コーパスシステム Co-Chu の検索比較機能を使った研究事例

山本裕子（愛知淑徳大学)、 小森早江子（中部大学)、本間妙（中部大学）
ラニガン マシュー

A case study using the result comparison feature of the corpus system Co-Chu

Hiroko YAMAMOTO, Aichi Shukutoku University　Saeko KOMORI, Chubu University
Tae HOMMA, Chubu University　Matthew LANIGAN

要旨：コーパスシステム Co-Chu はコンピューターに苦手意識がある人も平易に使えるウェブアプリケーションである。研究者や教師が自分で収集したオリジナルデータをコーパス化して利用できる。【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。

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観察できる。また発話データの文字化に不可欠な笑い声や聞き取れなかった部分などに使用する記号は形態素解析には都合が悪いが、Co-Chu ではそれらの処理を自動でおこない、MI スコアの算出に支障のないようにしている。

本発表ではオリジナルデータを用いて、日本語母語話者と学習者の言語使用の実態を比較する研究事例を報告する。

キーワード：コーパスシステムCo-Chu　オリジナルデータ処理　検索比較機能

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について説明し、特に検索比較機能を使った分析事例を紹介する。

2. コーパスシステム Co-Chu

2.1 概要と位置づけ

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、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、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、コーパスシステム Co-Chu を開発した。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。

2.2 Co-Chu によるコーパス構築の手法

ここでは、Co-Chu の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に関する部分について説明する。Co-Chu 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ことができる。まず、データを取り込む作業として、【Build】【Import】【Edit】の機能を使用する。データ取り込み後、【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。以下、簡単にそれぞれの主な機能をまとめる。

1. 【Build】: コーパスの階層構造設定、コーパスメタデータ作成
2. 【Import】: データ確認と取り込み
3. 【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、検索結果を見て気付いた解析の不備や誤りの修正など
4. 【Analyze】: 検索、頻度分析、コロケーション分析、検索比較機能

これら 1. ～3. の機能を順に使用し、容易にオリジナルデータを取り込み、さらに 4. によりさまざまな検索および分析を行うことができる。

2.3 検索比較機能

ここでは、検索比較機能について概要を説明する。いくつかの結果を比較することが研究の基本の一つであり、コーパス日本語学にも不可欠な手法であるが、場合によって、非常に手間がかかる作業でもある。この重要な作業を簡単にできるように検索比較機能を開発した。

検索比較機能では、「複数データセット比較」と「複数検索比較」という二種類の比較方法が使用できる。複数データセット比較では、同じ検索をいくつかのデータセットにかけ、それぞれの結果を比較する。また、複数検索比較では、一つのデータセットにいくつかの検索をかけ、それぞれの結果を比較する。

どちらの検索比較方法を用いても、結果は同一画面に出力される。例えば、複数データセット比較を用いて、「中国人日本語学習者」と「韓国人日本語学習者」という二つのデータセットに同じ検索をした場合、出力した結果の表示は、「中国人日本語学習者」のみ、「韓国人日本語学習者」のみ、あるいは「両方」のどれであっても、容易に切り替られる。複数検索比較も同様に切り替えられる。

このように、Co-Chu では、簡単に検索結果を比較することができる。

3. 分析例

本発表ではオリジナルデータを対象に、日本語母語話者（NS）と日本語学習者(NNS)の言語使用の実態について、複数データセット比較機能を用いて分析した結果を紹介する。

3.1 分析対象データ

分析には Co-Chu 雑談コーパス [1)]を使用した。Co-Chu 雑談コーパスは、上級日本語学習者と日本人大学生の雑談を文字化したもので、2017 年 2 月時点で 20 ファイル（録音時間は約 660 分）から成っている。本発表では、この「親しい友人同士の会話（以下「友人」）」と「初対面の会話（以下「初対面」）を使用する。

表 1　Co-Chu 雑談コーパス　サブコーパス

サブコーパス	データの内容	ファイル数	ライン数
友人	中国人上級学習者（CS）と日本人大学生（NS）の雑談	11	10152
	韓国人上級学習者（KS）と日本人大学生（NS）の雑談	5	5325
初対面	中国人上級学習者（CS）と日本人大学生（NS）の雑談	4	1589

3.2 MI スコアと記号の処理

MIスコアは語と語の結びつき、つまりコロケーションの強さを示す指標の一つとして使われている。MIスコアの算出は、一般的にそれぞれの語の頻度と2語の共起頻度（2グラム）を用いて計算するが、Wei and Li (2013)では英語のコーパスを例に複数の語（6グラムまで）の共起頻度を算出する手法を提案している。日本語のコーパスでも同様の手法が使えるかどうか、小森他（2015）で検討した。日本語では語の区切り方など独特の問題もある。日本語の公開されている書き言葉や話し言葉のデータを用いて、高頻度表現のMIスコアの結果を観察し、コロケーション表現の特徴について考察した結果、Wei and Li (2013)で提案された 2 語以上のコロケーション分析が日本語でも利用できることを示した。さらに日本語で分析する場合の問題点も指摘した。

ここでは検索比較機能を用いて、MIスコアを算出する際の日本語の記号の取り扱いについて考える。話し言葉を分析する場合、会話などの発話を書き越した文字データには句読点などの記号が含まれている。たとえば、以下の発話例は、Co-Chu雑談コーパスの中の日本語学習者(CS-O)と母語話者（JS-H）との友人会話の一部分である。こうした会話データには、実際の発話内容のほかに、日本語として読みやすくする句読点や、漢字語彙の読みを表す（　）、オーバーラップを表す／／、非言語情報を示す{笑}など、さまざまな記号が含まれている。このような情報は、発話を理解するためには必要であるが、通常、MIスコアの分析には不要である。

【CS-O】	あのー、なんか、んー、あのー、なんか、|何|(なん)だろうね、
【CS-O】	病院↑、病院の中の
【JS-H】	うん、そうそう。／／
【CS-O】	／／|何|(なに)これ
【JS-H】	心電図だ、心電図。
【CS-O】	ああ、心電図。／／そうだよね｛笑｝。

図 1　発話例：日本語学習者(CS-O)と母語話者（JS-H）の会話データからの抜粋

表2は、上記の日本語学習者（CS-O）と母語話者（JS-H）の会話データを対象に、上記のような記号を含んだままの未処理の「記号あり」データと記号を取り除いた「記号なし」データでMIスコアを産出し、頻度の高い順に並べ替えて比較できるようにしたものである。「記号あり」と「記号なし」のサブコーパスの総語数は、それぞれ14,818語、7,262語である。

表から、「…だよね」「…そうだね」「…そうだよ」のような文末表現が高頻度で使われていることが観察できる。また、語句や文節末に続く「…んだけど」「…っていうん」「…じゃなくて」のような表現も多用していることがわかる。

表2　「記号あり」と「記号なし」での比較（3グラム）

N グラム	N	あり頻度	あり MI スコア	なし頻度	なし MI スコア
だ　よ　ね	3	24	5.473672889	25	4.503213834
そう　だ　ね	3	19	5.574752643	19	4.404901254
そう　だ　よ	3	17	5.895717434	18	4.896286626
ん　だ　けど	3	16	5.759685926	16	4.738098046
って　いう　ん	3	14	4.28377716	14	3.207372266
じゃ　なく　て	3	10	5.609035138	10	4.585340806

同じデータで「記号あり」と「記号なし」のMIスコアを比較すると、多少数値の違いはあるが、どの表現でも「記号あり」のMIスコアは「記号なし」のと比べて約1ポイント高くなっていることがわかる。MIスコアは通常3以上の語彙で興味深いといわれている(Church and Hanks 1990)ため、「記号あり」では実際よりも語彙の結びつきが強く出てしまうことになる。

上記の例は一つのファイルで 3 グラムの場合のみの比較であるが、この結果からコロケーション分析のための MI スコアの計算には、記号の取り扱いに注意する必要があると言える。調べたい語や状況などに合わせて、記号の処理をしなければならないことがわかった。今後はさらにファイル数を増やし、4 グラム以上のコロケーションでも検証する。

3.3 検索機能の分析例

この節では、検索機能を使って実際にどのような分析が行えるのかを示す。まず初対面と友人という関係に注目したものとして 3.3.1 で文体、ついで「共話」の観点から 3.3.2 で補助動詞表現、3.3.3 で条件表現、3.3.4 で言いさし文について述べる。

3.3.1 文体

初対面と友人という関係で、大きな違いがあると想定されるのは文体である。そこで、Co-Chu 雑談コーパスを用いて、NS と CS の友人間、初対面の会話でそれぞれ「です・ます」がどの程度用いられているか、また終助詞の使用について調べた。

表 3　Co-Chu 雑談コーパスでの「です・ます」の使用数

	友人		初対面	
	NS	CS	NS	CS
ます	42	72	5	68
です	174	172	16	135
（内「でしょう」の数）	131	50	9	5
ライン数	8452	5784	791	796
ライン数に対する%	2.6%	3.4%	2.7%	25.5%

その結果、表 3 に示したように、CS は初対面で「です・ます」を多く用いているが、NS は初対面であっても「です・ます」は用いていないことが示された。まず、表 3 には友人間であっても「です・ます」を使用することがわかるが、この「です」には「でしょう」が相当数含まれていることが指摘できる。NS の友人間の「です」174 回のうち、「でしょう」は 131 回で 75.3%を占めており、「です・ます」と言っても、「丁寧体」として用いられているわけではない。また使い方を見てみると、NS、CS ともに他人の発話の引用として「です・ます」を用いている他、NS には次の(1)〜(3)のように冗談を言ったり、照れ隠しとして「です・ます」を用いる例が目立つ。

（1）今の嘘でーす。
（2）そうだよ、彼女いて残念でした。
（3）勉強し直します、すいません。

これらでは「です・ます」を用いることで、冗談であることが示されており、親しい関係ならではの発話となっている。

また、表 4 のように、CS は初対面では友人間に比べて終助詞「ね」「よ」の使用が少ないが、NS はどちらも大きな違いがないことがわかった。

表 4　Co-Chu 雑談コーパスでの終助詞の使用量

	友人		初対面	
	NS	CS	NS	CS
ね	1081（36.2%）	332(28.9%)	111（38.5%）	46(19.7%)
よ	435(14.5%)	235(20.5%)	35(12.2%)	33(14.2%)
か	507(17.0%)	204(17.8%)	30(10.4%)	70(30.0%)
助詞総数	2990	1147	288	233

CS には初対面では終助詞はあまり用いず、「です・ます」を用いるという傾向があると言えるだろう。ただし、終助詞「か」については NS と NNS で使用傾向が大きく異なる。CS は初対面で「か」を多く用いているが、「か」は「〜か？」と質問を投げかける際に用いられている。一方で NS は「か」よりも「ね」を用いて相手への働きかけをしているようである。今回のデータは NS と CS は「初対面」ではあるが、この会話のあと行動を共にするような状況で採取したものであり、お互いに親しくなりたいという気持ちを持っていたと考えられる。距離を縮めるためのストラテジーが NS と CS では異なっていることが窺える。

3.3.2　補助動詞表現

補助動詞は、「動詞(V1)のて形＋動詞(V2)」の形で用いられる後項の動詞（V2）であるが、その（V2 の）意味は単独で用いられる時とは異なり、実質的ではなく付加的なものとなっている。この形態をとることが可能な動詞は、「いる」「くる」「いく」など 10 数語に限られる。これらのほとんどを日本語教育では初級で学習する。しかし、近藤他（2010）、水谷（2015）、山本（2015）など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。ただし、これらの研究は学習者の補助動詞の使用をアンケート調査や、OPI のインタビューから分析したものであり、母語話者と日本語学習者の自然な談話の中で補助動詞がどのように使われているかの実態を調査したものではない。そこで、Co-Chu 雑談コーパス（友人）を用いて、それぞれ補助動詞表現がどの程度用いられているか調べたところ、表 5 のようであった。

表 5　Co-Chu 雑談コーパス（友人）での補助動詞の使用数

	友人		
友人	NS	CS	KS
補助動詞数	899	265	139
動詞数	4536	2098	1227
動詞に対する％	19.8%	12.6%	11.3%

このように、NS は CS や KS よりも補助動詞を多く用いている。次に内訳を表 6 に示す。表には、それぞれの縮約形や敬語形も含んだ総数を記載した。

表 6　補助動詞使用内訳　(%)

	いる	くる	いく	ある	おく	みる	しまう	もらう	あげる	くれる	動詞計
NS	13.2	0.93	0.79	0.35	0.24	0.42	2.84	0.37	0.04	0.62	4536
CS	7.57	0.43	0.19	0.05	0	0.38	3.29	0.10	0	0.62	2098
KS	7.74	0.65	0.89	0.08	0	0.49	0.81	0.16	0.08	0.41	1227

日本語は立場志向であるという指摘（水谷 2015、池上・守屋編 2009 他）を踏まえ、ここでは立場に直接関わる「いく」「くる」「あげる」「もらう」「くれる」を見る。母語で立場志向的な表現を取らないとされる中国人学習者 CS では NS ほど使用が多くない。「くれる」については CS と NS の使用率に違いがないが、CS の「くれる」は、「x 先生に『～てください』って言われた」のように他者の発話を引用して「ください」を用いているものであり、NS の使用と同列に扱えるわけではない。一方で、母語でも授受表現に類する表現を持ち、方向性をより重視するとされている韓国人学習者 KS は NS より多くこれらの表現を用いている場合もある。具体的な使用法はさらに検討が必要であるが、このように対照研究で指摘されている傾向が示された。

3.3.3 条件表現

日本語の条件表現には「たら」「なら」「ば」「と」のような表現があり、日本語学習者にとって使い分けが難しい。Co-Chu 雑談コーパス（友人）を用いて、これらの使用数を見たところ、表 7 のようであった。CS と KS の使用傾向にはかなりの相違がある。KS は「たら」に偏っていること、CS は「なら」の使用が多いことが目立つ。

表 7　Co-Chu 雑談コーパス（友人）に見られる条件表現　(%)

	たら	なら	と	ば	計	対ライン数%
NS	116(37.2%)	9(2.9%)	143(45.8%)	44 (14.1%)	312	3.7%
CS	35(35.7%)	15(15.3%)	28(28.6%)	20 (20.4%)	98	1.7%
KS	59(70.2%)	2(2.4%)	12(14.3%)	11 (13.1%)	84	3.2%

学習者は「親しい関係だったら敬語を使う」のようにいわゆる仮定をする場合に条件表現を使っている。一方、NS は「字違ってたら、ごめんね」「そんなこと言ったらダメ」のように、先行する状況や発話を受けて、条件表現を用いる場合も多く見られた。NS の談話の展開スタイルと併せて検討する必要がありそうである。

3.3.4 「言いさし文」

日本語の発話には、形式上最後まで言い切らずに複文の主節を省略していても、情報伝達においては完全文と同じ発話機能を果たしている「言いさし文」がある（朴 2010）[2)]。

言いさし文には、「けど」「から」「て」「し」「みたいな」など様々な形で言いさしているものが見られる。白川（2009）では、「けど」で終わる「ケド節」の言いさし文を「言い尽くしのタイプ」に分類している。白川（2009）によると、言い尽くしのタイプとは、関係づけられるべき事態が文脈上に存在せず、話し手が伝えようとする内容が従属節のみで言い尽くされているもので、その場合、節の内容は聞き手に持ちかけられ、帰結は聞き手の判断に委ねられるという。つまり、言いさし文「けど」（以下「けど。」）は、「共話」において成立するのである。さらに白川（2009）は、文レベルでの「文型」を教える従来の日本語教育文法では、言いさし文の持つ談話機能を十分に理解させられない可能性があると述べている。

ここでは「けど。」について、Co-Chu 雑談コーパス（友人）を用いて学習者と母語話者の使用実態を分析する。

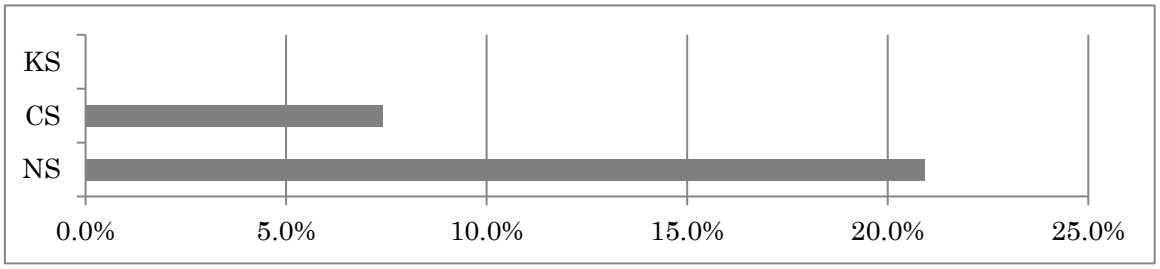

図 2　　Co-Chu 雑談コーパス（友人）に見られる終助詞を伴う「けど。」

総ライン数に対する「けど。」の割合は、NS が 86/8452 で 1.02%、CS が 27/5784 で 0.47%、KS が 25/2590 で 0.97%であり、NS の使用率が高い。「けど。」には、「けどね。」「けどなあ。」等のように終助詞を伴う形も含まれる。NS は 86 例の「けど。」のうち 18 例（「ね」12、「なあ」5、「ねえ」1）、つまり 20.93%が終助詞を伴う形である。これに対し、CS は 27 例中 2 例（「ねえ」2）の 7.41%であり、KS には終助詞を伴う形が見られなかった（図 2）。

また、NS は「～書いてたけど。」、「思うんだけど。」等のように、「けど。」に前接する表現として「～ている」、「～んだ」を多用する傾向がある。さらに「思ってたんだけど。」のように「～ている」と「～んだ」を共に前接させる場合もある。NS には、これらを「けど。」に前接させた表現が 27 例（31.4%）あり、CS、KS に比べ「けど。」の発話中に占める割合が高い（図 3）。一方、CS は「～んだ」を前接させておらず、「～ている」の前接も 1 例のみであり、KS も「～ている」、「～んだ」の前接はそれぞれ 2 例で、「～ている」と｛～んだ｝を共に用いた例はない。

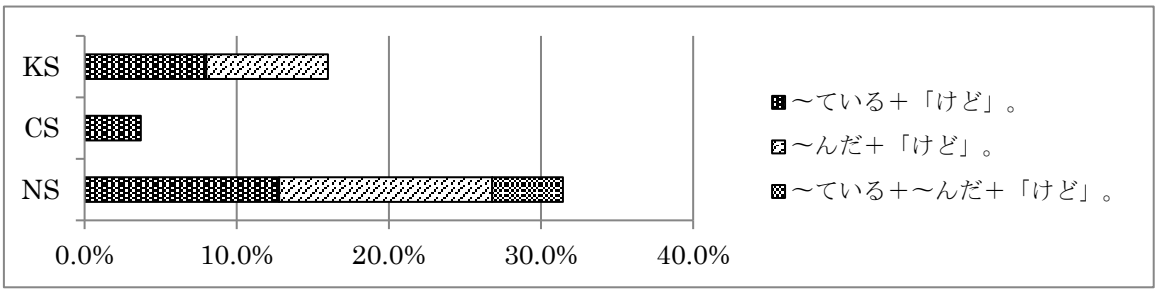

図 3　Co-Chu 雑談コーパス（友人）に見られる全「けど。」文中の「前接する表現」

NS は、「けど。」の使用に「終助詞」の「ね」「ねえ」、「〜ている」、「〜んだ」を伴うことで、既定性、継続性を示し、聞き手に事態の共有を働きかけているものと考えられる。

このように、補助動詞、条件表現、言いさし文のいずれについても NS は使用頻度、共起表現の豊富さの点で NNS の使用とは異なっていることが示された。NS はこれらを相手の発話を受け止め、働きかけるためにも用いており、これらの表現は「共話」に貢献していると考えられる。日本語教育での扱いも検討する必要があるのではないか。

4.　まとめと今後の Co-Chu 開発

本発表では、開発中の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。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、今後も Co-Chu の開発を続け、日本語教育・研究に貢献できるような様々な機能を加えようと考えている。

注

1) Co-Chu 雑談コーパスの構築に当たっては、平成 26 年度中部大学学部長裁量教育研究支援費および平成 28 年度中部大学特別研究費 I による助成を受けた。

2) 朴（2010）では「言いさし文」のことを「言いさし表現」として論じている。

参考文献

庵功雄・山内博之編.2015.『データに基づく文法シラバス』東京： くろしお出版.

池上嘉彦・守谷三千代編. 2009.『自然な日本語を教えるために　認知言語学をふまえて』東京：ひつじ書房.

小森早江子・山本裕子・本間妙・ラニガン・マシュー. 2015. 「日本語教育・研究のためのテキスト分析システム“Co-Chu”」 The proceedings of CASTEL/J in Hawaii 2015

近藤安月子・姫野伴子・足立さゆり．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ー」JCLA2010, pp.690-706
白川博之. 2009.『「言いさし文」の研究』東京: くろしお出版.
朴仙花. 2010.「OPI データにみる日本語学習者と日本語母語話者による文末表現の使用―接続助詞で終わる言いさし表現を中心に」,『言葉と文化』217–235. 愛知: 名古屋大学大学院・国際言語文化研究科.
水谷信子.2015. 『感じのよい英語　感じのよい日本語　日英比較コミュニケーションの文法』東京： くろしお出版.
山本裕子. 2015.「『感じのよさ』と補助動詞　−日本語母語話者と非母語話者の調査結果を比較して−」『人文学部研究論集』第34号，pp.1-18. 愛知： 中部大学人文学部.
Church, K., & Hanks, P. 1999. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16:1, 22-29.
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

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/