，日本語教師や研究者が自分で収集したオリジナルデータを，コーパスシステムに取り込んで使用する方法を紹介する。まず，データを取り込む作業として，以下の４点について説明する。ここでは Co-Chu の【Build】【Import】【Edit】の機能を使用する。

1. CSV ファイルの作成方法：話者情報の入力方法，実際の発話をラインごとに入力する方法，その他数字を含む語や固有名詞など入力上の注意事項など
2. 【Build】：コーパスの階層構造の設定の仕方など
3. 【Import】：データ取り込み手順，データ取り込み時の注意事項など
4. 【Edit】：UniDic・MeCab に基づく形態素解析，N グラム解析と MI スコアの算出，形態素解析後のエラー処理等の自動処理，取り込み時に気付いた解析の不備や誤りの修正の仕方，検索結果を見て気付いた解析の不備や誤りの修正の仕方，その注意事項など

このようにしてデータを取り込んだ後，【Analyze】の機能を使用して，検索，頻度分析や

コロケーション分析を行う。今回の発表ではオリジナルデータである，中国人上級日本語学習者，韓国人上級日本語学習者と日本語母語話者の雑談のデータを用いて，日本語母語話者と学習者の言語使用の実態を比較した結果を紹介する。一見，こなれた会話をしているように思われる上級学習者であっても，分析してみると，日本語母語話者と比べて機能的要素の使用が少なく種類も限られることや，話し言葉としては不自然な表現を用いたりすることが示された。このように母語話者と学習者のコーパスを比較し，中間言語の特色を明らかにする試みを通して，Co-Chu の利点を紹介したい。

図 1　Co-Chu の Analyze モード：検索項目と結果

引用文献・関連 URL
庵功雄・山内博之編(2015)『データに基づく文法シラバス』くろしお出版
Wei, N., and Li, J. (2013). A new computing method for extracting contiguous phraseological sequences from academic text corpora. *International Journal of Corpus Linguistics*, 18:4, 506-535.
MeCab　https://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
UniDic　http://www.ninjal.ac.jp/corpus_center/unidic/