様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　２３　日現在

機関番号：　６２６１８
研究種目：　基盤研究(C)
研究期間：　2012～2014
課題番号：　２４５２０５２２

研究課題名（和文）近世口語文を対象とした形態素解析辞書の開発

研究課題名（英文）Development of a dictionary for morphological analysis of Late Middle Japanese text

研究代表者

小木曽　智信（OGISO, Toshinobu）

大学共同利用機関法人人間文化研究機構国立国語研究所・言語資源研究系・准教授

研究者番号：２０３３７４８９

交付決定額（研究期間全体）：（直接経費）　　3,900,000円

研究成果の概要（和文）：今日、日本語の歴史を研究するためのコーパスの構築が望まれている。このコーパス構築のためには、さまざまな歴史的な資料に単語の情報を自動で付与する形態素解析を行う必要があるが、これまでに近世（江戸時代）の資料を対象としたものはなかった。本研究の目的は、主として近世の口語資料を、研究に必要な精度で解析できるようにすることである。既存の辞書を元に見出し語を増やし、専用のコーパスを用いて機械学習を行った結果、狂言台本や洒落本のテキストを高い精度で解析することが可能になった。この辞書は「日本語歴史コーパス」の構築に利用され成果を上げている。

研究成果の概要（英文）：It is expected to construct a corpus to study the history of Japanese language today. To construct this corpus, it is necessary to perform automatic morphological analysis to annotate word information to various historical documents. However, there was no dictionary fitting for documents of Late Middle Japanese (in the Edo era) so far. Our purpose was to make it possible to analyze documents of the Edo era with a high accuracy necessary for the linguistic study. Expanding the entries of the existing dictionary and using a specific corpus for training data, we could analyze texts of kyogen and Sharebon with a high degree of accuracy. These dictionaries are used for the construction of the Corpus of Historical Japanese now.

研究分野：　日本語学

キーワード：　近世語　形態素解析　コーパス　日本語史　狂言台本　江戸語　上方語

１．研究開始当初の背景

現代日本語の大規模コーパス（BCCWJ：『現代日本語書き言葉均衡コーパス』）の完成を受け、日本語の歴史的な研究を行うことのできる歴史的コーパスの構築が期待されている。国立国語研究所では「通時コーパスの設計」プロジェクトがその準備作業を行ってきた。海外ではオックスフォード大学のフレレスビック教授の下で日本語の歴史コーパス（Oxford VSRPJ corpus）の構築が進められており、国際的にもその必要性は高まっている。

研究代表者らは、これまでに現代語コーパス（BCCWJ）のための形態素解析辞書「UniDic」の開発に携わった。また、現代語の UniDic を元に、近代の文語論説文を対象とした「近代文語 UniDic」、中古の和文系資料を対象とした「中古和文 UniDic」を開発し公開してきた。これらの辞書により、中古以降の和文系の資料や、近代の文語論説文などについては形態素解析済みのコーパスを構築する目処が立っている。

しかし、日本語の通時コーパスの中でかなりの量を占める近世の口語文については、いまだ高精度な解析を行う目処が立っていない。日本語の通時的な研究が可能なコーパスの構築には、近世口語資料の解析の問題は避けて通れない。本研究課題は、この問題を解決しようとするものである。

２．研究の目的

本研究課題では、一般的な校訂済みの近世語の資料をコーパス構築に十分な精度で解析することが可能な形態素解析辞書を作成することを第一の目標とする。この辞書には現代語・近代文語・中古和文と同様、UniDic の枠組みを採用し、語彙素・語形・書字形に階層化された、言語研究に適した電子化辞書とする（UniDic の構造については、伝ほか「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22 号）。また、語の区切り方は「短単位」という規定に基づいて揺れのない斉一な単位となるように配慮し、他の同時代資料の解析結果との比較を可能にする。さらに現代語や近代文語・中古和文の UniDic との互換性を可能な限り確保し、通時的な比較も可能とする。この辞書と、既存の形態素解析器（プログラム）である「MeCab」（工藤拓）とを組み合わせ、解析システムとして日本語研究者に利用しやすい形で公開する。

この形態素解析システムが利用可能となる後半には、この解析辞書の評価のため、解析結果を用いたコーパス言語学的手法による近世語研究を行う。特に形態素解析が特に威力を発揮すると考えられる分野を中心に、テキスト検索では用例を採集しづらい助動詞などの付属語や、資料中の全語彙の集計結果を用いた文体比較等を対象とする。いずれもデータベース上で品詞情報付きのコーパスを用いて資料の全ての語彙を取り扱う新しい研究手法によるものである。これにより近世語の研究における形態素解析を用いた方法の有効性を示す。

近世語の研究においてはいまだ目視による用例収集が主流で、テキストデータの利用がようやく始まったところにすぎず、本格的なコーパスを用いた研究は行われていない。本研究で形態素解析が可能になることではじめて、近世語研究にコーパス言語学的な手法を本格的に導入することが可能になる。多数の資料が残されている近世語の研究では、現代語で行われているコーパス言語的な手法による研究が効果を上げるものと考えられ、本研究による成果は、今後の近世語研究の発展に資することになると考えられる。

また、本格的な通時コーパスの構築においては、日本語の歴史的資料のなかで大きな分量を占める近世語資料の解析が大きな問題になっている。本研究で近世口語文の形態素解析が可能になることによって、はじめて「通時的」なデータが可能になる。この意味で、コーパスを用いた日本語の通時的研究の発展全体に対しても大きな貢献を行うことができる。

３．研究の方法

(1)全体の流れ

既存の形態素解析用辞書『UniDic』に対して語彙を追加し、近世語の学習用コーパスを整備し、近世口語文を一定の精度で解析できる形態素解析辞書を開発する。この辞書を用いて近世口語資料を新規に解析して未登録語の発見・追加を行い、学習用コーパスの量を拡大する。

こうして辞書の整備を継続して、最終的に目標精度 95%の最終版の形態素解析辞書を公開する一方、作成したコーパスと解析データを用いた新しい手法にもとづく近世語研究を行った。

(2)研究分担

次のような分担で形態素解析辞書の作成を行った。

小木曽：近世語資料をより高い精度で解析するための技術開発。形態素解析辞書の公開。

岡部：江戸語の文法項目整備、辞書見出し語選定。コーパスを活用した江戸語の研究。

村上：上方語の文法項目整備、辞書見出し語選定。コーパスを活用した上方語の研究。

市村：辞書見出し語のデータベース登録とコーパスの人手修正。

(3)形態素解析辞書の整備

形態素解析辞書の構築とその精度向上のため、次の①～④を行って辞書と学習用コーパスの拡充に努めた。

①　近世語資料の解析と未登録語の抽出

これまでに作成した洒落本・人情本のテキストデータを、既存の形態素解析辞書（『中古

和文 UniDic』『近代文語 UniDic』で解析し、その解析結果にみられる未登録語・誤解析例から、辞書データベースに登録する候補を選び出す。

② 辞書データベースの整備

すでに構築されている UniDic 用の形態論情報データベース・システム（サーバと見出し語入力・編集用ソフト）を用いて、近世語資料で用いられる見出し語（活用表を含む）を登録してゆく。データベースへの登録は UniDic の構造にあわせて「語彙素」「語形」「書字形」の階層ごとに行う。語の区切り方は国語研で設計された「短単位」に従い、『近代文語 UniDic』『中古和文 UniDic』等既存の辞書との互換性を維持する。

③ 解析結果の整備、学習用コーパスの作成

解析対象としたテキストのなかから代表的な資料を選び、これを特に高精度なコーパスとして整備し、学習用コーパス（形態素解析システムが利用する語の連接コストなどを統計的に学習することのできる資料）とする。

④ 近世語資料を対象とした解析辞書の試作

語彙を増補した辞書（②）と、新たな学習用コーパス（③）を元にして、近世口語資料を対象とした解析辞書を作成する。この際、『近代文語 UniDic』『中古和文 UniDic』等の学習用コーパスも活用し、適切な機械学習が行えるように努めた。

(4)解析結果を用いた近世語研究

辞書の学習用に整備した学習用コーパスとテキストの形態素解析結果を活用し、コロケーション強度などの統計分析の手法を取り入れた近世語の記述的研究をおこなった。

４．研究成果

形態素解析辞書は、狂言用の形態素解析辞書と、洒落本等の近世口語資料汎用の形態素解析辞書の 2 つの辞書を作成した。前者の狂言用の辞書は、単語分割で 98.9%、品詞認定で 97%、語彙素認定で 96%という高い精度を達成し、国立国語研究所『日本語歴史コーパス』「室町時代編Ⅰ 狂言」の構築に活用された。後者の汎用辞書は、品詞認定でおおむね 90%程度にとどまるが、地の文と会話文を分けて解析を行うプログラムを開発することにより、より高い精度で解析することが可能になった。この辞書により洒落本のデータ計 15 作品について形態素解析を施し、人手修正作業を行った。

研究論文としては、近世語の形態素解析とコーパス開発を扱ったもの、形態素解析によるアノテーションを含む近世語のコーパスを活用した研究を含め、合計 12 編を発表した。

学会発表は、海外の国際学会での発表 3 件を含め 12 件行った（2015 年度に発表が確定している国際学会 1 件を含む）。

本研究で開発した形態素解析辞書を用いることで、中世から近世の口語資料について形態素解析を行うことが可能になった。これにより、室町時代語や江戸時代（上方・江戸）語の資料に形態素解析を施して『日本語歴史コーパス』を拡張する基盤が整った。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 12 件）

① 市村太郎、副詞「ほんに」をめぐって―「ほん」とその周辺―、日本語の研究 10(2) 、査読有り、2014、pp.1-16
② 小木曽 智信，小町 守，松本 裕治、歴史的日本語資料を対象とした形態素解析、自然言語処理 20(5)、査読有り、2013、pp.727-748
③ 岡部嘉幸、モダリティに関する覚え書き、語文論叢 28、査読なし、2013、pp.96-75
④ 小木曽智信、歴史コーパスにおける形態素解析と辞書整備、日本語学 33(14)、査読なし、2014、pp.83-95
⑤ 岡部嘉幸、近世江戸語のハズダに関する一考察－現代語との対照から－、『日本語文法史研究 2』（ひつじ書房）、査読なし、2014、pp.173-193
⑥ 村上謙、近世語研究の学史的展開―戦前における「対立」の思想を中心に―、『近代語研究』第 18 集（武蔵野書院）、査読なし、2015、pp.227-244
⑦ 村上謙、近世後期上方における待遇表現化のコロケーション、日本語学 33(14) 、査読なし、2014、pp.152-161
⑧ 市村太郎、近世口語資料のコーパス化―狂言・洒落本のコーパス化の過程と課題、日本語学 33(14)、査読なし、2014、pp.96-109
⑨ 市村 太郎，河瀬 彰宏，小木曽 智信、近世口語テキストの構造化とその課題、情報処理学会研究報告．人文科学とコンピュータ研究会報告、査読無、2012-CH-96、2012、pp.1-8
⑩ 小木曽智信、旧仮名遣いの口語文を対象とした形態素解析辞書、じんもんこん 2012 論文集、査読なし、2012、pp.25-32
⑪ 村上謙、明治大正期関西弁資料としての曾我廼家五郎喜劇脚本群、埼玉大学国語教育論叢 16、査読なし、2013、pp.1-15
⑫ 村上謙、ジャからヤへ―明治大正期関西弁指定表現体系における「標準語化」の影響―、『近代語研究』第 17 集、査読なし、2013、pp.97-114

〔学会発表〕（計 18 件）

① Toshinobu OGISO, Tomoaki KONO, Taro ICHIMURA、Morphological Analysis of Japanese Kyogen Text、DH2015、2015 年 06 月 29 日～2015 年 07 月 03 日、ウェスタンシドニー大学（オーストラリア・シ

ドニー）
② 小木曽智信，鴻野知暁，市村太郎、狂言台本の形態素解析、日本語学会 2015 年度春季大会、2015 年 05 月 24 日、関西学院大学（兵庫県・西宮市）
③ 市村太郎，渡辺由貴，鴻野知暁，河瀬彰宏，小林正行， 山田里奈，堀川千晶，村山実和子，小木曽智信，田中牧郎、『虎明本狂言集』コーパスの公開、日本語学会 2015 年度春季大会、2015 年 05 月 24 日、関西学院大学（兵庫県・西宮市）
④ 渡辺由貴，市村太郎，鴻野知暁、『虎明本狂言集』のコーパスデータにおける短単位認定の諸問題、第 7 回 コーパス日本語学ワークショップ、2015 年 03 月 11 日、国立国語研究所（東京都・立川市）
⑤ Akihiro KAWASE, Taro Ichimura, Toshinobu OGISO、Problems in Encoding Documents of the Early Modern Japanese、DH2014、2014 年 07 月 09 日、ローザンヌ大学（スイス・ローザンヌ）
⑥ 小木曽智信、歴史コーパスの構築と日本語研究、国立国語研究所合同研究発表会「コーパスに見る日本語のバリエーション －方言コーパス・会話コーパス・歴史コーパス・学習者コーパスから－」、2014 年 12 月 06 日、国立国語研究所（東京都・立川市）
⑦ 岡部嘉幸、洒落本における格助詞「に」と「へ」について－洒落本コーパスを資料として、国立国語研究所共同プロジェクト「通時コーパスの設計」研究発表会、2014 年 11 月 09 日、国立国語研究所（東京都・立川市）
⑧ 渡辺由貴，市村太郎、『虎明本狂言集』における濁点無表記箇所について－コーパス整備の過程から－、日本語学会 2014 年度秋季大会、2014 年 10 月 19 日、北海道大学（北海道・札幌市）
⑨ Toshinobu OGISO，Yoshiyuki OKABE、Design and Compilation of the Sharebon Corpus、14th International Conference of European Association for Japanese Studies、2014 年 08 月 29 日、リュブリャナ大学（スロベニア・リュブリャナ）
⑩ 河瀬彰宏，市村太郎，小木曽智信、『虎明本狂言集』における会話文の計量分析、言語処理学会第 20 回年次大会、2014 年 03 月 19 日、北海道大学（北海道・札幌市）
⑪ 鴻野知暁，小木曽智信、見出し語の時代情報を付与した電子化辞書の構築、言語処理学会第 20 回年次大会、2014 年 03 月 18 日、北海道大学（北海道・札幌市）
⑫ 河瀬彰宏，市村太郎，小木曽智信、TEI P5 に基づく近世口語資料の構造化とその問題点、じんもんこん（PNC/ECAI 合同開催）、2013 年 12 月 12 日、京都大学（京都府・京都市）
⑬ Akihiro KAWASE and Toshinobu OGISO、The Current Situation and Role of TEI P5 as an XML Standard for the Corpus of Historical Japanese、国際シンポジウムデジタル時代の人文学と仏教学の役割について（招待講演）、2013 年 11 月 17 日、東京大学（東京都・文京区）
⑭ Toshinobu Ogiso、Design and Compilation of the Corpus of Historical Japanese、国際ワークショップ・TEI と日本語歴史コーパス、2013 年 09 月 17 日、国立国語研究所（東京都・立川市）
⑮ 小木曽智信，市村太郎，鴻野知暁、近世口語資料の形態素解析の試み、第 4 回コーパス日本語学ワークショップ、2013 年 09 月 05 日、国立国語研究所（東京都・立川市）
⑯ 小木曽智信,伝康晴、UniDic2: 拡張性と応用可能性にとんだ電子化辞書、言語処理学会第 19 回年次大会、2013 年 03 月 15 日、名古屋大学（愛知県・名古屋市）
⑰ 市村 太郎,河瀬 彰宏,小木曽 智信、洒落本コーパスの構造化 －仕様と事例の検討－、第 3 回 コーパス日本語学ワークショップ、2013 年 03 月 01 日、国立国語研究所（東京都・立川市）
⑱ 岡部嘉幸,村上謙、「デハナイ、デナイ、ジャナイ －近世における否定表現一斑－」、NINJAL「通時コーパス」プロジェクト・OxfordVSARPJ プロジェクト合同シンポジウム「通時コーパスと日本語史研究」、2012 年 07 月 31 日、国立国語研究所（東京都・立川市）

〔その他〕
ホームページ等
http://www2.ninjal.ac.jp/lrc/index.php?UniDic

６．研究組織

(1)研究代表者
小木曽　智信　（OGISO Toshinobu）
国立国語研究所・言語資源研究系・准教授
研究者番号： 20337489

(2)研究分担者
村上　謙　（MURAKAMI Ken）
埼玉大学・教育学部・准教授
研究者番号： 20431728

岡部　嘉幸　（OKABE Yoshiyuki）
千葉大学・人文社会科学研究科・准教授
研究者番号： 80292738

市村　太郎　（ICHIMURA Taro）
国立国語研究所・コーパス開発センター・プロジェクト非常勤研究員
研究者番号： 10701352

(3)研究協力者
鴻野　知暁　（KOUNO Tomoaki）
国立国語研究所・コーパス開発センター・
プロジェクト非常勤研究員
研究者番号： 30751515