〔2018 年度日本語教育学会秋季大会（プラサ ヴェルデ，2018.11.25）発表・ポスター発表【予稿集】

話し言葉や誤用の含まれたテキストに対応可能なコーパス分析システムの開発

山本裕子（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）・本間妙（中部大学）
＜共同研究者＞ラニガン・マシュー（LendingHome）

1. はじめに

本研究は、話し言葉や誤用が含まれたテキストの分析が可能なシステムの開発を目的としている。近年、ICT 技術の進歩に伴い、日本語教育の分野でも多くのコーパスが提供され、様々な活用法が提案されている（庵・山内編 2015 他）。また、話し言葉用 UniDic（岡 2017）も公開され、形態素解析システムに組み入れることによって、話し言葉を研究する環境も整いつつある。だが、自ら集めたデータをもとに言語分析が行えるというツールは少ない。特に、縮約形、方言、言い間違えなどが含まれている話し言葉や、誤用が含まれているテキストを解析できるツールは提供されていない。筆者らはコンピューターに苦手意識がある人でも自らのデータをもとに言語分析ができるツールとして、コーパス分析システム Co-Chu の開発を進めている。本システムでは、話し言葉や誤用を含むテキストであっても形態素解析を適切に行い、言語分析ができるようにするための仕組みが整えられている。本発表においては、【Edit】機能を中心にシステムを紹介するとともに、この仕組みの運用実験の結果について報告する。

2. Co-Chu の概要

テキストの分析を適切に行うには、基礎データ、形態素解析を可能にするタグ付け、形態素解析、の３つを統合する必要がある。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。利用者は、まず、【Build】機能で新たに作成するコーパスの階層構造を決め、【Import】機能で自ら集めたオリジナルデータをシステムにアップロードする。その後、【Edit】機能を活用して、MeCab-UniDic による形態素解析を行い、解析結果に問題があれば適宜修正を行い、話し言葉や誤用を含むデータの適切な形態素解析が行えるように編集する。その後、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を【Analyze】機能で行う。

3. 誤解析に対する Co-Chu の対応策

話し言葉や日本語学習者のデータを形態素解析した場合、データをそのまま用いると解析結果に問題が生じることがある。例えば、発話データを文字化する際、上昇イントネーショや、相槌、同時発話など会話の状況を伝えるための記号を用いるが、形態素解析には妨げとなる。Co-Chu では、こうした記号を用いても、解析結果に悪影響を及ぼさないよ

うに記号を自動的に排除して形態素解析を行っている。その一方で、話し言葉の分析の際には、これらの記号を活用できるような工夫を施している。例えば、「うん」という相槌が会話の間に挿入された場合、(　) を用いて（うん）のような文字化を行っても、形態素解析の際には（うん）を存在しないものとして解析を行う。

また、「話し言葉」には言い間違い、言い直し、言い淀み、フィラーなど、不規則なものが多く含まれ、これらも形態素解析で誤解析されてしまうことがある。そのため、これらに対しても、Co-Chu では【Edit】機能を活用して対応する。以下に、その詳細について述べる。

3-1. 単語情報の変更

形態素解析において、単語の区切りが適切に解析されない場合がある。このような場合、Co-Chu では他の区切り方の候補を提示し、適切なものを選択することが可能である。（以下の各例において、誤解析のあった箇所は下線で示す。）

（例１）それが「どやばい」とかも使うし・・→ 解析結果：どや（名詞）＋ばい（助詞）
→ 修正した解析結果：ど（接頭辞）＋やばい（形容詞）

例１では「どやばい」が、「どや」と「ばい」に区切られ誤解析された。しかし、MeCab-UniDic によって提案された他の候補の中から「接頭辞＋形容詞」を選択して、単語情報が修正できる。ただし、適当な候補がない場合もある。また区切りが適切であっても、「読み」がおかしい場合も多い（例：「清水、清水寺に行って・・」と言い淀みを含んだ発話において、一つ目の「清水」には「しみず」の読みしか候補にない）。このようにこの方法では十分に解決できない場合もある。

3-2. タグ付け

誤用、言い直し、方言、若者ことばなどが含まれているために誤解析が起きた場合には、単語情報の変更では対応できない。このような場合、Co-Chu ではデータに

｜正しい語形｜（種別　実際に用いられた語形）

という形式でタグ付けを行ったうえで、形態素解析を行うこととした。現時点では、以下の６つのタグを用いて誤解析に対応している。

① 話し言葉タグ：「話」

現在の UniDic は話し言葉にもかなり対応が進んでおり、「ほんと（本当）」「やだ（嫌だ）」「っす（です）」等、問題なく解析できるものもある。しかしながら、次の例のように、UniDic の単語辞書に含まれていない単語等では、誤った形態素解析がなされる場合がある。そこで、話し言葉を示すタグとして「話」を用いることにした。

（例２）：よっし、あ、でもね、　→ 解析結果：よっ（動詞　因る）＋し（助詞）
タグ付け　｜よし｜（話　よっし）→ タグ付け後の解析結果：よし（感動詞）

例２では、「よっし」が、動詞「因る」の活用形と誤解析されたが、これは感動詞「よし」の意で用いられたものである。タグを付けることによって、正しく解析できるようになる。

② 方言タグ：「方」

方言も話し言葉ではあるが、誰もが用いるわけではなく、発話者の特徴を表すものであるため、一般の話し言葉とは別のタグ付けを行い、方言タグとして「方」を用いることにした。また、否定の「ない」が「ん」になったり、「〜ている」が「とる」になるといった広範囲にみられる方言は適切に解析される場合もあるが、正しく解析されない場合も多い。

（例３）どこやったっけな。→ 解析結果：どこ（代名詞）やっ（動詞）た（助動詞）

タグ付け　｜だっ｜（方　やっ）たっけ

→ タグ付け後の解析結果：だ（助動詞）＋た（助動詞）

③ 誤用タグ：「誤」

学習者のデータには誤用が含まれることが多い。また、母語話者であっても、会話の中では言い間違いや、コロケーションが不自然な場合等がある。そこで、誤用タグとして「誤」を用いて、誤用を示すことにした。

（例４）A：あー、いいだけど、でも少ない。

→ 解析結果：いい（形容詞）＋だ（助動詞）＋けど（助動詞）

これは、「兄弟がいる人はいるか」と聞かれたのに対する答えであるので、「いるけど」と答えるべきところである。次のようにタグを付けることによって、正しく解析できる。

タグ付け　｜いる｜（誤　いいだ）けど

→ タグ付け後の解析結果：　いる（動詞）＋けど（助動詞）

一方、誤用には、例４のように形態素解析に問題が生じるような誤用もあるが、解析自体には問題がない場合もある。助詞の誤用などがそれに当たる。

（例５）海兵隊の入隊しないと　→ タグ付け　　海兵隊｜に｜（誤　の）

例５では、「の」であっても「に」であっても格助詞であり、解析結果も問題は生じないが、誤用である。こうしたものについても、誤用タグを付けることでどのような誤用がどのくらい含まれているかを調べることも可能になる。

④ 言い淀みタグ：「淀」

話し言葉には多くの言い淀みがあるが、これらは誤解析を生じやすい。言い淀みによって誤解析が生じている箇所には、「淀」タグを付ける。

（例６）でも、おさ、お酒飲んでね→解析結果：おさ（名詞）＋お（接頭辞）＋酒（名詞）

タグ付け　｜お酒｜（淀　おさ、お酒）

→ タグ付け後の解析結果：お（接頭辞）＋酒（名詞）

このようにタグ付けすることで、言い淀みがあったことを示しながら、解析結果には問題が生じないようにすることが可能になる。

⑤ 上記以外のケースを示すタグ：「他」

①〜④以外の問題として、「言い直し（訂正）」「言いさし」あるいは①〜④が複合的に生じている場合をまとめて、「他」というタグを付加することにした。

（例７）：勝手に予約されとっ（笑）、間違えた

→ 解析結果：予約（名詞）＋さ（サ変）＋れ（助動詞）＋と（助詞）
タグ付け：予約され｜とって｜（他　とっ：言いさし）
→ タグ付け後の解析結果：予約（名詞）＋さ（サ変）＋れ（助動詞）
＋とっ（動詞）て（助詞）

また、このタグを用いることで、次の例のように、言いなおし等が行われているケースにも対応が可能となる。

（例８）名古屋弁を、なご、愛知県民に
→ 解析結果：名古屋（名詞）＋弁（名詞）＋を（助詞）＋なぐ（動詞　薙ぐ）
＋愛知（名詞）＋県民（名詞）
タグ付け：名古屋弁を、｜　｜（他　なご：名古屋と言いかけた？）愛知県民に
→ タグ付け後の解析結果：　名古屋（名詞）＋弁（名詞）＋を（助詞）＋愛知
（名詞）＋県民（名詞）

例８では｜｜の中に何も入れないため、タグ付け後の解析では「なご」部分はなかったものとして扱われる。しかし、画面上では元の発話を表示できるので、「なご」と言いかけて途中で止めていることが確認できる。また、（　）の中に書いたものは形態素解析には影響しない仕組みを応用し、スクリプト内にコメント等を書き入れることも可能である。

⑥ 単語情報に誤りがあることを示すタグ：「単」

上記①～⑤のタグでは問題が解消できず、解析された単語情報に誤りのあるものに「単」というタグを付ける。このタグは、誤解析は現時点では修正できないが、問題があることを示すためのものである。

（例９）　やっぱりそのまあ、さっき→ 解析結果：「その」は「連体詞」
タグ付け：やっぱり｜その｜（単　その：感動詞）まあ、

（例 10）　僕たちっていう、その、たち、っていうの→解析結果：「たち」は固有名詞
タグ付け：その、｜たち｜（単　たち：接尾辞）、っていうの

「そのー」のように長音であれば、感動詞（フィラー）と解析されるが、例９に含まれる「その」や「あの」は MeCab では基本的に「連体詞」として解析されてしまう。これらは次候補で「感動詞」が出せる場合もあるが、多くは「感動詞」が候補にならず、正しい単語情報に修正できない。また、「僕たち」の「たち」は接尾辞であるが、例 10 のように「たち」を話題にする場合、「たち」は接尾辞とは解析できない。これらでは、タグを付けても解析結果は変わらないが、タグによって問題があることは示すことができる。

4. 実験

4-1. 方法

実際にどのくらい誤解析が生じ、タグ付けによってどのくらい解消できるのかを検証するために運用実験を行った。実験用に話し言葉（母語話者同士の雑談３つ、母語話者と日本語学習者の雑談３つ、計６つのサブコーパス。録音時間 232 分）の文字化データを Co-Chu に取り込み、そのまま解析にかけるコーパス（タグなし）と、タグを付けて解析

するコーパス（タグ付き）を作成した。これを用いて、タグの有無で解析の結果がどのように異なるかの検証を行った。

タグ付きコーパスは以下の手順で作成した。まず、形態素解析の結果を【edit 画面】で確認し、3.1 で述べた方法により、単語情報の誤りを正すことができるものには、次候補を選択する処置を施した。そして、それ以外の解析の誤りに対して、タグ付けの処理を行った。なお、本実験でのタグ付けは、基本的には解析の誤りの要因となっている箇所のみを対象とし、言い直しや誤用と考えられる箇所であっても、例５のように解析に影響していない場合はタグ付けしないという方針で行った。

4-2. 結果

タグなしコーパスと、タグ付きコーパス間で、短単位数、タグ数、各品詞の出現数について比較を行ったところ、以下のようになった。

表 1　短単位数比較

サブコーパス	短単位延べ数（token）	短単位異なり数（type）
タグなし	42785	2199
タグ付き	42663	2122

表 2　誤解析数・タグ数

①話	②方	③誤	④淀	⑤他	⑥単	合計
150	106	20	64	49	272	671

表 2 に示したように、解析の誤りが合計 671 箇所あり、そのうち①～⑤のタグを付けた 397 については､タグを付すことによって誤解析を解消することができた｡①話し言葉や、②方言のタグが多いが、これは、友人同士の盛り上がった会話には、ら抜き言葉､「ちゃう（違う)」等の縮約形､「ぜってえ（絶対)」「やっすい（安い)」のような促音化した語､「～やん」「～やんか」のような方言の文末表現が多く用いられ､それらに誤解析が多かったことによる。また、現時点では解消できない問題である⑥の「単語情報の誤り」も多く見られた。話し言葉ではフィラーが多く、そのほとんどが誤解析となったことによる。

次に、タグの有無によって、品詞数にどのくらい相違があるかを検討する。Co-Chu の【Analyze】機能で品詞の出現頻度をカウントしたものを、表 3 に示す。

表 3 品詞別出現頻度比較

	助動詞	動詞	名詞	い形容詞	な形容詞	終助詞
タグなし	6097	4924	7756	1169	545	2754
タグ付き	6089(-0.1%)	4912(-0.2%)	7727(-0.4%)	1192(+2.0%)	547(+0.4%)	2572(-6.6%)

表 3 に示したように頻度数に多少の違いはあるが、終助詞がタグ付きではタグなしより 6.6%減になっているのが最も大きな相違であり、他はほとんど違いがないようにも見える。しかし、個々の表現に目を向けると、語によっては大きな相違のあるものもある。例えば、動詞の「違う」はタグなしコーパスでは 75 回であったが、タグ付きコーパスでは 104 回と 40%近く多くなっている。これは次のような「話し言葉タグ」のついた「違う」の存在の影響が大きい。

（例 12）うん。｜違う｜（話　ちゃ）、でかいの、だって持って帰るの大変じゃん。

（例 13）えっ、ちゃう｜違う｜（話　ちゃう）、何（なん）かね。

これらは、実際の発話では「ちゃ」、「ちゃう」と発音されている。例 12 の「ちゃ」は「助動詞　じゃ」と誤解析される。例 13 の「ちゃうちゃう」は一つ目の「ちゃう」は「違う」と解析されるが、二つ目の「ちゃう」は「〜てしまう」の縮約形「ちゃう」と誤解析される。このように、タグの付いているものを分析することによって、話し言葉でどのようなものが解析上、問題となりやすいかを見出すことができる。

5. おわりに

以上、Co-Chu では話し言葉や誤用を含むテキストであっても適切な形態素解析を行い、言語分析ができるような編集を行うことが可能である。誤解析の原因が、当該単語が UniDic の単語辞書に未登録であることによる場合には、UniDic に登録することさえできれば解析は可能になる。しかし、言い淀みや誤用など不規則に生じる現象に対しては、UniDic の単語辞書を修正しても解決できないため、タグ付けによる効果は大きい。ただし、現在の Co-Chu の機能では誤解析が解消できないケース（「単」タグをつけたもの）もあり、今後機能を拡充していく必要がある。また本研究の実証実験では、解析の妨げになるもののみにタグを付与したが、誤解析箇所に限らず、全ての誤用にタグを付す等、研究者の関心に応じて自由なタグ付けを行うことも可能である。Co-Chu では発話者の属性を指定した検索もできるため、誰に誤用が多いのか、どんな誤用が多いのか等を詳しく分析することもできる。また、方言等の分析も可能な仕組みであり、研究目的に応じてタグを工夫することで、広範な応用が可能となる。こうした Co-Chu の特性を生かして、さらにどのような活用方法が可能かについても研究を継続していきたいと考えている。

参考文献・関連 URL

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

(2) 岡照晃(2017)「CRF 素性テンプレートの見直しによるモデルサイズを軽量化した解析用 UniDic -unidic-cwj-2.2.0 と unidic-csj-2.2.0 - 」, 『言語資源活用ワークショップ 2017 発表予稿集』, pp.143-152.

(3) MeCab https://mecab.googlecode.com/svn/trunk/mecab/doc/index.html （2018 年 9 月 25 日閲覧）

(4) UniDic http://www.ninjal.ac.jp/corpus_center/unidic/ （2018 年 9 月 25 日閲覧）