2B1-2

表記のバリエーションを考慮した近代日本語の形態素解析

Handling Orthographic Variations in Morphological Analysis for Near-Modern Japanese

岡 照晃[*1] 小町 守[*2] 小木曽 智信[*1*3] 松本 裕治[*1]
Teruaki Oka Mamoru Komachi Toshinobu Ogiso Yuji Matsumoto

[*1]奈良先端科学技術大学院大学 Nara Institute of Science and Technology
[*2]首都大学東京 Tokyo Metropolitan University
[*3]国立国語研究所 National Institute for Japanese Language and Linguistics

National Institute for Japanese Language and Linguistics creates several morphological analysis dictionaries for historical texts, such as Kindai-bungo UniDic, which achieves high performance on analysis for texts written in near-modern literary Japanese language. However, the performance of morphological analyzers using the dictionaries deteriorates if the text is not normalized, because these dictionaries often lack orthographic variations such as mark-lacking characters. In this paper, we propose a morphological analysis method for historical texts which contain orthographic variations. First, we rewrote existing surface-forms in Kindai-bungo UniDic with some handcrafted rules and added them to the dictionary. Second, we implemented a dynamic expansion algorithm of orthographic variations for building a word lattice. We compared our method to the existing method on morphological analysis on Meiroku-Zasshi corpus, which is a corpus of near-modern literary Japanese language. As a result, about 18% of unknown words in the previous method was covered by the proposed method. We also confirmed that our method outperforms the existing method in accuracy of morphological analysis.

1. はじめに

現在，国立国語研究所では古代から近世までの大規模通時コーパス（**日本語歴史コーパス（CHJ）**）の整備が進められている [近藤, 12]．今年 3 月には平安時代の仮名文学作品を収録した「CHJ 平安時代編」の先行版が公開された [小木曽ら, 13]．また同研究所は通時コーパスと現代語のコーパスの間をつなぐ近代語コーパスの整備も行なっている．昨年 10 月にはそのモデルケースとして，明治期の雑誌「明六雑誌」をコーパス化した**明六雑誌コーパス**を公開した [近藤ら, 12]．

CHJ や明六雑誌コーパスの特徴の 1 つに形態素解析辞書 **UniDic**[伝ら, 07] の設計に基づく形態論情報が付与されていることが挙げられる．ただし古い時代の文献資料（**歴史的資料**）への形態論情報の人手付与は非常にコストが高い．そのためコーパス整備の現場では，予め形態素解析器で自動解析を行い，自動解析に失敗した箇所を人手でチェックして修正していく，という作業方針が採られている．形態素解析器には MeCab[Kudo et al., 04] が用いられ，形態素解析辞書は国語研で開発された歴史的資料専用の辞書 **近代文語 UniDic**[小木曽ら, 08] や **中古和文 UniDic**[小木曽ら, 10] 等が使用されている．しかし，これらの辞書は主として校訂済み資料の解析を念頭に単語登録が行われているため，未整備の資料の解析に使用した場合，解析結果の精度が低いという問題がある．これは未整備資料の中に校訂済み資料では現れないような**表記のバリエーション**（表記上の不徹底）が非常に多く含まれるためである[*1]．

例えば，未整備の資料の中には「及ひ（オヨビ）」の「ひ（ビ）」ように，濁点が付いていることが期待されるのに濁点の付いていない文字（**濁点無表記文字**）が多く含まれる．実際に明六雑誌第 1 号で調査を行った結果，濁点が期待される仮名文字の約 83%（368/446）が濁点無表記で書かれていた．この他にも表 1 に示したように，仮名遣や送り仮名が一貫していなかったり，省略記号である踊字を使った省略表記も頻繁に利用される．漢字片仮名交じり文で書かれた資料も多い．

表 1: 未整備歴史的資料に含まれる表記のバリエーション

	例
濁点無表記	及ひ（オヨビ）
仮名遣の不統一	用い（モチイ），用ひ（モチイ），用ゐ（モチイ）
送り仮名の不統一	限り，限ぎり，限（カギリ）
踊字による省略	こゝ（ココ），とゝまり（トドマリ），たゞ（タダ），及ば/ゝ（オヨバ/バ），出で/ゝ（イデ/テ），愉々快々（ユカイユカイ），恐る々々（オソルオソル），民主/々義（ミンシュ/シュギ），給は/ゝ（タマワ/バ），各ゝ（オノオノ），まに／＼（マニマニ），返す／＼（カエスガエス）
漢字片仮名交じり文	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

辞書に単語を登録する際，表記のバリエーションを考慮しておかないと，既知語であっても表 1 のような表記により未知語となってしまう．実際，明六雑誌コーパス中の単語表層形で近代文語 UniDic v1.1 に含まれていないものをランダムに 100 件重複なしに取り出して確認したところ，12 件は表 1 に示した表記のバリエーションによって未知語になったもので，表記さえ整えれば既に辞書に登録済みであることが分かった．

そこで本論文では，既存の辞書に登録されていない表記のバリエーションにも対応した形態素解析手法について述べる．表記のバリエーションとして，今回は [小木曽ら, 08] で問題として上げられた表 1 の 5 種類を扱った．提案手法は基本的に従来手法である UniDic を形態素解析辞書とした MeCab による解析と同じである．異なるのは，コスト最小の単語分割を探索する際に，表記のバリエーションを考慮した単語もラティスのノードに追加する点である．これは辞書ベースの形態素解析の異表記対応として素直な方法であり，[勝木ら, 11] でも同様の手法を用いて Web 上の崩れた表記に対応した形態素解析を実現している．しかし，歴史的資料中の表記のバリエーションへの適用はこれまでになかった試みである．

連絡先: 岡 照晃，奈良先端科学技術大学院大学 情報科学研究科，奈良県生駒市高山町 8916-5，teruaki-o@is.naist.jp

*1 これは日本語の正書法が公的に整えられ始めたのが明治時代以降であることに由来する．

2. 従来手法: 近代文語 UniDic v1.1 における表記のバリエーションへの対処

近代文語 UniDic や中古和文 UniDic は基本的に，校訂済み資料の解析を念頭に開発されている．しかし [小木曽ら, 08] によると，近代文語 UniDic では，以下のように濁点無表記などの表記のバリエーションに対してもある程度対策がなされていた．

濁点無表記，仮名遣の不統一，送り仮名の不統一： 表記のバリエーションを考慮した表層形を辞書に登録することで対処[*2]．

踊字による省略： 直前の文字列の繰り返しを表す踊字は現代語では**同字点**（々）が用いられるだけである．これに対し，歴史的資料の中では**一字点**（ゝ，ゞ，ヽ，ヾ），**二字点**（ゟ），**くの字点**（／＼，／″＼）も出現する．踊字についても基本的には辞書登録で対応する．ただし，一字点は大体において一字点列の直前の仮名 1 文字の繰り返しを表すため，**展開**（省略前の表記への置き換え）を以下のような単純なルールで自動的に実施することができる．そのため，一字点に関しては前処理で対処する．

a. **濁点なし一字点**：当該一字点の直前の文字が濁点の付いた仮名文字であった場合，その文字から濁点を外し，当該一字点と置き換える（e.g., 出でゝ→出でて）．それ以外の場合，当該一字点を直前の文字に置き換える（e.g., こゝ→ここ）．

b. **濁点付き一字点**：当該一字点の直前の文字が濁点の付き得る仮名文字であった場合，その文字に濁点を付与し，当該一字点と置き換える（e.g., たヾ→ただ）．それ以外の場合，当該一字点を直前の文字に置き換える（e.g., 御出でゞすか→御出でですか）．

漢字片仮名交じり文： カタカナ文字をすべて平仮名文字に置き換える前処理で対処．

上に示したように，近代文語 UniDic v1.1 には濁点無表記，仮名遣・送り仮名の不統一，踊字による省略を考慮した表層形も登録されている．しかし，辞書の煩雑化・肥大化を避けるため，登録されているのは実際に用例として採集された表記のみである．つまり，想定され得る全てのバリエーションを辞書中に網羅しているわけではなく，未知の表記には一切対処できないという問題がある．また一字点も上のようなルールでは誤った文字に置換してしまう恐れがある．例えば未整備資料の中では「思はゝ（オモワバ）」や「及ばゝ（オヨババ）」のように，一字点が濁点無表記で記述されている場合も多い．つまり，未整備資料での濁点文字直後の「ゝ」「ヽ」の置き換えには，直前の仮名文字から濁点を外した文字だけでなく，直前の仮名文字も候補として存在する．また同様に，「た」のように濁点の付き得る仮名文字直後の「ゝ」「ヽ」の置き換えにも，直前の文字に濁点を付けた文字が候補として存在する．片仮名を平仮名に置き換える前処理も，単純に片仮名文字を全て平仮名文字に置換してしまうと，「ゴツト」や「トーマス」など平仮名では表記しないような外来語まで平仮名表記になってしまうという問題がある．実際，漢字片仮名交じり文で書かれた明六雑誌中にはカタカナのままで表記されるべき箇所が 509 件含まれていた．

提案手法でも，濁点無表記，仮名遣・送り仮名の不統一には，辞書登録で対処する．ただし，用例に基づいた表記に限らず，既存の表層形をルールベースで自動的に書き換え生成した表層形を網羅的に追加する．また，踊字による省略や漢字片仮名交じり文には，前処理や辞書登録でなく，デコード時に動的に入力文字列を書き換えることで対処する．これにより前処理時に誤った文字に置換する問題を解消できる．

*2 近代文語 UniDic v1.2 からは濁点無表記の登録が中止されている．

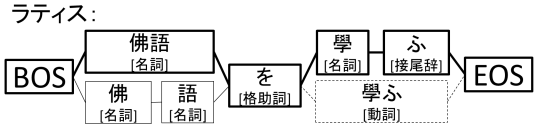

図 1: ラティス構造の例

3. 提案手法：表記のバリエーションを考慮した形態素解析

提案手法の基本的な部分は従来手法の UniDic を辞書とした MeCab による形態素解析と同じである．異なるのは，コスト最小の単語列を探索する際に表記のバリエーションを考慮し，辞書に未登録の表層形もラティスのノードへと追加する点である．

MeCab+UniDic による形態素解析は通常，以下のような手順で行われる．

手順 1 入力された文を先頭から 1 文字ずつ読み進め，辞書引き（辞書に登録されている表層形とのマッチング）により各位置から開始する単語を列挙する．

手順 2 手順 1 で列挙した単語からグラフ構造 (ラティス) を作成する．

手順 3 文として最も確からしい単語の並びをラティス上のコスト最小のパスとして出力する．コストは予め CRF[Lafferty et al., 01] を使って求められ，UniDic 中に記述されている．

例えば，「佛語を學ふ」という文が入力された場合，図 1 に示すラティスが作られ，従来手法では，最終的に太線で記されている単語列が出力される．点線部分は，提案手法によって追加されるパスである．通常，手順 1 では辞書に登録されている表層形のみが列挙されるが，提案手法では，以下の方法で候補の追加を行う．

送り仮名の不統一，濁点無表記，仮名遣の不統一： 辞書に新たに表層形を追加することで対応．追加する表層形は，以下の方法で辞書に既存の表層形から自動的に生成する．

送り仮名の不統一： 漢字直後の平仮名を読み飛ばすことで，送り仮名が縮退した表層形を生成する（e.g., 基づく→基く）．ただし，「限（カギリ）」や「定（サダム）」など，送り仮名が漢字に完全に引っ込んでいるような表層形を登録してしまうと，「限定」のような二字熟語や四字熟語が「限/定（カギリサダム）」のように細かく分割され過ぎるようになる．そのため「限り」の「り」のような表層形最後尾の平仮名文字の読み飛ばしは行わないことにした．また送り仮名の飛び出した表層形は，漢字読みの平仮名表記末尾を漢字と既存の送り仮名との間に挿入することで生成する（e.g., 志し→志ざし）．漢字読みの仮名表記は動的計画法を使って辞書中の表層形と仮名の対応付けをとることで得た．表層形を生成する際，送り仮名の縮退は最大で 1 文字，飛び出しは最大で 2 文字までとした．

濁点無表記，仮名遣の不統一： 送り仮名の伸縮を考慮した表層形へ表 2 の文字列書き換えルールを適用し，濁

い→ひ, い→ゐ, う→ふ, う→ゆ, え→へ, え→ゑ, お→を, おほ→あふ, かう→こう, が→か, ぎ→き, ぐ→く, げ→け, ご→こ, さう→そう, さう→そふ, ざ→さ, じ→し, じ→ち, じ→ぢ, ず→す, ず→つ, ず→づ, せう→しゃう, せう→しやう, せう→しょう, せう→しよう, ぜ→せ, そう→さう, そふ→さう, ぞ→そ, たう→とう, たう→とふ, たふ→とう, だ→た, ちゃう→てう, ちやう→てう, ちゅう→ちう, ちゆう→ちう, ぢ→し, ぢ→じ, ぢ→ち, つ→っ, づ→す, づ→ず, づ→つ, で→て, とう→たう, ど→と, なう→のう, なう→のふ, のふ→なふ, は→わ, はう→ほう, ば→は, ばう→ほう, ばう→ぼう, ひ→い, ひ→ゐ, び→ひ, ふ→う, ふ→ほ, ふ→ゆ, ふ→を, ぶ→ふ, へ→え, へ→ゑ, べ→へ, ほ→う, ほ→ふ, ほ→を, ぼ→ほ, ぼふ→はふ, ぼふ→ばふ, まう→もう, まう→もふ, まふ→もう, やう→よう, やう→よふ, ゆ→う, ゆ→ふ, よう→やう, よう→やふ, らう→ろう, れう→りゃう, れう→りやう, ろふ→らう, ろふ→らふ, わ→は, ゐ→い, ゐ→ひ, ゑ→え, ゑ→へ, を→お, を→ほ, ん→む, ゞ→ゝ, 々→／＼, 々→〳〵, 〳〵→〳〵

図 2: 書き換えルール（辞書登録表記→未整備資料中の表記）

点無表記，仮名遣の不統一を考慮した表層形を網羅的に生成する．このルールは [小木曽, 02] の仮名遣正誤表に，濁点文字を濁点無表記文字へ置き換えるルール（e.g., だ→た）を追加し，さらに経験的に設定したルール（e.g., っ→つ，々→／＼）を数個加えたものである．

踊字による省略： 踊字も濁点無表記などと同じく網羅的に登録することも可能だが，長さ 1 文字の語を「ゝ」や「々」で登録してしまうと，辞書が煩雑になるだけでなく，コスト最小のパスを求める際の曖昧性も増加する．そのため辞書に網羅せず，入力文を受け取った時に当該踊字直前の文字列を見て，動的に踊字の展開を実施することにした．

踊字の展開は以下のルール a〜d に従って行う．

a. **一字点**：当該一字点の直前が仮名文字であり，直後に一字点が現れていない場合に限り，次のルール (i) もしくは (ii) を適用する．

(i) **濁点なし一字点**：直前の文字が濁点文字であれば，当該一字点を直前の文字か，直前の文字から濁点を外した文字のいずれかに置換する（e.g., 及ばゝ→及ばば，出でゝ→出でて）．それ以外の場合，当該一字点を直前の文字に置き換る（e.g., こゝ→ここ）．

(ii) **濁点付き一字点**：直前の文字が濁点文字であれば，当該一字点を直前の文字に置き換える（e.g., 御出でゞすか→御出でですか）．濁点は付いていないが「た」のように濁点が付き得る文字の場合は，直前の文字に濁点を付与した文字に置換する（e.g., たゞ→ただ）．それ以外の場合，当該一字点を直前の文字に置き換える．

b. **同字点**：

(i) **同字点が連続しない場合**：当該同字点の直前の文字が漢字である場合，当該同字点を直前の漢字と置換するか（e.g., 民主々義→民主主義），もしくは同字点が単語先頭でない場合は当該同字点を読み飛ばす（e.g., 愉々快々→愉快）．

(ii) **同字点が連続する場合**：同字点列と同長の文字列が直前にあれば同字点列をその文字列と置き換える（e.g., 恐る々々→恐る恐る）．

c. **二字点**：二字点の用法は一字点の用法と同字点の用法を合わせたものであるため，置換ルールもその 2 つを合わせたものを使用する．

d. **くの字点**：当該くの字点が単語先頭でなく，直後にくの字点が現れない場合に限り，当該くの字点を読み飛ばす（e.g., 繰り返し／＼→繰り返し），もしくは当該くの字点を辞書引き中の文字列の先頭から当該くの字点直前までの文字列に置き換え（e.g., まに／＼→まにまに）．

上記の置換ルールは一意に適用できるものではない．そのため，各ルールをそれぞれ適応する場合とどれも適応しない場合，考え得る全ての可能性を試しながら辞書引きを実施する．

漢字片仮名交じり文： 漢字片仮名交じり文に対応するため，辞書引きにおいて片仮名を平仮名と同一視することとした．具体的には，辞書引きの際，片仮名文字を平仮名文字に置き換えた文字列でも辞書引きを行う．ただしこれも踊字と同じく一意には行わず，1 文字ずつ平仮名に置き換えた場合と置き換えない場合，考え得る全ての可能性で辞書引きを実施していく．

4. 形態素解析性能評価実験

提案手法の有効性を確認するため，[小木曽ら, 08] の手法と提案手法の比較実験を行なった．ここでは各手法でのコーパス中未知語数と，実際の形態素解析精度を比較した．評価時の単語同定基準は [小木曽ら, 08] に従い，表層形 or 境界，品詞，語彙素の 3 段階とした．ただし，品詞には品詞大分類〜細分類及び活用型・活用形を含めている．また語彙素は語彙素+語彙素読みとした．

評価用コーパスには，明六雑誌コーパス[*3]を利用した．これは明六雑誌コーパスが現在一般に公開されている歴史的資料に基づくコーパスの中で唯一 UniDic に基づく形態論情報が付けられ，XML 形式で未整備状態での表記も保持しているためである．明六雑誌コーパスはほぼすべての記事が文語体で書かれているため，形態素解析辞書には近代文語 UniDic v1.1 を使用する．評価のため，タグの情報を頼りに明六雑誌コーパスを未整備状態に戻し，口語を含まない文を抽出した（総文数 : 9,525 文中，9,139 文を取得）．抽出した文中の総単語数は 169,555 であった．異なり単語数は，単語同定基準の表層のみで 15,953，品詞までで 17,382，語彙素を含めて 17,669 であった．

4.1 実験設定： 従来手法

デコード時に動的に踊字を開く提案手法に対して，[小木曽ら, 08] では前処理としてあらかじめ一字点を展開する．今回は小木曽らの展開ルールに修正を加えた以下のルールで一字点の展開を実施した．

当該一字点の直前が仮名文字であり，直後に一字点が現れていない場合に限り，次のルール (i) もしくは (ii) を適用．

(i) **濁点なし一字点**：直前の文字が濁点文字であれば，直前の文字から濁点を外した文字に置き換える．それ以外の場合，直前の文字に置き換える．

(ii) **濁点付き一字点**：直前の文字が濁点文字であれば，直前の文字に置き換える．濁点は付いていないが「た」のように濁点が付き得る文字の場合は，直前の文字に濁点を付与した文字に置き換える．それ以外の場合，置き換えは行わない．

4.2 実験設定： 提案手法

近代文語 UniDic 中の表層形より，濁点無表記，仮名遣・送り仮名の不統一を考慮した表層形を生成し，辞書に追加登録した．この際，書字形出現形のフィールドは元のままとした．その結果，辞書登録単語数 887,800 が 5,745,175 と約 6.5 倍に増大した．

辞書中のコストは表層形を書き換える前と同一とした場合（再学習なし）と，既存のコストを破棄し，近代文語 UniDic

*3 http://www.ninjal.ac.jp/corpus_center/cmj/meiroku/meiro-ku_xml.zip（2013 年 3 月のデータ）

表 2: 明六雑誌コーパス内の未知語数の割合 (総単語数 : 169,555)

	従来手法		提案手法	
	のべ語数	異なり語数	のべ語数	異なり語数
表層形	3.9% (6,599)	25.4% (4,057)	3.2% (5,420)	21.7% (3,468)
品詞	6.7% (11,414)	29.1% (5,062)	6.0% (10,177)	26.0% (4,518)
語彙素	11.2% (19,013)	31.5% (5,557)	10.3% (17,402)	28.5% (5,036)

表 3: 形態素解析性能の比較

	従来手法			提案手法					
				再学習なし			再学習あり		
	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F
境界	91.23	**94.71**	92.94	**92.07**	94.35	**93.20**	91.67	94.33	92.98
品詞	85.42	88.68	87.02	85.38	87.49	86.42	**86.35**	**88.85**	**87.58**
語彙素	80.96	84.05	82.48	80.98	82.98	81.97	**81.97**	**84.35**	**83.14**

v1.1 の学習に使用されたコーパスで再度学習を行い直した場合（再学習あり）の 2 通りを用意した．再学習時には，CRF のハイパーパラメータ C(=σ^2) はデフォルトの 1.0 のままとした．また学習の際には辞書とコーパスの表層形の平仮名文字を全て片仮名文字に置き換え，さらにコーパスからは濁点をすべて取り除いた．

提案手法では片仮名を平仮名と同一視するため，片仮名の未知語処理は平仮名と同じとしている．

4.3 実験結果

各手法におけるコーパス内全単語中での未知語数の割合を表 2 に示す．表 2 を見ると，提案手法を用いることで従来手法よりも未知語数が減少したことが確認できる．のべ語数で見ると，表層形・品詞で 0.7%，語彙素で 0.9%の減少が見られた．異なり語数では，表層形で 3.7%，品詞で 3.1%，語彙素で 3.0%の減少であった．また従来手法で未知語であった語の内，提案手法で既知語になった割合は，のべ語数で見ると，表層形で 17.9%，品詞で 10.8%，語彙素で 8.5%であり，異なり語数では，表層形で 14.5%，品詞で 10.7%，語彙素で 9.4%であった．また，提案手法において表層の段階で未知語となった語を調査した．その結果，未知語として残っているのはほとんどが「壯勇」のような辞書に未登録の二字熟語や「エツキスピーヂエンシイ」のような外来語であった．しかし「衰微」や「纘ク（ツグ）」のように異体字によって未知語となっているものも含まれていた．

各手法の形態素解析性能を表 3 に示す．ここでは適合率 [%]・再現率 [%]・F 値を比較した．各値の式は [Kudo et al., 04] と同じものを使用している．結果として再学習を行なった場合において，境界を除き，従来手法よりも提案手法の方が高い適合率と再現率が得られた．境界は F 値の比較では提案手法の方がよくなったものの，再現率は従来手法の方が高かった．これは提案手法では表記のバリエーションを考慮したことで分割候補数が増え，曖昧性が向上したためだと考えられる．例えば，「他國ノ長ヲ」は「他國/ノ/長/ヲ」と分割されるべきであるが，提案手法では「ふ→を」というルールで「長ふ（ナゴー）」から生成された「長を」という単語が辞書中に存在したため，上手く分割が行えていなかった．

また，再学習を行なわなくとも，境界の適合率・F 値は従来手法よりも高くなることが確認できた．これは，未知語数が減ったことで過分割が抑制できたためだと考えられる．

5. おわりに

本論文では，表記のバリエーションに対応するため，辞書と辞書引きの拡張による形態素解析手法を提案した．明六雑誌コーパスを対象に評価実験を行ったところ，従来手法では未知語であった単語の約 18%が既知語となり，形態素解析性能も境界の再現率を除いて改善することが確認できた．

提案手法で新たに辞書に追加した単語は，表層形の欄だけを書き換え，書字形出現形の欄は書き換える前と同じにした．このため，再学習を行なった場合でも表記のバリエーションを考慮した単語としない単語のコストが同じになっている．しかし，例えば仮名遣の書き換えルールの基とした [小木曽, 02] の正誤表を見てみると，「正：ふ，誤：を」というペアの観測例はわずか 1 件しかない．これより「長を」という表層形は「長ふ」よりも出現コストを高く設定すべきと考えられる．また仮名遣以外にも，濁点の抜け落ちやすさや，送り仮名の伸縮しやすさなどを考慮して，それぞれのコストが別になるようにすることで，提案手法の再現率低下を防ぐ事が出来ると考えられる．

また，今回は扱わなかったが，未整備資料中に含まれる表記のバリエーションには他にも異体字がある．そのため，異体字を考慮することも今後の課題として挙げられる．

謝辞

本研究は，国立国語研究所の共同研究プロジェクト「統計と機械学習による日本語史研究」による研究成果の一部である．

参考文献

[Lafferty et al., 01] Lafferty, J., McCallum, A. and Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proc. the 18th International Conference on Machine Learning* (*ICML 2001*), pp. 282-289 (2001).

[Kudo et al., 04] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. the 2004 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2004*), pp.230-237 (2004).

[小木曽, 02] 小木曽智信：仮名遣いについて，雑誌「太陽」による確立期現代語の研究一「太陽コーパス」研究論文集，国立国語研究所報告 122，pp. 351-376，博文館新社 (2002).

[小木曽ら, 08] 小木曽智信，小椋秀樹，近藤明日子：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 (2008).

[小木曽ら, 10] 小木曽智信，小椋秀樹，田中牧郎ほか：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，Vol.2010-CH-85, No.4, pp.1-8 (2010).

[小木曽ら, 13] 小木曽智信，須永哲也，冨士池優美ほか：「日本語歴史コーパス 平安時代編」先行公開版について，第 3 回コーパス日本語学ワークショップ予稿集，pp.269-276 (2013).

[勝木ら, 11] 勝木健太，笹野遼平，河原大輔ほか：Web 上の多彩な言語表現バリエーションに対応した頑健な形態素解析，言語処理学会第 17 回年次大会発表論文集，pp.1003-1006 (2011).

[近藤ら, 12] 近藤明日子，小木曽智信，須永哲矢ほか：『明六雑誌コーパス』の開発 —近代語コーパスのモデルとして，第 2 回コーパス日本語学ワークショップ予稿集，pp.329-334 (2012).

[近藤, 12] 近藤泰弘：日本語通時コーパスの設計，NINJAL「通時コーパス」プロジェクト・Oxford VSARPS プロジェクト合同シンポジウム 通時コーパスと日本語史研究予稿集，pp.1-10 (2012).

[伝ら, 07] 伝康晴，小木曽智信，小椋秀樹ほか：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 (2007).