旧仮名遣いの口語文を対象とした形態素解析辞書

小木曽 智信
国立国語研究所 言語資源研究系

旧仮名遣いで書かれた口語文のテキストを形態素解析する場合、既存の形態素解析辞書では不十分な点があった。発表者は既存の形態素解析辞書 UniDic をベースに見出し語の追加やコストの再学習を行い、旧仮名遣いの口語文を解析するのに適した新しい UniDic を開発した。本稿では、この旧仮名遣いの口語文を対象とした形態素解析辞書の作成方法とその解析精度について述べる。

Morphological analysis dictionary for colloquial style Japanese written in old kana usage

Toshinobu Ogiso
Department of Corpus Studies, National Institute for Japanese Language and Linguistics

The existing morphological analysis dictionaries were unsuitable for the analysis of colloquial style text written in old kana usage. Therefore, I developed a new morphological analysis dictionary especially for the colloquial style Japanese in old kana usage, based on UniDic, the existing morphological analysis dictionary. In this report, I discuss about how to make this new dictionary and show the analysis precision of this dictionary.

１．まえがき

今日の書き言葉は、短歌や俳句などの例外を除き、ほぼ全てが現代仮名遣いによっている。しかし、やや時代をさかのぼると、文法的には現代語と変わらない内容であっても、歴史的仮名遣いなどの旧い仮名遣いで書かれている資料が少なくない。たとえば、現行の日本国憲法でさえ原文は歴史的仮名遣いで書かれている。明治期の言文一致よりのち戦後の国語改革の定着までの間に書かれたテキストの多くはこのような旧仮名遣いの口語文でかかれている。こうした資料を電子化してコンピュータ上で利用していく場合、表記の違いは処理上の問題を引き起こす。日本語の自然言語処理の基礎であるところの形態素解析においても、思わぬ解析エラーを引き起こし、全体の精度を低下させることになる。文字単位で対応表を用意すればことたりる旧漢字については比較的容易に常用漢字に置き換えることができる。しかし、仮名遣いについては、語を単位とするものであるため、このような単純な置換によるわけにはいかない。

発表者は、このような旧仮名遣いによって書かれているテキストの形態素解析を可能にすることを目標に、既存の形態素解析辞書「UniDic」[1][2]をベースとして、旧仮名遣いに対応した新たな UniDic を作成した。これにより、従来の現代語用の形態素解析辞書や、近代の文語文のために作られた形態素解析辞書「近代文語 UniDic」[3][4]よりも高い精度で解析を行うことが可能になった。

２．旧仮名遣いの口語文の性格

明治時代後半の言文一致により文法的には現代語に近い口語文体が成立した。しかし、当時のテキストは歴史的仮名遣いか旧来の慣用（旧仮名遣い）によっており、今日の仮名遣いとは異なっている。このような旧仮名遣いの口語文体（以下、旧仮名口語）のテキストは、戦後の現代仮名遣い（「現代かなづかい」）の定着までの期間にかなりの量が残されており、資料の質の面でも、先述した日本国憲法などの法令・公文書から近代の文学作品、新聞等まで幅広いジャンルの重要なテキストが残されている。

これらの旧仮名口語テキストは、再出版されたり電子化されたりする機会に現代仮名遣いに直されるものも少なくないが、旧仮名遣いのままで残されているものも多い。また、できる限り一次資料に基づくべきであるという点からも、旧仮名口語テキストを直接扱わなければならない場合は多い。

すでに電子化されたデータのうち、大規模なものとしては、国立国語研究所で構築された『太陽コーパス』[5]に含まれる口語記事がある。『太陽コーパス』は約 3,400 記事、約 1445 万文字という大規模なテキストからなるが、このうちの約半分が旧仮名遣いの口語記事である。このほか、文学作品

の多くも原典は旧仮名口語で書かれており、旧仮名遣いのまま電子化されたものも少なくない。「青空文庫」[6]に収録された作品の中でも、たとえば芥川龍之介の作品では 374 作品中 209 作品が旧仮名遣いであり、青空文庫における近代の作家のテキストのかなりの部分を旧仮名口語テキストが占めている。

一般的な現代語の文章は現代仮名遣いの口語文であるから、旧仮名口語との違いは、基本的には仮名遣いという表記法の一部分に過ぎないはずである。しかし、旧仮名遣いが主に利用された時代は、すでに 60 年以上前のことである。そのため、対象となる旧仮名口語テキストは、現代語テキストと比べた場合、語彙や文法など表記以外の点でも違いを生じている。特に、太陽コーパスの口語文のように明治期の口語資料となると、語彙的には近代文語 UniDic で利用されるような、現代語では用いられない古い見出し語が必要とされる。

また、表記の違いにおいても、仮名遣いの違いだけでなく、使用される漢字の面でも大きな違いがある。旧仮名口語文では、単に対応する旧漢字が利用されるだけでなく、今日では用いられない幅広い漢字が利用される傾向にある。たとえば、次に示すのは近代文語 UniDic の見出し語として登録されている「見る」の表記（書字形）である。

「*みる」「*見る」「*看る」「*観る」「*視る」「*診る」「*看護る」「試る」「睹る」「瞥る」「覧る」「瞰る」「瞻る」「覩る」「覽る」「觀る」「閲る」

このうち、現代語コーパスで用いられているのはアスタリスクを付けた最初の 7 書字形だけであり、残りは近代語用として追加されたものである。この 17 書字形のうち、太陽コーパスでは新字形である「観る」「覧る」と「看護る」を除く全てが実際に出現しており、さらにその全てが口語記事にも現れている。このように、旧仮名口語では多様な書字形が現れる可能性がある。

３．形態素解析辞書「旧仮名口語 UniDic」の作成

旧仮名口語テキストに形態素解析を施す場合、現代語用の辞書を用いると仮名遣いが異なる平仮名が現れる部分で解析に失敗することになる。一方、近代語の文語文を対象として開発された「近代文語 UniDic」を利用すれば、こうした旧仮名遣いや古い時代特有の語彙には比較的良く対応できる。しかし、近代文語 UniDic は文語文を対象として開発されたものであるため、口語特有の表現の解析に失敗する場合が少なくない。たとえば、活用語の場合には文語活用の語として登録されているために、口語の音便形に対応できない場合があるほか、口語の助動詞などが正しく解析できない場合がある。また、UniDic では口語文法と文語文法に対応して、動詞の活用型を区別しているが、近代文語 UniDic では文語形を優先して利用するため、現代語の解析結果とは異なる結果を返すという問題もある。

このような問題に対処するため、今回開発した新しい形態素解析辞書「旧仮名口語 UniDic」では、現代語用 UniDic と近代文語 UniDic のフルセットの見出し語を利用し、さらに不足する表記（書字形）を追加した。そして、現代語用 UniDic と近代文語 UniDic の学習用コーパスから必要と考えられるコーパスを流用し、新たなコーパスを加えて生起コスト・連接コストの学習を行った。形態素解析器には「MeCab」[7]を利用しており、コスト学習のアルゴリズムの面では既存のものと変わりはない。

見出し語の拡充

見出し語には、現代語用の最新の UniDic の見出し語に加え、近代文語 UniDic の見出し語を全て利用した。先述したとおり、旧仮名口語文は古い時代のテキストであるため、近代語特有の表記・語彙を含むためである。これに加えて、次に示す活用形の問題に対処するために、活用表を整備して見出し語の拡充を行った。

UniDic において、たとえば動詞「買う」の場合には、活用型が口語では「五段-ワア行」、文語では「四段-ハ行」のように区別されている。そして、口語では終止形「買う」、文語では終止形「買ふ」の形だけが見出し語として登録されていた。そのため、旧仮名口語文で「買ふ」が現れる場合、文語形である「四段-ハ行」として解析されてしまう。しかし、旧仮名遣いであるからといって口語形であることには変わりないわけであるから、他と同じ「五段」活用として認定されることが望ましい。

そこで、「買ふ」「買ひ」「買へ」（「かふ」「かひ」「かへ」）の活用形（書字形）を、口語の「五段-ワア行」の動詞から派生させるように活用表を拡充した。このとき、「買ふ」のようにハ行に活用しているものを「ハ行」とせず「ワア行」とすることの是非が問題になるが、ここでは「買う」

が仮名遣いにかかわらず同語として容易に抽出できることと、現代語コーパスとの互換性を優先して、これらも「ワア行」の一活用形とした。

また、旧仮名口語文が必ずしも歴史的仮名遣いによらないことを踏まえ、連用形ウ音便では「買う」（「かう」「こう」）だけでなく、「買ふ」（「かふ」「こふ」）も派生させている。特にウ音便で「ふ」表記がなされる頻度が高いためである。さらに、促音の「っ」が小書きされないことに対応するため「買つ」（「かつ」）を促音便形として派生させた。

これにより、口語活用の「買う」全体で表 1 のように多数の活用形表記形を持つこととなった。「新規追加分」に○印を付けた行が、新たに追加した活用形である。ここでいう意志推量形とは、未然形に助動詞「う」が付いた形を指す。音変化で「う」を切り出せない場合が少なくないことから UniDic では活用形の一つとして扱っている。

表 1　活用表の例（動詞：五段-ワア行）

活用形	活用語形	活用書字形	新規追加分
未然形-一般	カワ	買は	○
未然形-一般	カワ	買わ	
連用形-一般	カイ	買ひ	○
連用形-一般	カイ	買い	
連用形-ウ音便	コウ	買う	
連用形-ウ音便	コウ	買ふ	○
連用形-促音便	カッ	買っ	
連用形-促音便	カッ	買つ	○
終止形-一般	カウ	買ふ	○
終止形-一般	カウ	買う	
連体形-一般	カウ	買ふ	○
連体形-一般	カウ	買う	
仮定形-一般	カエ	買へ	○
仮定形-一般	カエ	買え	
命令形	カエ	買へ	○
命令形	カエ	買え	
意志推量形	カオウ	買はう	○
意志推量形	カオウ	買おう	
意志推量形	カオッ	買おっ	
意志推量形	カオッ	買おつ	○
意志推量形	カオ	買お	

なお、意志推量形のウ音便形については、このほかにも「買はふ」「買おふ」「買わう」などの形も考えられるが、現時点では追加していない。今後、実際の使用頻度などを踏まえて追加を検討する予定である（形容詞では追加を行っている）。

学習用コーパスの追加

先述したとおり、旧仮名口語は、現代語と比べて単に表記が違うだけでなく、語彙や文法の面でも違いが見られる。したがって、本来であれば専用の学習用コーパスを大量に用意して、コスト学習を行うことが望ましい。しかし、コーパスを人手で修正して整備するためには多大なコストを要するため、旧仮名口語の大量の学習用コーパスを用意することは困難である。そこで、現代語用の UniDic のために整備されたコーパスと近代文語用に整備されたコーパスを一部用いつつ、これに旧仮名口語専用の学習用コーパスを加えて MeCab による機械学習を行うこととした。

このために、後述する評価用データとあわせて約 81,400 語分の旧仮名口語テキストに対して UniDic による形態素解析を施したのちこれに人手による修正を加えて、正解となるコーパスを作成した。このうち、表 2 に示す約 58,000 語のコーパスを学習に利用した。テキストは、青空文庫と太陽コーパスから選定した。

現代語のコーパスは、全てのコーパスを学習に利用するのではなく、ターゲットとなる文体と違いが大きく、不要と思われるものは除くこととした。そのため、Web のブログと掲示板のデータは利用していない。その結果、書籍を中心に約 2,116,400 語を学習に利用することとなった。近代語のコー

パスからは、文体的に口語に近いもの、文語と口語の両方を含むものなどを中心に約 144,500 語分を選定して学習に利用した。

表 2　学習用コーパス（旧仮名口語分）

出典	タイトル（著者）	語数
青空文庫	蟲の聲（永井荷風）	2283
青空文庫	計畫（平出修）	9455
青空文庫	運動会の風景（葉山嘉樹）	1231
太陽コーパス	192501-07_近代兵器の進歩並に将来の趨勢	5753
太陽コーパス	192501-114_日本の記念切手と時価	3878
太陽コーパス	192501-116_漫画小説　握り損ねた玉	1658
太陽コーパス	192501-122_卓上私語	649
太陽コーパス	192501-12_鼻で見、指で聞く少女	2700
太陽コーパス	192501-13_政界太平記	2066
太陽コーパス	192501-15_歴代の総理大臣（一）	1826
太陽コーパス	192501-20_最近に於ける飛行機の発達	6130
太陽コーパス	192501-28_貸金庫とはドンなものか	2369
太陽コーパス	192501-40_予の実験したる熱湯浴若返法	2367
太陽コーパス	192501-57_最近X光線療法の進歩	5574
太陽コーパス	192501-65_長篇小説　蛇人（第一回）	4786
太陽コーパス	192501-72_世界的の大発明として推称すべきゴ氏の汚物焼却炉	2745
太陽コーパス	192501-78_放送無線電話の沿革と現状及其将来	2568
計		58038

※太陽コーパスを出典とするデータのタイトルの数字は、たとえば「192501-07_」の場合、1925 年の第 1 号にある 7 番目の記事であることを示す。

なお、この近代語・現代語のコーパスは、入手可能なデータから旧仮名口語テキストに近いと思われるものを選定しただけであり、必ずしも最適なバランスを考慮していない。また、旧仮名口語専用の学習用コーパスは、単に学習用に追加するだけでなく、効果的な手法を取り入れることが望ましいが、今回は単純な利用にとどまっている。

４．「旧仮名口語 UniDic」の解析精度

このようにして作成した「旧仮名口語 UniDic」と MeCab（ver. 0.993）で旧仮名口語テキストを解析した場合の精度を調査した。比較対象の UniDic は現時点での最新の公開版（現代語用は ver.1.3.12、近代文語は ver.1.2 ）を利用している。

解析精度の評価には、表 3 に示す約 23,400 語の人手修正済みコーパスを利用した。学習用コーパスと同様、評価用コーパスも青空文庫と太陽コーパスから選定したものである。

表 3　評価用コーパス

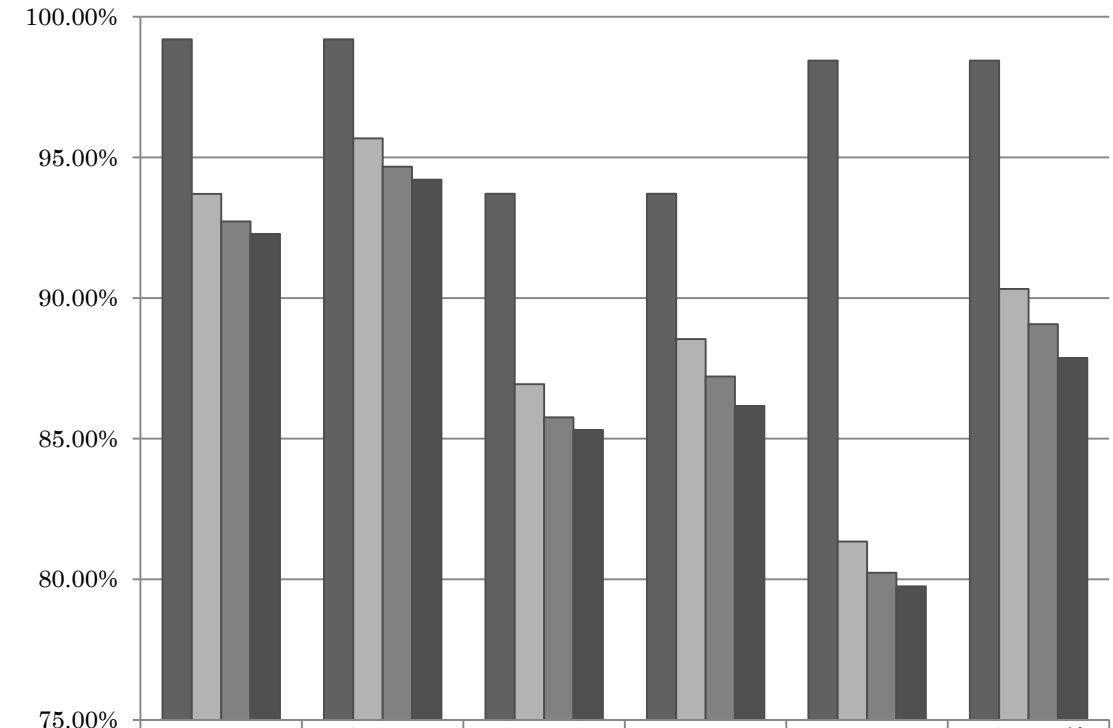
出典	タイトル（著者）	語数
青空文庫	井戸の底に埃の溜つた話（葉山嘉樹）	1331
青空文庫	幽霊の足（相馬御風）	647
青空文庫	硯友社の沿革（尾崎紅葉）	7470
太陽コーパス	192501-02_近代文明と発明	4989
太陽コーパス	192501-16_現代の女性美	2296
太陽コーパス	192501-55_帝都の復興に際して偉人星亨氏を想ふ	2364
太陽コーパス	192501-75_麻疹の予防の急務	4342
計		23439

解析精度

図 1 は、評価用コーパス全体を対象に各種の UniDic で解析した結果の精度をグラフにしたものである。数字は F 値（適合率と再現率の調和平均）をパーセント表示した値である。各辞書の評価区分で、

「L1 境界」としたのは、単語境界の認定が正しく行われているかを指す。「L2 品詞」は、L1 に加えて品詞の認定が正しいかどうか、「L3 語彙素」は、L1，L2 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。「L4 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、L1～L3 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。数字が大きいほど評価基準が厳しくなっている。

それぞれの辞書による通常の解析結果に加え、今回は解析結果を補正した結果の評価も調査した。これは、活用型が文語であるか口語であるかの違いによって誤りとされる例が多いため、この違いを無視した場合にどの程度の解析精度が出るかを調査したものである。たとえば「書く」は口語では「五段-カ行」だが、文語では「四段-カ行」で定義されている。そのため、正解が「五段-カ行」である場合に「四段-カ行」を出力すると誤りと見なされる。しかし両者は特に旧仮名遣いの場合には区別が困難であるだけでなく、本質的には同語であるといってよい。また、出力結果を置換することで容易に変換することもできる。そこで、こうした口語・文語の活用型の対については、いずれを出力した場合にも正解と見なすことにしたものが補正した値で、「旧仮名口語補正」のように各辞書の名前に「補正」を付けたものである。具体的には、「文語形容詞-ク」と「文語形容詞-シク」を「形容詞」と同一視したほか、「文語四段」は「五段」、「文語サ行変格」は「サ行変格」、「文語下二段」は「下一段」、「文語上二段」は「上一段」と同一視している。このほか、文語の「-ハ行」「-ワ行」を「-ア行」と同一視する処理も行っている。これにより、全ての辞書で補正版のほうでは精度が上がることになる。



	旧仮名口語 UniDic	旧仮名口語 補正	(現代語) UniDic	(現代語)補正	近代文語 UniDic	近代文語補正
L1境界	99.20%	99.20%	93.71%	93.71%	98.44%	98.44%
L2品詞	93.70%	95.67%	86.93%	88.54%	81.34%	90.32%
L3語彙素	92.73%	94.67%	85.76%	87.21%	80.23%	89.07%
L4発音形	92.28%	94.21%	85.31%	86.16%	79.75%	87.87%

図 1　解析精度比較（全体）

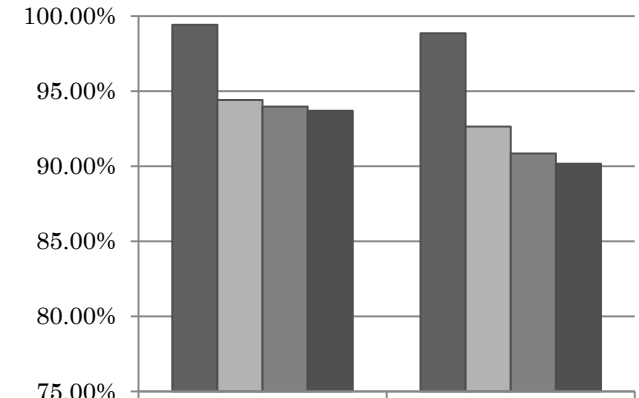
図 1 から分かるとおり、補正の有無にかかわらず、旧仮名口語 UniDic の精度が他の辞書を大きく上回っている。UniDic の評価精度で基準として用いてきた「L3 語彙素」では、補正なしの場合、現代語用 UniDic に対して約 7%ポイント差、近代文語 UniDic に対しては約 12%ポイント差以上に及ぶ。

補正を行った場合でも、現代語用 UniDic に対して約 7%ポイント以上、近代文語 UniDic に対しては約 5%ポイント以上の差を付けている。補正した場合にも他の辞書を大きく上回っていることから、単に活用型の文語・口語選択の選択によって精度が向上しているだけではなく、全体として解析がうまくいっていることが分かる。

旧仮名口語 UniDic の解析精度は、「L3 語彙素」では、補正なしで 92.7%、補正ありで 94.7%程度となっており、この精度は現代語の UniDic や近代文語 UniDic 等と比べるとやや低い。活用型以外における口語文法・文語文法による揺れが存在することが影響している可能性がある。また、このような基準の難しさもあり、コーパスの人手修正が万全でないために、コーパスの方に誤りが残っていたため誤りと誤判定された可能性も否定できない。

現代語と近代文語を比較すると、補正なしでは現代語用 UniDic が近代文語 UniDic を上回っているのに対し、補正を行うことによって近代文語 UniDic が現代語用の UniDic を上回るようになることが注目される。

図 2 は、コーパスを青空文庫由来のものと太陽コーパス由来のものに分けて、旧仮名口語 UniDic の精度を比較したものである。ここでは、補正なしの数値のみを掲出した。全体に太陽コーパスよりも青空文庫のテキストで解析精度が低くなっている。



	太陽コーパス	青空文庫
L1境界	99.42%	98.87%
L2品詞	94.42%	92.65%
L3語彙素	93.99%	90.86%
L4発音形	93.71%	90.17%

図 2　コーパス別の解析精度比較（旧仮名口語 UniDic 補正なし）

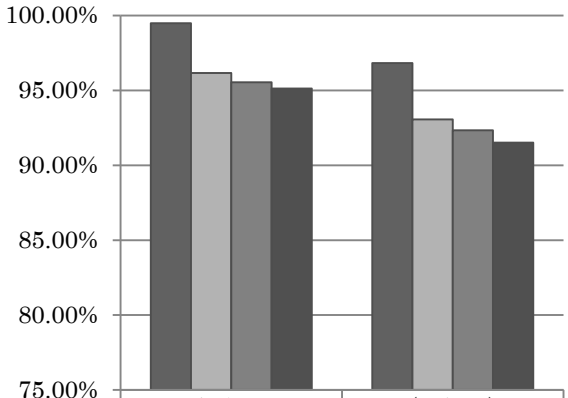
現代仮名遣いと旧仮名遣いが混在するテキストの解析精度

「旧仮名口語 UniDic」の実際の利用を考えた場合、青空文庫の多数のテキストを一括して形態素解析する場合など、現代仮名遣いと旧仮名遣いのテキストが混在しているデータに対して適用することが少なくないと考えられる。そこで、現代語用 UniDic と旧仮名口語 UniDic の二つの辞書について、現代語のテキストと旧仮名口語のテキストをほぼ同じ分量含めたテキストを対象として、仮名遣いが混在する場合の精度調査を行った。

現代語 UniDic のコスト学習に全ての人手修正済みの現代語コーパスを利用しているため、評価専用の現代語コーパスが入手できないため、今回は、辞書の学習に利用した、BCCWJ の書籍コアデータから約 25,000 語を評価に使用した。このデータは現代語用 UniDic と旧仮名口語 UniDic の両方で学習に利用したものである。評価コーパスは、旧仮名口語と現代語の書籍コアデータを合わせ、全体で約 48,400 語である。結果を図 3 に示す。数値はいずれも補正なしのものである。

図 3 から明らかなとおり、現代仮名遣いと旧仮名遣いが混在するテキストの場合であっても旧仮名口語 UniDic が現代語用の UniDic の解析精度を上回っている。実は、現代語のテキストについても旧仮名口語 UniDic は高い解析精度を示し、書籍コアデータの解析精度では現代語用の UniDic とほぼ同等の解析精度を示している。そのため、混在する場合にも旧仮名口語におけるアドバンテージがほぼそのまま反映される結果となっている。

現代仮名遣いテキストの評価データが、学習に利用されたデータであるため、今後より精確な調査が望まれるが、旧仮名口語 UniDic が、仮名遣いが混在する場合にも有効であることは確かである。



	旧仮名口語 UniDic	(現代語) UniDic
L1境界	99.49%	96.82%
L2品詞	96.17%	93.07%
L3語彙素	95.54%	92.34%
L4発音形	95.13%	91.52%

図 3　現代仮名遣いと旧仮名遣いが混在するテキストの解析精度

５．エラー分析

精度評価を踏まえ、旧仮名口語 UniDic ではどのような場合に誤りが多いのかエラーを調査した。

「L1 境界」のエラーでは、活用形が整備不足により促音便が「つ」で表記される場合が目立つ。また、連体詞「その」「この」などを代名詞「そ」「こ」と格助詞「の」に誤ったものが多い。これは、近代文語と現代語とで単語認定基準にずれがあるために起きていることであり、学習用コーパスを事前に現代語側の基準にあわせておくことで避けられる問題である。その他は同表記の語が現れやすいという各見出し語固有の事情によるものが多い。このように L1 のエラーは、今後の対応で改善が見込めるものが多かった。

「L2 品詞」で問題となるエラーは、連体形と終止形を誤る例、未然形と連用形を誤る例が非常に多かった。この中には評価コーパス側に問題があった例も含まれている可能性がある。「L3 語彙素」で問題となるエラーは、同表記となる語がある「今日（きょう）」と「今日（こんにち）」、「昨日（きのう）」と「昨日（さくじつ）」、「門（もん）」と「門（かど）」などの対が目立った。「L4 発音形」で問題となるエラーは、「一（いち）」と「一（いっ）」などの数詞の誤りが目立つ。

全体として、L3・L4 のエラーは避けがたいものが多いが、L1・L2 のエラーは今後の対策で改善できる可能性がある。特に L2 で大きく精度を落としていることから、活用形の問題を中心に対策を行いたいと考えている。

６．おわりに

国立国語研究所では、共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘）の下で、日本語の通時コーパスの設計と実装に関する研究を行っている。通時コーパスの構築のためには、各時代の各種の日本語テキストの形態素解析を可能にすることが必要であり、そのための辞書開発を行っている[8]。旧仮名遣いの口語文の解析も近代以降のコーパス開発に必要とされるものである。今後、「旧仮名口語 UniDic」の整備を進めてコーパス開発に利用するとともに、一般公開を行う予定である。

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 号 pp.101-122
[2]「UniDic ダウンロードサイト」 http://download.unidic.org
[3] 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
[4]「近代文語 UniDic」 http://www2.ninjal.ac.jp/lrc/index.php?UniDic
[5] 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
[6]「青空文庫」 http://www.aozora.gr.jp/
[7]「MeCab: Yet Another Part-of-Speech and Morphological Analyzer」 http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[8] 小木曽智信（2011）「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」『情報処理学会研究報告. 人文科学とコンピュータ研究会報告』 2011-CH-92(6), pp.1-4