〈受賞紹介〉　中古和文を対象とした形態素解析辞書の開発

著者	小木曽 智信
雑誌名	国語研プロジェクトレビュー
号	7
ページ	31-34
発行年	2012-02
URL	http://doi.org/10.15084/00000692

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
National Institute for Japanese Language and Linguistics

国語研プロジェクトレビュー　NINJAL Project Review
No. 7 pp. 31–34 (February 2012)

〈受賞紹介〉

情報処理学会では，研究会やシンポジウムで発表された論文の中から特に優れた論文に対し，「山下記念研究賞」を授与しています。小木曽氏の論文は，中古（平安時代）の文章を高い精度で単語に分割できる電子化辞書「中古和文 UniDic」を構築した点が高く評価され，平成 22 年度の「山下記念研究賞」を受賞しました。

受賞論文　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: 1–8.

中古和文を対象とした形態素解析辞書の開発

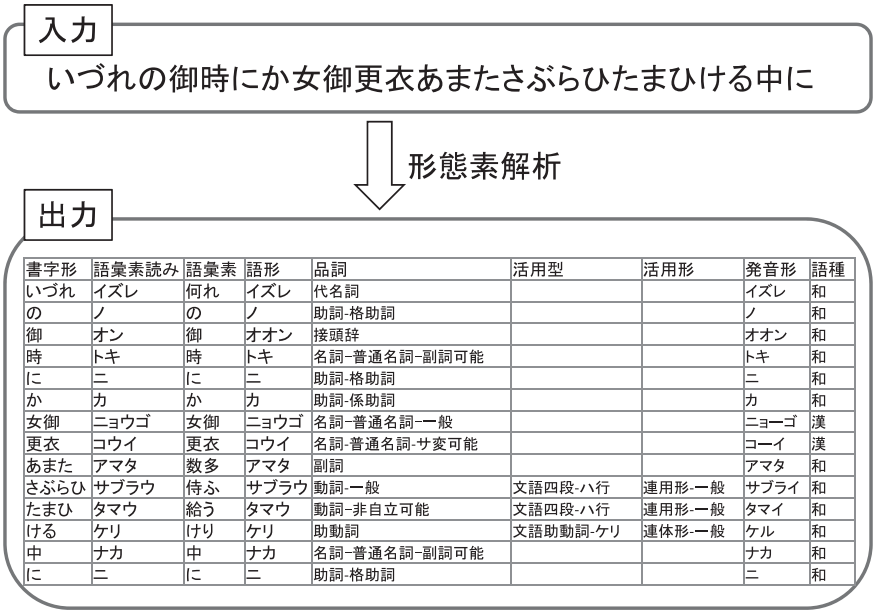
小木曽智信

国立国語研究所　言語資源研究系　准教授

1. 研究の背景

国立国語研究所コーパス開発センターでは日本語研究に役立つ「コーパス」の開発を行っています。コーパスとはコンピュータで利用可能な大規模な言語データベースのことです。先頃完成した『現代日本語書き言葉均衡コーパス』には 1 億語分以上の文章が収録されています。その全ての文章について，単語の切れ目・読み・品詞などの形態論情報が付けられており，これによって高度な検索や集計を行うことができます。1 億語もの単語に読みや品詞の情報を付けるために「形態素解析」と呼ばれる自然言語処理技術を活用し，コンピュータによる処理を行っています（図 1）。

センターでは，現代語のコーパスの完成を受けて，新たに「通時コーパス」の構築を計画しています。これは奈良・平安時代から江戸時代までの様々な日本語の文章をコーパス化しようというものです。通時コーパスの構築のためには，古文についても形態素解析を可能にする必要があります。すでに現代語については高い精度で形態素解析を行うことが可能になっていましたが，古文については十分なものがありませんでした。そこで，新たに古文を解析するための形態素解析の実現に取り組んだのが今回の研究です。一口に古文と言っても非常に多様なテキストがありますが，その中でも最も代表的な平安時代の仮名文学作品（中古和文）を対象としています。



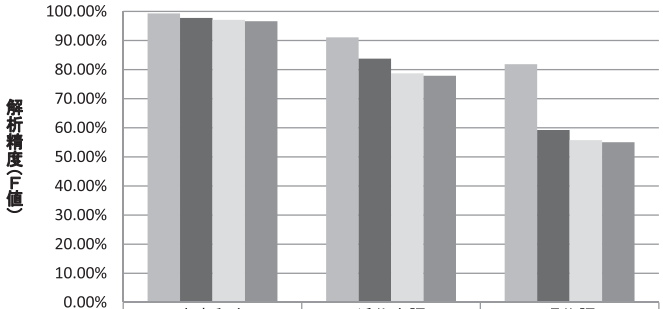
書字形	語彙素読み	語彙素	語形	品詞	活用型	活用形	発音形	語種
いづれ	イズレ	何れ	イズレ	代名詞			イズレ	和
の	ノ	の	ノ	助詞-格助詞			ノ	和
御	オン	御	オオン	接頭辞			オオン	和
時	トキ	時	トキ	名詞-普通名詞-副詞可能			トキ	和
に	ニ	に	ニ	助詞-格助詞			ニ	和
か	カ	か	カ	助詞-係助詞			カ	和
女御	ニョウゴ	女御	ニョウゴ	名詞-普通名詞-一般			ニョーゴ	漢
更衣	コウイ	更衣	コウイ	名詞-普通名詞-サ変可能			コーイ	漢
あまた	アマタ	数多	アマタ	副詞			アマタ	和
さぶらひ	サブラウ	侍ふ	サブラウ	動詞-一般	文語四段-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

図 1　形態素解析のイメージ

2. 研究の成果

この研究では，源氏物語や伊勢物語のような仮名文学作品を研究利用に堪える高い精度で解析できるようにすることを目標にしました。形態素解析のプログラムは一般に公開されているフリーソフトの「MeCab」を利用し，新たに中古和文専用の MeCab 用の辞書「中古和文 UniDic」を開発しました。そのために，現代語コーパス用に開発された「UniDic」と，これを元に開発を行っていた明治期の文語論説文向けの「近代文語 UniDic」を基礎として，古文用の見出し語や活用表を整備しました。そして解析の手本となる約 6 万語（最新版では約 27 万語）分の機械学習用データを作成して，形態素解析用の辞書を作成しました。

その結果，従来の形態素解析辞書では歯が立たなかった中古和文のテキストを十分に高い精度で解析することが可能になりました。図 2 は，文単位でランダムサンプリングした中古語のテキストを，中古和文用・近代文語用・現代語用の辞書でそれぞれ解析してその精度を評価した結果です。現代語用では 50 ～ 60% という全く実用にならない精度しか出ていなかったものが，中古和文 UniDic では 97%以上という高い精度で解析できるようになっています。現代語用の辞書で現代語のテキストを形態素解析した場合でもおよそ 98%程度の精度ですから，それと比べても遜色ないレベルに達しています。



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

図 2　中古和文の形態素解析精度

この形態素解析辞書「中古和文 UniDic」は国立国語研究所のホームページで一般公開を行っており，誰でもダウンロードして利用することができます。形態素解析を多くの人文系の研究者に使ってもらうために，パソコンに簡単にインストールできるようにしたほか，マウス操作だけで容易に形態素解析を行うことのできるプログラム「茶まめ」を開発して同梱しています（図 3）。

図 3　形態素解析補助ツール「茶まめ」実行画面

3. 今後の展望

研究用として実用になる形態素解析辞書を作るという所期の目的はおおよそ達成しましたが，見出し語の充実や精度向上など，まだまだ改善の余地があります。また，先にふれた通時コーパスの構築のためには，中古和文だけでなく過去の様々な種類の文体を解析できるようにしていく必要があります。特に，中世の軍記物や近世の戯作などは現在の辞書では歯が立たない状況なので，これら各時代のテキストにあわせた辞書を用意して形態素解析が利用できる資料の幅を広げてゆく予定です。

同時に，形態素解析結果を研究に活用できる環境を整備し，コーパスと統計的手法を活用した新しい方法による研究を進めていく必要があります。2010 年 10 月より開始した国立国語研究所共同研究プロジェクト「統計と機械学習による日本語史研究」などの場でこうした研究に取り組んでいます。

参照 Web サイト

MeCab　http://mecab.sourceforge.net/
形態素解析辞書 UniDic　http://download.unidic.org/
中古和文 UniDic　http://www2.ninjal.ac.jp/lrc/index.php?UniDic

小木曽智信（おぎそ・としのぶ）

国立国語研究所言語資源研究系准教授。修士（文学）。東京大学大学院人文社会系研究科博士課程単位取得満期退学。明海大学専任講師，独立行政法人国立国語研究所研究員を経て 2009 年 10 月より現職。コーパス開発センター兼任。

主な著書・論文：『講座 IT と日本語研究 6　コーパスとしてのウェブ』（共著，明治書院，2011），『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』（共著，博文館新社，2005），「明治大正期における補助動詞「去る」について」（『近代語研究』15，2010）．