様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月　 ８ 日現在

機関番号：３３９２１
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ００７２３

研究課題名（和文）マルチコーパスシステムCo-Chuの開発と運用実験

研究課題名（英文）The Development of Multi-Corpus Analysis System Co-Chu

研究代表者
山本　裕子（YAMAMOTO, Hiroko）

愛知淑徳大学・交流文化学部・教授

研究者番号：２０４１０６５７

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究では、研究者が自らの関心に基づいて収集したオリジナルデータを容易にコーパス化して分析できるコーパスシステムCo-Chuを開発した。本システムは1)オリジナルデータを扱うことができる、2)独自のタグおよびメタ情報が検索できる仕組みを備え、独自の視点からの分析が可能である、3)操作が平易である、の３点を満たすウェブアプリケーションであり、これによりコンピューターが苦手な研究者や日本語教師であっても容易に目的に応じたコーパス分析が可能になった。完成したシステムは、無償で公開し、日本語教育関係者が各々のニーズに応じて利用できるため、言語研究・教育への様々な活用が期待できる。

研究成果の学術的意義や社会的意義
本研究の学術的および社会的意義は次の2点に集約できる。1)誰でも使える平易なツールの開発：一つのインターフェイスで、データ取り込みから、形態素解析、分析まで行えるウェブアプリケーションとし、できる限り平易なシステムを目指して開発した。2)コーパス分析手法の提案による独創的な研究の進展：本システムはオリジナルデータの取り込みができるだけでなく、研究の目的に応じてさまざまな分析が可能である。システムの無償公開とともに、運用実験を通して得られた知見を基に、新たな分析手法や教育への各種の応用方法を提案した。これにより、一般の日本語教師の教育への活用や言語研究者の幅広い分野での研究活動が可能になった。

研究成果の概要（英文）：This study has developed a multi-corpus system, Co-Chu, which enables researchers to easily convert original data collected based on their own interests into a corpus for analysis. This system is a web application that satisfies the following three requirements: 1) it can handle original data, 2) it has a system for searching original tags and meta-information, which enables analysis from an original perspective, and 3) it is easy to operate. This makes it possible for researchers and Japanese language teachers who are not good with computers to easily analyze corpora according to their purposes. The completed system is open to the public free of charge and can be used by those involved in Japanese language education according to their own needs. It is expected to be used in a variety of ways in language research and education.

研究分野：日本語教育

キーワード：マルチコーパスシステム　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　メタ情報検索機能

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1) 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形でコーパスの活用が提案されるようになった。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きている。荻野・田野村(2011)[1]、石川(2012)[2]など日本語のコーパスを扱う入門書も多く出版されているが、これらはコンピューターに関する一定以上の知識を備えた研究者を想定したものであり、コンピューターに関する知識の乏しい日本語教師や文系研究者にはハードルの高いものとなっている。李・石川・砂川(2012)[3]は、一般の日本語教師を念頭に置いて、自らのデータをコーパス化し、分析する手法を紹介した入門書である。しかし、コーパスの構築、形態素解析、検索等、それぞれ別々のソフトが紹介されており、実際の活用にまでつなげるのは容易ではない。一方、国立国語研究所の「中納言」や「NINJAL-LWP」は、多様な機能を持った強力なツールであるが、データを含んだ形で構築されており、自らの収集したデータを分析することはできない。したがって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、自ら集めたデータをコーパスとして活用できない日本語教育関係者が多くいるという状況であった。

(2) そこで、本研究においては、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はどうあるべきかという問いをもとに、これまでに開発したシステム(Co-Chu)の発展形として、一般公開を前提にした試作版を開発し、運用実験を経て、コーパスシステム Co-Chu（以下 Co-Chu）を完成させることにした。

２．研究の目的

平易な操作で、知りたいことを分析できるコーパスシステム Co-Chu を開発し、公開するとともに、コンピューターが苦手な研究者や日本語教師であってもコーパス分析ができるよう研究目的に応じた言語分析や日本語教育への応用方法を提案することを目的とした。

３．研究の方法

本研究では、コンピューターを熟知した技術者と言語教育・言語研究の専門家とが協働して研究を進め、個々の研究者が自らの関心に基づいて収集したオリジナルデータを容易にコーパス化して分析できる汎用性の高いシステムの開発を行った。

具体的には a.汎用性の高いシステムの開発、 b.運用実験とそれに伴うシステムの改良、c.システムの公開と活用方法の提案という３つの面から研究を進めた。

４．研究成果

研究成果を、(1)システム開発のプロセス、 (2) 開発したシステムの特徴、 (3)システム及び成果の公開という３点に分けて、以下に述べる。

(1) システム開発のプロセス

多様な検索機能を搭載するとともに幅広いジャンルのデータに対応させるために、メタデータ（誤用や語用論的情報）の処理機能の拡充を中心に開発を進めた。

まず、日本語教育関係者がコーパス分析の際にニーズの高い領域と考えられる、書きことば（作文)、話しことば（会話データ、アニメのスクリプト）といったジャンルごとにデータの収集と各種のタグやメタ情報に対応可能な機能の拡充を行った。その上で、研究目的に応じて開発された機能を用いて、データを分析し、システムの問題点等を開発担当者にフィードバックし、必要な改良を経て仕様を決定した。当初、動画の取り込みも可能な仕様を想定していたが、動画を含めるとデータ量が非常に多くなる。しかし、公開後、各ユーザーの取り込む分析データの総容量を想定すると、公開システムでは動画の取り込みは避ける必要がある。また、さまざまなタイプのデータをもとに運用実験した結果、メタ情報を活用することで動画そのものを含めなくても十分に多様な検索が可能であることが実証できた。そこで、公開システムで扱うデータは、テキストデータに限ることとした。仕様決定後、ユーザー管理用システムの構築を経て、コーパス分析システム Co-Chu として完成させた。

(2) 開発したシステムの特徴

Co-Chu の特徴は、自ら収集したデータ（csv あるいは txt 形式で作成）をもとにコーパスを構築し、各自のニーズに応じた多様な検索が、平易な操作で可能に行える点にある。

①　平易な操作性：

図１は、Co-Chu のトップページを示している。画面上部のメニューバーが示す通り、Co-Chu は、コーパスの構築（【Build】)、データ取り込みと形態素解析（【Import】)、データの編集（【Edit】)、分析（【Analyze】）という一連の作業が、一つのインターフェイスで行えるウェブアプリケーションシステムになっている。

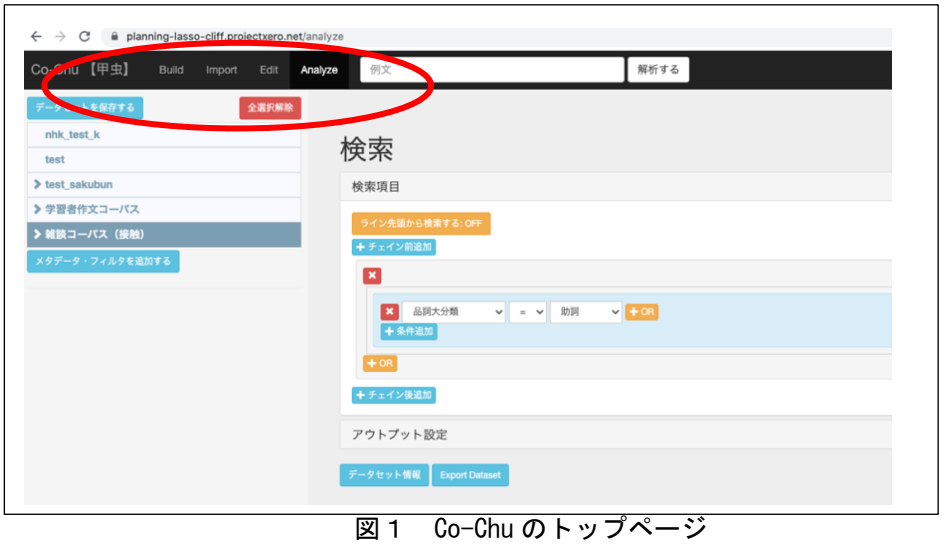

図１　Co-Chu のトップページ

② 形態素の誤解析への対応

a. タグ付けによる対応

本研究で分析対象として想定しているデータは、会話のスクリプトや日本語学習者の作文等、形態素解析に適さない表現等も含まれているデータである。Co-Chu では、こうした形態素解析の際に誤解析の生じやすい誤用や話し言葉を含むテキストについても、タグ付けによって、適切に分析することが可能である。

形態素解析で誤解析が生じた場合には、Co-Chu の【Edit】機能を活用し、データに、以下のようにタグ付けを行う。

｜適切な語形｜（タグ種別　実際に用いられた語形）

例１　残念だね、｜本当｜（話　っと）にね。

例１は、「残念だね、っとにね」という発話にタグを付したものである。「っとに」は「と（助詞）＋に（助詞）」と解析されてしまうが、「｜本当｜（話　っと）に」とタグ付けすることによって、「本当（名詞）＋に（助詞）」と適切に解析できるようになる。

例２　元の発話：それ忘れなく、やっぱり次の子とか

タグ付け：それ忘れ｜ないで｜（誤　なく：活用）、やっぱり次の子とか

この例では、「忘れないで」の活用が誤っていたため、「誤」タグを付け、「タグメモ」には「活用」と記載している。このように、タグメモにどのような誤用だったかなどの情報を記載しておくことで、後の分析がスムーズに行えるようになる。

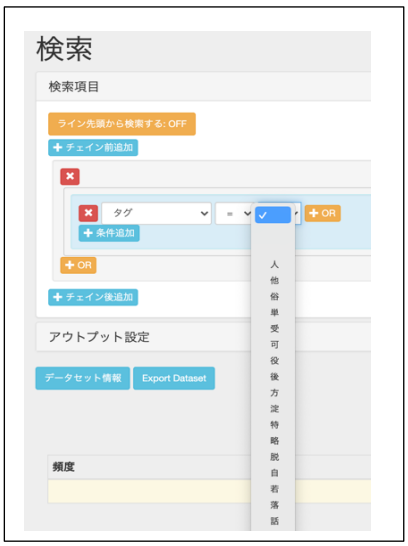

図２　タグを検索項目として指定する

タグおよびタグメモという形で記載した情報は、【Analyze】機能を用いることによって、タグの種別ごとに検索ができる仕様になっている。分析方法はユーザーごとに異なっているので、タグは各ユーザーが個別に研究目的に応じて自由に設定することができる。

また、Co-Chu では、新たに設定したタグをシステムが自動でタグの種類として追加する仕様になっている。

図２はタグ検索の画面である。画面中央左の検索項目欄で「タグ」を指定すると、図２のようにプルダウンによってタグの種類が表示される。そこから検索したいタグ項目を指定することによって、コーパス内の当該タグが付与された文を抽出することが可能になる。このようにタグ検索機能は、誤用分析にも有用である。

b. 品詞情報の書き換えによる対応

タグ付けによって、誤解析の多くを解消することができるが、誤解析の中には、単語の切り出しは正しくても品詞情報が誤っているという場合もある。その場合にはタグ付けを次のようにした上で、品詞情報を書き換えて対応できるようにした。

例３　で、えっ、その XX 大学で、→ 解析結果：「その」は「連体詞」
タグ付け：｜その｜（単　その：感動詞）XX 大学で、

例３において、「その」が「そのー」のように長音で発話されれば、感動詞（フィラー）と正しく解析されるが、「その」のように長音でない場合、MeCab は「その」を「連体詞」として解析する。こうした場合は、【Edit】画面で解析結果を編集することができる。図３左は形態素解析結果の画面であるが、図３右のように、修正したい箇所にカーソルを当て、適切な内容に編集することが可能である。この修正は形態素解析の結果そのものを書き換えるわけではなく、結果画面の表面的な修正に過ぎないが、書き換え作業をしなければ分析結果も誤ったままになってしまうため、この作業は正確な分析には不可欠である。

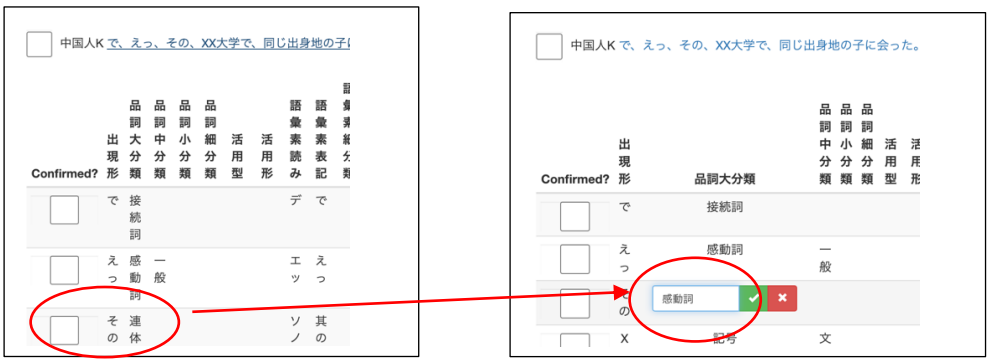

図3　形態素解析の結果の直接編集画面

③　文レベルの問題への対応

話し言葉や学習者の作文を分析しようとすると、問題点が複数の形態素にまたがっているため、形態素単位では対応できない場合もある。典型的な例は、日本語学習者の作文などに見られる次のような「ねじれ文」である。

例４　よかったのは、皆がやさしくて、いろいろ助けてくれた。

例４のような場合に、Co-Chu では、取り込む前のデータに当該の文（あるいは発話）に関する情報を図４のように記すことで、検索に使用できる。

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

図４　文レベルでメタ情報を付したデータ

④　メタ情報の活用

上記②③は、当初、形態素解析の誤りへの対応や、学習者の作文などに見られる「誤り」を記述・検索できるように開発した機能であるが、これを応用することで、幅広い研究目的に対応可能である。システム使用者が自身の目的に合わせて自由にメタ情報を付すことができるので、a. 語（表現）の運用情報の正確な収集、b. 同一形態素で意味が異なるものの区別、c. 存在しない要素（必要な語句の非用等）の検索なども可能になる。

a. 語（表現）の運用情報の正確な収集

タグメモの活用によって、どのように語（表現、文法項目など）が使用されているかをより正確に把握することができる。例えば、日本語学習者の助詞の誤用は上級になっても多く見られるが、誤用の内容を例５、例６のようにタグメモに記載することで、詳しい情報を収集できる。

例５　中国語｜に｜（誤　で：助詞で）曖昧表現はほとんどないが
例６　そんなこと｜は｜（誤　では：助詞で、助詞過剰）ないはずだ。

タグメモによって、助詞「で」のどのような誤用が多いのか、詳細かつ正確に把握できる。

作文に見られる誤用には個人差が大きい。個人差が大きいからこそ個人の傾向に即した指導が求められる。Co-Chu の活用は個人の傾向を掴むのに非常に有効であろう。

b. 同一形態素で意味が異なるものの区別：例　受け身形と可能形の区別

受け身形と可能形は形式上の区別がないため、従来のコーパスではそのまま検索することはできない。しかし、Co-Chu では、タグ付けを活用することで区別できる。

例７　ここでは、おいしい焼肉が｜食べられ｜（可　食べられ）るよ。
例８　あー、とっといたケーキが｜食べられ｜（受　食べられ：持ち主）ちゃった。

例７、例８のようにタグ付けすることで、「可能」の「られ」と「受け身」の「られ」を区別して検索することが可能である。さらに、例８のようにタグメモに「受け身の種類」を記載しておくこともできる。このようにすれば、どのような種類の受身文がどの程度用いられているかを容易に検索できる。

c. 存在しない要素の検索

さらに、発話されていないものもタグによって検索可能になる。例えば、自然な話しことばには助詞が用いられない場合も多く見られる。このような「助詞の脱落」には例９のようにタグを付して検索対象にすることができる。

例９　さっきあいつ｜が　｜（脱　）電話してきたんだけど、、

また、いわゆる「ら抜きことば」も例 10 のようにタグをつけることで検索が可能になる。

例 10　あの子と同じ風景が｜見られる｜（可　見れる：ら抜き）から。

このようにデータ内の一段活用動詞の可能形に「可」タグを付し、タグメモにら抜きかどうかを記載しておけば、「ら抜きことば」の使用の様子を容易に確認することができる。図５は実際にアニメスクリプトにおいてタグ「可」を検索し「ら抜き」の有無を比較した結果画面の一部である。

ライン	サブコーパス名	ライン番号	タグメモ
唯一　起きて｜いられ｜（可　られ：非ら抜き、音変化）たぜ。	4月_ep2	149	非ら抜き、音変化
慌てて花を買って渡した今日のことは　忘れ｜られ｜（可　られ：非ら抜き）ないよ。	4月_ep2	186	非ら抜き
そんな生活｜に｜（脱　）｜は｜（脱　）普通　耐え｜られ｜（可　られ：非ら抜き）ないわよ。	4月_ep4	15	非ら抜き
あの子と同じ風景が見｜られる｜（可　られる：非ら抜き）から。	4月_ep9	35	非ら抜き
僕らは　誰かと出会った瞬間から一人では　い｜られ｜（可　られ：非ら抜き）ないんだ。	4月_ep22	22	非ら抜き
忘れ｜られ｜（可　られ：非ら抜き）ない風景がこんな　ささいなことなんておかしいよね。	4月_ep22	148	非ら抜き
指｜を｜（脱　）くわえて待って｜られ｜（可　られ：非ら抜き）ませんよ！	Conan_81	389	非ら抜き
｜来られる｜（可　来れる：ら抜き）よな？	Conan_81	543	ら抜き
これから|おじさん｜（人　おじさん：１人称として）の言うとおりに｜来られる｜（可　来れる：ら抜き）よな？	Conan_82	51	ら抜き

図５　タグ「可」でアニメスクリプトを検索し、タグメモを表示させた画面

このように、従来のツールでは検索できなかった存在しないものも検索が可能になり、自然発話の実態を解明するのに大きく寄与できると考えられる。

(3) システム及び成果の公開

システム開発の各過程で運用実験を行い、その結果を国内外の学会で発表し、論文としてまとめた。最終年度においては、完成したシステムを、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た。公開にあたっては、学会等で利用希望者を募るとともに、Co-Chu の具体的な活用方法を伝えるためにセミナーを開催した。また、運用実験を通して得られた研究成果をもとに、言語分析の手法や、研究への応用方法をマニュアルの形でまとめ、一般に配布するとともに、関連する国内外の学会で発表した。現在、すでに、日本語研究者、日本語教育関係者等がシステムの活用を開始している。今後、Co-Chu の利用法や活用方法に関して、ホームページを活用して情報提供を継続していく予定である。

<引用文献>

[1]荻野綱男・田野村忠温編(2011)『コーパスの作成と活用』(講座 IT と日本語研究)明治書院.
[2]石川慎一郎(2012)『ベーシックコーパス言語学』ひつじ書房.
[3]李在鎬・石川慎一郎・砂川有里子(2012)『日本語教育のためのコーパス調査入門』くろしお出版.

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　6件）

１．著者名 山本裕子	４．巻 11
２．論文標題 コーパス分析システムを用いたアニメのスクリプト分析－助詞の脱落に注目して－	５．発行年 2021年
３．雑誌名 愛知淑徳大学論集　交流文化学部篇	６．最初と最後の頁 77-92
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 山本裕子・本間妙・川村よし子	４．巻 43
２．論文標題 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－	５．発行年 2020年
３．雑誌名 中部大学人文学部研究論集	６．最初と最後の頁 1-24
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 山本裕子	４．巻 10
２．論文標題 中国人日本語学習者は「よい会話」をどのように捉えているかー日本語母語話者との会話データに基づいてー	５．発行年 2020年
３．雑誌名 愛知淑徳大学論集 交流文化学部篇	６．最初と最後の頁 1-17
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 浅井優介・北村達也・川村よし子	４．巻 13-1
２．論文標題 小学生向け教育番組の音声に用いられる語彙の予備調査	５．発行年 2020年
３．雑誌名 甲南大学紀要知能情報学編	６．最初と最後の頁 67-75
掲載論文のDOI（デジタルオブジェクト識別子） 10.14990/00003648	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 小森早江子	４．巻 44
２．論文標題 英語母語上級日本語学習者の話し言葉における助詞の脱落について	５．発行年 2020年
３．雑誌名 中部大学人文学部論集	６．最初と最後の頁 69-82
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 山本裕子・ラニガン　マシュー	４．巻 24
２．論文標題 アニメの日本語教育への活用－コーパス分析システムCo-Chuを用いて－	５．発行年 2020年
３．雑誌名 ヨーロッパ日本語教育	６．最初と最後の頁 629-631
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計11件（うち招待講演　0件／うち国際学会　5件）

１．発表者名 北村達也・松本侑也・川村よし子
２．発表標題 小学生低学年向け教育番組の音声に用いられる語彙の調査
３．学会等名 第56回日本語教育方法研究会
４．発表年 2021年

１．発表者名 山本裕子・本間妙・川村よし子・小森早江子
２．発表標題 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
３．学会等名 2020年度日本語教育学会秋季大会
４．発表年 2020年

１．発表者名
山本裕子

２．発表標題
コーパス分析システムを用いたアニメの分析－助詞の脱落に注目して－

３．学会等名
第11回日本語実用言語学国際会議（国際学会）

４．発表年
2020年

１．発表者名
本間妙・山本裕子

２．発表標題
コーパス分析システムを活用した作文指導の可能性

３．学会等名
第31回　第二言語習得研究会全国大会

４．発表年
2020年

１．発表者名
山本裕子

２．発表標題
中国人日本語学習者の共感的言語行動-日本語母語話者との会話データから比較して－

３．学会等名
対照言語行動学研究会

４．発表年
2019年

１．発表者名
山本裕子・川村よし子・本間妙・小森早江子・ラニガンマシュー

２．発表標題
誤用や話し言葉に対応可能なコーパス分析システムにおけるタグ検索機能

３．学会等名
CASTEL/J2019（国際学会）

４．発表年
2019年

１．発表者名
山本裕子・ラニガンマシュー

２．発表標題
アニメの日本語教育への活用-テキスト分析システムCo-Chuを用いて-

３．学会等名
第23回ヨーロッパ日本語教育シンポジウム（国際学会）

４．発表年
2019年

１．発表者名
本間妙・山本裕子・川村よし子・小森早江子

２．発表標題
コーパス分析システムCo-Chuの作文指導への活用

３．学会等名
日本語教育方法研究会

４．発表年
2019年

１．発表者名
ラニガン・マシュー,山本裕子,本間妙

２．発表標題
日本語テキスト分析システムCo-Chuの開発と分析事例

３．学会等名
ヴェネツィア2018日本語教育国際研究大会　（国際学会）

４．発表年
2018年

１．発表者名
山本裕子,小森早江子,ラニガン・マシュー

２．発表標題
アニメは日本語教育に使えるか　-Co-Chuを使ったアニメの分析-

３．学会等名
ヴェネツィア2018日本語教育国際研究大会　（国際学会）

４．発表年
2018年

１．発表者名
山本裕子,川村よし子,小森早江子,本間妙

２．発表標題
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

３．学会等名
2018年度日本語教育学会秋季大会

４．発表年
2018年

〔図書〕　計1件

１．著者名	４．発行年
Zimmerman, E. and McMeekin. A. (eds)	2019年
２．出版社	５．総ページ数
Multilingual Matters	384
３．書名	
Technology Supported Learning In and Out of the Japanese Language Classroom	

〔産業財産権〕

〔その他〕

研究成果公開用ホームページ
https://cochu.org

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	川村　よし子 (KAWAMURA Yoshiko) (40214704)	東京国際大学・言語コミュニケーション学部・教授 (32402)	

６．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	小森　早江子 (KOMORI Saeko) (60221248)	中部大学・人文学部・教授 (33910)	

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	本間　妙 (HOMMA Tae)		
研究協力者	ラニガン　マシュー (LANIGAN Matthew)		

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関