様式Ｃ－１９

# 科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 22 年　5 月 31 日現在

機関番号：62618
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：21520492
研究課題名（和文）：和文系資料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 an Electronic Dictionary for Morphological Analysis of Classical Japanese
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

研究成果の概要（和文）：平安時代の仮名文学作品を主たる対象とする和文系資料を対象とした形態素解析辞書「中古和文 UniDic」を一般公開した。この辞書により、校訂済みの仮名文学作品本文を高い精度（約 97%）で解析し研究に利用することが可能になった。「中古和文 UniDic」は、すでに日本語史の研究に用いられはじめているほか、国立国語研究所で構築準備が行われている通時コーパスの形態論情報付与に活用される予定である。

研究成果の概要（英文）：We have developed a dictionary for morphological analysis of classical Japanese: “UniDic for Early Middle Japanese”. This dictionary can analyze Japanese classical texts with high accuracy (approx. 97%). UniDic-EMJ is now used for linguistic research of classical Japanese texts and it will be employed in the construction of the diachronic corpus that is currently being planned at NINJAL.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2009 年度 | 1,400,000 | 420,000 | 1,820,000 |
| 2010 年度 | 900,000 | 270,000 | 1,170,000 |
| 2011 年度 | 1,000,000 | 300,000 | 1,300,000 |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 3,300,000 | 990,000 | 4,290,000 |

研究分野：
科研費の分科・細目：言語学・日本語学
キーワード：日本語史　形態素解析　コーパス　中古和文

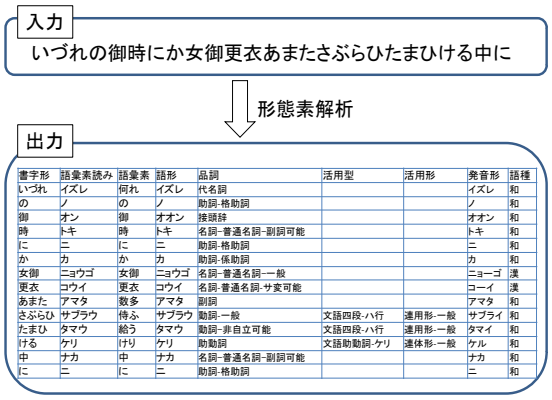
## １．研究開始当初の背景

近年、電子化された大規模な言語資料を用いて大量の用例をもとに研究を進めるコーパス言語学が大きな成果を上げつつある。しかし、日本語の歴史的な研究においては、テキストデータの集積は進みつつあるものの、高度で大規模な通時的コーパスの構築には至っていない。その理由の一つとして、文章を語に区切って品詞付けなどを自動で行う形態素解析が古典語においては十分に実現していないことがあげられる。こうした状況をふまえ、研究代表者等は近代文語文の形態素解析辞書『近代文語 UniDic』を開発、公開した（2008 年）。これにより、近代の一般的な論説文であれば 96%以上の精度で解析することが可能となり、近代から現代にかけての形態論情報付きコーパスの構築のめどがたった。しかし『近代文語 UniDic』が主たる対象とするのは近代の文語論説文であり、研究者の要望が強い文学作品や、近世以

前の資料については、充分な精度で解析を行うことができなかった。特に、語彙・表記・文法にわたって近代論説文との違いが大きい和文系資料では精度が充分ではない。そのため、和文系の資料を充分な精度で解析することのできる形態素解析辞書が望まれていた。『近代文語 UniDic』が漢文訓読系資料をある程度解析できるものであることからも、その対極的位置にある和文系資料を対象とする解析辞書の必要性は高かった。

## ２．研究の目的

本研究の目的は、中古の仮名文学作品を中心とする和文系資料を十分な精度で解析することができる形態素解析辞書を開発することである。形態素解析とは、分かち書きされない本文を自動で語に分割し、品詞や読みなどの情報を付与する技術である（下図）。




| 書字形 | 語彙素読み | 語彙素 | 語形 | 品詞 | 活用型 | 活用形 | 発音形 | 語種 |
|---|---|---|---|---|---|---|---|---|
| いづれ | イズレ | 何れ | イズレ | 代名詞 | | | イズレ | 和 |
| の | ノ | の | ノ | 助詞-格助詞 | | | ノ | 和 |
| 御 | オン | 御 | オオン | 接頭辞 | | | オオン | 和 |
| 時 | トキ | 時 | トキ | 名詞-普通名詞-副詞可能 | | | トキ | 和 |
| に | ニ | に | ニ | 助詞-格助詞 | | | ニ | 和 |
| か | カ | か | カ | 助詞-係助詞 | | | カ | 和 |
| 女御 | ニョウゴ | 女御 | ニョウゴ | 名詞-普通名詞-一般 | | | ニョーゴ | 漢 |
| 更衣 | コウイ | 更衣 | コウイ | 名詞-普通名詞-サ変可能 | | | コーイ | 漢 |
| あまた | アマタ | 数多 | アマタ | 副詞 | | | アマタ | 和 |
| さぶらひ | サブラウ | 侍ふ | サブラウ | 動詞-一般 | 文語四段-ハ行 | 連用形-一般 | サブライ | 和 |
| たまひ | タマウ | 給う | タマウ | 動詞-非自立可能 | 文語四段-ハ行 | 連用形-一般 | タマイ | 和 |
| ける | ケリ | けり | ケリ | 助動詞 | 文語助動詞-ケリ | 連体形-一般 | ケル | 和 |
| 中 | ナカ | 中 | ナカ | 名詞-普通名詞-副詞可能 | | | ナカ | 和 |
| に | ニ | に | ニ | 助詞-格助詞 | | | ニ | 和 |

平安時代の仮名文学作品を高精度に解析することが可能な辞書をまず開発し、これをもとにして後代の和文系資料（中世の紀行文や擬古物語、和文系説話集、近世・近代の擬古文等）にも対応することを考えた。

和文系資料は既に多くのものが電子化され公開されている。これらに形態素解析を施すことが可能になれば、解析結果を活用した新たな研究が可能になる。また、和文系資料の解析が可能になれば、既に実現している現代語・近代文語の形態素解析システムとあわせ、斉一な単位に基づく通時的な日本語コーパスの構築のための基盤整備が大きく前進することになる。

## ３．研究の方法

開発する形態素解析辞書の主たる対象を、漢字仮名交じりに校訂済みの中古仮名文学作品（物語・日記文学等）とし、これらについて精度 95%以上で解析することが可能な形態素解析辞書を作成した。また、中古仮名文学作品に加えて、中世の紀行文や擬古物語、和文系説話集、近世・近代の擬古文、和歌等の韻文についても解析を可能としている。

この辞書には現代語・近代文語と同様に「UniDic」（千葉大学・伝康晴他）の枠組みを採用し、語彙素・語形・書字形に階層化して研究目的に応じた見出し付与を可能とした。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、他の資料の解析結果との比較を可能にした。「短単位」の規定は現代語や近代語の 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した。

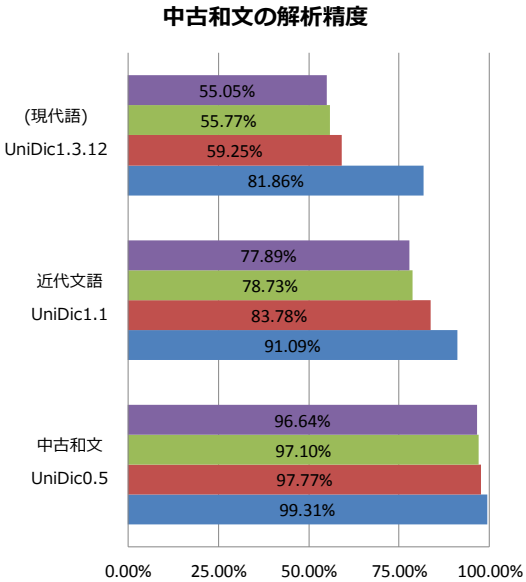
辞書への語彙登録と同時に、語の生起コスト・連接コストを統計的に取得するための機械学習用コーパスの整備を行った。中古和文テキストの解析結果に対して人手により修正を行い、最終的に他のプロジェクトによる成果とあわせて約 27 万語ぶんの中古和文の学習用コーパスを作成した。

これらの辞書・コーパスをもとにして、既存の形態素解析器「ChaSen」,「MeCab」によって形態素解析辞書を作成した。

また、この形態素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研究を行った。データベース上で統計的分析手法を用いて資料の全語彙を取り扱う新しい研究手法によるものである。こうした応用研究のために、奈良先端科学技術大学院大学で開発されたコーパス管理ツール「茶器」を用いて研究用の環境を整備した。

## ４．研究成果

上記の方法により、和文系資料を対象とした形態素解析辞書「中古和文 UniDic」を作成した。下図はその解析精度を、現代語向け、近代語向けの UniDic と比較したものである。




| | 中古和文 UniDic0.5 | 近代文語 UniDic1.1 | (現代語) UniDic1.3.12 |
|---|---|---|---|
| 発音形認定 | 96.64% | 77.89% | 55.05% |
| 語彙素認定 | 97.10% | 78.73% | 55.77% |
| 品詞認定 | 97.77% | 83.78% | 59.25% |
| 単位境界 | 99.31% | 91.09% | 81.86% |

中古和文テキストを解析した場合、現代語用の辞書では品詞認定で約 60%の精度しか出なかったものが、「中古和文 UniDic」によって約 97%という高い精度で解析を行うことが可能になった。

この形態素解析辞書は、解析用のユーザインターフェイスプログラム「茶まめ」(下図)とともにパッケージ化し、日本語研究者に使いやすい形で公開を行った。


この辞書の見出し語の整備にあたって作成した中古和文用の「短単位」の規程については『中古和文 UniDic 短単位規程集』にまとめ公開した（小椋・須永 2012)。

学会発表としては、「中古和文 UniDic」を広く研究者に紹介するため日本語学会等の国内学会でデモンストレーションを含む発表を行ったほか、EAJS ヨーロッパ日本学会においてもこの辞書の紹介を含む研究発表を行った（Ogiso 2011)。

「中古和文 UniDic」を活用した研究としては、当該辞書を用いた総索引の作成システムを開発して中世の擬古物語「恋路ゆかしき大将」を例として実際に総索引を作成して発表した（小木曽・須永 2010）ほか、この辞書による解析結果をコーパス管理ツール「茶器」で活用する研究を行った（小木曽ほか 2011)。

また、この辞書を日本語史研究に応用したものとして、コロケーション強度を用いた「あり」「なし」を後項とする複合語に関する研究発表（須永 2011)、勅撰集序文の語彙に関する研究発表（冨士池 2011）を行った。

これらの研究活動により、日本語史研究に利用可能な中古和文資料のための形態素解析辞書を開発して公開するとともに、その有効性を示すという当研究の目的は達成された。

当研究の成果の一部である「中古和文を対象とした形態素解析辞書の開発」（情報処理学会研究報告 人文科学とコンピュータ Vol.2010/CH-85、2010 年 2 月）は、情報処理学会「山下記念研究賞」を受賞しており、学会からも高い評価を得ている。また「中古和文 UniDic」は、一般向けの古文の電子辞書（iPhone・iPad アプリ『旺文社 全訳古語辞典（第三版)』物書堂）の開発に用いられるなど一般社会での活用も行われている。

## 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 5 件）


① 須永哲矢（2011）「コロケーション強度を用いた中古語の語認定」『国立国語研究所論集』 第 2 号 2011 年 11 月 pp. 91-106, 査読あり

② 小木曽智信・岡照晃・小町守・松本裕治（2011）「コーパス管理ツール「茶器」による単語情報付き古典語コーパスの活用」『人文科学とコンピュータシンポジウム論文集』 2011(8) pp. 255 - 260, 査読なし

③ 小木曽智信（2011）「通時コーパスの構築に向けた古文用形態素解析辞書の開発」『人文科学とコンピュータ』(情報処理学会研究報告) Vol.2011/CH-92: pp. 1-4, 査読なし

④ 小木曽智信・須永哲矢（2010）「「近代文語 UniDic」「中古和文 UniDic」を利用した総索引作成システムの開発」『人文科学とコンピュータシンポジウム論文集』 2010(15) pp.119-124, 査読なし

⑤ 小木曽智信・小椋秀樹・田中牧郎・近藤明日子・伝康晴（2010）「中古和文を対象とした形態素解析辞書の開発」『人文科学とコンピュータ』(情報処理学会研究報告) Vol.2010/CH-85: pp.1-8, 査読なし


〔学会発表〕（計 5 件）


① Toshinobu Ogiso（2011）「歴史的資料を対象とした形態素解析辞書によるテキスト解析」 The 13th international conference of the European association for Japanese Studies (EAJS) Tallinn, Estonia （2011 年 8 月 25 日, タリン大学)

② 須永哲矢・小木曽智信（2011）「コーパスとコロケーション強度を用いた中古語の語認定 ―名詞＋よし／あし／あり／なしを例に―」日本語学会 2011 年度春季大会（2011 年 5 月 29 日, 神戸大学）(予稿集 pp.275-280)

③ 冨士池優美（2011）「中古・院政期勅撰集序文の語彙」日本語学会 2011 年度春季大会（2011 年 5 月 29 日，神戸大学）（予稿集 pp.155-162）
④ 小椋秀樹・須永哲矢・小木曽智信・近藤明日子・田中牧郎（2011）「「中古和文 UniDic」における言語単位の設計」言語処理学会第 17 回年次大会(2011 年 3 月 8 日，豊橋技術科学大学）（発表論文集 pp.312-315）
⑤ 小木曽智信・小椋秀樹・近藤明日子・須永哲矢（2010）「形態素解析辞書「中古和文 UniDic」とその活用例」日本語学会 2010 年度秋季大会(2010 年 10 月 24 日，於愛知大学）（予稿集 pp.243-248）

〔図書〕（計 2 件）
① 小木曽智信ほか（2012）『『和文系資料を対象とした形態素解析辞書の開発』研究成果報告書』国立国語研究所（全 126 ページ）
② 小椋秀樹・須永哲矢(2012)『中古和文 UniDic 短単位規程集』国立国語研究所（全 102 ページ）

〔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）
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

(2)研究分担者
小椋　秀樹（OGURA HIDEKI）
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：00321547

田中　牧郎（TANAKA MAKIRO）
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：90217076

須永　哲矢（SUNAGA TETSUYA）
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・コーパス開発センター・プロジェクト奨励研究員
研究者番号：10589584

(3)連携研究者
傳　康晴（DEN YASUHARU）
千葉大学・文学部・教授
研究者番号：70291458

近藤　明日子（KONDO ASUKO）
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・コーパス開発センター・プロジェクト奨励研究員
研究者番号：30425722