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†
†国立国語研究所, ‡千葉大学
{togiso,naberyo,kondo}@kokken.go.jp, den@cogsci.l.chiba-u.ac.jp

1.はじめに

1.1.近代文語 UniDic の概要

発表者等は近代文語文を対象とした形態素解析辞書「近代文語 UniDic」を開発している。この辞書は主として日本語学分野での応用を意図し、近代の文語論説文（主として明治普通文）を対象とした解析辞書である。現代語用の UniDic をベースに、近代語用の見出し語約4.6万語(書字形)を追加している。現在 ver.0.9 を一般公開中である。

公開中の辞書の総見出し語数は、語彙素:14万語、語形：15.7万語、書字形：23.7万語。解析精度は、境界認定:99.43%、品詞認定:98.42%、語彙素認定:97.90%（いずれもF値）となっている。

なお、ここでいう語彙素とは UniDic の見出し語階層のうち、辞書見出しに相当するレベルで、例えば「工場（こうじょう)」と「工場（こうば)」を正しく選択できることを意味する。

1.2.近代語の学習用コーパス

近代文語 UniDic では、近代の文語文約38万語分を人手により整備して学習・評価に用いている。コーパスの資料別の内訳は次の通りである。

表 1　近代文語 UniDic の学習用コーパス

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	(福澤諭吉)	4.86
青空文庫	(山路愛山)	2.43
青空文庫	(北村透谷)	4.42
青空文庫	(陸羯南)	3.20
青空文庫	(その他)	4.80
法令・公文書		5.21
近代詩		0.18
計		37.87

現在の38万語のコーパスは、必ずしも十分な量であるとは言えない。しかし、近代語コーパスの人手修正は、近代語・文語文に関する知識を要することから、現代語コーパスの整備と比較しても極めて手間と費用がかかり、追加することは容易でない。

1.3.近代語と現代語の語彙の共通性

ところで、近代語と現代語とでは、文法上の違いは大きいものの、語彙の面ではかなり共通する部分がある。極端な例になるが、2004年に文語体から口語体に改正された民法を見ると、内容が変わらなければ、文体の違いにもかかわらず語彙(特に内容語)の違いは極めて小さいことがわかる（下線部が共通な語)。

【口語民法】
第一条　私権は、公共の福祉に適合しなければならない。２　権利の行使及び義務の履行は、信義に従い誠実に行わなければならない。３　権利の濫用は、これを許さない。

【文語民法】
第一条　私権ハ公共ノ福祉ニ遵フ　②　権利ノ行使及ヒ義務ノ履行ハ信義ニ従ヒ誠実ニ之ヲ為スコトヲ要ス　③　権利ノ濫用ハ之ヲ許サス

このような特殊例に限らず、一般的にも近代語と現代語との語彙の共通性は高い。図 1・図 2は近代の総合雑誌を収めた『太陽コーパス』の論説文と、『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）の書籍データ（文学作品を除いたもの）に形態素解析を施し[1]、その語彙を UniDic の語彙素のレベルで比較したものである(助詞・助動詞・記号類を除く)。異なり語で見た場合にも近代語だけで用いられる語は多くなく、大多数の語が現代語でも共通に用いられているが、特に延べ語数でみた場合には90%以上が共通となっている。表面上感じられる大きな違いは、助詞・助動詞、複合辞等の違いによるところが大きい。

[1]BCCWJ は UniDic1.3.10i,『太陽コーパス』は近代文語 UniDic 0.91 で解析した（いずれも非公開バージョン)。解析器は MeCab 0.97 を使用した。

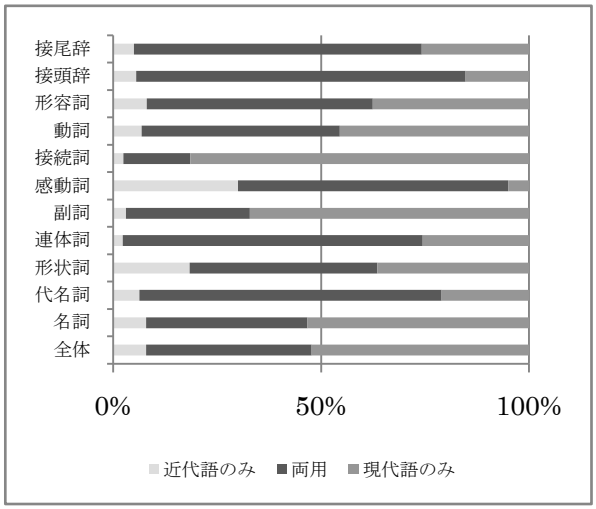

図 1　近現代語彙の共通性（異なり語数）

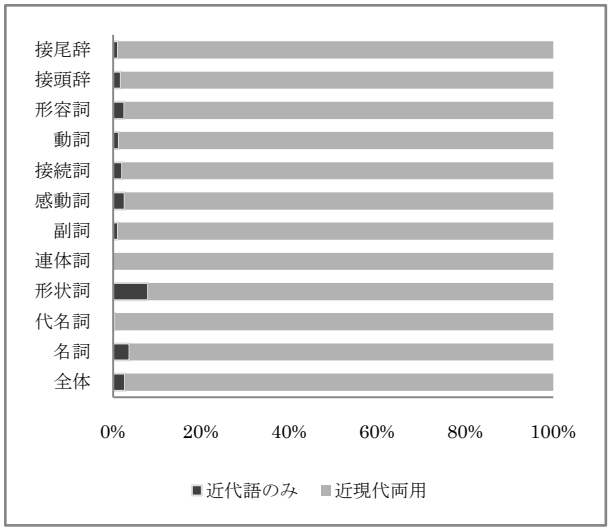

図 2　近現代語彙の共通性（延べ語数）

1.3.現代語コーパスと近代語解析

このような近･現代の語彙の共通性に着目すると、現代語のコーパスを近代語の形態素解析の精度向上に利用することが考えられる。現代語の人手修正コーパスとして､現代語 UniDic の開発に用いられている「BCCWJ コアデータ」が利用可能である[2]。近代文語 UniDic は現代語 UniDic をもとにしており、単位認定の基準も同じ「短単位」にそろえていることから、見出し語に互換性がある。本発表では現代語コーパスである BCCWJ コアデータ 60 万語（書籍 20 万語、新聞 20 万語、白書 20 万語）を用いて、近代語文語 UniDic の解析精度を向上させる方法を検討する。

2.現代語コーパスを利用した解析精度の向上

2.1.ベースライン

本発表の目的は、現代語コーパスを利用することによって、近代語コーパスのみで学習した場合よりも高い精度で形態素解析を行うことである。基準となる近代語コーパスのみを利用した場合の精度（ベースライン）は表 2 の通りである。評価コーパス約 4.5 万語を残した 33.5 万語で学習を行った。解析器に MeCab 0.97 を用いた。

表 2　解析精度（ベースライン）

レベル	精度（F 値）
1	0.994364
2	0.984215
3	0.979050
4	0.976131

表中のレベルは、1 が境界、2 が品詞（活用型・活用形を含む）、3 が語彙素、4 が発音形の認定を正しく行えることを意味する（以下同じ）。

2.2.現代語コーパスの単純利用

現代語コーパスを利用するもっとも単純な方法として、近代語の学習用コーパスに現代語のコーパスを追加して学習する方法が考えられる。この方法により、BCCWJ のコアデータ 60 万語を追加して作成した辞書の評価結果が表 3 である。

表 3　現代語を単純に利用した場合の解析精度

レベル	精度（F 値）
1	0.993904
2	0.959820
3	0.954611
4	0.951603

ベースラインと比較して大きく精度を下げてしまっているが、この原因として次のことが考えられる。

一つは、助詞・助動詞などの機能語のコストが近代語と現代語で大きく違う点である。近代語と現代語では、語彙の共通性が高いとはいえ、文法の違いにより機能語のコストは大きく異なる。機能語は語彙化して学習しているため影響が大きい。

また、文語と口語で活用型が変わった活用語の影響が考えられる。たとえば文語における四段活用の動詞「読む」は口語では五段活用になる。したがって現代語コーパスに引きずられて「読む」を五段活

[2]現在「BCCWJ モニター公開データ（2008 年度版）」が公開されている。http://www.kokken.go.jp/kotonoha/ex_8.html

用と解析すると誤りとなってしまう。このほかにも文語でク活用・シク活用に分かれていた形容詞の活用型が現代語では一つになっているなどの活用型の違いがある。このため、活用型認定を含むレベル 2 以降の精度が大きく悪化している。

このような問題があるため、近代語の形態素解析の精度向上に現代語コーパスを利用するためには、現代語コーパスをそのまま用いることはできない。現代語コーパスで学習したコストから悪影響のないものを適切な方法で反映させる必要がある。

2.3.現代語の生起コストの利用

そこで、現代語コーパスを近代語コーパスに追加するのではなく、現代語コーパスだけで学習したコストをベースラインの辞書に反映させることで解析精度を高める方法をとることとした（図３）。語彙表は近・現代語の見出し語を全て含む共通のものを使用する。コストの取得には MeCab 標準のプログラムをそのまま用い、混合処理のみを別途行った。

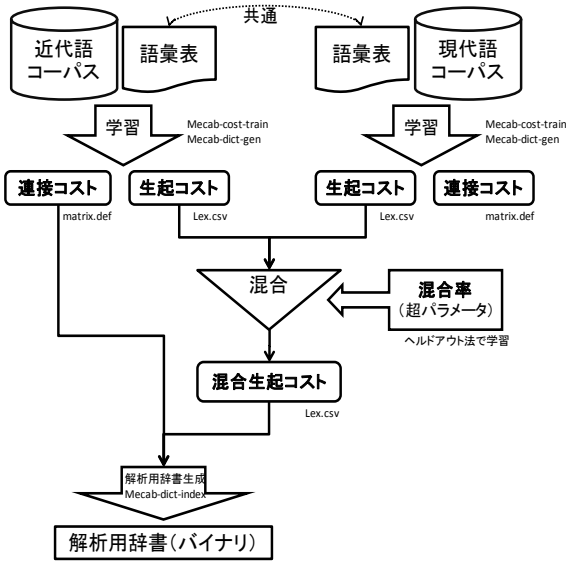

図 3　現代語の生起コスト利用の流れ

現代語コストを反映させるやり方には様々な方法が考えられる。今回は、近代語コーパスでは学習できていない低頻度語の出現コストを補うという方針の下、表 5 の４通りの方法を試行することとした。条件はいずれも現代語コーパスの生起コストが反映されすぎないよう制限をかけるものである。いずれの場合も、文法の違いが大きく影響すると考えられる連接コストは利用しない。

表 5　正規コストの反映方法

	無条件	近代語優先
全品詞	a	A
活用語・助詞を除く	b	B

近代語優先とは、近代語コーパスで学習した生起コストが現代語コーパスで学習した場合よりも大きい場合にのみ、現代語のコストを一定の割合で混合することを示す。

いずれの場合も現代語の影響が大きすぎると精度が下がるため、現代語のコストの割合（混合率）を調整する必要がある。

2.4. ヘルドアウト法による混合率の学習

現代語コストを反映させる適切な混合方法と最適な混合率を求めるため、ヘルドアウト法を用いて学習を行った。ベースラインの学習コーパスのうち 3.5 万語をヘルドアウトコーパスとして取り置き、残り約 30 万語で学習した辞書を用いた。

表 4 は各方式で混合率を変化させた辞書について、ヘルドアウトコーパスで評価した精度（レベル 3 の F 値）である。表の混合率は、0.1 の場合に現代語 1：近代語 9 の割合で混合することを示す。現代語が近代語の割合を超えない 0.5 までを 0.025 刻みで調査した（表は 0.05 刻み）。

この結果から B 方式の混合率 0.3 を最適な混合方法・最適混合率として選択した。この混合率は、活用語・助詞以外の近代語の語彙と現代語の語彙の共通性に基づくものであり、今回の学習コーパスに限らず利用可能な数値であると考えられる。

2.5. 現代語コーパスと最適混合率を用いた補正

ヘルドアウトコーパスを除く近代語 30 万語で学習した辞書と、この辞書に最適混合率で現代語コーパスの生起コストを混合した辞書を作成し、その評価結果を比較した（表 6）。

表 4　混合率の学習結果（レベル３）

混合率	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
A	0.978654	0.978639	0.978839	0.978739	0.978739	0.978979	0.978879	0.978879	0.978438	0.978295
B	0.978682	0.978626	0.978909	0.978823	0.978837	**0.979007**	0.978822	0.978751	0.978366	0.978393
a	0.978368	0.978167	0.978532	0.976434	0.975662	0.974482	0.973232	0.972065	0.970671	0.968693
b	0.978626	0.978668	0.978822	0.977988	0.977460	0.976792	0.976053	0.975513	0.974644	0.973990

表 6　混合辞書の解析精度（ヘルドアウトコーパスを含まない 30 万語で学習）

レベル		近代語のみ	混合
1	評価語数	44551	44551
	出力数	44499	44495
	正	44261	44274
	誤	290	277
	F 値	0.994070747	0.994407385
2	評価語数	44551	44551
	出力数	44499	44495
	正	43800	43806
	誤	751	745
	F 値	0.983717013	0.983895964
3	評価語数	44551	44551
	出力数	44499	44495
	正	43557	43575
	誤	994	976
	F 値	0.978259405	0.978707634
4	評価語数	44551	44551
	出力数	44499	44495
	正	43428	43440
	誤	1123	1111
	F 値	0.975362156	0.975675494

つづいてベースラインの辞書（混合率を学習したヘルドアウトコーパスを含む 33 万語で学習）に対して最適な混合率を用いて現代語コーパスの生起コストを反映させた辞書を作成した。表 7 はこの辞書とベースラインの評価結果を比較したものである。

表 7　混合辞書の解析精度（ベースラインとの比較）

レベル		ベースライン	混合
1	評価語数	44551	44551
	出力数	44519	44513
	正	44284	44291
	誤	267	260
	F 値	0.994363983	0.994588161
2	評価語数	44551	44551
	出力数	44519	44513
	正	43832	43831
	誤	719	720
	F 値	0.984214663	0.984258511
3	評価語数	44551	44551
	出力数	44519	44513
	正	43602	43607
	誤	949	944
	F 値	0.979050185	0.97922842
4	評価語数	44551	44551
	出力数	44519	44513
	正	43472	43475
	誤	1079	1076
	F 値	0.976131133	0.976264259

この方法により、差はわずかであるが、ベースラインの精度を上回ることを得た。

3.おわりに

現代語コーパスの生起コストを利用することにより、わずかではあるが近代語形態素解析の精度を向上させることができた。近代語の学習用コーパスが限られた種類のテキストであるのに対し、BCCWJ コアデータはランダムサンプリングされたさまざまな種類のテキストを含んでいる。多様なテキストから得た生起コストを反映させたことにより、近代語の解析においても幅広い種類のテキストに対応できるようになったと考えられる。

今回の方法をおし進めて、より精度を向上させるために、次のような方法が考えられよう。

・混合学習の割合をクロスバリデーションによって決定する
・現代語コーパスのうち近代語とは共通性の低いテキストを調査し、学習対象外とする
・近・現代語間で変化した活用型を対応させることにより動詞・形容詞の生起コストを利用する
・近・現代語で共通性の高い品詞の組み合わせについて連接コストを利用する

こうした方法の検討は今後の課題としたい。

参考文献

国立国語研究所（2005）国立国語研究所報告 122『太陽コーパス』博文館新社
国立国語研究所（2006）『近代女性雑誌コーパス』
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 pp.101-122
小木曽智信ほか（2008）「近代文語文を対象とした形態素解析辞書の開発」『言語処理学会第 14 回年次大会発表論文集』 pp.225-228
国立国語研究所（2008）『BCCWJ 領域内公開データ（2008 年度版）』

Web サイト

近代文語 UniDic
http://www.kokken.go.jp/lrc/index.php?UniDic
近代女性雑誌コーパス
http://www.kokken.go.jp/lrc/index.php?近代女性雑誌コーパス
MeCab　オリジナル辞書/コーパスからのパラメータ推定
http://mecab.sourceforge.net/learn.html（工藤拓）

本研究は、科学研究費補助金・若手 B（課題番号 19720110）、および文部科学省科学研究費補助金・特定領域研究「日本語コーパス」による成果の一部を含むものである。