KH Coderで何ができるか
日本語習得・日本語教育研究利用への示唆

佐野香織・李在鎬

1. はじめに

本稿の目的は、データ・コーパス分析に優れた機能を備えるツール＝KH Coder[1] を紹介し、日本語習得･日本語教育研究への利用可能性を示すことである。

KH Coder は、日本学術振興会(大阪大学)の樋口耕一氏によって製作され、本来は社会学の分野での利用が想定された内容分析およびテキストマイニング用のソフトウェアであるが、言語研究のためのコーパス分析にも優れた機能を発揮するツールである。

日本語・日本語教育研究においては言語データ処理にコンピュータ･ツールを利用した研究はまだ多いとは言えない(山内 2004)。その背景には、多くの日本語・日本語教育研究者のコンピュータ・リテラシーの不足(滝沢 2006)や、研究に適したツールがないという問題がある(大曾 2006)。

KH Coder は上記のような問題に応える機能を持つフリーソフトウェアである。日本語習得・教育研究分野においても大きな可能性を持つと思われる。以下にこの KH Coder の特徴を述べ、紹介を行う。

2. KH Coderの特徴

このソフトウェアの特徴として主に次の 3 点が挙げられる。

① 無償でウェブサイトから入手でき、インストールから実際の分析にいたるまで簡単なマウス操作のみで使用可能である。
② 優れたデータ抽出機能を備えており、テキストから様々な情報を簡単に取り出すことができる
③ 他のデータ分析処理アプリケーション(Excel 等)との連携力に優れている。

以下この 3 点について順次解説する。

2.1 コンピュータ・リテラシー

コーパスを使用して分析を行うためにはコマンド使用を含め、ある程度のコンピュータ・リテラシーが必要であるが(grep 等)、KH Coder には形態素解析ソフト「茶筌[2]」の解析情報に基づくデータ抽出機能が組み込まれており、ごく簡単なマウス操作でほとんどの操作が可能である。

2.2 データ抽出機能

コーパスに基づく研究では言語データから自分が知りたい基本情報をスムーズに得られることが重要である。しかし例えばある語の使用頻度を調査したいという場合、その語の品詞認定をどうするのかという問題やどのようにして人為的なミスなく調査するのかという問題がある。

KH Coder では自然言語処理の分野で高い精度と有効性が示されている前出の「茶筌」を組み込んでおり、上述の問題を解消している。またキーワード入力による方法(以後 KWIC(Keyword in Context の略))で分析データが抽出できExcel などで読み取り可能なデータ形式として出力できる。

また、①文字列に基づく検索②品詞に基づく検索③文字列と品詞の両方に基づく検索④検索の③に加え、前後のパターンに基づく検索、の 4 つの検索オプションが用意されている。特に注目すべき点として③や④を使用した場合、大曾(2006)が指摘する問題が解消できるということがある。例えば文字列のみによる検索の限界として「味噌汁を飲む」で検索を行った場合、「味噌汁を飲んだ」「味噌汁を飲んでいる」等の活用形の検索は困難であると指摘されているが、KH Coder では茶筌による解析情報を手掛かりにデータ抽出を行うため、すべての活用形を 1 度のキーワード入力で抽出可能である。

表 1 ではKY コーパス中の韓国語母語話者学習者発話(コーパス 1 : 延べ 76,702 語)と英語母語話者学習者発話(コーパス 2 : 延べ 77,996 語)の中から上述のオプションに対応する KWIC の簡単な実例を示した。

表 1　KWIC の分析結果実例

検索ターゲット	コーパス 1	コーパス 2
①文字列としての「が」	1,643	1,205
②格助詞の全体	5,125	4,159
③格助詞の「が」	1,403	979
④動詞の直前の「が」	179	88

紙幅の都合上、表 1 の解釈の詳細には立ち入らないが、KH Coder を使用すれば、瞬時に表 1 の

ような集計が行えるという利点がある。

2.3 アプリケーションとの連携

KHCoder はデータを抽出するのみならず、そのデータを Excel で自動的に整形可能にする Excel アドインや SPSS のような統計解析用アプリケーションと強い連携関係にある。このため抽出データの二次的な利用においても非常に便利なツールと言える。

3. 日本語習得・教育研究利用への示唆

佐野(2006)では長期定住ブラジル人の発話データを分析しているが、KH Coder を使用すれば以下のような使用頻度の高い動詞や、動詞「行く」の前後パターンを見る検索が可能であり、使用実態を詳細に見られるという利点がある。

表 2　動詞使用頻度上位 4 語

1	分かる	60
2	違う	40
3	行く	29
4	ある	28

表 3 ある動詞が用いられるパターン調査(例:行く)

前の文脈			ｷｰﾜｰﾄﾞ	後の文脈	
左 3	左 2	左 1		右 1	右 2
	どこ	に	行き	ません	
	デパート	に	行く		
を	見	に	行く		

これらの機能は日本語教育研究にも多いに有効活用できると考えられる。例えば「期待されてならない」の「〜てならない」のようなある言語形式を対象に実際の言語使用を調査する研究(杉村 2005 等)や、コーパスを使用してコロケーションから学習者の誤用を分析するというような研究にも使用可能であろう(大曾・滝沢 2003)。

4. おわりに

KH Coder は言語・言語教育を念頭において開発されたツールではないため、もちろん使用には限界もある。例えば、KH Coder におけるすべての処理は茶筌の解析情報をベースにしているため、調査の精度は茶筌の精度に依存することになる。茶筌は機械翻訳などの工学的な利用を目的としており、大量のデータ(例えば日刊紙の 10 年分の記事)を効率良く処理するために設計されたものであることから、茶筌の規定自体が不自然な場合もあることは否めない。最終的には人間によるチェックが必要な場合もある。このようなツールの限界も考慮にいれつつ、研究に活かし使用することが望まれる。

こういった点を含めても KH Coder は日本語・日本語教育研究の発展に貢献可能な優れたツールであるといえよう。また、このようなツールを無償で公開し、かつ掲示板でユーザーの様々な要求に耳を傾けてくれる製作者、樋口耕一氏の人柄の良さが KH Coder の使い心地を更に良くしてくれることも強調したい。

注

1. KH Coderの詳細はhttp://khc.sourceforge.net/参照。
2. 奈良先端大学院大学の松本研究室で開発された形態素解析ソフト。http://chasen.naist.jp/hiki/ChaSen/ 参照。

主要参考文献

大曾美恵子(2006)「日本語コーパスと日本語教育」『日本語教育』130, 3-10.

佐野香織(2006)「日本社会で生活する成人ブラジル人の言語使用－用法基盤モデルの観点から－」日本語教育国際研究大会（コロンビア大学）発表資料

滝沢直宏(2006)「コーパス利用のためのコンピュータ・リテラシー」『日本語教育』130, 22-31.

山内博之(2004)「語彙習得研究の方法―茶筌と N グラム統計」『第二言語としての日本語の習得研究』7, 141-162.

さの　かおり／お茶の水女子大学大学院　人間文化研究科　　り　じぇほ／(独)情報通信研究機構