コーパス分析システムの公開と日本語教育への活用

山本　裕子（愛知淑徳大学），本間　妙（愛知淑徳大学）
Matthew Lanigan（LendingHome），川村　よし子（元東京国際大学）
小森　早江子（中部大学）

要旨

日本語研究や日本語教育において，さまざまな形でコーパス活用の有効性が指摘されている。しかし，手元にある学習者のデータ等を実際に活用するのは容易ではない。本研究グループでは，汎用性が高く，かつ，コンピューターが苦手な研究者や教師でも利用可能なツールとして，コーパス分析システム Co-Chu（以下 Co-Chu）を開発し，一般公開を開始した。Co-Chu は，研究者や日本語教師が自ら収集したオリジナルデータを取り込んでコーパスを作成し，分析できるウェブ・アプリケーションである。本稿では，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を紹介する。

【キーワード】　コーパス分析システム，オリジナルデータ，分析の多様性，ウェブ・アプリケーション，タグ検索機能

Keywords:　corpus analysis system, original data, various analyses, web application, tag search function

1　はじめに

近年，言語研究や教育の現場において，さまざまな形でコーパスの活用が提案されるようになっている。コーパスによって個人の内省だけでは容易に知ることができない様々な指標を得ることが可能になり，日本語や日本語教育研究の分野においても多くの新たな貢献が可能であると注目されている。しかし，コーパスの活用が日本語研究や日本語教育に有益だとわかってはいても，コンピューターに関する知識の乏しい日本語教師には，自分の手元にある学習者のデータなどをコーパスとして活用するこ

とは容易ではない。

そこで，筆者らは，汎用性が高く，かつ，コンピューターが苦手な研究者や教師でも利用可能なツールとして，コーパス分析システム Co-Chu（以下 Co-Chu）を開発し，一般公開を開始した[1]。Co-Chu は，研究者や日本語教師が自ら収集したオリジナルデータを取り込んでコーパスを作成し，分析できるウェブ・アプリケーションである。本稿では，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を紹介する。

2　システムの概要

Co-Chu は，日本語のテキストを分析するためのウェブ・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でコーパスを構築し，多種多様な検索が可能な点である。以下では，データの作成から，コーパス構築，分析までのプロセスを概観する。ここでは紙幅の都合上，ごく簡単に手順を紹介する。詳細は Co-Chu マニュアルを参照されたい。

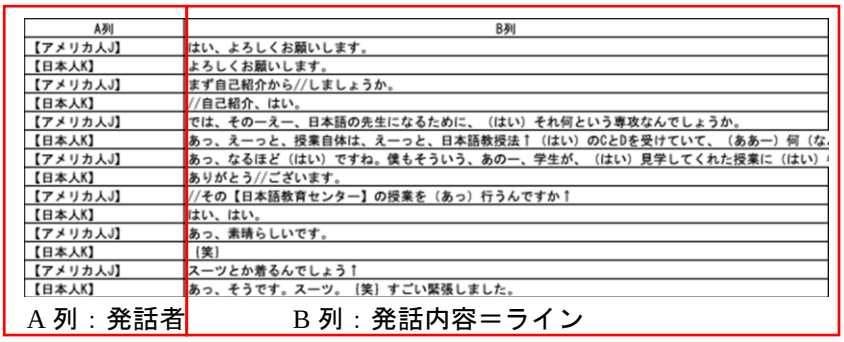
2.1　データ作成

はじめにコーパスに取り込むためのデータ作成について述べる。

Co-Chu では，CSV あるいは txt 形式で作成したデータを取り込むことができる。例えば，会話データの場合は，発話者の列と発話内容の列を分け，CSV 形式でデータを作成する。図 1 に例を示す。

図 1 では，A 列に発話者，B 列に発話内容を記載している。図 1 に示したように，B 列の発話内容部分（Co-Chu では「ライン」とする）には，必要に応じて記号を含むこともできる。図 1 では発話の重なりが「//」で示されている。また上昇イントネーションが「↑」，発話の途中での相づちが「(はい)」のように示されている。こうした記号の使用には，特にルールは設けていないので，文字化のルールは，利用者が自

由に決めることができる。形態素解析において，記号類は一般的には解析の妨げになることが懸念されるが，Co-Chu では，解析の際，記号は自動的に排除して解析する。よって，分析等に必要な情報を自由に入れておくことができる。



A列	B列
【アメリカ人J】	はい、よろしくお願いします。
【日本人K】	よろしくお願いします。
【アメリカ人J】	まず自己紹介から//しましょうか。
【日本人K】	//自己紹介、はい。
【アメリカ人J】	では、そのーえー、日本語の先生になるために、（はい）それ何という専攻なんでしょうか。
【日本人K】	あっ、えーっと、授業自体は、えーっと、日本語教授法１（はい）のCとDを受けていて、（ああー）何（な
【アメリカ人J】	あっ、なるほど（はい）ですね。僕もそういう、あのー、学生が、（はい）見学してくれた授業に（はい）
【日本人K】	ありがとう//ございます。
【アメリカ人J】	//その【日本語教育センター】の授業を（あっ）行うんですか？
【日本人K】	はい、はい。
【アメリカ人J】	あっ、素晴らしいです。
【日本人K】	〔笑〕
【アメリカ人J】	スーツとか着るんでしょう？
【日本人K】	あっ、そうです。スーツ。〔笑〕すごい緊張しました。

A 列：発話者　　B 列：発話内容＝ライン

図 1　会話データの例

また，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。こうしたものをデータとして用いる場合，これらは形態素解析において誤解析の原因となるため，注意が必要である。Co-Chu ではこれらに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ことができる。タグは次のような形で付す。

｜適切な語形｜（タグ種別　実際に用いられた語形：タグメモ）

例 1　わー，メンタル｜弱い｜（話　弱っ：音変化）。

例 2　色んな国の人が｜いる｜（誤　ある：単語）から，その，フランスとかドイツとか。

例 1 は，「メンタル弱っ」のように，「弱い」が「弱っ」と発音された発話である。このまま形態素解析すると，動詞「弱る」の連用形促音便と解析されてしまうが，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。そして，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，タグ種別に「話」，タグメモに「音変化」と記載している。例 2 は，「いる」を用いるべきところで，誤って「ある」と発話したものであるので，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「単語」と記載している。このタグ，タグメモはいずれも検索キーとして検索の際に活用可能である。

また，誤用の中には「ねじれ文」のように，個々の形態素にタグを付して示すことが困難なものもある。さらに，誤用ではないが，「倒置」も文レベルで観察されるものである。Co-Chu では，こうしたものに注目したい場合に，タグではなく，文ごとにメタ情報をつけることが可能である。例えば，表 1 のように該当する文の左の「列」に情報を記載しておけば，文レベルのメタ情報として，検索の際に活用できる。

表 1 文レベルのメタ情報の付与

メタ情報	対象となる文
ねじれ文	よかったのは，皆がやさしくて，いろいろ助けてくれた。
倒置	でも，わかりません，そんなことは。

こうした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，見たいものを探す際に効率のよい検索が可能になる。

2.2 コーパス構築

データの作成後，Co-Chu の【Build】【Import】【Edit】の 3 つの機能を用いてコーパスを構築する。まず，【Build】でコーパスの階層構造を決める。次に【Import】でデータをインポートする。その際に，発話者に関する情報も入れる。メタデータは，図 2 にある「Add Metadata」で必要なだけ入れることができるので，性別や母語やレベルなど，分析に活用したい情報を入れておく。

図 2 【Import】で話者に関するメタ情報を入力する

次に，【Edit】で取り込んだデータを形態素解析し，分析できる状態にする。【Edit】

では取り込んだデータを編集できるので，解析結果にミスがあった場合，ここで編集して修正を行うことも可能である。

2.3　コーパス分析

分析は【Analyze】機能を用いる。図 3 は【Analyze】での画面構成である。【Analyze】では，分析対象データ（A），検索項目（B），メタ情報の条件（C），結果の表示方法（D）が指定できるようになっている。

Co-Chu は形態素解析に MeCab-Unidic を使用している。そこで，Unidic の辞書項目を検索条件に活用する。Co-Chu のコーパス分析では，図 3 の D にあるように，「語を探す」，「語を数える」，「N グラムを見る」の 3 種類の分析ができる。また，検索条件，出力項目の設定にも自由度が高いため，他種多様な検索が可能となっている。以下，例を通して，具体的にどのようなことができるのかを紹介する。

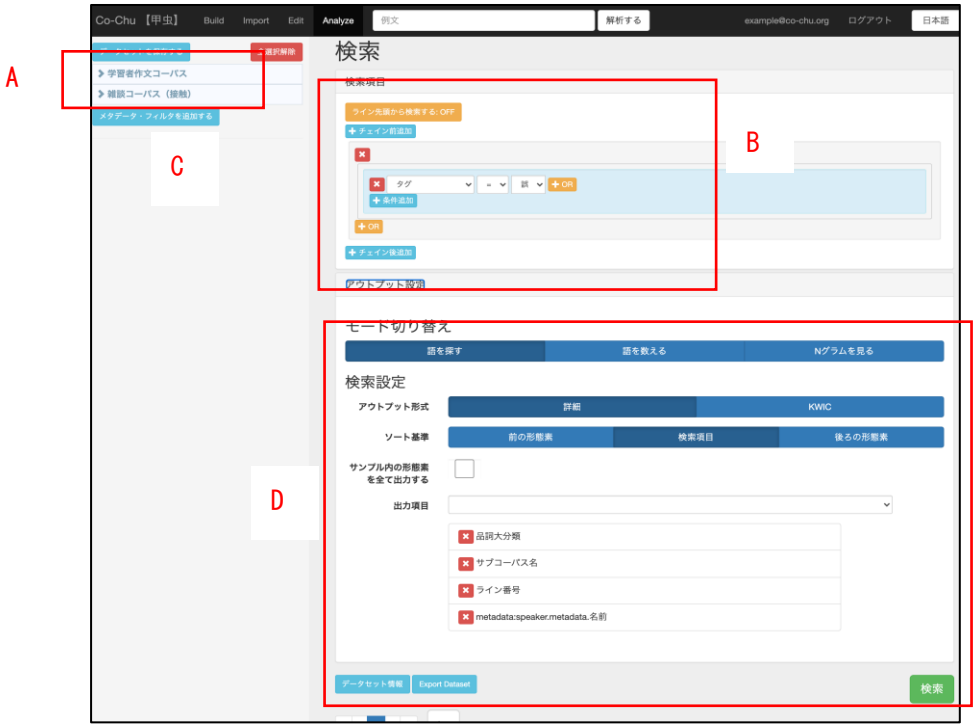


図 3　【Analyze】の画面構成

3　Co-Chu を用いた実践例（1）：語彙リストの作成

語彙リストの作成の実践例としてここではアニメを取りあげる。アニメには，教科

書で学習する以外の語彙も多く用いられている（熊野 2011，山本・小川 2021 等）が，Co-Chu を用いれば，語彙リストを作成することも容易にできる。

ここでは，学習者に人気のアニメの『鬼滅の刃』を例にする。『鬼滅の刃』1～26 話のスクリプトをもとにアニメコーパスを作成し，「語を数える」のモードで，高頻度の「動詞」「形容詞（イ形容詞・ナ形容詞)」「名詞」を検索する。

図 4　検索条件の指定

図 4 は検索条件を「動詞」を数えるように指定している。同様に，形容詞，名詞を検索し，頻度の高い上位 15 語をまとめた結果を表 2 に示す。

表 2 動詞・形容詞・名詞の高頻度の語

順位	動詞		形容詞（イ形容詞・ナ形容詞）		名詞	
1	する	728	ない	322	鬼	303
2	いる	384	いい	185	こと	265
3	なる	233	よう	120	禰豆子	209
4	言う	225	はやい	73	人	161
5	来る	179	強い	72	炭治郎	122
6	行く	169	大丈夫	70	呼吸	85
7	やる	147	すごい	55	匂い	84
8	ある	140	痛い	45	今	83
9	くれる	126	そう・悪い	33	血	79
10	できる	115			刀	70
11	わかる	114	怖い	26	伊之助	68
12	死ぬ	111	みたい・いや	23	時	63
13	切る	91			善逸	60
14	見る	89	かたい	21	体・もの	58
15	殺す	85	うるさい	19		
頻度累計	2,936(45.0%)		1,120(60.4%)		1,768（20.2%)	

表 2 に示したように，動詞および形容詞は上位 15 語で品詞ごとの頻度累計の 50%前後を占めるが，名詞では，上位 15 語の頻度数は 20%にしかならない。アニメを理解する際，名詞ではより多くの語彙を知る必要があることがわかる。

また，動詞は「する」「いる」が上位にあるが，これらは単独で用いられているとは限らず，「名詞＋する」や「〜ている」のような組み合わせで用いられるものも含む。そこで，より詳細に検索し，実際に用いられている語を把握する必要がある。ここでは「する」が「名詞＋する」の形で用いられている場合に，どのような名詞が前接しているかを検索した結果を表 3 に示す。728 回の「する」のうち，304 回（41.8%）は「名詞＋する」の形で用いられていること，さらに前接の名詞の種類と頻度がわかる。

表 3 「する」に前接する名詞 （N=304）

名詞	頻度
安心・邪魔	各 8
結婚	6
破裂，回復，説明	各 5
心配，失礼，移動，否定，理解，期待，攻撃，承知，怪我	各 4
油断，完治，顔，我慢，遭遇，失神，回転，証明，埋葬，信用，下山，拘束	各 3

これらの語のうち，「完治，遭遇，失神，埋葬，下山」は「旧日本語能力試験出題基準」では級外とされるものである。また級内の 22 語のうち，N2 以上が 13 語であり，全体的に難易度が高いこと，さらに『鬼滅の刃』の内容に応じた語彙（破裂，埋葬，攻撃，怪我，下山など）が多く含まれていることがわかる。このようにして，アニメを楽しむのに必要な語彙を抽出できる。

4 Co-Chu を用いた実践例（2）:「語」の使い方を示す（例文検索）

次に，一つの語形に注目し，それがどのような使い方がされているかを調べる，「語を探す」を用いた検索について説明する。表 2 の上位には，「よう」「こと」「もの」といった形式名詞がある。これらはアニメにかぎらず頻度が高いものだが，ここでは「よう」を例に，どのように使われているか調べることにする。

まず「語を数える」で前接する要素と後接する要素のそれぞれ頻度の高いものを検索し，どのようなつながりで用いられる場合が多いかを調べた。その結果，表 4 に示したように「〜のような」「〜のように」の形で使われる場合が多いことがわかった。

表 4 「よう」に前接する要素と後接する要素

順位	前接する要素	頻度	後接する要素	頻度
1	の	27	な・に	43
2	ない	11	だ	13
3	た	10	です	7
4	いる・できる・その	5	の	4
5	どの	4	-	3

次に，「〜のような〜」「〜のように〜」が実際にどのような台詞として現れているかを「語を探す」モードで検索する。以下がその場合の検索条件である。

検索条件　[出現形＝の]
＋
[語彙素表記＝様]　AND　[品詞大分類＝形状詞][2]
＋
[出現形＝な]　OR　[に]

このようにして検索した結果，「〜のような〜」「〜のように〜」は，全部で 22 例あった。以下に例を示す。

例 3　a. 今日は呼吸法と型のようなものを習う。
b. 耳に花札のような飾りをつけた鬼狩りの首を持ってこい。
c. 人間のまま鬼のように強くなれるの。

b の「耳に花札のような飾りをつけた鬼狩り」というフレーズは，4 回用いられていた。これは「鬼滅の刃」の主人公を指しており，アニメのストーリーを理解するのに重要な表現の一つであると言える。このように，アニメの中で実際にどのように用いられているかが，容易に検索できるのも Co-Chu の特徴の一つである。

5　Co-Chu を用いた実践例（3）：学習者の誤用傾向の抽出　（タグ検索）

Co-Chu では，見たいものが決まっている場合，そこに「タグ」を付して，効率的に抽出することが可能である。これを「タグ検索」という（詳細は山本・本間・川村 2020 を参照）。ここでは，タグ検索を活用して，学習者の作文における誤用傾向を抽出した実践例をもとに説明する。

作文指導では個々の学習者の誤用傾向を把握し，それに合わせた指導をすることが望ましい。誤用傾向を見るために，学習者の作文中に出現した全ての誤用に「誤タグ」を付し，それを用いてどのような誤用が多いかを検索した。

「誤タグ」は次のように付す。既に述べたようにタグやタグメモは，自由に付すことができる。ここでは，例 4〜6 に示したように，「タグ種別」は「誤用タグ」の「誤」とし，「タグメモ」には「誤用の種類」を記入した。

｜適切な語形｜（タグ種別　実際に用いられた語形：タグメモ）

例 4　元の文：5 年前の私がわかく，何もわからなかった。

→　5 年前のわたし｜は｜（誤　が：助詞が）わかく，何もわからなかった。

例 5　元の文：悔しいだが，それが真実かもしれない。

→　悔しい｜　｜（誤　だ：体）が，それが真実かもしれない。

例 6　元の文：時々うまくコミュニケーション取れない。

→　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。

例 4 は，「は」にすべきところを「が」にしている「助詞が」の誤用であり，例 5 は，「だ・である体」に問題があることを示している。一方，例 6 は助詞「が」が必要なところ，脱落しているので，タグメモには「助詞脱落が」と記入している。このように Co-Chu では，あるべき語句がデータ上存在しないような誤用の場合にもタグを付すことができる。例 4〜6 のように「誤タグ」を付した上で，次のように「誤タグ」を検索項目として検索を実行すると，全ての誤用を表示することができる。

検索条件　　タグ＝誤

図 5　学習者を指定して誤用を検索する

さらに図 5 のように id(speaker)情報を指定して個別の誤用を検索すると，その学習者の誤用を全て表示できる。またタグを付す際に誤用の原因を記載したタグメモを活用して，誤用の種類を「時制」「活用」などのように絞り込んで種類別に検索すれば，さらに詳細な形で誤用の傾向を抽出することができる。

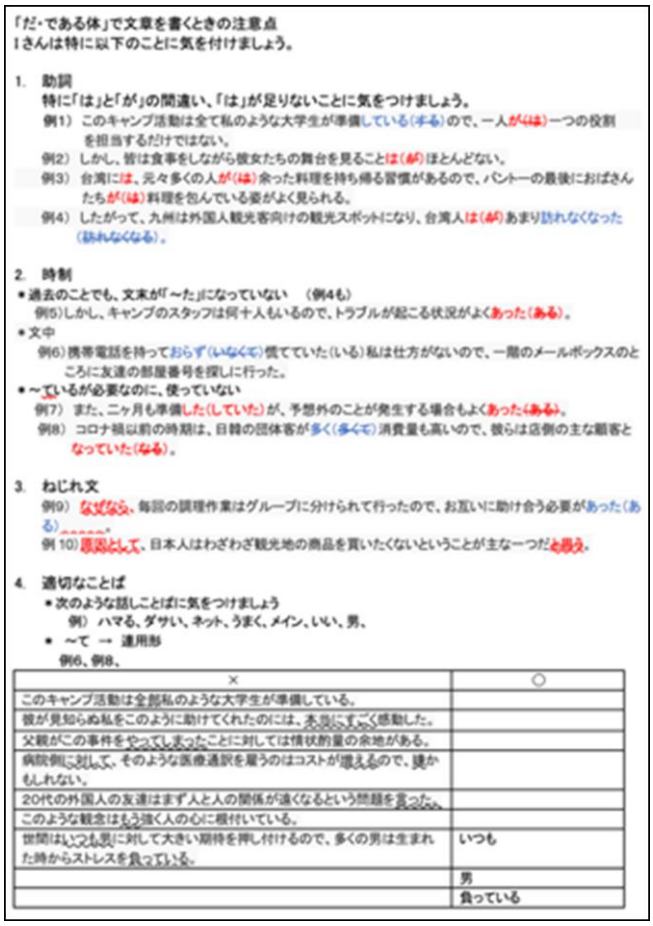
教師は学習者の作文を添削しながら，「助詞の誤用が多い」などのように大方の傾向は把握できる。しかし，Co-Chu を活用することによって，例えば「に」と書くべきところを「で」にする助詞の誤用が著しく多いなど，より詳細な傾向を見出すこと

ができ，個々の学習者に対してより具体的な指導が可能になる。

6　Co-Chu を用いた実践例（4）：「タグ検索」の活用による練習問題の作成

作文中の誤用について指導する場合，個々の学習者に目立つ誤用を指摘する他，間違いやすい項目を取り上げてクラス全体で確認したりする。ここでは，実際の学習者の誤用例を元に作成した練習問題を活用した実践例を紹介する。

図 6 は，学習者 I の誤用傾向から作成した「指導シート」である。シートの下部に話し言葉など不適切な語句を修正する練習問題を添えている。この練習問題は，図 7 のように検索条件として「タグメモに『書』を含むもの」という指定をおこない，「書きことばを使用していないための誤用」を抽出することによって作成した。


「だ・である体」で文章を書くときの注意点
I さんは特に以下のことに気を付けましょう。

1. 助詞
特に「は」と「が」の間違い、「は」が足りないことに気をつけましょう。
例1）このキャンプ活動は全て私のような大学生が準備している(する)ので、一人が(は)一つの役割を担当するだけではない。
例2）しかし、皆は食事をしながら彼女たちの舞台を見ることは(が)ほとんどない。
例3）台湾には、元々多くの人が(は)余った料理を持ち帰る習慣があるので、バントーの最後におばさんたちが(は)料理を包んでいる姿がよく見られる。
例4）したがって、九州は外国人観光客向けの観光スポットになり、台湾人は(が)あまり訪れなくなった(訪れなくなる)。

2. 時制
・過去のことでも、文末が「～た」になっていない　(例4も)
例5）しかし、キャンプのスタッフは何十人もいるので、トラブルが起こる状況がよくあった(ある)。
・文中
例6）携帯電話を持っておらず(いなくて)慣てていた(いる)私は仕方がないので、一階のメールボックスのところに友達の部屋番号を探しに行った。
・～ているが必要なのに、使っていない
例7）また、二ヶ月も準備した(していた)が、予想外のことが発生する場合もよくあった(ある)。
例8）コロナ禍以前の時期は、日韓の団体客が多く(多くて)消費量も高いので、彼らは店側の主な顧客となっていた(なる)。

3. ねじれ文
例9）なぜなら、毎回の調理作業はグループに分けられて行ったので、お互いに助け合う必要があった(ある)。
例 10）原因として、日本人はわざわざ観光地の商品を買いたくないということが主な一つだと思う。

4. 適切なことば
・次のような話しことばに気をつけましょう
例）ハマる、ダサい、ネット、うまく、メイン、いい、男、
・ ～て → 連用形
例6、例8、

×	○
このキャンプ活動は全部私のような大学生が準備している。	
彼が見知らぬ私をこのように助けてくれたのには、本当にすごく感動した。	
父親がこの事件をやってしまったことに対しては情状酌量の余地がある。	
病院側に対して、そのような医療通訳を雇うのはコストが増えるので、嫌かもしれない。	
20代の外国人の友達はまず人と人の関係が遠くなるという問題を言った。	
このような観念はもう強く人の心に根付いている。	
世間はいつも男に対して大きい期待を押し付けるので、多くの男は生まれた時からストレスを負っている。	いつも
	男
	負っている

図 6　学習者の誤用傾向と誤用に基づいた練習問題を示した「指導シート」

図 7　練習問題作成のための検索索項目指定

7　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，多様な試みをおこなうことができる。本システムは，コーパスシステム Co-Chu として一般への無償公開を開始している。日本語教育や日本語研究での幅広い活用を期待したい。

＜付記＞本研究は科学研究費基盤研究（C）18K00723（2018 年～2020 年度　研究代表者：山本裕子）の助成を受けたものである。

＜注＞

1. 詳細は Co-Chu のホームページ（https://cochu.org）を参照のこと。マニュアルもホームページから閲覧可能である。
2. 検索条件の「形状詞」は，日本語教育での「ナ形容詞」に相当する。

＜資料＞

鬼滅の刃　https://kimetsu.com/anime/ （2021 年 11 月 3 日）

＜引用文献＞

熊野七絵（2011）「アニメ・マンガの日本語～ジャンル用語の特徴をめぐって～」『広島大学国際センター紀要』1, pp.35-49, 広島大学国際センター.

山本裕子・小川満梨奈（2021）「アニメに用いられる日本語−スクリプト分析による語彙的・文法的特徴の抽出の試み−」『2021 年度日本語教育学会秋季大会予稿集』pp.319-323, 日本語教育学会.

山本裕子・川村よし子・小森早江子・本間妙（2020）「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 」『2020 年度日本語教育学会秋季大会予稿集』pp103-108, 日本語教育学会.

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.