日本語教育・研究のためのテキスト分析システム"Co-Chu"
Co-Chu: A TEXT ANALYSIS SYSTEM FOR JAPANESE LANGUAGE EDUCATION AND RESEARCH

小森早江子，山本裕子，本間妙，ラニガン マシュー*
中部大学，*中部大学大学院
Saeko Komori, Hiroko Yamamoto, Tae Homma, Matthew Lanigan*
Chubu University, *Chubu University Graduate School

概要：日本語教育・研究のためのテキスト分析システム Co-Chu の開発について紹介する。コーパスを利用した日本語研究のためのテキスト分析には、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１つの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、研究者や教師が集めたオリジナルデータを統一した形式に変換し、形態素解析にかけ、分析できるようにするシステム Co-Chu の開発を試みた。今回は、そのシステム（語彙頻度表作成、共起頻度表作成、MI スコア算出、語・句の検索など）とそれを使った研究事例について報告する。

キーワード：コーパス　データ化　母語話者と学習者　コロケーション　MI スコア

１．はじめに

日本語教育や研究に携わる研究者のための日本語テキスト分析システム Co-Chu (Chubu University Corpus System の略)を開発した。このシステムは研究者や教師が集めたオリジナルデータを統一した形式に変換し、形態素解析にかけ、分析できるようにデザインしたものである。ここでは、Co-Chu のシステムの概要、コロケーション検出の手法、システムを使った分析例について説明する。

２．コーパスシステム Co-Chu

２.１　概要と位置づけ

コーパス開発にはコンピューター技術が必要である。もちろん利用可能なコーパス分析ツールもあるが、コンピューターに関する知識の乏しい日本語教師や文系研究者が自分で集めたデータを自力で分析するのはかなり困難である。簡便に利用可能な日本語テキスト分析ツールを備えたシステムがあれば、日本語教育や日本語研究に大きく貢献できる可能性があると考えて、システムの開発をおこなった。

テキスト分析システム Co-Chu には、話し言葉データおよび書き言葉データのコーパス構築の手法と、ウエブ上で簡単に利用可能な分析機能の開発に関する部分がある。

２.２　Co-Chu によるコーパス構築の手法

Co-Chu のコーパス構築部分はサブコーパスとサンプルの２つで構成されている。サブコーパスは、Co-Chu の中ではグループと呼び、無制限に組み入れることができる。サブコーパスには「サンプル」があり、基本的に１つの発話、あるいは文章を示す。

コーパスを作成する際は、まずテキストデータを分析できるように統一したフォーマットに変換しなければならない。対応できるデータフォーマットは TXT、XLS と CSV の３つである。さらに、データに「読み選定タグ」「語形選定タグ」「辞書エントリータグ」などシステム用のタグを付ける必要がある（ラニガン 2015 参照）。データの階層構造に合わせてサブコーパスを作って、データを取り込む。データの取り込みの際には形態素解析およびコロケーションなどが自動的に計算される。

２.３　Co-Chu のコーパス分析機能

Co-Chu の開発では主に形態素解析ソフト MeCab と電子化辞書 UniDic を利用する。現時点で開発したシステムの分析ツールは形態素連鎖検索とコロケーション検索の２種類である。

１つ目の形態素連鎖検索は、形態素の連鎖に条件を付けて検索するものである。例えば、「行かない」の場合、

例）行か：品詞大分類＝動詞 AND 活用形＝未然形

＋

ない：品詞大分類＝助動詞 AND 語彙素読み=ナイ

のように動詞の未然形の後ろに助動詞の「ナイ」が続く連鎖を指定し、動詞の否定形を検索することができる。検索の条件は UniDic のフィールド全てが利用できるように作られている。

また、連鎖検索のアウトプットとして様々な設定ができる。アウトプット形式は、全ての検索結果のサンプルを出力する際、サンプル形態素詳細と KWIC の２種類から選択できる。また、条件を指定し、サンプルを数えたアウトプット形式も選択できる。例えば、結果の前に現れる語の品詞で数えることなども設定できる。

２つ目の分析ツールはコロケーション検索である。データを取り込む際に、形態素解析をおこなうが、同時に語彙頻度表と共起頻度表を作成し、n-gram に基づくコロケーションも計算するように設計されている。また、それぞれの分析ツールは全体のデータに加え、サブコーパス別分析もできる。コーパス研究において、語と語の結びつき、つまりコロケーションの強さを表す指標として MI スコアが使われている。例えば、現代書き言葉均衡コーパス（BCCWJ）の白書や雑誌などのサブコーパスを対象として表現を抽出する場合、MI スコアを算出し、比較することによってそれぞれのサブコーパスにおける語の結びつきを観察することができる。一般的に MI スコアの算出にはそれぞれの語の頻度と２語の共起頻度（2-gram）で計算するが、Wei and Li (2013)では英語のコーパスを例に 2-gram 以上算出する手法を提案している。日本語では語の区切り方が問題と

なるため、BCCWJ（DVD 版）の短単位と長単位のデータを用いて、高頻度表現の MI スコアの結果を比較し、それぞれの単位によるコロケーション表現の特徴について考察する。

３．分析例

この分析ツールを活用してどのような分析が可能か、実際の分析例を次節で紹介する。

３.１　補助動詞

文末形式の１つである補助動詞表現は、日常的に頻繁に使用され自然な日本語には欠かせないものである。しかし、近藤他（2010）、水谷（2011）、山本（2010）では補助動詞の使用が母語話者と比べて日本語学習者は少なく、使用法が異なっていることが指摘されている。また日本語母語話者にとって、補助動詞は事態の捉え方を表す有力な手段の１つであり、補助動詞の有無が自然さに大きく影響する。よって、補助動詞がどのように使われているのか実態を適切に把握することは日本語教育にとっても有益である。

本発表では、まず BCCWJ で 11 種類の補助動詞がどのように用いられているかを、頻度、承接関係、共起表現等の観点から詳細に分析する。そして、それを踏まえ、日本語母語話者の話し言葉、日本語学習者の話し言葉での補助動詞の使用を分析し、どのような特徴があるか示したい。

３.２　「ちょっと」の使用分析例

ここでは、日本語談話に頻出する「ちょっと」に注目して分析する。「ちょっと」は、程度・量副詞としてだけでなく、呼びかけ、伝達内容の和らげ、強調、フィラーなどの多くの用法がある（本間 2011、彭 2004）。しかし、それらの用法が用いられている量的な実態は把握されていない。さらに、多用な用法があるにも関わらず、日本語学習者が使用する用法は限られているようである。これは、学習者の母語の中の「ちょっと」に相当する語と、日本語の「ちょっと」の意味・用法が異なる（彭 2004）ためではないかと考えられるのだが、実際の学習者の「ちょっと」の使用実態も数的には明らかにされていない。

本発表では、３.１に倣い、まず BCCWJ で「ちょっと」のコロケーションの頻度、頻度の高いものの MI スコアなどから用法の特徴を捉える。次に、日本語母語話者、日本語学習者の話し言葉コーパスに見られる用法の特徴を比較する。これにより、日本語教育における接触場面指導への示唆を得られると考える。

４．まとめと今後の Co-Chu 開発

本発表では、今回開発した日本語テキスト分析システム"Co-Chu"の紹介と、システムを使った研究事例を報告した。今後も Co-Chu の開発を続け、様々な機能を加えようと考えている。現時点では、新しく本システムに取り込んだデータに関しては、MeCab による短単位形態素しか解析されないが、中・長単位形態素解析ソフト Comainu を使って中単位および長単位解析機能も追加す

る計画である。また、係り受けソフト CaboCha を利用し、コロケーション検索をより効果的にする。さらに、本システムで未解決の問題、例えば、同時発話や相槌など、話し言葉特有の問題の扱いなども考慮し、どのようにして解析可能なデータにするのかという解決策を見出さなければならない。これらについては、今後の課題としたい。システムは構築途中であるが、今後も日本語教育・研究に貢献できるようなシステムを目指して充実させるとともにシステムを活用して日本語母語話者と学習者の言語使用に関する研究も続けていきたい。

参考文献

近藤安月子他（2010）「中国語母語日本語学習者の事態把握－日本語主専攻学習者を対象とする調査の結果から－」JCLA2010, pp.690-706

本間妙（2011）「実質的意味を持つフィラーの研究談話的研究－特定のインタラクションに表出する『ちょっと』『なんか』『やっぱり』－」中部大学博士学位論文　未公刊

彭飛（2004）『日本語の「配慮表現」に関する研究』泉書院

水谷信子（2011）「補助動詞に見る立場志向表現－授受表現と『くる、いく』を中心に－」日本語教育世界大会 2011

山本裕子（2012）「話しことばにおける移動の補助動詞『～テイク』『～テクル』の使用傾向」, AATJ Annual Conference.

ラニガン　マシュー（2015）「コーパスシステム『Co-Chu』の開発―MeCab 拡張データ処理機能について―」『第 7 回コーパス日本語学ワークショップ予稿集』国立国語研究所

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, pp.506-535.

資料＆関連 URL

国立国語研究所『現代日本語書き言葉均衡コーパス（BCCWJ）』DVD 2011年度版

Co-Chu　http://www.co-chu.org/

MeCab　https://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

UniDic　http://www.ninjal.ac.jp/corpus_center/unidic/

Comainu　http://comainu.org/

CaboCha　https://code.google.com/p/cabocha/