様式 C-19

科学研究費補助金研究成果報告書

平成 21 年　6 月 16 日現在

研究種目：若手研究（B）
研究期間：2007～2008
課題番号：19720110
研究課題名（和文）近代文語文を対象とした形態素解析のための電子化辞書の作成とその活用
研究課題名（英文）Development and application of electronic modern-literary Japanese dictionary for morphological analysis

研究代表者
小木曽 智信（OGISO Toshinobu）
独立行政法人国立国語研究所・研究開発部門・研究員
研究者番号：20337489

研究成果の概要：
本研究では近代文語文の形態素解析を行うための電子化辞書「近代文語 UniDic」を作成し、日本語研究者に利用しやすい形にまとめ上げてインターネット上で一般公開を行った。この辞書は「現代日本語書き言葉均衡コーパス」の開発に用いられている UniDic をベースにしており、斉一な単位・階層化された見出し語などの設計を受け継いだ、言語研究に適した辞書となっている。また、近代文語 UniDic は現代語用の UniDic と見出し語の互換性があるため、近代語と現代語の比較研究に利用することが可能となった。
この形態素解析辞書の応用として、この辞書で『太陽コーパス』を解析した結果を用いて語彙頻度表を作成し、コーパス言語学的手法による近代語語彙の記述的研究を行った。また、これを『現代日本語書き言葉均衡コーパス』（モニター版）を現代語版の UniDic で解析した結果と比較することにより、近代語と現代語の語彙の比較研究を行った。
これらの研究成果については研究成果報告書『近代文語文を対象とした形態素解析のための電子化辞書の作成とその活用』（全 233 ページ）にまとめ、公開した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,800,000	0	1,800,000
2008 年度	1,100,000	330,000	1,430,000
年度			
年度			
年度			
総　計	2,900,000	330,000	3,230,000

研究分野：人文学
科研費の分科・細目：言語学・国語学
キーワード：近代語　形態素解析　電子化辞書　コーパス

１．研究開始当初の背景

今日、言語研究の分野では大規模な言語資料を用いて大量の用例を基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が、日本語学において、この方面の研究は未だ十分とはいえない水準にある。その理由の一つは、日本語の文章が分かち書きされず、単語の認定が容易でないために、単語ごとに区切られ情報が付与されたコーパスが十分にできていない点にある。分かち書きされない言語の文章を単語に区切り、品詞などの情報を付加するには、自然言語処理の分野で実

用化された自動形態素解析を用いることができる。これは、電子化された辞書と機械学習用のデータをもとにして、電子化された文章をコンピュータによって自動処理することを可能にするものである。現代語については、この自動形態素解析が実用化されているうえ、大規模なコーパスの開発計画が現在進行中であることから、コーパスの立ち後れの問題はやがて解決されるものと思われる。しかし、現代語以外の古い時代の日本語については形態素解析が実現されていないため、文字列での検索しか行えない状況である。このままでは、文字列検索が難しい語は研究テーマとして取り上げられないうえ、語に付けられた情報を利用した高度なコーパス言語学的研究は望めない。

本研究課題の代表者は『太陽コーパス』(日本語研究用の近代雑誌の本文データベース。約 1450 万字。国立国語研究所編。博文館新社刊）の構築に関わり、その設計や応用プログラムの作成、これを利用した研究を行ってきた。また、近年では同時期の資料を対象にした『近代女性雑誌コーパス』の作成に携わっている。これらのデータは文献学的な処理を行い、文体や引用情報なども付与された本格的な電子化言語資料であるが、形態素解析が行えないために、文字列検索での利用にとどまっている。また、近代語のテキストは著作権の問題が比較的少ないため「青空文庫」などインターネットのサイトで公開されているデータが多いが、これも十分に言語研究に活用されているとはいえない。これらのデータに形態素解析を施すことができるようになれば、近代日本語の記述的研究をより高精度に行うことが可能になる。さらに、形態素解析された現代語のコーパスと組み合わせて統一的に用いることで、現代語の確立過程を明らかにすることにもつながり、日本語史研究の面でも現代語研究の面でも大きな意味を持つことになる。

２．研究の目的

本研究の目的は、このような近現代語の通時的コーパスを構築するために、近代語のデータを十分な精度で解析できる形態素解析システムのための近代語電子化辞書を作成し、多くの日本語研究者に利用しやすい形で公開することである。

そして、この電子化辞書による解析結果を利用して近代語の記述的研究を行い、新しい研究手法を紹介することにより、日本語学、特に日本語史の分野におけるコーパス言語学の手法の普及・発展を図る。

３．研究の方法

本研究では、実用に耐える近代語の形態素解析のための電子化辞書を作成した。この辞書には『現代日本語書き言葉均衡コーパス』の構築に用いられている形態素解析辞書「UniDic」の枠組みを採用し、UniDic コンソーシアム（千葉大学・伝康晴代表）の協力のもと、現代語の UniDic に見出し語を補ってゆく形で行った。UniDic は、語彙素・語形・書字形・発音形に階層化され、斉一な解析単位をもつ、言語研究に適した辞書である。この現代語の UniDic をもとに、現代語と同じ「短単位」で近代語の見出し語を追加していくことにより、現代語と近代語の語彙を比較することが可能になった。

この解析システムは、近代語の文章のなかでも明治普通文とよばれる標準的な文語論説文を主要な対象としている。これは、文語論説文が質的に一定のまとまりがあるため解析対象として目標にしやすいこと、十分な量のデータが残っており今後の応用範囲が広いと考えられることからである。

見出し語の追加に際しては、『太陽コーパス』の開発時に作成された「スカウト式用例採集データ」を利用した。これは『太陽コーパス』に採録された年の雑誌『太陽』本文から、採集者が注目した用例をピックアップして作成したものである。このほか、『太陽コーパス』『近代女性雑誌コーパス』、『青空文庫』所収の近代文語文の作品、『文明論之概略』をはじめとする近代資料の解析結果から未登録語を 5 万語以上追加した。

解析システムの構築に必要な学習用コーパスについては、第２回 博報「ことばと文化・教育」研究助成を得て作成したデータを含む約 38 万語分を利用している（表 1)。

表１：学習用コーパス

資料名		語数（万語）
太陽コーパス	(1901 年)	7.42
近代女性雑誌	(1894 年)	1.08
文明論之概略	(福澤諭吉)	4.28
青空文庫		4.86
	(山路愛山)	2.43
	(北村透谷)	4.42
	(陸羯南)	3.20
	(その他)	4.80
法令・公文書		5.21
近代詩		0.18
計		37.87

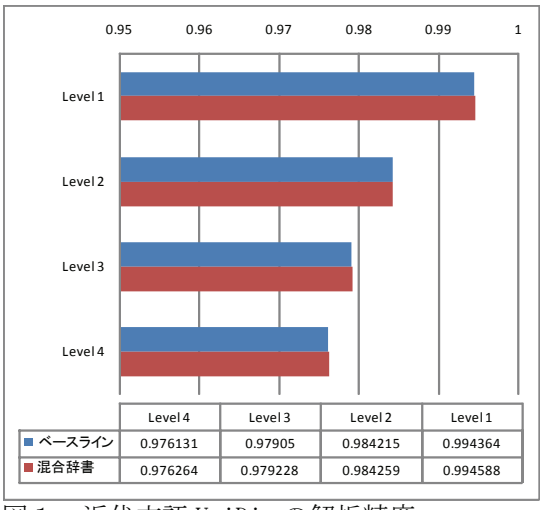
38 万語のデータは、現代語の形態素解析辞書の作成に利用される学習用コーパスと比較すると、充分な量であるとは言えない。そこで、この不足を補うために、現代語と近代語の語彙の共通性に着目し、「現代日本語書き言葉均衡コーパス」のコアデータ（人手修正済みデータ）を用いることによって、解析

精度の改善を図った。現代語のデータを単純に追加するだけでは精度が悪化するため、一部の品詞の生起コストに限って、近代語の生起コストを適切な割合で混合することにより、精度を向上させることを得た。

このような方法により、形態素解析辞書「近代文語 UniDic」を開発し、近代文語文の高精度な解析を実現した。

４．研究成果

図１に、近代語コーパスのみを使った辞書と、現代語コーパスのコストを利用した辞書の解析精度を示す（小木曽ほか 2009）。図の Level 1 は単語の境界の認定、Level 2 は品詞の認定、Level 3 は語彙素の認定（国語辞典の見出しに相当。「金」を「キン」であるか「カネ」であるか区別する）、Level 4 は発音形の認定の精度である。一般に形態素解析の精度比較で用いる品詞認定では 98.4%以上、語彙素の認定でも 98%に近い数字となっている。



	Level 4	Level 3	Level 2	Level 1
ベースライン	0.976131	0.97905	0.984215	0.994364
混合辞書	0.976264	0.979228	0.984259	0.994588

図１：近代文語 UniDic の解析精度

図１に示したのは、未知語（テキスト中に現れる辞書に登録されていない見出し語）が存在しない場合の解析精度であるが、未知語が存在する一般的な文章であっても充分な精度を示す。公開版の近代文語 UniDic の解析精度は、一般的な近代文語文であれば概ね 96～98%程度となっている。

96%以上の解析精度は、現代語の形態素解析辞書の解析精度とほぼ同等であり、日本語研究においても充分に実用的な精度であるといえる。テーマを選べば解析結果をそのまま利用することも可能であり、解析結果を修正して利用する場合にも大きな負担なく利用することができる。

このようにして開発した形態素解析辞書「近代文語 UniDic」と、既存の形態素解析プログラム「ChaSen※1」、「MeCab※2」を組み合わせて近代文語の解析システムとしてまとめ上げた。そのうえで、日本語研究者に利用しやすいように Windows 用のインストーラと解析用 GUI（「近代茶まめ」）、ユーザーズマニュアルを整備してインターネット上で一般公開した。

※1 茶筌（奈良先端科学技術大学院大学松本研究室）
http://chasen-legacy.sourceforge.jp/

※2 MeCab（工藤拓）
http://mecab.sourceforge.net/

「近代茶まめ」はコンピュータに不慣れな文系研究者でも利用可能なようにマウスクリックによる操作だけで形態素解析を行うことを可能にした解析補助用プログラムである（図２）。解析前処理として、踊り字を仮名に変換したり、漢字カタカナ交じり文を漢字ひらがな交じり文に変換したりする機能も備えている。

図２：近代文語 UniDic 利用画面（近代茶まめ）

一方、この辞書の活用の面では、『現代日本語書き言葉均衡コーパス』（モニター版）の書籍（現代語）と、『太陽コーパス』の文語記事（近代文語）を解析した結果をデータベースに取り込んで比較し、主に語種の変化に関する研究を行った。

図３・図４は、近代語現代語それぞれの語種比率を比較したものである。

近代語では漢語が特に活発に用いられているのに対し、現代語では和語とともに外来語が活発に用いられていること，延べと異なりにおける割合の違いから、和語は高頻度語が多く、漢語・外来語は比較的低頻度語が多いことなどがわかる。また、漢語の延べと異なりの差が現代語よりも近代語において大きいことから、現代語において漢語の高頻度語が多くなっていることが示唆される。このことは度数段階別に語種比率を集計することによって確認される（小木曽・近藤 2009）。

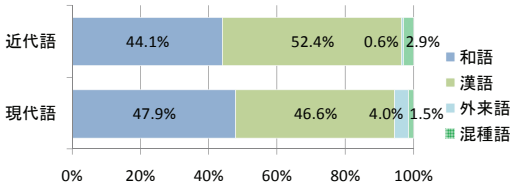


図３：近現代語の語種比率（延べ語数）

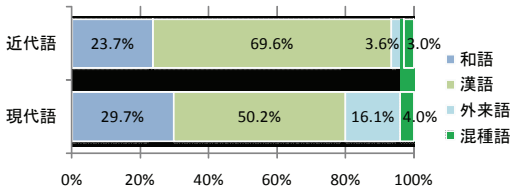


図４：近現代語の語種比率（異なり語数）

形態素解析辞書「近代文語 UniDic」（プログラム）は、下記のホームページにて、解析例や学会発表資料などとともに一般に公開している（図５）。

http://www.kokken.go.jp/lrc/index.php?UniDic

図５：近代文語 UniDic ホームページ

これまで日本語史の分野では形態素解析は利用されてこなかったが、本研究の成果をもとに、形態素解析とそれを用いて作られるコーパスを利用した新しい日本語研究が行われることを期待したい。

このほか、開発に関連して１件の雑誌論文を執筆したほか、デモンストレーション１件を含む４件の学会発表を行ったほか、応用研究の成果を含む研究成果報告書（全 233 ページ）を作成・公開した。

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　１件）

①小木曽智信「形態論情報の自動付与とその問題点」『国文学 解釈と鑑賞』74 巻 19 号，査読なし, pp. 35-43

〔学会発表〕（計　４件）

①「現代語コーパスの利用による近代語形態素解析の精度向上」小木曽智信・伝康晴・渡部涼子・近藤明日子 『言語処理学会第 15 回年次大会発表論文集』pp. 801-804（2009 年 3 月 5 日・於鳥取大学）

②「語種を観点とした近代語と現代語の語彙の比較―形態素解析辞書「近代文語 UniDic」「UniDic」を用いて―」近藤明日子・小木曽智信『言語処理学会第 15 回年次大会発表論文集』pp. 741-744 2009 年 3 月（2009 年 3 月 5 日・於鳥取大学）

③デモンストレーション「近代文語文を対象とした形態素解析辞書・近代文語 UniDic」小木曽智信・小椋秀樹・近藤明日子 『日本語学会 2008 年度春季大会予稿集』pp. 211-218 （2008 年 5 月 18 日・於日本大学）

④「近代文語文を対象とした形態素解析辞書の開発」小木曽智信・小椋秀樹・近藤明日子 『言語処理学会第 14 回年次大会発表論文集』 pp. 225-228 （2007 年 11 月 17 日・於沖縄国際大学）

〔図書〕（計　１件）

①研究成果報告書『近代文語文を対象とした形態素解析のための電子化辞書の作成とその活用』（小木曽智信），全 233 ページ，2009 年 3 月

〔その他〕

ホームページ等

http://www.kokken.go.jp/lrc/index.php?UniDic

（プログラム「近代文語 UniDic」 Ver. 1.0 を公開中）

６．研究組織

(1)研究代表者

小木曽 智信（OGISO Toshinobu）

独立行政法人国立国語研究所・研究開発部門・研究員

研究者番号：20337489