近代語テキストの形態素解析

近代語テキストの形態素解析

小木曽　智信　（国立国語研究所言語資源研究系）[1]

１．はじめに

国立国語研究所により 2005 年に公開された『太陽コーパス』（2005）は単語情報を含まず、文書構造や注記等をマークアップしただけのコーパスだった。一方、同じ年に公開された現代語の『日本語話し言葉コーパス』（CSJ）や 2011 年に公開された『現代日本語書き言葉均衡コーパス』（BCCWJ）では、単語の読みや品詞などの形態論情報が付与されている。この形態論情報を用いることで、活用形や表記の違いにとらわれず語としての検索や集計が可能となり、語がもつ情報を組み合わせた高度な処理も行うことができる。

『太陽コーパス』に単語情報が付与されていないのは、当時の技術では、現代語と大きく異なる近代語のテキストに形態素解析を施すことが困難であったことによる。しかし、その後「近代文語 UniDic」が整備されたことにより、近代語のテキストであっても実用的な精度で形態素解析を行うことが可能になってきた。これにより、新たに構築される近代語コーパスでは、BCCWJ と同様の単語情報付きのコーパスとすることができる。

本稿では、新たな近代語コーパスの試作データである『明六雑誌コーパス』における処理を例に、近代語テキストの形態素解析について述べる。

２．近代語の形態素解析

２．１　日本語の形態素解析

日本語の形態素解析は 1990 年代以降にコンピューターの処理性能の向上とともに技術開発が進展し本格的な利用が可能となった。今日では、形態素解析を行うプログラム（形態素解析器）として、京都大学言語メディア研究室の JUMAN（1992～）、奈良先端科学技術大学院大学松本研究室の茶筌〔ChaSen〕（1996～）、同研究室で生まれた和布蕪〔MeCab〕（2002～）、KyTea〔京都テキスト解析ツールキット〕（2009～）などが自由に利用可能なソフトウェアとして公開されている。形態素解析は、コンピューターによる日本語処理の基盤であり、インターネット上の多くのサービスなどで活用され、欠かすことのできない技術となっている。

CSJ や BCCWJ は、国立国語研究所が中心となり新たに開発した言語研究に適した形態素解析用の電子化辞書「UniDic」（伝ほか 2007）を用いてコーパス中のテキストの形態素解析を施した。BCCWJ では、MeCab と UniDic を用いて、およそ 98%の解析精度での形態論情報のアノテーションを実現している。

２．２　近代語の形態素解析

従来、形態素解析を行うことができるのは現代語の文章だけであり、文語文の形態素解析を行うことはできなかった。たとえば、既存の形態素解析辞書（ChaSen 標準の IPADIC 2.7.0）によって文語文を解析すると図１のような結果となる（例文「こゝに漢字の利害と題するは、即ち聊か袈裟の眞價を問はんとするなり。」『太陽コーパス』「漢字の利害」より）。現代語向けの辞書によるものであるから当然の結果ではあるが、多くの誤りがあり、この解析結果を研究に利用することはできない。近代語のテキストを解析するためには、近代語向けの形態素解析辞書を作成する必要があるのである。

[1] togiso@ninjal.ac.jp

IPADIC 2.7.0/ChaSen 2.4.2

出現形	読み	品詞	活用型	活用形
こ	コ	名詞-一般		
ゝ	ゝ	記号-一般		
に	ニ	助詞-格助詞-一般		
漢字	カンジ	名詞-一般		
の	ノ	助詞-連体化		
利害	リガイ	名詞-一般		
と	ト	助詞-並立助詞		
題	ダイ	名詞-一般		
する	スル	動詞-自立	サ変・スル	基本形
は	ハ	助詞-係助詞		
、	、	記号-読点		
即ち	スナワチ	副詞-一般		
聊か	イササカ	副詞-一般		
袈裟	ケサ	名詞-一般		
の	ノ	助詞-連体化		
眞	マコト	名詞-固有名詞-人名-名		
價		未知語		
を	ヲ	助詞-格助詞-一般		
問	トイ	名詞-一般		
はん	ハン	名詞-接尾-人名		
と	ト	助詞-格助詞-一般		
する	スル	動詞-自立	サ変・スル	基本形
なり	ナリ	名詞-一般		
。	。	記号-句点		

図 1 従来の形態素解析辞書による近代文語文の解析結果

３．近代文語 UniDic

一方、図 2 に示すのは近代語向けに新たに開発した形態素解析辞書「近代文語 UniDic」（小木曽ほか 2008, 2009）による解析結果である（近代文語 UniDic 1.2 と MeCab 0.99 で解析）。この結果からわかるように、文語の活用・歴史的仮名遣い・旧漢字・踊り字などに対応しており、文語文を正しく解析することが可能になっている。ここで、この「近代文語 UniDic」について説明する。

近代文語UniDic 1.2 /MeCab 0.99

出現形	発音形	代表形	代表表記	品詞	活用型	活用形	語種
こゝ	ココ	ココ	此処	代名詞			和
に	ニ	ニ	に	助詞-格助詞			和
漢字	カンジ	カンジ	漢字	名詞-普通名詞-一般			漢
の	ノ	ノ	の	助詞-格助詞			和
利害	リガイ	リガイ	利害	名詞-普通名詞-一般			漢
と	ト	ト	と	助詞-格助詞			和
題する	ダイスル	ダイスル	題する	動詞-一般	文語サ行変格	連体形-一般	混
は	ワ	ハ	は	助詞-係助詞			和
、			、	補助記号-読点			記号
即ち	スナワチ	スナワチ	即ち	接続詞			和
聊か	イササカ	イササカ	些か	副詞			和
袈裟	ケサ	ケサ	袈裟	名詞-普通名詞-一般			外
の	ノ	ノ	の	助詞-格助詞			和
眞價	シンカ	シンカ	真価	名詞-普通名詞-一般			漢
を	オ	ヲ	を	助詞-格助詞			和
問は	トワ	トウ	問う	動詞-一般	文語四段-ハ行	未然形-一般	和
ん	ン	ム	む	助動詞	文語助動詞-ム	連体形-撥音便	和
と	ト	ト	と	助詞-格助詞			和
する	スル	スル	為る	動詞-一般	文語サ行変格	連体形-一般	和
なり	ナリ	ナリ	なり-断定	助動詞	文語助動詞-ナリ-断定	終止形-一般	和
。			。	補助記号-句点			記号

図 2 近代文語 UniDic による解析結果

３．１　近代文語 UniDic の作成

形態素解析を行うには、解析に用いる語（見出し語）のリストに、語の出現しやすさ（生起コスト）、語・品詞間のつながりやすさ（連接コスト）の情報を付けた形態素解析用の辞書が必要である。ChaSen や MeCab などの現在使われている主な形態素解析システムでは、生起コスト・連接コストを機械学習と呼ばれる方法によって統計的に取得する。その

ため、形態素解析辞書を新たに作成するには、解析に用いる語の一覧（辞書データ）と、その辞書の内容にあわせて文章に正しく情報を付与した手本となる文章のデータ（学習用コーパス）が必要となる。辞書データと学習用コーパスから、プログラム（学習器）によって形態素解析辞書が作られる（図 3）。なお、辞書データは活用表によって各活用形に展開できるようにしておく必要がある。

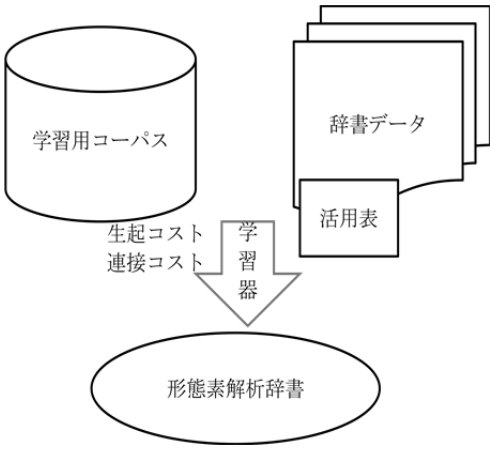

図 3 形態素解析辞書作成の流れ

３．１．１　辞書見出し語の整備

現代語とは異なるテキストを解析できるようにするためには、まず辞書データへの見出し語の追加が必要である。近代語用に追加が必要な見出し語としては、現代語では使われなくなった語、文語形、旧字・旧仮名遣いの形などさまざまなものがある。

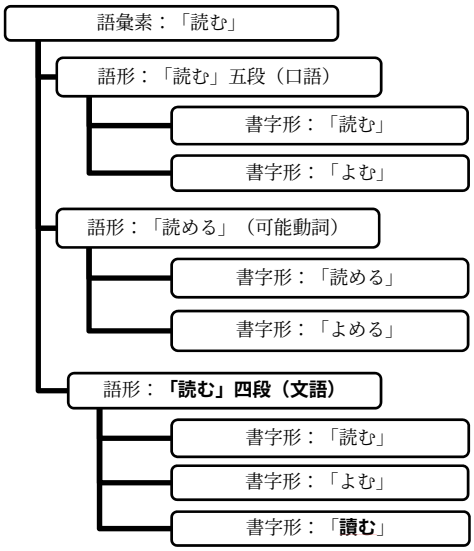

図 4 UniDic の階層（語彙素・語形・書字形）と文語形・旧字形

UniDic では見出し語を語彙素・語形・書字形・発音形の 4 段階で階層的に管理しているため、近代語解析に必要な語を各階層に整理して追加することができる。現代語としては使われなくなっている語は「語彙素」のレベルで、文語活用型の語は「語形」のレベルで、旧字形などは「書字形」のレベルで追加することになる（図 4）。これにより、現代語の語と統一的に管理することができるとともに、文語形と口語形、新字形と旧字形がそれぞれ関係を持つものであることを示すことができる。この方法により、近代語のテキストの

ためにおよそ数万語の見出し語を追加した。

この方式で、近代語のテキストのためにおよそ数万語の見出し語を辞書データベースに追加した。追加した見出し語は、当初は自動生成した文語形や旧字形を追加するところからはじめ、既存の辞書やデータ集の見出し語からも追加を行った。しかし、形態素解析辞書では、詳細な品詞や実際に現れる表記形を入力する必要があるため、単なる辞典の見出し語リストでは登録用のソースとして不十分な場合が少なくない。たとえば「名詞」といっても漢語サ変動詞の語幹としても使われるかどうかや、形容動詞の語幹や副詞としての用法を持つかどうかを区別する必要がある。また、表記の面では、辞典類の見出しに掲げられる代表的な表記ではなく、実際のテキストに現れる表記形を追加する必要がある。したがって、見出し語の追加にあたって最も効果的だったのは、実際に近代語のテキストを解析した結果から、未知語（見出し語（表記形）が形態素解析辞書にないために正しく解析されていない語）を見つけだして辞書データベースに登録することであった。

なお、辞書データでは、見出し語を追加登録してゆくとともに、活用語について活用表を整備して、必要な形を展開できるようにしておく必要がある。もともと UniDic は文語の活用型をもっていたが、近代文語 UniDic ではこれをさらに整備した。活用形の整備では、一般的な文語活用表にある活用形をそろえるほかに、特殊な表記に対応するための書字形を整備することも必要となる。たとえば、「讀て」（よみて）、「讀ず」（よまず）のように送り仮名が省略された表記等が多く用いられるためである。文法上の観点から作成される一般の活用表では問題とならないものだが、形態素解析辞書の活用表では表記上の違いについても活用表での対応が必要となる場合が少なくない。

３．１．２　学習用コーパスの整備

近代文語文の解析辞書を作るためには、辞書・活用表のほかに、機械学習を行うための学習用コーパスを整備する必要がある。近代文語文の解析のためには、辞書を拡充するとともに手本となる近代文語文の学習用コーパスを整備する必要がある。現在の近代文語 UniDic では表１に示したテキスト計約 46 万 6 千語を利用している。

表 1 近代文語 UniDic（1.2.1）の学習用コーパス

太陽	90604
女学雑誌	10802
文明論之概略	42800
法律	30868
青空文庫・論説	194364
青空文庫・小説	39294
文語詩	58377
総計	**467109**

３．１．３　テキストの解析前処理

近代語のテキストは、表記の上で、個々の語に揺れがあるにとどまらず、本文全体にわたって、仮名遣いの違い・漢字の新旧・踊り字使用の有無などのバリエーションがある。これらの問題に対処するためには、辞書に見出し（書字形）を追加して解析する方法と、あらかじめ本文の側を変換・修正してから解析する方法がある。近代文語 UniDic では、単純な置き換えが難しい仮名遣いや漢字の新旧については形態素解析辞書で対処した。一方、次の点については辞書での対応が困難であるため、解析の前に変換処理を行うことによって解析できるようにした。

漢字カタカナ交じり文

漢字カタカナ交じり文をそのまま解析できるようにするためには、仮名を含む書字形すべてについて、ひらがなとカタカナの二通りを用意する必要があり、現実的ではない。そこで、こうした本文については解析前に漢字ひらがな交じり文に変換したうえで解析することとした。「近代文語 UniDic」付属の解析用のアプリケーション「近代茶まめ」では、必要に応じて自動でカタカナをひらがなに変換させる機能を持たせている。

この処理では、漢字カタカナ交じりの文章中にカタカナとして残したい外来語等がカタカナで表れる場合、これのみを区別してカタカナのまま残すことはできない。したがって、完全な処理のためには人手による確認が必要になる。

濁点無表記

近代語のテキストでは濁点が表記されない場合も少なくないが、濁点無表記形を一々辞書登録していくことは無駄が大きい上に解析精度を低下させることにつながるため、これもあらかじめテキストを修正した後に解析を行うこととした。単純な変換処理は行えないため、原則として人手によって濁点付与を行うこととなる。

濁点付与作業を助け大量のテキストを処理するために、濁点の付与を自動で行うための研究とそのためのアプリケーション開発も行っている（岡ほか 2011）。

踊り字

「ゝ」「ゞ」などの仮名一字を単位とする踊り字については、これを含む一々の出現形を辞書登録するのではなく、解析前に踊り字を対応する文字に変換してから解析することとした。「近代茶まめ」では、この変換処理をボタン一つでできるようになっている。

しかし、くの字点（/ \）については繰り返される範囲が明瞭でないため自動変換は行わず、人手によるテキスト修正を経たのちに解析を行うか、またはそのままの形で解析することとした。そのままの形で解析される場合のために、くの字点は、「そろ / \」のように語の一部となっているものはその形を辞書に登録している。語や句を繰り返すものについては「/ \」全体を記号扱いの一単位として扱った。

なお、漢字を繰り返す「々」は今日でも「人々」のように用いられるため変換を行わず、その形を辞書登録している。しかし、近代語では「民主々義」のように語（短単位）の境界を跨いで繰り返される場合がある。近代文語 UniDic では、これらについて高い頻度で出現するものは辞書登録を行っているが、網羅的な対応は行っていない。また、漢字を繰り返す「〻」は「々」に置換している。

以上の解析前処理を完全な形で行うために、後述する『明六雑誌コーパス』の構築にあたっては、「漢字カタカナ交じり文中でカタカナをそのまま残す部分のアノテーション」や、「くの字点等の踊り字によって繰り返される範囲の明示」「濁点が期待される位置への濁点付与」の全ての作業を人手で行っている。修正を行った部分はすべてタグにより原文の状態を保持している。

３．２　解析精度

現在公開されている近代文語 UniDic（Ver.1.2.1）の解析精度は表 2（次ページ）に示す通りである。評価対象は、学習用のコーパスから約 10%を文単位でランダムサンプリングして学習対象から取り除いた人手修正済みのデータ 44587 語である。

表 2 で、「境界」とあるのは、最も基本的な評価基準で、解析結果において単語の境界が正しかったかどうかを意味する。「品詞」は境界が正しいことに加えて単語の品詞も正しく認定されていたかどうかを意味する。「語彙素」は境界と品詞に加えて語彙素（辞書見出し）としての認定も正しかったかどうかを意味する。たとえば「金」が「きん」でなく「かね」と正しく解析されているかどうかといった違いに相当する。「発音形」は、ここでは発

音というよりは語形の違いが正しく認定されているかどうかを評価するもので、境界・品詞・語彙素が正しいことに加え、さらに語形が正しいかどうかを意味する。たとえば、「言語」が文脈にあわせて「げんご」ではなく「ごんご」と正しく解析されているかどうかといった違いに相当する。表の右に行くほど評価基準が厳しくなっている。

表 2 近代文語 UniDic（1.2.1）の解析精度

	境界	品詞	語彙素	発音形
正解データ語数	44587			
出力語数	44573			
一致語数	44244	43594	43291	43162
再現率	99.23%	97.77%	97.09%	96.80%
適合率	99.26%	97.80%	97.12%	96.83%
F値	99.25%	97.79%	97.11%	96.82%

「正解データ語数」としたのは、評価データの語数である。評価データはあらかじめ人手による修正を経ているため、これが正解とみなされる。「出力語数」は形態素解析結果として出力されたデータの語数である。「一致語数」としたのは出力語数のうち評価データ（正解）に一致した語数である。たとえば、境界認定の場合、出力された 44573 語中、329 語は誤りだったことになる。

「適合率」「再現率」「F 値」は情報検索システムの性能評価でしばしば用いられる概念で、ここでは適合率（precision）は「一致語数／出力語数」（出力されたもののうちどれだけが正しかったか）に、再現率（recall）は「一致語数／正解データ語数」（正しいもののうちどれだけを出力できているか）に相当する。F 値は再現率と適合率の調和平均で「２×再現率×適合率／（再現率＋適合率）」で計算できる。一般に再現率を上げると適合率が下がり、適合率を上げると再現率が下がるため、システムの評価としては両方の値を加味する必要がある。そのため、一つの数値で精度を示す場合にはしばしば F 値が用いられる。

表 2 に示された精度は、すでに現代語の形態素解析の精度と比べても遜色ないほどのレベルに達している。しかし、これは「未知語なし」のデータに対する評価結果である。近代語のテキストでは多様な語が用いられるため、辞書に登録のない見出し語（未知語）が多く発生しがちである。近代文語 UniDic は、明治普通文と呼ばれるような比較的平易な文語論説文であれば高い精度で解析を行うことができるが、雅文調のテキストや口語的な内容を含むものではこれだけの精度は期待できない。また、もともと文語文を対象としたものであり口語文はうまく解析ができない。近代語のコーパスの中で口語文は大きな割合を占めるが、近代の口語文の解析のためには今後辞書の整備を行っていく必要がある。

４．近代語コーパスへの形態論情報付与（『明六雑誌』の場合）

『明六雑誌コーパス』の構築作業では、近代文語 UniDic で解析した結果を人手によって修正することで高い精度の形態論情報を付与した。明治初期の『明六雑誌』の語彙は、明治後期以降のデータを中心に整備してきた近代文語 UniDic の語彙とは異なる部分が大きく、登録されていない見出し語が多いため解析エラーも多くなっていた。

図 5 は『明六雑誌』の一部の修正済みデータを、公開中の近代文語 UniDic1.2.1 で解析した結果と比較して、明六雑誌コーパス構築開始時における、形態素解析の状況を示したものである（『明六雑誌』1874 年 1 号「洋字ヲ以テ国語ヲ書スルノ論」の一部で特に誤りの目立つ部分）。左側が正解となる人手修正済みのデータで、右側が 1.2.1 による自動解析結果であり、左端に「●」を付した部分が解析に誤りがあった語である。

	文境界	書字形	語彙素読み	語彙素	品詞	活用型	活用形	書字形	語彙素読み	語彙素	品詞	活用型	活用形	語種
	B	然る	シカル	シカリ	然り	動詞-一般	文語ラ行変格	然る	シカリ	然り	動詞-一般	文語ラ行変格	連体形-一般	和
	I	に	ニ	に	助詞-接続助詞			に	ニ	に	助詞-接続助詞			和
	I	如此き	カクノゴトシ	如此し	形容詞-一般	文語形容詞-ク	連体形-一般	如此き	カクノゴトシ	如此し	形容詞-一般	文語形容詞-ク	連体形-一般	和
	I	人民	ジンミン	人民	名詞-普通名詞-一般			人民	ジンミン	人民	名詞-普通名詞-一般			漢
	I	の	ノ	の	助詞-格助詞			の	ノ	の	助詞-格助詞			和
	I	愚	グ	愚	名詞-普通名詞-一般			愚	グ	愚	名詞-普通名詞-一般			漢
	I	も	モ	も	助詞-係助詞			も	モ	も	助詞-係助詞			和
●	I	左提	サテイ	左提	名詞-普通名詞-一般			左	サ	然	副詞			和
●	I	右挈	ユウケツ	右挈	名詞-普通名詞-一般			提	サゲル	下げる	動詞-一般	文語下二段-ガ行	連用形-一般	和
●	I	勞來	ロウライ	労来	名詞-普通名詞-一般			右	ミギ	右	名詞-普通名詞-一般			和
●								挈	タズサエル	携える	動詞-一般	文語下二段-ハ行	連用形-一般	和
●								勞	ロウ	労	名詞-普通名詞-一般			漢
●								來	ライ	来	接尾辞-名詞的-副詞可能			漢
	I	輔翼	ホヨク	輔翼	名詞-普通名詞-一般			輔翼	ホヨク	輔翼	名詞-普通名詞-一般			漢
	I	其	ソノ	其の	連体詞			其	ソノ	其の	連体詞			和
	I	苗	ナエ	苗	名詞-普通名詞-一般			苗	ナエ	苗	名詞-普通名詞-一般			和
	I	を	ヲ	を	助詞-格助詞			を	ヲ	を	助詞-格助詞			和
●	I	攫	ヌク	抜く	動詞-非自立可能	文語四段-カ行	連体形-一般	攫			名詞-普通名詞-一般			
●	I	冂	コト	事	名詞-普通名詞-一般			冂		冂	補助記号-一般			記号
	I	なく	ナイ	無い	形容詞-非自立可能	文語形容詞-ク	連用形-一般	なく	ナイ	無い	形容詞-非自立可能	文語形容詞-ク	連用形-一般	和
	B	去	サル	去る	動詞-非自立可能	文語四段-ラ行	連用形-一般	去	サル	去る	動詞-非自立可能	文語四段-ラ行	連用形-一般	和
	I	て	テ	て	助詞-接続助詞			て	テ	て	助詞-接続助詞			和
	I	転ら	クサギル	転る	動詞-一般	文語四段-ラ行	未然形-一般	転ら	クサギル	転る	動詞-一般	文語四段-ラ行	未然形-一般	和
	I	ざる	ズ	ず	助動詞	文語助動詞-ズ	連体形-補助	ざる	ズ	ず	助動詞	文語助動詞-ズ	連体形-補助	和
	I	冂	コト	事	名詞-普通名詞-一般			冂	コト	事	名詞-普通名詞-一般			和
	I	なく	ナイ	無い	形容詞-非自立可能	文語形容詞-ク	連用形-一般	なく	ナイ	無い	形容詞-非自立可能	文語形容詞-ク	連用形-一般	和
	I	時宜	ジギ	時宜	名詞-普通名詞-一般			時宜	ジギ	時宜	名詞-普通名詞-一般			漢
	I	を	ヲ	を	助詞-格助詞			を	ヲ	を	助詞-格助詞			和
	I	制し	セイスル	制する	動詞-一般	文語サ行変格	連用形-一般	制し	セイスル	制する	動詞-一般	文語サ行変格	連用形-一般	混
	I	て	テ	て	助詞-接続助詞			て	テ	て	助詞-接続助詞			和
	I	漸次	ゼンジ	漸次	副詞			漸次	ゼンジ	漸次	副詞			漢
	I	開明	カイメイ	開明	名詞-普通名詞-一般			開明	カイメイ	開明	名詞-普通名詞-一般			漢
	I	の	ノ	の	助詞-格助詞			の	ノ	の	助詞-格助詞			和

図 5 近代文語 UniDic による解析結果

図 2 のように、『明六雑誌』の解析では、多くの未知語が発生するため、新たに辞書登録を行いながら修正作業を行った。『明六雑誌コーパス』全体の語数はのべ語数で約 180500 語・異なり語数で約 15500 語である（記号を含む）。このコーパスを整備するために新たに約 3700 語を辞書に登録する必要があった。新たに追加した語は語彙素（辞書見出し相当）のレベルから追加したものもあれば、すでにある見出し語に書字形（表記形）を新たに追加したものもある。

新規登録語のうち 2834 語は頻度が 1 であり、471 語は頻度 2 であった。つまり、新規に追加した語の大部分は非常に使用頻度の低い語であった。のべ語数では約 5600 語が未知語であり、逆に約 174900 語は既知語であった。すなわち、『明六雑誌コーパス』全体の 96.89%（174900/180500）は既存の近代文語 UniDic の語彙でカバーされていたことになる。

未知語を含まないデータで評価した近代文語 UniDic の解析精度は語彙素認定で約 97%であった（表 2）。この解析精度を加味すると、既存の近代文語 UniDic による当初の『明六雑誌コーパス』の解析精度は次のように推定できる。すなわち、未知語部分の 5600 語は全て誤りと見なし、既知語部分が 97%の精度で解析されていたとすると、正しく解析されていた語数は約 169700 語（174900*0.97）であることから、概算で全体の解析精度は約 94%（169700 / 180500）であったといえる（これは再現率ベースでの計算だが、適合率・F 値でもほぼ同じ数字である）。

表 3 に『明六雑誌コーパス』のために新たに辞書登録した、コーパスにおける頻度が 8 以上の新規追加語（60 語）を挙げる。新規追加語の中では高頻度の語だが、総じて一般的でない語や表記であることがわかる。

表 3　明六雑誌コーパスの語数と近代文語 UniDic への新規追加語数

語彙素	語形	書字形	品詞	頻度
如何	イカ	何	名詞-普通名詞-一般	10
易直	イチョク	易直	名詞-普通名詞-形状詞可能	8
曰く	イワク	云	名詞-普通名詞-副詞可能	17
置く	オク	舎く	動詞-非自立可能	13
思う	オモウ	謂ふ	動詞-一般	12
思えらく	オモエラク	以爲く	副詞	12
開交	カイコウ	開交	名詞-普通名詞-一般	15
変える	カユ	易ゆ	動詞-一般	10
関渉	カンショウ	關渉	名詞-普通名詞-サ変可能	9
気学	キガク	氣學	名詞-普通名詞-一般	9
議者	ギシャ	議者	名詞-普通名詞-一般	8
議法	ギホウ	議法	名詞-普通名詞-一般	9
下観	ゲカン	下觀	名詞-普通名詞-サ変可能	12
下民	ゲミン	下民	名詞-普通名詞-一般	13
限制	ゲンセイ	限制	名詞-普通名詞-サ変可能	12
孤陰	コイン	孤陰	名詞-普通名詞-一般	9
好和	コウワ	好和	名詞-普通名詞-一般	13
国中	コクチュウ	國中	名詞-普通名詞-一般	9
試み	ココロミ	嘗み	名詞-普通名詞-一般	8
国君	コックン	國君	名詞-普通名詞-一般	10
異	コト	特	名詞-普通名詞-一般	39
今時	コンジ	今時	名詞-普通名詞-一般	8
裁成	サイセイ	裁成	名詞-普通名詞-一般	11
三聖	サンセイ	三聖	名詞-普通名詞-一般	11
三宝	サンポウ	三寶	名詞-普通名詞-一般	50
シビリゼーション	シビリゼーション	シヴヰリゼーション	名詞-普通名詞-一般	14
者流	シャリュウ	者流	名詞-普通名詞-一般	24
習	シュウ	習	名詞-普通名詞-一般	14
上観	ジョウカン	上觀	名詞-普通名詞-サ変可能	13
シロシ	シロシ	素	名詞-固有名詞-人名-名	22
信紙	シンシ	信紙	名詞-普通名詞-一般	8
人主	ジンシュ	人主	名詞-普通名詞-一般	14
数百	スウヒャク	數百	名詞-数詞	9
少しく	スコシク	少く	副詞	8
大宝	タイホウ	大寶	名詞-普通名詞-一般	11
タバコ	タバコ	烟	名詞-普通名詞-一般	9
治刑	チケイ	治刑	名詞-普通名詞-一般	9
忠諒	チュウリョウ	忠諒	名詞-普通名詞-形状詞可能	8
蝶鋏	チョウコウ	蝶鋏	名詞-普通名詞-一般	8
つく	ツク	付く	動詞-一般	20
妻	ツマ	婦	名詞-普通名詞-一般	22
無い	ナシ	無し	形容詞-非自立可能	101
パッション	パッション	パツシヨン	名詞-普通名詞-一般	8
独り	ヒトリ	獨	名詞-普通名詞-副詞可能	28
ベーコン	ベイコン	倍根	名詞-固有名詞-人名-一般	25
邦	ホウ	邦	名詞-普通名詞-一般	8
磨する	マス	磨す	動詞-一般	8
先ず	マズ	先	副詞	11

間々	ママ	間	副詞	13
魅する	ミス	魅す	動詞-一般	17
アメリカン	メリケン	米利堅	名詞-普通名詞-一般	9
最も	モットモ	尤	副詞	12
基づく	モトヅク	本づく	動詞-一般	11
止む	ヤム	息む	動詞-一般	11
容忍	ヨウニン	容忍	名詞-普通名詞-サ変可能	11
与聞	ヨブン	與聞	名詞-普通名詞-サ変可能	9
濫出	ランシュツ	濫出	名詞-普通名詞-サ変可能	9
リバティー	リボルチー	リボルチー	名詞-普通名詞-一般	14
ルーサー	ルーサー	路惕	名詞-固有名詞-人名-一般	9
論無い	ロンナシ	論なし	形容詞-一般	11

５．おわりに

以上、近代語テキストの形態素解析について、近代文語 UniDic の解説と『明六雑誌コーパス』の構築時の解析結果修正作業を中心に述べた。

『明六雑誌』は近代文語 UniDic の主たる対象からはずれたテキストであるため、多数の未知語を追加しながら自動解析結果を修正して対処する必要があった。もっとも『明六雑誌コーパス』のように、（把握できる範囲で）誤りを全て修正したコーパスを公開するようなケースは稀であると思われる。一般的な研究利用であれば必要とされる部分についてのみ修正を行えば良いし、94%程度の解析精度があれば十分な場合も少なくないだろう。また、ここでの精度評価は、単語境界・品詞認定・語彙素認定（代表表記・読み・語種を含む）の全てが正しい場合のみを正解と見なすという、非常に厳しい評価基準によっている。読みや語種、品詞といった一部についてだけの精度であればこれを上回ることは確実である。

単に稀例を探すような場合には文字列検索で事足りるが、調査対象がテキスト全体の中でどのような位置を占めるのかを把握するためには、データ全体に対して形態素解析が施されている必要がある。形態素解析がなされたコーパスは、単に検索の手間が少なく、索引ではできなかったような組み合わせ検索ができるだけでない。テキストを、順序を持った語の集合として扱って、データベース上で自由に集計し、統計的な処理を行うことが可能になるのである。今後、近代語の研究においてもこうした本格的な語彙研究等のコーパスを活用した研究が行われることに期待したい。

なお、今回追加した『明六雑誌』の語彙を含む新しい近代文語 UniDic を近く公開する予定である。

文　献

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―』（CD-ROM、博文館新社）

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 号 pp.101-122.

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」『日本語学会 2008 年度春季大会予稿集』 pp.211-218

小木曽智信ほか（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』科研費若手研究(B) 研究成果報告書（課題番号 19720110）（http://dl.dropbox.com/u/73297026/report/unidic-MLJ_report2009.pdf）

岡照晃・小町守・小木曽智信・松本裕治（2011）「機械学習による近代文語文への濁点の自動付与」『情報処理学会 自然言語処理研究会報告』Vol.2011-NL201, No.6

URL

形態素解析辞書 UniDic ダウンロードサイト：http://download.unidic.org/
近代文語 UniDic：http://www2.ninjal.ac.jp/lrc/index.php?UniDic
形態素解析器 MeCab ホームページ：http://mecab.sourceforge.net/