情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2011-CH-92 No.6
2011/10/8

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

小木曽智信†

国立国語研究所で計画している「通時コーパス」の構築には，様々な時代・文体のテキストに対する形態素解析を実現することが必要となる．本発表ではその基礎となる各種の歴史的資料を対象としたUniDicについて報告する．

Development of Dictionaries for Morphological Analysis of Pre-Modern Japanese Aiming at Construction of the Diachronic Corpus

Toshinobu Ogiso†

National Institute for Japanese Language and Linguistics is planning to construct the Diachronic Corpus of Japanese. In order to compile the diachronic corpus, it is necessary to implement morphological analysis of various texts in different times and writing styles. In this paper, I report the UniDic dictionaries for various historical Japanese texts.

1. はじめに

国立国語研究所では日本語の通時コーパスの構築を目指してその準備を行っている．本格的な日本語コーパスの構築には形態素解析が必要になるが，通時コーパスが対象とする過去の日本語テキストは一般的な形態素解析辞書では解析できない．そのため，各時代・文体に対応した形態素解析辞書が必要となる．本発表では，通時コーパスの構築に向けて開発を行っている，歴史的資料を対象とした形態素解析辞書・UniDicシリーズについて報告し，今後の拡張計画について述べる．

2. 日本語コーパスと形態素解析

2.1 コーパスと形態素解析

現代日本語のコーパスは，諸外国語に比較してその整備に後れを取り，近年になってようやく完成を見たところである．この後れには様々な理由が考えられるが，日本語の形態素解析が可能になるまでに時間を要したこともその原因の一つである．

高度な言語研究のための本格的なコーパスでは，すべてのテキストについて単語ごとに品詞や読みなどの用法を付けることが求められる．大規模なコーパス構築において，この作業は人手ではとうてい対処できない量であり，コンピュータによる自動処理が必須となる．また，量だけでなく質の面からも，均質なタグ付けを行うために自動処理が必要である．

さらに，日本語では分かち書きがなされないうえに，漢字仮名交じりの複雑な表記法によるため，プレーンテキストでは対処できない多くの問題があり，その解決のためにも形態素解析が必要とされる．たとえば，同じ語が漢字・ひらがな・カタカナによって様々に表記されるため，送り仮名の異同も含めるとおびただしい数の異表記形が存在するが，これは語彙調査にあたって大きな問題となる．逆に，仮名書きされる場合などには別語が同一に表記される場合も少なくないため，その曖昧性の解消が必要となる．

自然言語処理技術の発達により日本語の形態素解析が高い精度で行えるようになったのは1990年代後半のことである．現代日本語のコーパス構築が2000年代になって行われたのはこうした理由を含んでいる．

2.2 形態素解析辞書 UniDic

1990年代後半から，ChaSen，MeCabなどの形態素解析器がフリーソフトとして公開され広く使われてきた．これらの解析器はIPADICという辞書とともに用いられて

† 国立国語研究所
National Institute for Japanese Language and Linguistics

いたが，この辞書は，1.もっぱら新聞記事のデータを元に作られているため他ジャンルのテキストでは解析精度が低下する 2．解析単位（語）の認定基準が必ずしも明確でない 3．異表記であれば別語と見なされるためそのままでは語彙調査に利用できない といった難点があった．

そこで，「現代日本語書き言葉均衡コーパス」（BCCWJ）の構築にあたり，国立国語研究所が中心となって，言語研究に適した新しい電子化辞書 UniDic が開発されることとなった（伝ほか 2007）．UniDic は ChaSen または MeCab と組み合わせて利用する形態素解析用の辞書である．UniDic の特長として次の点が挙げられる．

1.「短単位」という揺れが少ない斉一な単位を見出し語に採用している．
2．語彙素・語形・書字形・発音形の階層構造を持ち，表記の揺れや語形の変異にかかわらず同一の見出しを与えることができる．
3．話し言葉のテキストの解析に対応しているほか，アクセントや音変化の情報を付与することができ，音声処理の研究に利用することができる．
4．BCCWJ に納められた多様なジャンルのテキストが高い精度で解析できる
5．語種など日本語研究に有用な多くの情報が付与できる

このような特長により，先に挙げたような従来の辞書が抱えていた問題を解決している．「短単位」については単位の認定方法を規程集に詳細に定めており，揺れを防いでいる（小椋・小潮ほか 2011）．最新版の UniDic では，現代語の様々なジャンルのテキストを 98%以上の精度で解析することが可能になっている．

2.3 UniDic の古文への応用

上記の UniDic は種々の文体に対応しているとはいえ，あくまでも現代語用の形態素解析辞書であり，そのままでは古文を解析することはできない．しかし，この辞書を元にして古文用の見出し語を追加し，学習用のコーパスを準備することにより，古文用の形態素解析辞書を作成することは可能である．UniDic がもつ斉一な単位や階層化された見出し構造は，古文の形態素解析辞書の作成時にもたいへん有用である．

古文の見出し語についても短単位を採用していくことにより，テキストの解析結果を用いた語彙比較が可能になる．同時代のテキスト間の比較ができるだけでなく，時代の違いをも超えて各種のテキストを相互に比較することか可能になる．

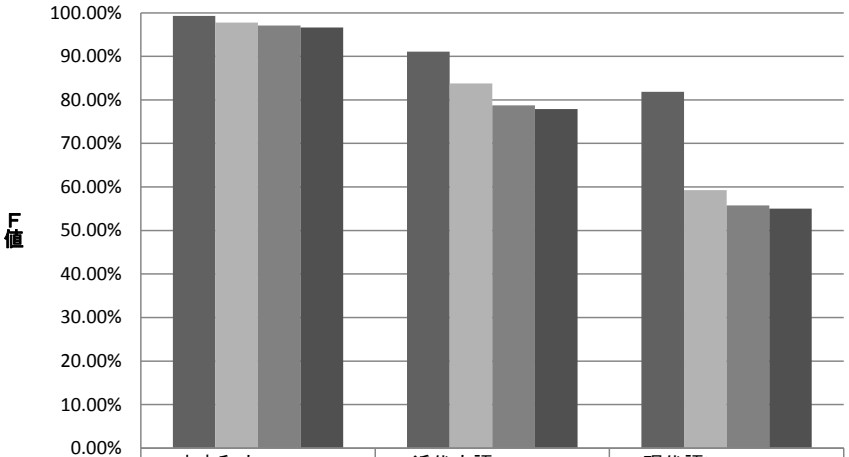
また，階層化された見出しを用いることで，文語形や旧字・旧仮名遣いの表記語を同一見出しの元にまとめることができるため，さまざまな時代のテキストに出現する語形・表記を統一的に扱うことができる．

このような目的で，発表者らは 2006 年から古文を対象とした形態素解析辞書の開発をはじめ，これまでに「近代文語 UniDic」と「中古和文 UniDic」の 2 種類の辞書を一般公開している．見出し語は現代語の UniDic の互換性に配慮して設計している（小椋・須永ほか 2011）．

2.4 近代文語 UniDic と中古和文 UniDic

現代語用の UniDic をもとにした，最初の古文用形態素解析辞書として「近代文語 UniDic」を開発，公開した（小木曽・小椋・近藤 2008）．これは主として近代の文語論説文（明治普通文）を対象とした解析辞書であり，文語の活用・旧仮名遣い・旧漢字などに対応し，文語文を正しく解析することが可能になっている．解析精度は，現代語版の UniDic には及ばないものの，おおむね 96%以上を達成している．これにより，「太陽コーパス」（国立国語研究所 2005）の文語記事など，近代文語文で書かれたテキストを解析して研究に利用することができるようになった．

これに続き，中古の仮名文学作品を中心とする和文系資料を対象とした「中古和文 UniDic」を開発した（小木曽ほか 2010）．現在，おおむね 97%以上の精度で解析可能になっている．同じ古文といっても近代文語 UniDic が対象とするものとは大きく異なる文体であるため，専用の辞書を用いない場合には大きく解析精度が下がる．図 1 は同一の中古和文のテキストを各種の UniDic で解析したときの精度である．通時コーパスのように多様なテキストを解析する場合にはテキストにあった辞書を利用する必要がある．



	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 1　中古和文データの解析精度の比較（形態素解析辞書別）

Vol.2011-CH-92 No.6
2011/10/8

3. 通時コーパスと UniDic 拡張計画

3.1 通時コーパスとテキストの多様性

国語研究所で計画中の通時コーパスには，小学館『新編日本古典文学全集』所収のテキストをはじめとする，多様なテキストが含まれる予定である．すでに『新編全集』の次のテキストは入手済みであり，整備をはじめている．

古今和歌集，竹取物語・伊勢物語・大和物語・平中物語，土佐日記・蜻蛉日記，源氏物語，枕草子，和泉式部日記・紫式部日記・更級日記・讃岐典侍日記，落窪物語・堤中納言物語，今昔物語集，平家物語，近松門左衛門集

このほかに近世の洒落本等のテキストも対象としていく予定となっている．

これらのテキストの文体は，文法・語彙・表記にわたって極めて多様であって，単に「古文」としてひとくくりにできるものではない．表 1 に上記のテキストと，近代の太陽コーパスのテキストの例を挙げる．

表 1　通時コーパスに含まれるテキストの例

A	源氏物語 1（夕顔）	白き袷、薄色のなよよかなるを重ねて、はなやかならぬ姿いとらうたげにあえかなる心地して、そこととりたててすぐれたることもなけれど、細やかにたをたをとして、ものうち言ひたるけはひあな心苦しと、ただいとらうたく見ゆ。
B	更級日記	つとめて舟に車かき据ゑて渡して、あなたの岸に車ひきたてて、送りに来つる人々これよりみなかへりぬ。上るはとまりなどして、行き別るるほど、行くもとまるも、みな泣きなどす。幼心地にもあはれに見ゆ。
C	今昔物語集 1（高野姫天皇造西大寺語第十八）	今昔、高野姫天皇ハ聖武天皇ノ御娘ニ御マス。女ノ身ニ御マスト云ヘドモ、心ニ智リ広クシテ文ノ道ヲ極メ給タリケリ。亦、法ノ道ヲ知テ、「何カデ道場ヲ建立セム」ト思食ケル。未ダ位ニモ不即給シテ姫宮ニテマシマシケル時ニ
D	平家物語 1（源氏揃）	蔵人衛門権佐定長、今度の御即位に、違乱なくめでたき様を、厚紙十枚ばかりにこま〴〵と記いて、入道相国の北の方、八条の二位殿へ参らせたりければ、ゑみをふくんでぞよろこばれける。かやうにはなやかにめでたき事どもありしかども、世間は猶しづかならず。
E	近松門左衛門集 1（五十年忌歌念仏）	実ぢや／＼と言ひ中ければ．詞それが定なら、晩に寝所へござんすか．地色オヽなるほど／＼、ハル忝い．それについて、今ちよつと問ふことありと言ひけれども．それも寝所で色しつぽりとハル聞きませう。かならず欺しにさんすなえ．
F	洒落本（月花余情）	佐右衛門こりや。喜八ニさいぜんのくもわたを。ちよつとすましにして上ケませいといふてこひ。扨。マア。あがつてごろふじませ。むかいの京升屋が。此間京へ。仕かへものについてゆかれましてその土産に。若狭のたらの。雲わたを。けふもらひまして。御座ります。きつと。じまんで上ケます。客そりやよかろ。
G	太陽コーパス（1895 年 11 号・狂言娘）	少時は泣きしが、忽ち又顔を上げて、屹度見し目には冷笑を含みぬ。『家爺さんが心配してお出でだツて。餓死んだら如何するツて、ほゝゝ、家爺さんも口ばツかしさ。私が死んだら喜ぶだらう、お玉は喜ぶだらうね。
H	太陽コーパス（1895 年 03 号・文学上の新事業）	我が社會の事、不整頓なるもの少なからぬが中に、文學の如きは、恐くはその最も亂雜なるものゝ一ならん。我が文界は今尚ほ過渡の時代にありと云はざるべからず。
I	太陽コーパス（1925 年 03 号・長篇科學小説　生ける死『第三回』）	『『夢ぢやないか？』とトムスンは云ひました。『いや、夢ぢやない、抓つてみたら痛かつたから。』 『極樂だ。』とハムデンは呟きました。 『さうですよ、リヴイングストン大佐、極樂を發見したんですよ、あなたは――ああ、こんな凍つた大陸の眞中で！』
J	太陽コーパス（1925 年 02 号・歴代の総理大臣（二））	黒田内閣は、大隈外相が實權を握り、攻撃も之に集つた。前の順序で、大隈が大久保の後を繼ぐべきであり、薩長聯合で排斥せられたのが、恰も大久保が大隈を用ゐた如く、黒田が之を用ゐようとし、大久保程押が利かなかつたのである。

3.2 テキストの多様性と UniDic の対応

このような種々のテキストに形態素解析を施す場合，どのような形態素解析辞書をどれだけ開発する必要があるだろうか．

テキストに大きな違いがある以上，個々のテキストに最適の辞書を作成することができれば望ましいが，少量のテキストのために個別に辞書を作成することは現実的ではない（その手間で当該テキストをすべて人手で整備できてしまう）．したがって，文体的に近いテキストをグループにまとめ上げ，グループごとに適した形態素解析辞書を用意することが適切であると考えられる．

図２は，試みに，各時代の日本語の文体についてごく大まかにまとめたものである．図中の①現代語と②近代文語（表１のH），③和文（表１のＡ・Ｂ）については，すでにそのための UniDic を開発・公開を行った．③は後の時代にも擬古文・雅文といった文語として使われ続けるが，後世のものも同一の辞書で十分に解析が可能である．近代口語④（旧仮名遣いの口語文，表１のＩ・Ｊ）についても，現代語の UniDic を元に旧仮名遣いの見出し語を加え，ほぼ十分な精度で解析できる辞書をすでに作成している．

以上①②③の文体（図２で実線のボックスで示したもの）については，既存の辞書の拡充を行いながら，通時コーパスのテキストに対応していく．

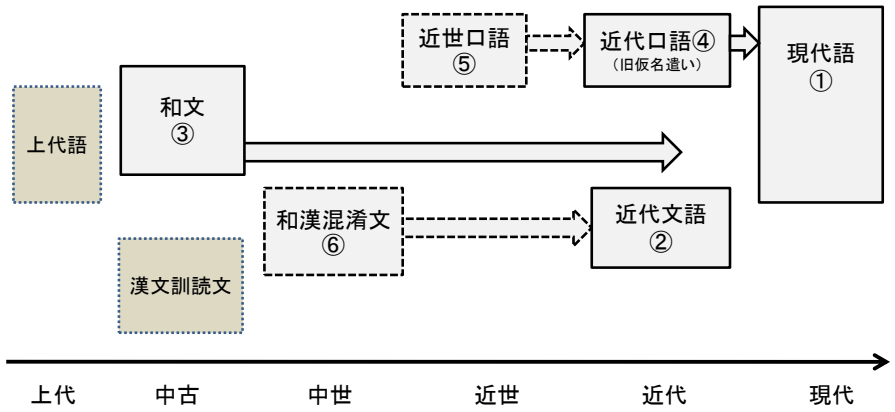


図 2 各時代の資料・文体

中古末から現れる和漢混淆文（表１のＣ・Ｄ，図２中の⑥）は，漢語を多く含み漢文訓読由来の語法が目立つ点では近代文語に近い．しかし近代文語文は文法が簡略化され固定的な言い回しが多くなっている上に，両者には語彙的にも非常に大きな違いがあるため，別途辞書を作成する必要があると思われる．和漢混淆文の内部での差異も大きいが，テキスト量の観点からも一つのグループとしてまとめ，新しく和漢混淆文 UniDic を開発する計画である．なお，今昔物語集などは漢字カタカナ交じり文で書かれているうえに返読を要する文字列も含まれるが，こうした部分は漢字ひらがな交じりの通常のテキストに変換した上で，形態素解析を行う予定である．

近世の口語文（表１のＥ・Ｆ，図２中の⑤）は，これまでに開発した UniDic では全く精度が出ない文体である．今後，近世口語 UniDic として新たな辞書を整備していく予定である．近世前期の上方語と，後期の江戸語とで別の辞書を用意することも考えられる．なお，表１のＧに見るように，近代の小説の会話文にもこれに近いものが多いほか，狂言の文体とも比較的近いことから，この辞書を元にしてより広い範囲のテキストに応用することができるのではないかと考えられる．

なお，図２にも示した上代語・漢文訓読文については，通時コーパスへの収録について未確定であることから今回の検討の対象外とした．

3.3 UniDic の分野適応

上述のように多数の辞書を作成する場合，学習用コーパスを一から整備するのでは手間がかかりすぎるという問題がある．そこで，既存の「近代文語 UniDic」「中古和文 UniDic」を活用し最小限の学習用コーパスで新しい辞書を用意することを予定している．現代語のジャンル別のデータを用いた実験で，ターゲットの文体に合わせた少量の学習用コーパスを汎用の辞書と組み合わせることにより，ターゲットを高い精度で解析することが可能になることがわかっている（小木曽ほか 2009）．こうした方法を活用することで多様なジャンルのテキストに適合した辞書を作成していく．

3.4 適切な辞書の自動選択

表 1 の G に典型的に見られように，近世・近代のテキストでは，地の文と会話文で文体が大きく違う場合が少なくない．この場合には，地の文と会話文とで利用する形態素解析辞書を切り替えることにより全体の解析精度を向上させることが可能である．コーパスに付けられたタグを活用し適切な辞書を文単位で選択することで精度向上を図る予定である．

4. おわりに

国語研究所で計画中の通時コーパスに合わせて進めている UniDic の整備計画について紹介した．今後も様々な形で通時コーパス構築のための基盤となる技術の開発と言語資源の整備を行っていきたいと考えている．

参考文献

1) 伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「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
2) 小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・原裕（2011）『『現代日本語書き言葉均衡コーパス』形態論情報規程集第４版』国立国語研究所　特定領域研究「日本語コーパス」平成２２年度研究成果報告書
3) 国立国語研究所（2005）『太陽コーパス―雑誌「太陽」日本語データベース』博文館新社
4) 小木曽智信・小椋秀樹・近藤明日子（2008）「近代文語文を対象とした形態素解析辞書・近代文語 UniDic」『言語処理学会第 14 回年次大会予稿集』pp.225-228
5) 小木曽智信（2009）『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』国立国語研究所・科研費報告書 19720110
6) 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」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, pp.1-8
7) 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 回年次大会発表論文集』pp.312-315
8) 小木曽智信・伝康晴・渡部涼子（2009）「ジャンル別 UniDic 作成の試み」特定領域研究「日本語コーパス」平成 20 年度公開ワークショップ（研究成果報告会）予稿集. pp. 17-22
9) 「形態素解析辞書 UniDic」　http://download.unidic.org
10) 「近代文語 UniDic」「中古和文 UniDic」http://www2.ninjal.ac.jp/lrc/index.php?UniDic