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
{togiso,ogura,kondo}@kokken.go.jp

独立行政法人 国立国語研究所

1. 近代文語と形態素解析

1.1. コーパス日本語学と史的資料

昨今、日本語学・国語学の分野でもコーパス利用の気運が高まりつつある。この分野では、以前より古い時代の資料を扱う歴史的研究が大きな位置を占めている。しかし、歴史的な研究においてコーパス言語学的な研究を行うことは困難であった。特に、形態素解析が行えないため、語のレベルでの調査が行えないことが問題となっている。

そこで、伝康晴氏（千葉大学）と国立国語研究所が中心となって開発中の形態素解析辞書 UniDic を拡張し、文語文の解析が可能な解析辞書の開発を行うこととした。

1.2. なぜ近代文語か

文語文の解析辞書の最初の目標として、近代の文語論説文を対象とすることとした。これは次のような理由による。

一つは、この文体で書かれた資料が非常に多いことである。明治普通文とも呼ばれる近代の文語論説文は、明治以降、戦前まで安定的に用いられた文体であり、公文書から新聞・雑誌まで各種の資料がこれによって書かれている。解析辞書の利用を考えたとき、対象となる資料が多いことは何よりも重要である。

もう一つの理由は、すでに電子化されテキストとして公開されている資料が多いことである。近代文語で書かれた資料の多くは著作権保護期間を過ぎているため、比較的自由に利用することが可能であり、青空文庫でも多くの近代語資料が公開されている。法律などの公文書も多くが電子化されて公開されている。さらに国立国語研究所で作成された『太陽コーパス』[i]という近代総合雑誌の大規模なデータがある。このように、すぐに利用可能な資料を十分に確保できることもポイントである。

このほか、解析結果の利用に関する面で、近代語は現代語に直接つながる時代の資料であるため、解析結果を比較して利用しやすいという点が挙げられる。今回作成した辞書は解析単位を現代語とそろえ、通時的な比較が行えるように配慮している。これによって近代語から現代語への変化を捉えることが可能になる。

2. 近代語語彙の増補

2.1. UniDic の階層構造と近代語語彙

近代文語文を解析するためには、近代語向けの語彙を大幅に増補する必要がある。現時点で近代語用に追加した語は、約 37,000 語（書字形）である。この登録は以下のように行った。

UniDic は「斉一な単位」「階層的な見出し」「アクセントなどの音声情報」を特徴とする新しい日本語形態素解析辞書である（伝ほか 2007）。階層構造をもつため、近代語特有の語形・書字形（表記形）を現代語の語形・書字形とともに統一的に扱うことが可能になっている。たとえば動詞「読む」では、**図 1** に示すように文語形（四段活用）や旧字体の書字形（「讀む」）を一つの見出し（語彙素「読む」）の下に登録することができる。

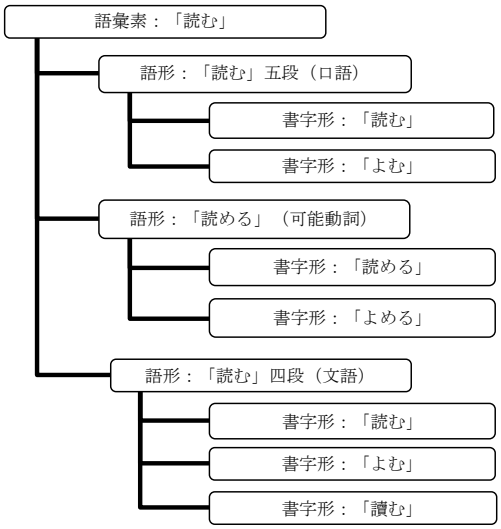

図 1 UniDic の階層構造と近代語

近代語には、旧字体・旧仮名遣いなどの表記違いの語や、現代語とは活用が異なる語がきわめて多いが、これらを同一見出しの下にまとめて、必要に応じて区別して利用できるのは UniDic の利点である。

登録に当たっては現代語と同じく「短単位」（小椋

2007）を採用している。そのため、斉一な単位による解析が可能であり、現代語と近代語の解析結果を相互に比較することが可能となっている。

2.2. 文語形・旧字形の生成

文語文の解析には活用語の文語形を整備する必要がある。その多くは、たとえば五段活用の動詞を四段活用にするように機械的に生成することができる。しかし、UniDic に登録したすべての口語活用語から機械的に文語活用を生成すると、現代の極めて口語的な語や、活用型が変化した動詞などで、近代語としてありえない不適切な語形が生じることになる。たとえば次のようなものである。

五段活用「死ぬ」→誤：四段活用
正：ナ行変格活用
形容詞「やばい」→不要：「やばし」

そのため、いったん機械的に文語活用を生成した後、人手で確認して近代語として不適当な語を削除した。また、ア行に活用する語をハ行・ワ行に割り振るなどの整理を行った。

また、近代語の資料では基本的に漢字は旧字体で書かれているが、それをそのまま電子化したテキストが多いため、辞書に旧字形を用意する必要がある。これも大部分の語は機械的に生成できるが、次のような場合には過剰に生成されたり、あるいは生成できなかったりする。よって、これらも手作業によって整備した。

「弁」 語によって「瓣･辯･辮」に分かれる
（例：「花瓣」「辯護」「辮髪」）
「連」 語によって「聯」になるものがある
（例：「聯合」）

「仏国」「佛國」「佛国」「仏國」のように旧字新字の組み合わせを全て取り入れれば膨大な数に上ることになるが、両者が組み合わさることはまれなので、はじめに漢字を全て旧字体に変換したもののみを整備し、これ以外の組み合わせは実例から追加することとした。

2.3. 資料からの語彙追加

近代語の文章の実例からも語彙を追加した。後述する学習用コーパスに現れた未知語のほかに、国語研究所で『太陽コーパス』を開発した際に作成された「スカウト式用例採集データ」を整備し直し、ここから追加を行っている。これは雑誌『太陽』から人手で採集した語のリストで、1901 年分だけで約 43 万件にのぼる。現在、形容詞・副詞などの登録を終えたところであり、今後、名詞などの追加を行なう予定である。

2.4. 近代語彙の問題点

近代語のテキストには現代語には見られないさまざまな特徴があり、それが解析辞書作成上の問題となることが少なくない。主な問題とそれに対する対処の方法を示す。

無濁点

近代語では濁点が表記されない場合が少なくない。したがって、無濁点の書字形も辞書登録した。この中には助動詞「ず」のような活用語も含まれる。

踊り字

近代語では「ゝゞ」のような踊り字（繰り返し記号）が多く用いられている。したがって、これを含む書字形を辞書登録しておく必要がある。しかし、「思は/ゞ」「民主/々義」のように語（短単位）の境界をまたいだ踊り字も存在するため、辞書で完全に対応することは困難である。そこで、必要に応じて踊り字を普通の仮名に置き換える前処理を行うこととした。

しかし、複数の文字を繰り返す「くの字点」については、繰り返しの範囲が一定しないため自動的な変換が難しい。そこで、「そろ/＼」のように語の一部をなすものはその形を辞書に登録し、「進め/＼」のように語句そのものを繰り返すものは記号として処理した。

カタカナの変換

近代語ではしばしば本文がカタカナで書かれており、そのままでは解析ができない。これについては辞書で対応するのは困難であるため、カタカナをひらがなに変換する前処理によって対処することとした。

踊り字と仮名変換の前処理は XSLT で実装し、解析用の GUI（「茶まめ」）から容易に利用できるようにした。この処理により、解析対象の本文はたとえば次のように変換されることになる。

処理前「裁判官ハ刑法ノ宣告又ハ懲戒ノ処分ニ由ルノ外其ノ職ヲ免セラルヽコトナシ」
処理後「裁判官は刑法の宣告又は懲戒の処分に由るの外其の職を免せらるることなし」

下線を付した無濁点部分は、解析辞書で対応することになる。

仮名遣い

近代語資料の多くは現代語とは異なる仮名遣いで書かれているが、それは必ずしも歴史的仮名遣いとは一致しない。たとえば「かおる」は歴史的仮名遣いでは「かをる」であるが「かほる」も頻繁に用いられている。そこで、仮名遣いのバリエーションの追加は単に既存の国語辞典等によるのではなく、実際の用例を基にして行っている。

送り仮名の省略

近代語の活用語は、語形のレベルでは学校文法の活用表がきれいにあてはまるので比較的問題は少ない。しかし、表記のレベルでは「有（あら）む」「読（よめ）り」のようなさまざまな省略表記が行われる。これらは活用表を利用して、書字形を展開する際に対応した。

このほか、ク語法（「宣はく」「庶幾はく」）などの現代語では生産的でなくなった語形についても活用表を利用して整備した。

なかには、UniDic の「短単位」とどうしても対応がとれない近代語特有の表記がある。たとえば「加之（しかのみならず）」「由是観之（これによりてこれをみるに）」などの漢文的表現である。こうしたものについては全体で一つの接続詞として扱うなど、例外的な処理を行っている。

3. 学習用コーパス

当時の標準的な文語論説文と考えられるものを中心に、すでに電子化され利用可能なデータの中から資料を選び、解析結果に手修正を加えて学習用のコーパスとして整備した。現在利用している学習用コーパスは以下に挙げるもので、総語数約 10 万語である。作業の進捗状況などに左右されるため、現時点では必ずしも十分にバランスが考えられているわけではない。

公文書

『大日本帝国憲法』『教育勅語』『民法（第一編）』ほか計 8 編

青空文庫

『人生に相渉るとは何の謂ぞ』（北村透谷）

『小説総論』（二葉亭四迷）ほか評論計 9 編

文明論之概略[ii]

「緒言」「巻之一第一章」「巻之一第二章」

太陽コーパス

「教育時評」（大町桂月）、「宗教時評」（龍山学人）ほか 1901 年 1 号より計 12 記事

これ以外に、島崎藤村・与謝野晶子らの近代詩を数編加えている。漢文訓読調の論説文だけでは、「けむ」「らむ」などの専ら和文脈で用いられる助動詞が出現しないため追加したものである。

4. 解析辞書と精度

現代語版の UniDic 1.3.7[iii]に 2 で示した近代語語彙の増補を行った辞書と、3 の学習用コーパスを利用して、ChaSen 版と MeCab 版の辞書を作成した[iv]。MeCab での学習に際しては、現代語の UniDic と同じ素性を用いている。

表 1・表 2 は、学習に使用していない手修正データを用いて ChaSen 版と MeCab 版それぞれの解析精度を測定したものである。テストデータは次の通りである。

福澤諭吉：「経世の学、また講究すべし」「物理学の要用」

山路愛山：「北村透谷君」「透谷全集を読む」

太陽：「明治三十四年の経済界」「昨年の経済問題」「経済時評」（いずれも 1901 年 1 号）

民法：第三編（後半部分）

なお、「民法」以外のデータはそれぞれ数語ずつの未知語を含んだ状態で計測している。

表 1 近代文語 UniDic 0.6－ChaSen 2.4.2 の解析精度（アウトサイド）

		福沢諭吉	山路愛山	太陽	民法	ALL
語数		4194	3048	6181	10854	24277
出力数		4238	3068	6203	10876	24385
Level1（境界）	正解数	4120	3015	6092	10821	24048
	再現率	0.982356	0.989173	0.985601	0.99696	0.990567
	適合率	0.972157	0.982725	0.982105	0.994943	0.98618
	F 値	0.97723	0.985939	0.98385	0.99595	0.988369
Level2（品詞）	正解数	3977	2898	5928	10632	23435
	再現率	0.948259	0.950787	0.959068	0.979547	0.965317
	適合率	0.938414	0.944589	0.955667	0.977565	0.961042
	F 値	0.943311	0.947678	0.957364	0.978555	0.963175
Level3（語彙素）	正解数	3940	2853	5886	10629	23308
	再現率	0.939437	0.936024	0.952273	0.97927	0.960086
	適合率	0.929684	0.929922	0.948896	0.977289	0.955834
	F 値	0.934535	0.932963	0.950581	0.978279	0.957955

表 2 近代文語 UniDic 0.6－MeCab 0.96 の解析精度（アウトサイド）

		福沢諭吉	山路愛山	太陽	民法	ALL
語数		4194	3048	6181	10854	24277
出力数		4211	3048	6182	10857	24298
Level1（境界）	正解数	4142	3005	6119	10827	24093
	再現率	0.987601	0.985892	0.989969	0.997512	0.992421
	適合率	0.983614	0.985892	0.989809	0.997237	0.991563
	F 値	0.985604	0.985892	0.989889	0.997375	0.991992
Level2（品詞）	正解数	3991	2892	5967	10708	23558
	再現率	0.951598	0.948819	0.965378	0.986549	0.970383
	適合率	0.947756	0.948819	0.965222	0.986276	0.969545
	F 値	0.949673	0.948819	0.9653	0.986412	0.969964
Level3（語彙素）	正解数	3954	2850	5926	10704	23434
	再現率	0.938969	0.935039	0.958589	0.985908	0.964442
	適合率	0.942775	0.935039	0.958745	0.98618	0.965276
	F 値	0.940869	0.935039	0.958667	0.986044	0.964858

総じて解析精度は良好であり、十分に実用に耐えるレベルを実現している。特に、民法のようによく整備された本文では、現代語の解析結果に引けをとらない精度を達成している。

ただし、これが未知語をほぼ取り除いた条件下での数値であることには注意を要する。近代語の論説文には難解な語句が多く、未知語が多く発生する傾向があるため、全く未知の文章では現代語の場合以上に精度が低下する可能性がある。一方で、現段階のテストデータや学習用コーパスは誤りを含むと考えられるので、整備することにより精度の一層の向上も期待できる。

おわりに

近代文語文を対象とした形態素解析辞書を開発し、日本語研究に利用可能な精度で文語文を解析することが可能になった。解析単位が同じであるため、現代語と近代語の解析結果を直接比較することも可能である。今後、コーパス言語学の手法を応用した歴史的な研究が進むものと期待される。

同時に、日本語研究以外の方面での応用にも期待したい。UniDic の情報を使うことで、濁点の自動付与や新旧仮名遣いの変換、表記の統一などが可能になる。将来的には、文語文の現代語訳も可能になると思われる。文語民法と口語民法とを解析した結果は対訳コーパスとしても使用可能である。

今後、さらなる語彙の追加と、学習用コーパスの整備を進め、解析精度の向上をはかる必要があるのは言うまでもないが、同時に UniDic により多くの情報を付与することで応用の幅を広げていきたい。

また、今回対象とした近代文語の論説文は基本的に漢文訓読調の文章であり、従来日本語学の資料として多く用いられてきた小説類とは異なる文体である。この辞書をてがかりとして、小説や和文系の文語文にも対応していく必要があると考えている。

なお、この形態素解析辞書「近代文語 UniDic」は本年度末に一般公開予定である。ライセンスはベースとなる現代語の UniDic を踏襲する。

参考文献

国立国語研究所編（2005）国立国語研究所報告 122『太陽コーパス』博文館新社

小椋秀樹（2007）国立国語研究所内部報告書『『現代日本語書き言葉均衡コーパス』 短単位規程集 version 1.2』

伝康晴ほか（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

付記

本研究にあたっては、第 2 回 博報「ことばと文化・教育」研究助成の援助を受けた。

[i] 総合雑誌『太陽』（1895～1928 年）の一部を構造化テキストにしたもので、文字数約 1450 万字。約 3400 記事のうち約半数を文語が占める。

[ii] 『文明論之概略』は上田修一氏（慶応大）により公開されている電子化テキストを利用させていただいた。
http://www.slis.keio.ac.jp/~ueda/

[iii] UniDic は 1.3.5 が下記にて一般公開中。1.3.7 は非公開。
http://download.unidic.org/

[iv] ChaSen 版の作成には浅原正幸氏（奈良先端大）・伝康晴氏（千葉大）による学習プログラムを利用させていただいた。また MeCab 版の作成には中村純平氏（東京農工大）の協力を得た。