〔2020 年度日本語教育学会秋季大会（オンライン開催，2020.11.29 口頭発表⑦〕【予稿集】

コーパス分析システムの公開と日本語教育・日本語研究への活用

山本　裕子（愛知淑徳大学）・本間　妙（愛知淑徳大学）
川村　よし子（東京国際大学）・小森　早江子（中部大学）
＜共同研究者＞ラニガン・マシュー（Lending Home）

1. はじめに

筆者らは、自ら収集したデータを自由に取り込んで活用できる操作の平易なコーパス分析システム Co-Chu（以下 Co-Chu）を開発し、今回、一般公開を開始した。近年、ICT 技術の進歩に伴い、日本語教育の分野でも様々なコーパスが提供され、多様な活用法が提案されている（庵・山内編 2015)。しかし、自ら収集したデータを分析できるツールは限られている。Co-Chu は教師自らが集めた日本語学習者の作文や会話等のデータを取り込み、キーワード検索、コロケーション分析、複数データの比較等のテキスト分析が可能なウェブ・アプリケーションシステムである。データには必要に応じてタグを付すことができ、そのタグの検索も可能である。また、一般に学習者の作文や話し言葉を含んだテキストをそのまま形態素解析すると誤解析を生じることも多いが、Co-Chu はタグの活用によりそれらに対処する仕組みも整っている。

本発表では、前半でシステムの概要とメタ情報の付与について解説し、後半は、日本語教育や日本語研究での活用例として、作文教育、日本語教師養成課程、日本語研究における 3 つの実践例を紹介する。

2. Co-Chu の概要

2.1 システムの概要

Co-Chu は、日本語テキスト分析のためのウェブ・アプリケーションであり、【Build】【Import】【Edit】【Analyze】の 4 つの機能が 1 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。Co-Chu 最大の特徴は自ら集めたデータ（csv あるいは txt 形式で作成）でコーパスを構築し、多種多様な検索が可能な点である。（詳細は山本他 2020 を参照）

2.2　メタ情報の付与

日本語学習者の発話や作文などに限らず、日本語母語話者の話し言葉にも、言い間違い、言い直し、言い淀み、フィラー、縮約形など、不規則なものが多く含まれている。これらは形態素解析において誤解析の原因となるが、Co-Chu では【Edit】機能を活用し、メタ情報としてタグを付し、誤解析が生じないようにする。タグは次のような形で付す。

｜正しい語形｜（タグ種別　実際に用いられた語形：タグメモ）
例１　へえ、あいつ、いいところ｜ある｜（話　あん：音変化）な。

例２　それ忘れ｜ないで｜（誤　なく：活用)、あの、次に。

例１は、「へえ、あいつ、いいところあんな」のように、「ある」が「あん」と発音された発話である。これをそのまま形態素解析すると、「あんな（形状詞[(1)])」と解析されてしまう。そこで、例１のようにタグを付すことで、正しく解析できるようになる。また、これは話しことば特有の音変化によるものであることを、タグ種別「話」とタグメモ「音変化」で示している。例２は、「忘れないで」を、活用を誤って「忘れなく」と発話したものである。例２には、誤用であることを示す「誤」タグを付し、タグメモには誤用の種類を「活用」と記載した。このタグ、タグメモはいずれも検索の際に活用することができる。

また、誤用には「ねじれ文」のように、形態素にタグを付して示すことが困難なものもある。Co-Chu では、こうした場合にタグではなく、文ごとにメタ情報をつけることで問題があることを示すことが可能である。メタ情報は、取り込み前のエクセルデータに必要な情報を加えることができる。図１のように該当する文の左の「列」に情報を記載すると、文レベルのメタ情報も、検索の際に活用できる。

メタ情報	発話内容/作文内容
	私は去年日本に来たばかりだ。
ねじれ文	私は、日本は果物がおいしい。

図１　文レベルのメタ情報の付与

このようにタグや文レベルのメタ情報は、形態素解析の上で問題になる箇所や誤用だけでなく、必要に応じて自由に付すことが可能である。これらを利用することで、誤解析が避けられるだけでなく、見たいものを探すための効率のよい検索が可能になる。

次に、Co-Chu を用いた実践例を紹介する。

3. Co-Chu の活用実践

3-1 作文教育での活用

Co-Chu では、目的に応じてタグを付すことで見たい箇所を分析できるという特質に加えて、データ上に存在しない語句や表現にタグを付して検索することもできる。助詞の誤用を例に説明する。一般のコーパスでは、例えば助詞の「が」が使われた文を検索することによって、正用と誤用の例を探しだすことはできる。しかし、助詞の誤用には、使うべきところで使っていない（非用)、使い過ぎている（過剰使用）のようなものもある。Co-Chu ではこれらにもタグを付すことで検索できるので、誤用を網羅的に記述、検索することが可能である。ここでは、これらを利用して、指導中の学習者の作文に見られる誤用の傾向を読み取り、指導に活用した試みについて報告する。まず、作文上の全ての誤用にメタ情報として、以下のような形でタグを付した。

｜正しい語形｜（タグ種別　実際に用いられた語形：タグメモ）

例 3　その言葉は｜どのような｜（誤　どんな：書）意味を持っているのか、
例 4　時々うまくコミュニケーション｜が｜（誤　：助詞脱落が）取れない。

タグ種別には「誤用」を示す「誤」、タグメモには誤用の種類を記載している。例 3 の誤用の種類は「どんな」を使用したことで書き言葉のルールに違反していることを示すために「書」と記載した。また、例 4 は、助詞「が」が入るべきなのに脱落しているため、タグメモに「助詞脱落が」と記載した。このように、学習者の作文に見られる全ての誤用に「誤」タグを付すとともに、タグメモに誤用の種類を 18 種類に分類して記載した。また、文レベルの誤用である「ねじれ文」については、メタ情報として記載した（図 1 参照）。これらを用いて、学習者がそれぞれどのような誤用を犯しているかを検索した。図 2 に検索結果の一部を示す。この結果をもとに、個々の学習者に対し指導項目を決定し、図 3 のような指導シートを作成し、それを用いて個別に指導した。

図 2　タグ「誤」で検索した結果画面

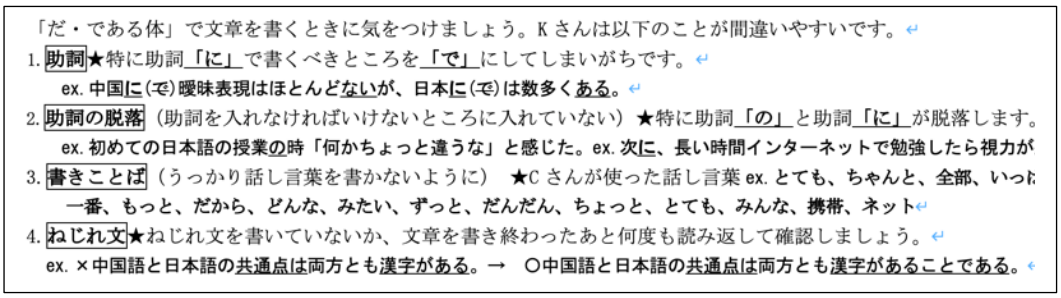
「だ・である体」で文章を書くときに気をつけましょう。K さんは以下のことが間違いやすいです。
1. 助詞★特に助詞「に」で書くべきところを「で」にしてしまいがちです。
ex. 中国に(で)曖昧表現はほとんどないが、日本に(で)は数多くある。
2. 助詞の脱落（助詞を入れなければいけないところに入れていない）★特に助詞「の」と助詞「に」が脱落します。
ex. 初めての日本語の授業の時「何かちょっと違うな」と感じた。ex. 次に、長い時間インターネットで勉強したら視力が
3. 書きことば（うっかり話し言葉を書かないように）　★C さんが使った話し言葉 ex. とても、ちゃんと、全部、いっぱ
一番、もっと、だから、どんな、みたい、ずっと、だんだん、ちょっと、とても、みんな、携帯、ネット
4. ねじれ文★ねじれ文を書いていないか、文章を書き終わったあと何度も読み返して確認しましょう。
ex. ×中国語と日本語の共通点は両方とも漢字がある。→　○中国語と日本語の共通点は両方とも漢字があることである。

図 3　指導シートの一例（抜粋・誤用例は全て学習者 K の作文中に出現したもの）

このように「誤用」に特化したタグを利用することによって、具体的に学習者の誤用の傾向を把握し指導することが可能になる。また、この指導により、学習者は自らの問題点を適確に理解し、作文を書く際にそれらを意識するようになった。

3-2 日本語教師養成課程での活用

Co-Chu の用例検索機能は、コンテキストの提示も容易にできるため、日本語教育だけでなく日本語教師養成課程においても有用である。学習者と日本語母語話者の会話や、アニメやドラマのスクリプトをデータとして取り込むことで、誤用例やコンテキストが想起しやすい身近な用例を授業で提示し、臨場感のある教室活動の実践が可能になった。教科書等で示されている例文と実際の用例には乖離がある場合があるが、養成課程を受講している学生は一般に学習者との接触経験に乏しく、そうしたことに気づきにくい。

例えば、教科書では次のように、構文が理解しやすい例文が示される。ここでは使役を例にする。

例 5　教科書の例文　A：お子さんに何かうちの仕事をさせていますか。
B：ええ。食事の準備を手伝わせています。
A：そうですか。いいことですね。」(『みんなの日本語 48 課練習 C』)

しかし、実際にはこのように格関係が明確な例ばかりが用いられるわけではない。会話やアニメのスクリプトからは次のような使役の用例が検索できる。

例 6　アニメからの例文
a.「雫のしたいようにさせようか、母さん」(『耳をすませば』)
b.「あの、ここで働かせてください！」(『千と千尋の神隠し』)

例 7　会話からの例文
a.「私はそれは言わせねえぞ！って思ってるんだけど・・(以下、略)」
b.「遊んでそうに見られるからこそ、安心させてあげようみたいな気持ちになるというか・・」

これらの例のように、「誰が誰に働きかけているのか」は文脈から読み取る必要がある。運用に結びつく教室活動のためには、実際の使用例や誤用例を知ることが不可欠である。BCCWJ 等の大規模コーパスでも、検索した用例の前後の文も確認はできるものの、学生にとってはなじみのない文脈も多い。Co-Chu では、学生たちがよく知っているアニメやドラマをコーパスとして活用することも可能である。実際に、アニメやドラマからの用例を用いたところ、例 6 のように、短文であってもコンテキストが想起しやすいものも多く、有用であることがわかった。また、必要があれば、図 4 のように、データ全体の確認も容易にでき、各例文が使われたコンテキストを示すこともできる。

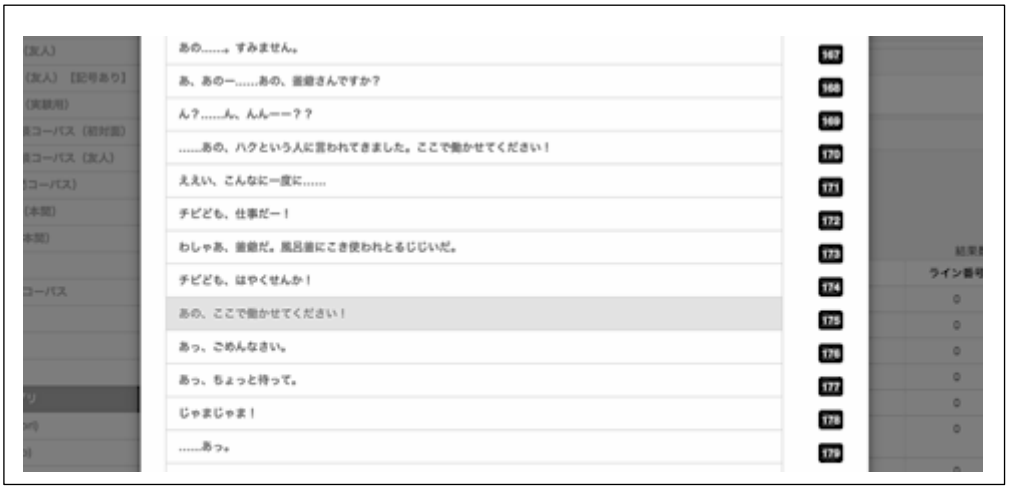

図 4 検索結果から、コンテキストを確認した画面

3-3 日本語研究での活用例

こうした授業での活用のほか、Co-Chu は日本語研究にも大きく貢献できる。ここでは、従来のツールでは検索しにくいものを検索し、日本語研究に活用した試みを紹介する。従来のコーパスでは、可能形や受け身形は、形態上の区別がしにくいため、検索が難しい。また、助詞の脱落のような存在しないものを検索することも困難である。しかし、本システムでは「ら抜き言葉」や「受け身形」にタグを付すことで、実際に会話の中にどのくらいの「ら抜き言葉」が用いられたのかを「ら抜きでない可能形」と比較して抽出したり、どのくらい受け身が用いられ、その中に「間接受け身」がどのくらいあるかを抽出したりすることも容易にできる。

例 8　ら抜きのタグ付け：元の文　全部は食べれない。
→　全部は｜食べられ｜（可　食べれ；ら抜き）ない

例 9　受け身のタグ付け：元の文　ずっとバカにされてきて・・
→　ずっとバカに｜されて｜（受　されて：直接）きて…。

例 10 助詞の脱落のタグ付け：元の文　そんなことないよ。
→　そんなこと｜は｜（脱　）ないよ。

上記のようなタグを付すことによって、たとえばアニメ（『バクマン』8 話分、約 200 分）でら抜きになりうる動詞の可能形において、「非ら抜き」が 10 回に対し、「ら抜き」は 1 回であったのに対して、大学生の雑談（5 組分、約 180 分）では「非ら抜き」が 6 回に対し、「ら抜き」も 6 回使われていることが分かった。図 5 に示したように、誰の発話であるかも「id」欄で容易に確認ができるので、「ら抜き」が個人的な傾向であるのか否かといったことに関しても分析することが可能である。

ライン	サブコーパス名	タグメモ	ライン番号	出現形	id
みんな決め｜られ｜（可　られ：非ら抜き）ずに	B_ep1	非ら抜き	236	られ	秋人
そんなに簡単に話しかけ｜られる｜（可　られる：非ら抜き）くらいならこんな｜に｜（脱　）苦しくないって。	B_ep2	非ら抜き	140	られる	最高
次の試験｜は｜（脱　）受け｜られ｜（可　られ：非ら抜き）そうか？	B_ep2	非ら抜き	235	られ	秋人
せっかく来たんだからネーム｜を｜（脱　）　｜見られる｜（可　見られる：非ら抜き）だけ見る。	B_ep3	非ら抜き	95	見られる	秋人
よく　同じ教室で知らんぷりしてい｜られる｜（可　られる：非ら抜き）よ。	B_ep4	非ら抜き	381	られる	秋人
そんなにすぐ｜寝られる｜（可　寝られる：非ら抜き）かよ。	B_ep5	非ら抜き	253	寝られる	最高
なんかいても立っても｜いられ｜（可　いられ：非ら抜き）なくて…。	B_ep6	非ら抜き	44	いられ	最高
そう考えたら　急に嫌になったっていうか　｜耐えられ｜（可　耐えられ：非ら抜き）なくなって。	B_ep6	非ら抜き	93	耐えられ	秋人
その３話は　いくらでも時間がかけ｜られる｜（可　られる：非ら抜き）。	B_ep17	非ら抜き	162	られる	最高
こんな事って信じ｜られる｜（可　られる：非ら抜き）？	B_ep26	非ら抜き	264	られる	亜豆
僕は恥ずかしくて顔｜が｜（脱　　）｜見られ｜（可　見れ：ら抜き）ないけど。	B_ep26	ら抜き	358	見れ	最高

図５　タグ「可」でアニメスクリプトの「非ら抜き」と「ら抜き」を検索した結果の画面

4. おわりに

このように、Co-Chu では、利用者が各自のデータを目的に合わせて活用し、多様な試みを行うことが可能である。今回、このシステムの一般公開(2)を開始した。Co-Chu が日本語教育関係者の教育や研究活動の一助になることを願っている。

注

（1）　形状詞は、形容動詞のことを指す。

（2）　https://co-chu.org　を参照のこと。

参考文献・参考資料

（1）　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.

（2）　山本裕子・本間妙・川村よし子（2020）「コーパス分析システム Co-Chu におけるタグ検索機能とその活用-誤用や話し言葉にどのように対応するか-」『中部大学人文学部研究論集』43,1-24.中部大学人文学部.

（3）　BCCWJ　https://pj.ninjal.ac.jp/corpus_center/bccwj/　2020 年 6 月 11 日閲覧

（4）　『みんなの日本語初級Ⅱ』スリーエーネットワーク.

アニメ

「千と千尋の神隠し」「耳をすませば」http://www.ghibli.jp/works/　　2020 年 9 月 10 日閲覧

「バクマン」https://www.nhk.or.jp/anime/bakuman/1st/character/index.html

2020 年 9 月 10 日閲覧