Vol.2014-NL-216 No.8
Vol.2014-SLP-101 No.8
2014/5/22

# 形態素解析との同時最適化による歴史的資料の自動表記整理

岡 照晃[1,a)]　松本 裕治[1,b)]

**概要：**
日本語の歴史的資料の中では表記規範が確立していないための表記揺れが著しい．そういった表記揺れを含んだ文は現代人にとって読み辛く，資料をコーパス化した際の検索性も低い．そのため，歴史コーパスを整備する際には原文表記を整える作業（表記整理）が実施される．しかしこの作業は人手のコストが非常に高い．そこで統計的機械学習を用いた自動表記整理として，単語辞書を用いた辞書ベースの手法と，周辺文字列の情報だけで表記整理を行う文字ベースの手法が提案されている．辞書ベースの手法は形態素解析と同時に表記整理を実施する．そのため表記整理時に単語境界や品詞の情報が利用できる．しかし学習に品詞タグ付きコーパスが必要であり，学習用コーパスが限られるという問題がある．一方，文字ベースの手法は学習に品詞タグ付きコーパスを必要としない．そのため学習用コーパスは辞書ベースの手法よりも多く確保できる．しかし表記整理時に単語境界や品詞の情報が使えないため，部分文字列にマッチする単語に引かれて誤った表記整理を行う問題がある．そこで本論文では，辞書ベースの手法と文字ベースの手法のそれぞれの欠点を互いの利点で補い合わせるために，2 つをハイブリッドした自動表記整理手法を提案する．提案手法は辞書ベースの表記整理と同様に形態素解析のフレームワークを利用するが，Augmented-Loss Training とよばれる学習アルゴリズムを採用することで，単語境界や品詞情報を持たない表記整理済みコーパスも学習に利用することができる．性能評価実験では，提案手法を用いることで，近代の雑誌「太陽」に対して F1 値 85.3 と，従来法（F1 値:74.8）よりも高い精度で表記整理が行えることが分かった．

## 1. はじめに

国立国語研究所では現在，古い時代の文献資料（**歴史的資料**）を基にした**歴史コーパス**の整備が進められている [15][17][18]．例えば，明治〜昭和初期の雑誌「太陽」をコーパス化した**太陽コーパス** [15] がその成果の 1 つである．

雑誌「太陽」のような歴史的資料の中では，表記規範が確立していないための表記揺れが著しい．例えば，歴史的資料の中には「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**）が多い．実際，太陽コーパスの文語記事の中では，濁音仮名文字（e.g., ガ，ザ，バ...）中の濁点無表記（e.g., か，さ，は...）使用率が通年で 3.3%，一番古い 1985 年の記事では 5.9%と報告されている [16]．濁点無表記の他にも，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現れる．漢字片仮名交じり文で書かれた資料も多い．**表 1** に歴史的資料中の代表的な表記揺れを示す．

**表 1**　未整備歴史的資料に含まれる代表的な表記揺れ．括弧内は読み，「/」は単語（短単位）境界をそれぞれ表している．

| | 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），<br>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），<br>たゞ（タダ），及ば/ゝ（オヨバ/バ），<br>出で/ゝ（イデ/テ），<br>恐る々々（オソルオソル），<br>民主/々義（ミンシュ/シュギ），<br>給は/ゟ（タマワ/バ），各ゟ（オノオノ）<br>まに／＼（マニマニ），<br>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 |

表 1 のような表記揺れを残したまま歴史的資料のコーパス化を行なった場合，コーパスユーザの可読性・検索性が損なわれる恐れがある．また単語分割・品詞付与のための自動形態素解析の結果の精度も低くなる．これは形態素解

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology, Takayama, Ikoma, Nara 630–0192, Japan
[a)] teruaki-o ｛at｝ is.naist.jp
[b)] matsu ｛at｝ is.naist.jp

Vol.2014-NL-216 No.8
Vol.2014-SLP-101 No.8
2014/5/22

析に用いられている単語辞書が，表記を人手で整えた後の文の解析を念頭に単語登録されていることが原因である．国語研で整備されている歴史コーパス [17][18] には単語境界や品詞のアノテーションも行われているが，人手の作業ではコストが高い．そのため，一度自動で形態素解析した結果を人手修正することでコストの削減を図っている．しかし，自動形態素解析の結果の精度が低いと，今度はその修正にかかるコストが高くなってしまう．

以上の理由から，歴史コーパス整備の際には，あらかじめ原文表記を人手で整える作業（**表記整理**）が行われる．例えば太陽コーパスでは，濁点無表記への濁点付与，仮名遣の統一，踊字の展開 [*1]が行われている．しかし人手による表記整理もコストは高く，以前より自動化を望む声が上がっていた．

そこで文献 [8] では，統計的機械学習を用いた濁点の自動付与手法を提案している．この手法は文字単位の識別学習を採用しており，各文字に対して独立に，濁点を付けるか否かの分類を実施していく（**文字ベースの表記整理**）．分類時の素性にも周囲の文字列の情報だけを使い，単語境界や品詞の情報は一切必要としない．これは，表記整理前の資料に対する自動形態素解析の結果の精度が低いこと．そして，歴史的資料には単語分割済みかつ品詞タグ付きのコーパス（以下，**品詞タグ付きコーパス**）は少ないが，太陽コーパスのような表記整理を行なっただけのコーパス（以下，**表記整理済みコーパス**）ならば大規模に利用できたためである．しかしながら，文字ベースの濁点付与では単語境界や品詞の情報を使用しないため，以下のように，部分文字列にマッチする単語に引かれ，誤った濁点付与を行うといった問題があった．

いなどいひて（正解：いな（否）といひて）

文献 [8] ではまた，辞書ベースの濁点付与手法も提案している（**辞書ベースの表記整理**）．この手法は単語辞書を用いた日本語形態素解析 [3]（以下，単に形態素解析）のフレームワークを利用したもので，単語ラティスを構築する際に，濁点無表記を考慮して辞書引き[*2]した単語もラティスへと追加する．これにより形態素解析結果として濁点付与後の文を得ることができる．辞書ベースの濁点付与の具体例を**図 1** に示した．図 1 の例の場合，通常の辞書引きでラティスに載せられる単語は実線のものだけである．しかし濁点無表記を考慮した辞書引きにより，「學ふ」という文字列で辞書中の「學ぶ」という単語がマッチングし，

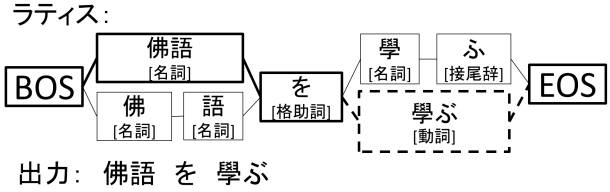


**図 1** 辞書ベースの表記整理．太線はスコア最大のパス．点線は表記揺れを考慮した辞書引きによって追加されたノード．

ラティスへと加えられる．その上でスコア最大のパスを求め，出力された単語列「佛語 | を | 學ぶ」は入力文に濁点付与を行なった結果となっている．辞書ベースの濁点付与では形態素解析と濁点付与を同時に行うため，文字ベースのように部分文字列に引かれて誤った濁点付与を行うことは少ない．しかしながらこのモデルの学習には，表記整理済みコーパスよりも作成コストの高い品詞タグ付きコーパスしか使えない．そのため，学習に利用可能なコーパスが限られる．学習に使えるコーパスが多い分，文献 [8] では，文字ベースの濁点付与が辞書ベースよりも高い性能を示している．

本研究では，以下の文字ベースと辞書ベースそれぞれの自動表記整理手法の利点と欠点を相互に補完するため，その 2 つをハイブリッドした手法を提案する．

- 文字ベースの自動表記整理：
  - **利点**：学習には表記整理済みコーパスを使用．学習用コーパスに単語境界や品詞のアノテーションは不要．
  - **欠点**：部分文字列にマッチする単語に引かれて誤った表記整理を行いやすい．

- 辞書ベースの自動表記整理：
  - **利点**：形態素解析との同時解析により，部分文字列にマッチする単語に引かれて誤った表記整理を行うことが少ない．
  - **欠点**：品詞タグ付きコーパスしか学習に使えない．

提案手法では辞書ベースの表記整理手法と同じく文献 [3] の形態素解析のフレームワークを利用しつつ，そのモデルに文字ベースの手法で用いた素性を追加し，それらを同時に最適化していく．学習には Augmented-Loss Training [2] と呼ばれるオンライン学習アルゴリズムを採用した．これにより，品詞タグ付きコーパスだけでなく，表記整理済みコーパスからの学習も可能となった．具体的には，表記整理済みコーパスを学習に利用する場合，gold の形態素解析結果でなく，その下流タスクとなる表記整理の gold に向けた最適化を行なっていく．

また提案手法で辞書引き時に考慮する表記揺れは濁点無表記に限らず，表 1 に挙げたすべてを対象とする．すなわち，提案手法で扱える表記整理は以下の全 5 種類である．

[*1] 「恐る々々」のように踊字を用いて省略表記された文字列から「恐る恐る」のような省略前の表記を復元する作業．

[*2] 辞書の検索キー（辞書に登録されている単語の表層形）とのマッチングによる辞書内単語検索と列挙．文献 [8] では，濁点の一部〜すべてが抜け落ちた表層形を辞書の検索キーとして登録することで濁点無表記を考慮した辞書引きを実現している．

(1) 濁点付与
- e.g., 及ひ → 及び

(2) 仮名遣の統一
- e.g., 用い，用ひ，用ゐ → 用ゐ

(3) 送り仮名の統一
- e.g., 限り，限ぎり，限（カギリ） → 限り

(4) 踊字の展開
- e.g., 恐る々々 → 恐る恐る

(5) 片仮名の変換（漢字片仮名交じり文の漢字平仮名交じり文への書き換え）
- e.g., 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

→ 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ぜらるゝことなし

## 2. 表記整理済みコーパスも学習に利用可能な辞書ベースの表記整理

提案手法は，形態素解析のフレームワークを利用した辞書ベースの自動表記整理である．

形態素解析は通常，以下のような手順で行われる．

**手順 1** 入力文を先頭から 1 文字ずつ読み進め，各位置から開始する単語を辞書引きにより列挙する．

**手順 2** 手順 1 で列挙した単語からラティスを作成する．

**手順 3** 単語スコアと単語連接スコアを計算し，文として最も確からしい単語の並びをラティス上のスコア最大のパスとして出力する．

辞書ベースの表記整理では，手順 1 で表記の揺れを考慮した辞書引きを行い，手順 3 で得られる出力が入力文に表記整理を行なった結果となるようにする．文献 [8] では濁点付与をのみを扱ったため，辞書引きで考慮する表記揺れを濁点無表記に限定していた．これに対し文献 [9] では，表 1 に挙げた表記揺れすべてを考慮した辞書引き手法を提案している．具体的には，濁点無表記，仮名遣・送り仮名の不統一に対しては，ルールベースの書き換えによる辞書検索キーの追加．踊字による省略表記と漢字片仮名交じり文に対しては，入力文の動的な書き換えでこれを実現している．提案手法では文献 [9] の辞書引きに若干の修正を加えて使用することで，濁点付与だけでなく，仮名遣の統一，送り仮名の統一，踊字の展開，片仮名の変換，全 5 種類の表記整理を同時に行えるようにする．

### 2.1 歴史的資料中の表記揺れを考慮した辞書引き

通常の形態素解析では，ラティスを作成する際に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，以下の方法でラティスに載せられる単語の追加を行う．

#### 2.1.1 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（キー）を追加することで対応する．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．

(1) **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，「限り」→「限（カギリ）」のように漢字直後の平仮名文字列をすべて読み飛ばすことを許してしまうと，「愛くるしい」→「愛（アイクルシイ）」といった表層形まで生成されてしまう．そこで，漢字直後の平仮名文字列を完全に読み飛ばしてもよいのは「限り」→「限（カギリ）」のように漢字直後のひらがな文字列長が 1 の場合に限ることにした．また送り仮名の飛び出した表層形は，当該漢字の仮名表記末尾を当該漢字直後に挿入することで生成する（e.g., 志（こころざ）し→志ざし）．各漢字の仮名表記は，文字列同士の多対多アライメントツール mpaligner[14] を使用し，辞書中の表層形と仮名形出現形の対応付けをとることで得た．

(2) **濁点無表記，仮名遣の不統一**：(1) で作成した送り仮名の伸縮を考慮した表層形へ付録 A.2 の文字列書き換えルールを適用し，濁点無表記，仮名遣の不統一を考慮した表層形を網羅的に生成する．このルールは文献 [11] 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．

#### 2.1.2 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，スコア最大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 (1)〜(4) に従って行う．

(1) **一字点（ゝ，ゞ，ヽ，ヾ）**：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，当該一字点の種類の応じて次のルール (a) もしくは (b) を適用する．

(a) **濁点なし一字点（ゝ，ヽ）**：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ら濁点を外した文字に置換する（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（ゞ，ヾ）**：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．

(2) **同字点（々）**:
  (a) **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）.
  (b) **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）.
(3) **二字点（ゞ）**：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.
(4) **くの字点（／＼，／″＼）**：当該くの字点が単語先頭でなく，直前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る（e.g., まに／＼→まにまに）. また，当該くの字点を 1 単語と認め，ラティス作成時に左連接した単語と置換する（e.g., 薄い/／＼→薄い/薄い）.

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.

#### 2.1.3 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.

### 2.2 識別モデルによる形態素解析の定式化と素性

文献 [8] の辞書ベースの自動表記整理手法では，文献 [3] の形態素解析のフレームワークを用いていた．文献 [3] では，単語ラティスのノード（単語）のスコア及び，エッジ（単語連接）のスコアを識別モデルを用いて表現している．提案手法でも同様に識別モデルを使用する．つまり，入力文 $C = c_1c_2...c_{|C|}$（$c$ は文字）が与えられたとき，スコア最大の形態素解析結果 $\hat{W} = \hat{w}_1\hat{w}_2...\hat{w}_{|\hat{W}|}$（$w$ は単語）は以下のようにして求められる.

$$\hat{W} = argmax_W \left\{ \sum_{i=2}^{|W|} \left( \sum_k \lambda_k^u \phi_k^u(C, w_i) + \sum_l \lambda_l^b \phi_l^b(w_{i-1}, w_i) \right) \right\}$$

ここで $\phi^u$ は単語 Unigram 素性，$\lambda_k^u$ は学習によって得られた $\phi_k^u$ の重みである．また $\phi^b$ は単語 Bigram 素性であり，$\lambda_l^b$ は学習によって得られた $\phi_l^b$ の重みである．$w_i$ は当該ノードが相当する単語であり，$w_{i-1}$ はラティス上で $w_i$ に左連接する単語を表している．また $w_1$ と $w_{|W|}$ はそれぞれ文頭を表すダミーの語（BOS）と文末を表すダミーの語（EOS）である.

単語 Unigram 素性および，単語 Bigram 素性には形態素解析で従来より用いられているバイナリ素性をそのまま使用する．またそれらに加えて，表記揺れの有無や種類に応じて単語のスコアを調整するための単語 Unigram 素性も追加する．これには文献 [8] で濁点付与に用いた**文字 n-gram 素性**（**図 2**）を使用する.

文字 n-gram 素性は，表記整理の対象となる文字を target として，その前後 N 文字の window 内に存在する文字 1〜n-gram をバイナリ素性としたものである．素性関数は以下の通り.

$$\phi(window) = \begin{cases} 1: & window \text{ 内の位置 } pos \text{ に} \\ & \text{文字列 } c_{pos}c_{pos+1}\cdots \text{が出現} \\ 0: & \text{それ以外} \end{cases}$$

ここで $pos$ は target からの相対位置を表している．また $window$ は文字列 $c_{-N}$, $c_{-(N-1)}$, ⋯, $c_0$, ⋯, $c_{N-1}$, $c_N$ であり，$c_0$ が target である．また文献 [8] と同様，文字 n-gram 中の濁点を一部〜すべて外した文字列（**疑似濁点無表記文字 n-gram**）も文字 n-gram 素性として同時に発火させる．例えば，位置 $pos$ における文字 3-gram が「がぎぐ」であるとするなら，$pos$/がぎぐ だけでなく，$pos$/がぎく，$pos$/がきぐ，$pos$/かぎぐ，$pos$/がきく，$pos$/かぎく，$pos$/かきぐ, $pos$/かきく の素性も同時に値 1 とする．疑似濁点無表記文字 n-gram は図 2 ではハイライトで示してある.

文字 n-gram 素性は各表記整理ごとに別個に用意し，それぞれ区別して利用する．つまり，濁点付与用の “文字 n-gram 素性：-3/がぎぐ”と，仮名遣用の “文字 n-gram 素性：-3/がぎぐ”がそれぞれ別に用意されている（詳細は付録 A.3 を参照のこと）.

濁点付与，片仮名の変換では，表記整理を施す対象文字を target とした．また，仮名遣の統一，踊字の展開では，表記整理を施す文字列の先頭文字を target とし，送り仮名の伸縮では，当該送り仮名の直前の漢字を target とした.

### 2.3 Augmented-Loss Training

提案手法の基礎となる識別モデルを用いた形態素解析の学習には通常，品詞タグ付きコーパスのみが利用される．しかし歴史的資料の品詞タグ付きコーパスは数も量も少ない．そこで提案手法では，単語分割や品詞タグ付けの行われていない表記整理済みコーパスからでも学習が行えるよう，Augmented-Loss Training [2] (以下，ALT) と呼ばれるオ

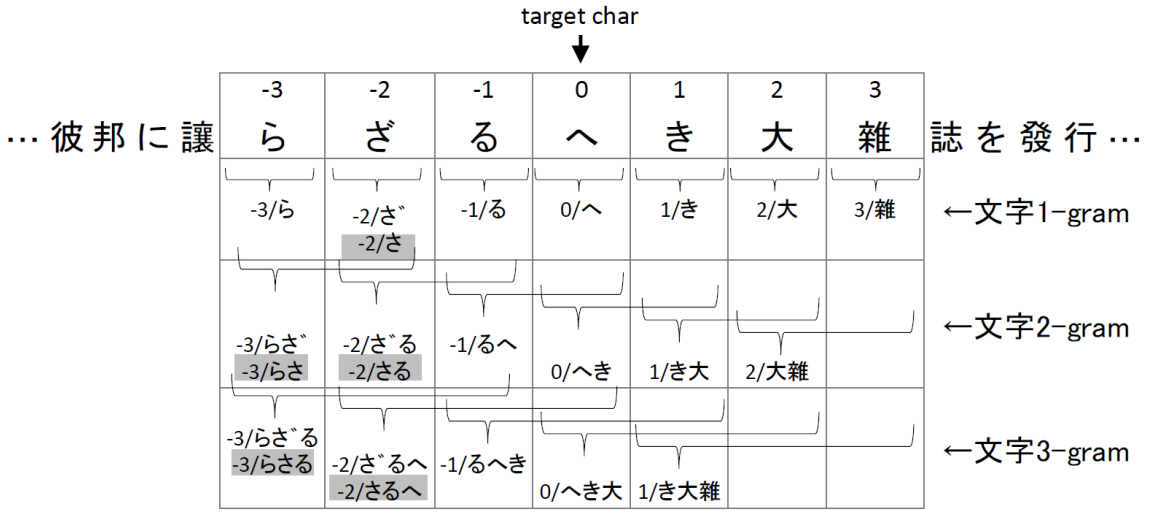


**図 2**　文字 n-gram 素性．疑似濁点無表記文字 n-gram は編み掛けで表している．

ンライン学習アルゴリズムを採用した．

ALT では，上流タスク（e.g., 依存構造解析）と，上流タスクの出力を用いる下流タスク（e.g., 統計的機械翻訳の Reordering）が与えらえれたときに，下流タスク側で設定したコスト関数を最小とするような上流タスクのモデルを学習する．この学習アルゴリズムの利点は，上流タスク側の gold アノテーションを持たない下流タスクのデータも上流タスクの学習に使えるところにある．これは Inline Reranker という仕組みによって実現される．つまり，上流タスク側の gold を持たないデータであっても，学習中の上流タスクのモデルを使って上流タスクの k-best を求め，その中から下流側で設定したコストが最も低くなる候補を選んで，上流タスクの gold として重みの更新に使用する．ただし，この学習によって得られるモデルは下流タスクで有効な上流タスクのモデルであり，それは必ずしも上流タスクで有効なモデルとは限らない．

提案手法では，上流タスクとして形態素解析，下流タスクに表記整理を設定した．訓練用データセットには上流側の学習用コーパスとして品詞タグ付きコーパス $D^W$ と，下流側の学習用コーパスとして原文情報を保持した表記整理済みコーパス $D^T$ を使用する．ALT を用いた提案手法の学習アルゴリズムを**図 3** に示す．ここで $C^W$ は 1 文，$W^W$ は $C^W$ に対する gold の形態素解析結果である．また $C^T$ は原文，$C'^T$ は $C^T$ に表記整理を行なった結果である．下流側コスト関数には，自動表記整理の結果と $C'^T$ との最小編集距離を設定した．

文献 [2] では学習モデルに構造化パーセプトロンを使用している．しかし文献 [2] によると，ALT のフレームワーク自体はオンライン学習に複数の目的関数を組み込むための一般的な体系であり，学習モデルは構造化パーセプトロンでなくてもよい．そこで今回は学習モデルとして，オンライン学習版の CRF [4] を使用した．また図 3 では簡単のため正則化は省略しているが，実際には FOBOS [1] による L1 正則化を行なっている．文献 [2] では重みの更新タイミングを様々な設定で試しているが，今回の評価実験では，1 種類の品詞タグ付きコーパスと 2 種類の表記整理済みコーパスで 1:1:1 の更新とした．ただし，図 3 では簡単のため，1 種類の品詞タグ付きコーパスと 1 種類の表記整理済みコーパスによる 1:1 の更新を示した．実際には，Inline Relranker を用いたアップデートの直後に，もう 1 つの表記整理済みコーパスを用いた Inline Reranker アップデートを行なっている．

Augmented-Loss Training

▷ 品詞タグ付きコーパス
$D^W \leftarrow \{(C_1^W, W_1^W), ..., (C_{|D^W|}^W, W_{|D^W|}^W)\}$
▷ 表記整理済みコーパス
$D^T \leftarrow \{(C_1^T, C'^T_1), ..., (C_{|D^T|}^T, C'^T_{|D^T|})\}$
▷ インデックスの初期化
$s^W \leftarrow 1,\ s^T \leftarrow 1$
▷ 重みの初期化
$\Lambda \leftarrow \overrightarrow{0}$
$i \leftarrow 0$
**repeat**
　▷ 通常のオンライン CRF のアップデート
　$\Lambda \leftarrow update(\Lambda, lattice(C_{s^W}^W), W_{s^W}^W)$
　$s^W \leftarrow [(s^W \equiv |D^W|)?1 : s^W + 1]$
　▷ Inline Reranker を用いたアップデート
　　▷ k-best の形態素解析結果の取得
　　$\{\hat{W}_1, ..., \hat{W}_k\} \leftarrow kbest(\Lambda, C_{s^T})$
　$\hat{W} \leftarrow min_{\hat{W} \in \{\hat{W}_1, ..., \hat{W}_k\}} minEditDist(toString(\hat{W}), C'^T_{s^T})$
　$\Lambda \leftarrow update(\Lambda, lattice(C_{s^T}^T), \hat{W})$
　$s^T \leftarrow [(s^T \equiv |D^T|)?1 : s^T + 1]$
　$i \leftarrow i + 1$
**until** $i \equiv Max\ number\ of\ iterations$
**Return** $\Lambda$

**図 3**　Augmented-Loss Training

## 3. 表記整理性能の評価実験

提案手法の有効性を確認するため，濁点付与のタスクを取り上げ，文献 [8] で辞書ベースの手法よりも高い性能を示した文字ベースの手法と，提案手法の性能を比較した．文献 [8] と同じく近代文語論説文（**図 4** a のような文体）で評価を行い，濁点付与の適合率，再現率，F1 値を調査した．

### 3.1 実験設定

提案手法が使用する単語辞書には**近代文語 UniDic** [10] の v1.2 を使用する．

品詞タグ付きコーパス $D^W$ には近代文語 UniDic の学習用コーパス（MLJ-Train）を用いた（22,575 文）．このとき未知語学習のため，語彙素（語彙素読み lForm+語彙素 lemma）レベルでコーパス中に 1 度しか出現しないような単語は未知語[*3]とし，辞書からもその語彙素を持つ単語はすべて取り除いた．また，提案手法の辞書引きを使用しても引き出せない単語もすべて未知語とした．さらに，MLJ-Train 中の文の表層形から濁点を取り除き，平仮名文字をすべて片仮名文字に置換した文も $D^W$ へと追加した（45,150 文 =22,575 文+22,575 文）．

原文情報を保持した表記整理済みコーパス $D^T$ には太陽コーパスを利用する．太陽コーパスには文体が文語体の記事と口語体の記事が含まれている．そこでまず，文体が文語かつ漢字平仮名交じり文だけを抽出した（312,893 文）．[*4]抽出した文からランダムに 9 割を学習（T-Train，281,603 文），残り 1 割を評価（T-Eval，31,290 文）に使用する．

太陽コーパスでは，原文情報を保持しつつ表記整理として濁点付与，仮名遣の統一，踊字の展開が行われている．そこで，T-Train 中の文を原文表記に戻したものを $C^T$，表記整理済み本文を $C'^T$ として，表記整理済みコーパス $D_1^T$ を作成した．また，$C^T$ 中の濁点をすべて外し，平仮名文字もすべて片仮名文字に書き換えた疑似原文も $D_1^T$ に加えた（563,206 文=281,603 文+281,603 文）．この際，疑似原文に対応する $C'^T$ は書き換え前の $C^T$ に対応するものと同一とした．

また T-Train から，濁点付き文字もしくは濁点の付き得る文字を含んだ表記整理済み本文を抽出した（272,956 文）．この本文を $C'^T$ とし，そこからすべての濁点を取り除いた文を $C^T$ とする表記整理済みコーパス $D_2^T$ を作成した．

評価用コーパスには T-Eval の他に，**明六雑誌コーパス** [17] の文語文を利用する（9,139 文）．明六雑誌コーパスも太陽コーパスと同じく原文情報を保持した表記整理済みコーパスであり，濁点付与，踊字の展開，漢字片仮名交じり文の漢字平仮名交じり文への書き換えが実施されている．評価にはタグに保持された情報から原文を再現して使用した（M6-Eval）．

| a) 表記整理前 |
|---|
| 而テ己知ラスシテ信スル者ヲ以テ人ヲシテ強ヰテ之ヲ信セシメムト欲セハ天下豈眞心是ニ服スル者アラムヤ況ヤカノ所謂神ナル者ト相並ヒ造物ノ元始ニ反リ萬有ノ窮極ニ達スル者ト其權ヲ均ウセムト欲スト云ハヽ誰亦其妄誕ヲ信スル者アラムヤ |
| b) 表記整理後 |
| 而して己知らずして信ずる者を以て人をして強ひて之を信ぜしめんと欲せば天下豈眞心是に服する者あらんや況やかの所謂神なる者と相並び造物の元始に反り萬有の窮極に達する者と其權を均ふせんと欲すと云はば誰亦其妄誕を信ずる者あらんや |

**図 4** 提案手法を用いた自動表記整理例

文字 n-gram 素性のパラメータは文献 [8] と同じく，文字 n-gram 素性作成時の窓幅:7（$N = 3$），$n$-gram の最大長を 3（$n = 3$）に設定した．学習時のハイパーパラメータは実験的に，CRF の学習率 $\eta$ =1.0，L1 正則化のパラメータ C=0.00005，Inline Reranker 内で求める k-best の k=100，最大イテレーション回数（Max number of iterations）を 100 万回に設定した．

また，ラティス構築時には文中の長さ 1〜6 の部分文字列も品詞大分類「未知語」としてラティスに加えた．形態素解析用の単語 Unigram 素性と Bigram 素性は近代文語 UniDic 付属の MeCab[*5]用素性テンプレートを選別して使用する（詳しくは付録 A.4 を参照）．また文献 [6] で使用された以下の文字単位の Unigram 素性（バイナリ素性）も使用した．

- 単語内文字数: 1, 2, 3, 4, 5, 6 以上
- 単語内の先頭文字列: 先頭 1 文字，先頭 2 文字
- 単語内の末尾文字列: 末尾 1 文字，末尾 2 文字
- 単語内の先頭文字種列: 先頭 1 文字，先頭 2 文字
- 単語内の末尾文字種列: 末尾 1 文字，末尾 2 文字
- 単語内の文字種遷移

提案手法を用いて表記整理を行なった結果の一例を図 4 b に示す．また，付録 A.1 には明六雑誌コーパス中の 1 記事に対して実際に表記整理を実施した結果の全体を示す．

### 3.2 ベースライン：文字ベースの濁点付与手法

ベースラインとして文献 [8] で辞書ベースの手法よりも高い濁点付与性能を示した文字ベースの濁点付与手法を設定した．具体的には，濁点付与を文字ごとの独立した 2 値分類問題として定式化し (+1:濁点を付ける，-1:濁点を付けない)，SVM を用いて識別を行なった．

分類の素性には，付録 A.3 の濁点付与用の素性を使用する．ここでは提案手法と同じく，文字 n-gram 素性作成時

[*3] 品詞大分類「未知語」とし，それ以外の品詞中分類〜細分類，活用や語彙素といった情報をすべて削除した．
[*4] 3 文字以上平仮名文字を含む文を漢字ひらがな交じり文として抽出した．
[*5] 文献 [3] の手法を実装した形態素解析器． `http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html`

の窓幅:7（$N = 3$），$n$-gram の最大長を 3（$n = 3$）に設定した．SVM の実装は liblinear[*6] v1.94 の L2 正則化 L2 ロス SVC を使用し，ハイパーパラメータは 5-fold のクロスバリデーションにより決定した．

文字ベースの濁点付与手法では，訓練用事例は濁点無表記文字を含まない表記整理済みコーパスから作成する．今回は，MLJ-Train の表層文字列と，T-Train の表記整理済み本文，および各それぞれの文中の平仮名文字をすべて片仮名文字に置換した文から訓練用事例を抽出した．訓練用コーパス中の濁点文字から正例，濁点の付いていない文字から負例を作成する．これらの事例は濁点を付けるか否か判断したい文字の分類に利用するため，事例作成時，濁点文字の事例では target の濁点を外しておく．表記整理済みコーパスからの訓練用事例の作成手順は以下の通り．

Step 1: 表記整理済みコーパスから濁点文字，もしくは濁点の付き得る文字を 1 つ取り出す．

Step 2: Step 1 で取り出した文字とその左右 N(=3) 文字を合わせて 1 つの事例とみなし，素性ベクトルを作成する．この際，Step 1 で取り出した文字が濁点文字であれば，その濁点を外した後，素性ベクトルを作成する．

Step 3: Step 1 で取り出した文字が濁点文字ならば正例，濁点の付き得る文字ならば負例とする．

抽出の結果，488,718 個の正例と，2,378,195 個の負例が得られた（計：2,866,913）．

### 3.3 濁点付与の性能評価実験結果

ベースラインである文字ベースの手法と，提案手法を用いてそれぞれの評価用コーパスに濁点付与を行い，評価を行なった．ここでは濁点付与の適合率，再現率，F1 値で評価した．結果を**表 2** に示す．実験設定が多少異なるため正しい比較にはならないが，参考までに文献 [8] に記載された辞書ベースの手法の濁点付与性能を併記した．また，濁点付与の適合率，再現率，F1 値の式は以下の通り．

$$Precision = \frac{\text{正しく濁点を付けた文字数}}{\text{濁点を自動付与した文字数}} \times 100[\%]$$

$$Recall = \frac{\text{正しく濁点を付けた文字数}}{\text{評価用コーパス中の濁点無表記文字数}} \times 100[\%]$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

T-Eval で評価を行なった場合，文字ベースの手法に比べて提案手法の適合率が大幅に向上し，F1 値で文字ベースよりも約 10 高い性能を示した．ただし，再現率は文字ベースよりも約 7%低下している．同様に，M6-Eval で評価を行なった場合も，適合率が向上したが，再現率が低下し，M6-Eval では文字ベースの方が高い F1 値を示している．しかし，M6-Eval に対する提案手法の形態素解析出力を調査したところ，コーパス中に頻出する以下の 2 つの語がいずれも辞書に最初から登録されている濁点無表記の表層形として出力されていることが分かった．

- 我が　連体詞
- 先づ　副詞

近代文語 UniDic には元々，濁点無表記を含んだ文の解析にも対応するため，少数ではあるが濁点無表記の表層形が登録されていた [10]．ただし，今回使用した v1.2 からは濁点無表記の表層形は廃止されている．しかし，その方針も厳密に適用されたわけではなく，上のような表層形が未だに残っている状態であり，これが今回再現率の低下を引き起こした原因の 1 つと考えられる．上記の 2 語に対し，後処理として濁点付き表層形への置換を実施したところ，提案手法の再現率が向上し，F1 値で文字ベースよりも高い性能を示すことがわかった．その結果を表 2 の「提案手法（補正）」に示す．

文字ベースの手法の欠点は部分文字列にマッチする単語に引かれて誤った濁点付与を行なってしまうことであった．実際に文字ベースの濁点付与結果を確認したところ，「ヘカラス（ベカラズ）」を「ベ**ガ**ラズ」と誤って濁点付与する例が非常に多くみられた．これは「カ」に濁点付与を行う際，「ヘカラス」の部分文字列「カラス」に引かれて，「ガラス」と誤って濁点付与を行なった結果と考えられる．これに対し，提案手法では単語分割と濁点付与を同時に行うため，「ヘカラ | ス」と正しく単語分割を行った上で「ベカラ | **ズ**」と正しく濁点付与を行うことができた．

以上より，提案手法は表記整理済みコーパスを学習に利用可能となったことで，従来の辞書ベースの手法よりも高い精度で濁点付与が可能になることが分かった．また形態素解析と表記整理の同時解析を行うことによって，文字ベースの手法よりも高い適合率で濁点付与が可能になることも確認できた．

### 3.4 濁点付与以外の表記整理結果の分析

太陽コーパスでは濁点付与以外に，仮名遣の統一，踊字の展開が実施されている．このアノテーションを基に，提案手法の仮名遣の統一結果と踊字の展開結果を分析した．

#### 3.4.1 仮名遣の統一結果の分析

仮名遣の統一結果を分析するため，まず提案手法が仮名遣の統一を実施した文を T-Eval からランダムに 100 文取り出し，人手で確認作業を行なった．その結果，101 事例中，97 事例が正しく仮名遣を修正できていたことが分かった．仮名遣を正しく直せていなかったのは以下の 4 事例であり，いずれも仮名遣のアノテーションが行われていなかっ

[*6] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

表 2　濁点付与性能の比較

| 評価用コーパス | 手法 | tp | tn | fp | fn | Prec.[%] | Rec.[%] | F1 |
|---|---|---|---|---|---|---|---|---|
| T-Eval | 辞書ベース | - | - | - | - | 50.9 | 91.8 | 65.5 |
| | 文字ベース | 906 | 106,737 | 554 | 56 | 62.1 | **94.2** | 74.8 |
| | 提案手法 | 840 | 107,123 | 168 | 122 | **83.3** | 87.3 | **85.3** |
| M6-Eval | 辞書ベース | - | - | - | - | 90.1 | 95.9 | 92.9 |
| | 文字ベース | 6,096 | 49,713 | 342 | 148 | 94.7 | **97.6** | 96.1 |
| | 提案手法 | 5,821 | 49,928 | 127 | 423 | **97.9** | 93.2 | 95.5 |
| | 提案手法（補正） | 5,905 | 49,928 | 127 | 339 | **97.9** | 94.6 | **96.2** |

た個所に修正を行うものであった．

- 全う（副詞）$\overset{\text{提案手法}}{\rightarrow}$ 全ふ
- 渝へ（動詞）$\overset{\text{提案手法}}{\rightarrow}$ 渝え
- 危うし（形容詞）$\overset{\text{提案手法}}{\rightarrow}$ 危ふし
- 超へ（動詞）$\overset{\text{提案手法}}{\rightarrow}$ 超え

次に，T-Eval から仮名遣のアノテーションが行われていた文をランダムに 100 文取り出し，提案手法の仮名遣の統一結果と比較を行なった．その結果，102 事例中，61 事例が正しく仮名遣を修正できていたことが分かった．仮名遣を正しく直せていなかった事例は，いずれも仮名遣を直すべき箇所を見逃したものであった．

以上より，提案手法の仮名遣の統一は，濁点付与の場合と同じく適合率が高く，再現率が低いものであるとわかった．再現率が低い原因も濁点付与の場合と同じく，辞書中に仮名遣の不統一が考慮された表層形 [10] があらかじめ登録されていたことが大きく，再現率を向上させるためには，辞書に登録済みの表層形を見直す必要があると考えられる．

### 3.4.2　踊字の展開結果の分析

踊字の展開結果を分析するため，まず提案手法が踊字の展開を実施した文を T-Eval からランダムに 100 文取り出し，人手で確認作業を行なった．その結果，102 事例中，96 事例が正しく踊字の展開を行なえていたことが分かった．踊字の展開に失敗していたのは以下の 6 事例であり，6 事例中 5 事例がくの字点の展開に失敗した例であった．

(1) そよ／＼（ソヨソヨ）$\overset{\text{提案手法}}{\rightarrow}$ そよよ
(2) くよ／＼（クヨクヨ）$\overset{\text{提案手法}}{\rightarrow}$ くよよ
(3) づか／＼（ヅカヅカ）$\overset{\text{提案手法}}{\rightarrow}$ づかか
(4) まに／＼（マニマニ）$\overset{\text{提案手法}}{\rightarrow}$ まにに
(5) こと／＼（コトゴト）$\overset{\text{提案手法}}{\rightarrow}$ ことこと
(6) つゝいて（ツツイテ）$\overset{\text{提案手法}}{\rightarrow}$ つづいて

(1)〜(4) はいずれもくの字点が直前の 1 字を繰り返していると誤った例であった．(5)(6) は濁点付与に失敗した例であった．また取り出した 100 文中でのくの字点の出現は上の (1)〜(5) を含めて全部で 7 箇所あり，以下の 2 箇所に関しては正しく踊字の展開を実施することができていた．

- 部屋／＼（ヘヤベヤ）$\overset{\text{提案手法}}{\rightarrow}$ 部屋部屋
- いよ／＼（イヨイヨ）$\overset{\text{提案手法}}{\rightarrow}$ いよいよ

次に，T-Eval から踊字の展開がアノテーションされていた文をランダムに 100 文を取り出し，提案手法の踊字の展開結果と比較を行なった．その結果，101 事例中，94 事例が正しく踊字を展開できていたことが分かった．踊字の展開に失敗した事例は，以下の 7 例であった．

(1) くど／＼（クドクド）$\overset{\text{提案手法}}{\rightarrow}$ くど／＼（未知語）
(2) うつら／＼（ウツラウツラ）$\overset{\text{提案手法}}{\rightarrow}$ うつら／＼（未知語）
(3) 夜な／＼（ヨナヨナ）$\overset{\text{提案手法}}{\rightarrow}$ 夜な／＼（未知語）
(4) 顔つく／＼（カオツクツク）$\overset{\text{提案手法}}{\rightarrow}$ 顔つく／＼（未知語）
(5) オヽシス（オオシス）$\overset{\text{提案手法}}{\rightarrow}$ オヽシス（未知語）
(6) 誠に／＼（マコトニマコトニ）$\overset{\text{提案手法}}{\rightarrow}$ 誠 | に | ／＼（記号）
(7) はか／〃＼しき（ハカバカシキ）$\overset{\text{提案手法}}{\rightarrow}$ はかがしき

8 事例中，(1)〜(5) の 5 事例が未知語と解釈され，踊字が展開されなかった例であった．(6) はくの字点が直前の文節を省略する例であり，現状の提案手法では扱うことができない．(7) はくの字点を直前 1 文字の繰り返しと誤って解析した例である．また取り出した 100 文中でくの字点の出現は全部で 8 箇所あり，以下の 2 箇所に関しては正しく踊字の展開を実施することができていた．

- おの／＼（オノオノ）$\overset{\text{提案手法}}{\rightarrow}$ おのおの
- それ／〃＼（ソレゾレ）$\overset{\text{提案手法}}{\rightarrow}$ それぞれ

以上より，提案手法の踊字の展開性能は高いものの，大多数を占める一字点や同字点の展開が上手くいくだけで，くの字点に限ると上手く展開できない場合が多いことが分かった．特に，直前の単語を省略しているくの字点を，直前の 1 文字の省略と誤る場合が多いことがわかった．これは 2.1.2(4) で述べた，ラティス作成時に当該くの字点をそれに左連接する単語に置き換えることで踊字の展開を実施した結果であった．ただしこの際，置換後のノードに左連接可能なノードを一意にしぼっていなかったため，冗長な曖昧性が発生したと考えられる．これに対処するための単純な方法として，当該くの字点に左連接する単語が当該くの字点置き換え後の単語と同一か否かを表す単語 Bigram 素性を追加することが考えられる．

## 4.　関連研究

日本語の歴史的資料の表記整理と類似のタスクに，日本

語で書かれた Web の文の正規化がある．歴史的資料中の表記揺れと同様に，Web のテキストの中には辞書未登録の崩れ表記が多く現れ（e.g., ぉぃしかったでーーす），形態素解析結果の精度を下げる原因となっている．

文献 [7] では，形態素解析の前処理として「ゎ → わ」「ナニ → た」のような書き換えルールを用いたブログテキストの正規化を提案している．このルールの適用は識別モデルによって判定され，素性の 1 つとして正規化前と後の文の形態素解析コストの差も利用している．ただしこの判定のための素性は 3 つのみであり，そのため各素性の重みはすべて人手で設定されている．文献 [21] では，マイクロブログテキストの正規化を文字単位の系列ラベリング問題として定式化している．文字ベースの手法であるが，事前に行なった形態素解析の結果を素性として使うことを試みている．しかし，この素性を導入したとき，正規化の性能は悪化することが報告されている．

文献 [5][12][13][19][20] では，提案手法と同じく，正規化処理を形態素解析と同時に行う手法が提案されている．文献 [5] では，Web の崩れ表記に頑健な形態素解析を提案している．本研究と同様，辞書引き時に表記揺れを考慮する手法であるが，崩れ表記に対する単語 Unigram コストはすべて人手で設定されている．文献 [12][13] では，同時解析のモデルを生成モデルとして定式化している．これに対し，文献 [19][20] は提案手法と同じく，形態素解析を識別モデルとして定式化しており，崩れ表記に対する単語 Unigram コストも文字単位の素性で表現している．しかし，文字単位の素性と単語単位の素性はそれぞれ別個のモデルとして最適化され，デコード時に MERT で重みづけした線形結合を行なっている．そのため，提案手法のように単一のモデルを同時に最適化しているわけではない．

## 5. おわりに

本研究では，文字ベースの手法と辞書ベースの手法をハイブリッドした自動表記整理手法を提案した．辞書ベースの表記整理に文字ベースの手法で使用していた素性を追加し，Augmented-Loss Training を採用することで，単語境界や品詞情報のない表記整理済みコーパスも学習に利用可能にした．その結果，提案手法の適合率が文字ベースの手法に比べて大幅に向上し，F1 値でも提案手法の方が高い性能を示すことが分かった．ただし，再現率はいずれの評価用コーパスでも文字ベースの手法の方が高くなった．これは辞書中に表記揺れを含んだ表層形が予め登録されていたことが原因であると考えられる．そのため，今後の課題として，辞書中の表層形を再確認し，選別を行うことが上げられる．また，くの字点の展開性能を向上させるため，素性やラティスの構築法の見直しも今後の課題である．

本稿の評価実験では近代文語論説文を対象に実験を行なったが，今後，平安時代の中古和文系資料や近世の資料に対しても提案手法を適用し，評価を行なっていきたいと考えている．

また今回は表記整理が完全に実施されたコーパスを学習に用いることを想定していたが，ALT ではコスト関数さえ用意すれば，部分的に表記整理されたコーパスや，初めから一部にのみ濁点が付いているようなコーパスであっても学習に利用可能である．実際，コーパス整備の現場ではそういった状態の資料が完全に表記整理済みの資料よりも多く手に入る．そのため，今後は部分的にアノテーション済みのコーパスからの学習も検討していきたい．

提案手法に能動学習を取り入れることも今後の課題である．これにより提案手法の出力に対し，人間が修正を行い，その結果を使ってモデルを再学習するといったユーザインタラクティブな自動表記整理ツールの開発に取り組んでいきたいと考えている．

### 参考文献

[1] Duchi, J. and Singer, Y.: Efficient Online and Batch Learning Using Forward Backward Splitting, *Journal of Machine Learning Research*, Vol.10, pp. 2899-2934 (2009).

[2] Hall, K., McDonald, R., Katz-Brown, J. and Ringgaard, M.: Training dependency parsers by jointly optimizing multiple objectives, *Proc. the 2011 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2011*), pp.1489-1499 (2011).

[3] 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).

[4] 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).

[5] Sasano, R., Kurohashi, S. and Okumura, M.: A Simple Approach to UnknownWord Processing in Japanese Morphological Analysis, *Proc. the 6th International Joint Conference on Natural Language Processing* (*IJCNLP 2013*), pp 162-170 (2013).

[6] 東藍，浅原正幸，松本裕治：条件付確率場による日本語未知語処理情報処理学会研究報告第 173 回自然言語処理研究会, Vol. 2006-NL-173, No.11, pp.67-74 (2006).

[7] 池田和史，柳原正，松本一則，滝嶋康弘：くだけた表現を高精度に解析するための正規化ルール自動生成手法，情報処理学会論文誌 データベース, Vol.3, No.3, pp.68-77 (2010).

[8] 岡照晃，小木曽智信，小町守，松本裕治：統計的機械学習を用いた歴史的資料への濁点付与の自動化，情報処理学会論文誌, Vol.54, No.4, pp.1641-1654 (2013).

[9] 岡照晃，小木曽智信，小町守，松本裕治：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，人工知能学会全国大会 (JSAI2013) 論文集, 2B1-2, pp. 1-4 (2013).

[10] 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 (NLP2008) 発表論文集，pp.225-228 (2008).

[11] 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—「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).

[12] 風間淳一，光石豊，牧野貴樹，鳥澤健太郎，松田晃一，辻井純潤一：チャットのための日本語形態素解析，言語処理

学会第 5 回年次大会 (NLP1999) 発表論文集，pp.509-512 (1999).
[13] 工藤拓，市川宙，David Talbot，賀沢秀人：Web 上のひらがな交じり文に頑健な形態素解析言語処理学会第 18 回年次大会 (NLP2012) 発表論文集，pp.1272-1275 (2012).
[14] 久保 慶伍，川波 弘道，猿渡 洋，鹿野 清宏：日本語の未知語に対する発音付与のための多対多アライメント情報処理学会論文誌, Vol.54, No. 2, pp.452-462 (2013).
[15] 国立国語研究所（編）：太陽コーパス，国立国語研究所資料集 15，博文館新社 (2005).
[16] 近藤明日子：濁点文字使用率から見る濁音表記，雑誌「太陽」による確立期現代語の研究—「太陽コーパス」研究論文集，国立国語研究所報告 122, pp.331-350, 博文館新社 (2002).
[17] 近藤明日子，小木曽智信，須永哲矢，田中牧郎：『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).
[18] 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).
[19] 斉藤いつみ，貞光九月，浅野久子，松尾義博：正規-崩れ表記のアライメントに基づく表記崩れパタンの抽出と形態素解析への導入，情報処理学会研究報告 第 214 回自然言語処理研究会, Vol.2013-NL-214, No.5, pp1-9 (2013).
[20] 斉藤いつみ，貞光九月，浅野久子，松尾義博：正規-崩れ表記のアライメントと文字種変換を用いた崩れ表記正規化に基づく日本語形態素解析，言語処理学会第 20 回年次大会 (NLP2014) 発表論文集，pp.777-780 (2014).
[21] 佐々木彬，水野淳太，岡崎直観，乾健太郎：機械学習に基づくマイクロブログ上のテキストの正規化，人工知能学会全国大会 (JSAI2013) 論文集, 4B1-4, pp.1-4 (2013).
[22] 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：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).

# 付　　録

## A.1 提案手法による表記整理の例

提案手法を用いて実際に表記整理を行なった結果を**表 A·1** と**表 A·2** に示す．表 A·1 は明六雑誌コーパスより抜粋した記事であり，表記は全て原文に戻してある．また，表 A·2 は表 A·1 に対し，提案手法を用いて表記整理を実施した結果である，

## A.2 書き換えルール

表記揺れを考慮した辞書引きにおいて，濁点無表記及び仮名遣の不統一等を含んだ表層形の作成に使用したルールを**表 A·3** に示す．

## A.3 文字 n-gram 素性の詳細

提案手法では 2.2 節で述べた文字 n-gram 素性を各表記整理ごとに別個で用意し，それぞれ区別して利用している．具体的には各表記整理の内容を表したラベルを文字 n-gram に併記することでこれを行なっている．**表 A·4** に文字 n-gram 素性の詳細を示す．

“OKURI_*”は辞書引き時に送り仮名の統一を行なった場合に用いられるラベルである．“OKURI_IN”は原文表記の送り仮名が縮退していた場合に用い，“OKURI_OUT”は原文表記の送り仮名が漢字から飛び出している場合に用いられる．送り仮名が全て縮退している場合に限り，“OKURI_IN (ALL)”というラベルが併用され，それ以外の場合は “OKURI_* (*len*)”が併用される．変数 *len* には伸縮した文字数が入る．

“RULE_*”は，2.1.1 (2) のルールで作成した表層形の場合に利用する．“(*before*, *after*)”有りのラベルと無しのラベルを常に併用する．変数 *before* は表記整理前の文字列（表 A·3 の “未整備資料中の表記”），*after* は表記整理後の文字列（表 A·3 の “辞書登録表記”）をそれぞれ表す変数である．

“ODORI_IN_WORD”は，展開した踊字が単語自体でなく，単語の部分文字列を省略していた場合に用いられる．“(*o_type*)”有りのラベルと無しのラベルを常に併用する．変数 “*o_type*”は踊字の種類を表す変数である．

“H2K”は片仮名文字を平仮名に置き換えた時に用いられる．当該文中に平仮名文字が 1 文字でも含まれていれば，“H2K_HAS_HIRAGANA”というラベルが併用される．

“ODORI_WORD”は，展開した踊字が単語自体を省略していた場合に用いられる．“(*o_type*)”有りのラベルと無しのラベルを常に併用する．変数 “*o_type*”は踊字の種類を表す変数である．このラベルには文字 n-gram 素性を併記しない．

また今回は，文字 n-gram だけでなく，以下 4 種類の n-gram も別立てで併用している．

**文字種 n-gram:** 文字 n-gram 中の各文字を文字種で置き換えた文字列．

**濁点可能性 n-gram:** 文字 n-gram と疑似濁点無表記文字 n-gram 中の各文字を 0:濁点は付けられない，1:濁点を付けられるが付いていない，2:濁点文字，と置き換えた文字列．

**片平文字 n-gram:** 文字 n-gram と疑似濁点無表記文字 n-gram 中の片仮名文字を平仮名に置き換えた文字列．

**片平文字種 n-gram:** 片平文字 n-gram 中の各文字を文字種で置き換えた文字列．

## A.4 近代文語 UniDic 付属の素性テンプレート

本論文の評価実験では，単語 Unigram 素性および，単語 Bigram 素性として，近代文語 UniDic v1.2 に付属の MeCab 用素性テンプレートを使用している．このテンプレートには約 300 種類もの素性が記述されているが，本論文では効率化のため，それらをすべて使用せず，最低限必要であると考えられるもののみを使用した．具体的には，文献 [3] で使用している素性に相当する素性と，その他数種類を残

**表 A·1** 提案手法を用いた表記整理例：表記整理前

煉火石造ノ説

西周

余曾テ歐洲ニ遊テ煉火石造ノ家屋ヲ見ル其高キヤ五層六層其廣キヤ二町三町而テ堅牢固結搖カス可ラス撓マス可ラス嶄然タル一片ノ石壁四合スル者ナリ近日我カ新橋以北ノ街區亦此法ヲ用フ其堅牢固結彼ニ如カサル者アリト雖亦觀ルヘキアリ嗚呼何ソ匠氏ノ巧ニシテ獨リ柄政者ノ是ニ類セサルヤソレ火石ハ其質堅緻ニシテ其形方正ナル者ナリ火石能堅緻能方正撓マス曲ラス而テ匠氏能ク之ヲ用ヒ正サニ其堅緻ト方正トヲ利シ累々層々以テ其高大ヲ致ス今苟モ匠氏ノ石ヲ製スル其質堅緻ナラス其形方正ナラス專ラカノ石灰ノ力ヲ恃ミ強壓力逼シテ以テ其累々層々ヲナサムト欲セハ其石盪磨跳轢正角漫磨シ稜消毀シ變シテ將ニ團々ナル者トナラムトス則チ匠氏亦何ヲ恃ミ以テカノ高堂大厦ヲ搆スルヲ望マムヤ今有司ニシテ其下ニ遇スル強壓力逼以テ各個人々ノ權分ヲ虧損ス則チ人々己カ權分ヲ存保スルコト能ハス遂ニ變シテ圓轉流活以テ俗ヲナスニ至ラムトス苟モ圓轉流活一タヒ俗ヲナスニ至レハ間亦剛毅強直ナル者アリト雖亦碌々世ト相推移セサルヲ得ス猶數十方正石ノカノ團々ナル者ト相盪磨スルカ如シ幾ハクカソレ其角ヲ存スルヲ得ムヤ況ヤ有司ノ人ヲ遇スル其權分ヲ虧損シ且姑ク我カ意ニ隨テ汝カ方正ヲ枉ケヨト云ハヽ猶匠氏ノ石ノ方正ヲ利セス擅マヽニ槌斷斧斫シテ以テ其用ニ適セムトスルカ如シ亦焉ソ其槌斷斧斫ノ處異日壞崩滅裂ノ地タルニ非サルヲ知ラムヤソレ堅緻方正ハ火石ノ性ナリ人民ノ權分ヲ守ルハ亦人ノ性ナリ今苟モ一旦其性ヲ變シ石ヲシテ團々ナラシメ人ヲシテ圓活ナラシメハ匠氏其巧ヲ盡シ柄政者其能ヲ窮ムト雖亦將ニ其力ヲ施スノ地亡カラムトス今夫數丈ノ壁間一個ノ火石脆疎ニシテ脱スレハ餘石ノ堅緻ナル者亦從テ陷ル則チ支離滅裂亦壁ノ一面ニ及ハムトス故ニ匠氏ハ一石ノ質ヲ輕ンセス柄政者ハ匹夫ノ權ヲ慢ラス以テ能ク其功ヲ成ス況ヤ本邦ノ如キ火石素ヨリ脆疎ニシテ人民ノ權分殊ニ薄弱ナルヲヤ維新ノ初制度簡疎人々頗ル振フノ氣アリ今日ニ至リ文恬武熙萎靡復風ヲ成ス況ヤ百度更張シ節目頗ル備ハル從テ強壓力逼ノ蔽生セサルヲ得ス煉火石造ノ説ヲ作ル

**表 A·2** 提案手法を用いた表記整理例：表記整理後

煉火石造の説

西周

余曾て歐洲に遊て煉火石造の家屋を見ル其高きや五層六層其廣きや二町三町而て堅牢固結搖かす可らず撓まず可らず嶄然たる一片の石壁四合する者なり近日我か新橋以北の街區亦此法を用う其堅牢固結彼に如かざる者ありと雖亦觀るべきあり嗚呼何ぞ匠氏の巧にして獨り柄政者の是に類せざるやそれ火石は其質堅緻にして其形方正なる者なり火石能堅緻能方正撓まず曲らず而て匠氏能く之を用ゐる正さに其堅緻と方正とを利し累々層々以て其高大を致す今苟も匠氏の石を製する其質堅緻ならず其形方正ならず專らかの石灰の力を恃み強壓力逼して以て其累々層々をなさんと欲せば其石盪磨跳轢正角漫磨し稜消毀し變じて將に團々なる者とならんとす則ち匠氏亦何を恃み以てかの高堂大厦を搆するを望まんや今有司にして其下に遇する強壓力逼以て各個人々の權分を虧損す則ち人々己が權分を存保すること能はず遂に變じて圓轉流活以て俗をなすに至らんとす苟も圓轉流活一たび俗をなすに至れば間亦剛毅強直なる者ありと雖亦碌々世と相推移せざるを得ず猶數十方正石のかの團々なる者と相盪磨するが如し幾ばくかそれ其角を存するを得んや況や有司の人を遇する其權分を虧損し且姑く我か意に隨て汝が方正を枉げよと云はば猶匠氏の石の方正を利せず擅ままに槌斷斧斫して以て其用に適せんとするが如し亦焉ぞ其槌斷斧斫の處異日壞崩滅裂の地たるに非ざるを知らんやそれ堅緻方正は火石の性なり人民の權分を守るは亦人の性なり今苟も一旦其性を變じ石をして團々ならしめ人をして圓活ならしめば匠氏其巧を盡し柄政者其能を窮むと雖亦將に其力を施すの地亡からむとす今夫數丈の壁間一個の火石脆疎にして脱すれば餘石の堅緻なる者亦從て陷る則ち支離滅裂亦壁の一面に及ばんとす故に匠氏は一石の質を輕んぜず柄政者は匹夫の權を慢らず以て能く其功を成す況や本邦の如き火石素より脆疎にして人民の權分殊に薄弱なるをや維新の初制度簡疎人々頗る振ふの氣あり今日に至り文恬武熙萎靡復風を成す況や百度更張し節目頗る備はる從て強壓力逼の蔽生せざるを得ず煉火石造の説を作る

**表 A·3** 濁点無表記・仮名遣の不統一を含んだ表層形を生成するための書き換えルール

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

表 A·4　文字 n-gram 素性詳細

| 表記整理の種類 | ラベル | n-gram |
|---|---|---|
| 送り仮名の統一 | OKURI_IN | $\phi$<br>文字 n-gram<br>文字種 n-gram<br>濁点可能性 n-gram<br>片平文字 n-gram<br>片平文字種 n-gram |
| | OKURI_IN (ALL) | |
| | OKURI_IN (*len*) | |
| | OKURI_OUT | |
| | OKURI_OUT (*len*) | |
| 仮名遣の統一 | RULE_KANA | |
| | RULE_KANA (*before*, *after*) | |
| 濁点付与 | RULE_DAKU | |
| | RULE_DAKU (*before*, *after*) | |
| 仮名遣の統一+濁点付与 | RULE_KANA_DAKU | |
| | RULE_KANA_DAKU (*before*, *after*) | |
| 踊字の統一 | RULE_ODORI | |
| | RULE_ODORI (*before*, *after*) | |
| 踊字の展開（単語内文字列の省略） | ODORI_IN_WORD | |
| | ODORI_IN_WORD (*o_type*) | |
| 片仮名の変換（文字単位） | H2K | |
| | H2K_HAS_HIRAGANA | |
| 踊字の展開（単語の省略） | ODORI_WORD | |
| | ODORI_WORD (*o_type*) | |

した．

次ページより近代文語 UniDic v1.2 に付属の素性テンプレートを示す．この中で，評価実験に用いた素性はハイライトで明示している．またテンプレート内の各記号の意味は以下の通り（詳細は文献 [22] を参照のこと）．

- pos1: 品詞大分類
- pos2: 品詞中分類
- pos3: 品詞小分類
- pos4: 品詞細分類
- cType: 活用型
- cForm: 活用形
- lForm: 語彙素読み
- lemma: 語彙素
- orth: 書字形出現形
- orthBase: 書字形基本形
- pron: 発音形出現形
- pronBase: 発音形基本形
- goshu: 語種
- iType: 語頭変化型
- iForm: 語頭変化形
- fConType: 語末変化結合型
- fType: 語末変化型
- fForm: 語末変化形
- iConType: 語頭変化結合型
- t: 文字種

テンプレート内の添え字 ‘R’ は当該単語を表し，添え字 ‘L’ は当該単語の左に連接する単語を表している．また，元々の近代文語 UniDic1.2 付属の素性テンプレートでは，iType, iForm, fConType と fType, fForm, iConType の添え字 L, R がそれぞれ逆に記述されていたため，ここでは修正を施してある．

$UNIGRAM\ G01 : pos1$
$UNIGRAM\ G02 : pos1,\ pos2$
$UNIGRAM\ G03 : pos1,\ pos2,\ pos3$
$UNIGRAM\ G04 : pos1,\ pos2,\ pos3,\ pos4$

$UNIGRAM\ C01 : cType$
$UNIGRAM\ C02 : cForm$
$UNIGRAM\ C03 : cType,\ cForm$

$UNIGRAM\ GC01 : pos1,\ cType,\ cForm$
$UNIGRAM\ GC02 : pos1,\ pos2,\ cType,\ cForm$

$UNIGRAM\ T01 : t$

$UNIGRAM\ GT01 : pos1,\ t$
$UNIGRAM\ GT02 : pos1,\ pos2,\ t$
$UNIGRAM\ GT03 : pos1,\ pos2,\ pos3,\ t$
$UNIGRAM\ GT04 : pos1,\ pos2,\ pos3,\ pos4,\ t$

$UNIGRAM\ GCT01 : pos1,\ cType,\ cForm,\ t$
$UNIGRAM\ GCT02 : pos1,\ pos2,\ cType,\ cForm,\ t$

$UNIGRAM\ O01 : orth$
$UNIGRAM\ O02 : orthBase$
$UNIGRAM\ O03 : orth,\ orthBase$

$UNIGRAM\ GO01 : pos1,\ orthBase$
$UNIGRAM\ GO02 : pos1,\ pos2,\ orthBase$
$UNIGRAM\ GO03 : pos1,\ pos2,\ pos3,\ orthBase$
$UNIGRAM\ GO04 : pos1,\ pos2,\ pos3,\ pos4,\ orthBase$

$UNIGRAM\ GCO01 : pos1,\ cType,\ cForm,\ orth$
$UNIGRAM\ GCO02 : pos1,\ pos2,\ cType,\ cForm,\ orth$

$UNIGRAM\ GL01 : pos1,\ lForm,\ lemma$
$UNIGRAM\ GL02 : pos1,\ pos2,\ lForm,\ lemma$
$UNIGRAM\ GL03 : pos1,\ pos2,\ pos3,\ lForm,\ lemma$
$UNIGRAM\ GL04 : pos1,\ pos2,\ pos3,\ pos4,\ lForm,\ lemma$

$UNIGRAM\ CL01 : cType,\ cForm,\ lForm,\ lemma$

$UNIGRAM\ GCL01 : pos1,\ cType,\ cForm,\ lForm,\ lemma$
$UNIGRAM\ GCL02 : pos1,\ pos2,\ cType,\ cForm,\ lForm,\ lemma$

$UNIGRAM\ LO01 : lForm,\ lemma,\ orthBase$

$UNIGRAM\ GLO01 : pos1,\ lForm,\ lemma,\ orthBase$
$UNIGRAM\ GLO02 : pos1,\ pos2,\ lForm,\ lemma,\ orthBase$
$UNIGRAM\ GLO03 : pos1,\ pos2,\ pos3,\ lForm,\ lemma,\ orthBase$
$UNIGRAM\ GLO04 : pos1,\ pos2,\ pos3,\ pos4,\ lForm,\ lemma,\ orthBase$

$UNIGRAM\ GCLO01 : pos1,\ cType,\ cForm,\ lForm,\ lemma,\ orth$
$UNIGRAM\ GCLO02 : pos1,\ pos2,\ cType,\ cForm,\ lForm,\ lemma,\ orth$

$UNIGRAM\ W01 : goshu$

$UNIGRAM\ GW01 : pos1,\ goshu$
$UNIGRAM\ GW02 : pos1,\ pos2,\ goshu$
$UNIGRAM\ GW03 : pos1,\ pos2,\ pos3,\ goshu$
$UNIGRAM\ GW04 : pos1,\ pos2,\ pos3,\ pos4,\ goshu$

$UNIGRAM\ GCW01 : pos1,\ cType,\ cForm,\ goshu$
$UNIGRAM\ GCW02 : pos1,\ pos2,\ cType,\ cForm,\ goshu$

$UNIGRAM\ OW01 : orthBase,\ goshu$

$UNIGRAM\ LW01 : lForm,\ lemma,\ goshu$

$UNIGRAM\ GCLOW01 : pos1,\ cType,\ cForm,\ lForm,\ lemma,\ orth,\ goshu$
$UNIGRAM\ GCLOW02 : pos1,\ pos2,\ cType,\ cForm,\ lForm,\ lemma,\ orth,\ goshu$

$UNIGRAM\ GLOP01 : pos1,\ lForm,\ lemma,\ orthBase,\ pronBase$
$UNIGRAM\ GLOP02 : pos1,\ pos2,\ lForm,\ lemma,\ orthBase,\ pronBase$
$UNIGRAM\ GLOP03 : pos1,\ pos2,\ pos3,\ lForm,\ lemma,\ orthBase,\ pronBase$
$UNIGRAM\ GLOP04 : pos1,\ pos2,\ pos3,\ pos4,\ lForm,\ lemma,\ orthBase,\ pronBase$

$UNIGRAM\ GCLOP01 : pos1,\ cType,\ cForm,\ lForm,\ lemma,\ orth,\ pron$
$UNIGRAM\ GCLOP02 : pos1,\ pos2,\ cType,\ cForm,\ lForm,\ lemma,\ orth,\ pron$

$BIGRAM\ G\_G01 : pos1_R\ /\ pos1_L$
$BIGRAM\ G\_G02 : pos1_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ G\_G03 : pos1_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ G\_G04 : pos1_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$
$BIGRAM\ G\_G05 : pos1_R,\ pos2_R\ /\ pos1_L$
$BIGRAM\ G\_G06 : pos1_R,\ pos2_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ G\_G07 : pos1_R,\ pos2_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ G\_G08 : pos1_R,\ pos2_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$
$BIGRAM\ G\_G09 : pos1_R,\ pos2_R,\ pos3_R\ /\ pos1_L$
$BIGRAM\ G\_G10 : pos1_R,\ pos2_R,\ pos3_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ G\_G11 : pos1_R,\ pos2_R,\ pos3_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ G\_G12 : pos1_R,\ pos2_R,\ pos3_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$
$BIGRAM\ G\_G13 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ pos1_L$
$BIGRAM\ G\_G14 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ G\_G15 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ G\_G16 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$

$BIGRAM\ C\_C09 : cType_R,\ cForm_R\ /\ cType_L,\ cForm_L$

$BIGRAM\ G\_C01 : pos1_R\ /\ cType_L,\ cForm_L$
$BIGRAM\ G\_C02 : pos1_R,\ pos2_R\ /\ cType_L,\ cForm_L$
$BIGRAM\ G\_C03 : pos1_R,\ pos2_R,\ pos3_R\ /\ cType_L,\ cForm_L$
$BIGRAM\ G\_C04 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ cType_L,\ cForm_L$

$BIGRAM\ C\_G01 : cType_R,\ cForm_R\ /\ pos1_L$

$BIGRAM\ C\_G02 : cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ C\_G03 : cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ C\_G04 : cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$

$BIGRAM\ G\_GC01 : pos1_R\ /\ pos1_L,\ cType_L,\ cForm_L$
$BIGRAM\ G\_GC02 : pos1_R\ /\ pos1_L,\ pos2_L,\ cType_L,\ cForm_L$
$BIGRAM\ G\_GC05 : pos1_R,\ pos2_R\ /\ pos1_L,\ cType_L,\ cForm_L$
$BIGRAM\ G\_GC06 : pos1_R,\ pos2_R\ /\ pos1_L,\ pos2_L,\ cType_L,\ cForm_L$
$BIGRAM\ G\_GC09 : pos1_R,\ pos2_R,\ pos3_R\ /\ pos1_L,\ cType_L,\ cForm_L$
$BIGRAM\ G\_GC10 : pos1_R,\ pos2_R,\ pos3_R\ /\ pos1_L,\ pos2_L,\ cType_L,\ cForm_L$
$BIGRAM\ G\_GC13 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ pos1_L,\ cType_L,\ cForm_L$
$BIGRAM\ G\_GC14 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ pos1_L,\ pos2_L,\ cType_L,\ cForm_L$

$BIGRAM\ GC\_G01 : pos1_R,\ cType_R,\ cForm_R\ /\ pos1_L$
$BIGRAM\ GC\_G02 : pos1_R,\ cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ GC\_G03 : pos1_R,\ cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ GC\_G04 : pos1_R,\ cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$
$BIGRAM\ GC\_G05 : pos1_R,\ pos2_R,\ cType_R,\ cForm_R\ /\ pos1_L$
$BIGRAM\ GC\_G06 : pos1_R,\ pos2_R,\ cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ GC\_G07 : pos1_R,\ pos2_R,\ cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ GC\_G08 : pos1_R,\ pos2_R,\ cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$

$BIGRAM\ C\_GC01 : cType_R,\ cForm_R\ /\ pos1_L,\ cType_L,\ cForm_L$
$BIGRAM\ C\_GC02 : cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L,\ cType_L,\ cForm_L$

$BIGRAM\ GC\_C01 : pos1_R,\ cType_R,\ cForm_R\ /\ cType_L,\ cForm_L$
$BIGRAM\ GC\_C02 : pos1_R,\ pos2_R,\ cType_R,\ cForm_R\ /\ cType_L,\ cForm_L$

$BIGRAM\ GC\_GC01 : pos1_R,\ cType_R,\ cForm_R\ /\ pos1_L,\ cType_L,\ cForm_L$
$BIGRAM\ GC\_GC02 : pos1_R,\ cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L,\ cType_L,\ cForm_L$
$BIGRAM\ GC\_GC05 : pos1_R,\ pos2_R,\ cType_R,\ cForm_R\ /\ pos1_L,\ cType_L,\ cForm_L$
$BIGRAM\ GC\_GC06 : pos1_R,\ pos2_R,\ cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L,\ cType_L,\ cForm_L$

$BIGRAM\ O\_O05 : orthBase_R\ /\ orthBase_L$

$BIGRAM\ G\_O01 : pos1_R\ /\ orthBase_L$
$BIGRAM\ G\_O02 : pos1_R,\ pos2_R\ /\ orthBase_L$
$BIGRAM\ G\_O03 : pos1_R,\ pos2_R,\ pos3_R\ /\ orthBase_L$
$BIGRAM\ G\_O04 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ orthBase_L$

$BIGRAM\ GO\_O01 : pos1_R,\ orthBase_R\ /\ orthBase_L$
$BIGRAM\ GO\_O02 : pos1_R,\ pos2_R,\ orthBase_R\ /\ orthBase_L$
$BIGRAM\ GO\_O03 : pos1_R,\ pos2_R,\ pos3_R,\ orthBase_R\ /\ orthBase_L$
$BIGRAM\ GO\_O04 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ orthBase_R\ /\ orthBase_L$

$BIGRAM\ O\_G01 : orthBase_R\ /\ pos1_L$
$BIGRAM\ O\_G02 : orthBase_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ O\_G03 : orthBase_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ O\_G04 : orthBase_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$

$BIGRAM\ O\_GO01 : orthBase_R$ / $pos1_L$, $orthBase_L$
$BIGRAM\ O\_GO02 : orthBase_R$ / $pos1_L$, $pos2_L$, $orthBase_L$
$BIGRAM\ O\_GO03 : orthBase_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $orthBase_L$
$BIGRAM\ O\_GO04 : orthBase_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $pos4_L$, $orthBase_L$

$BIGRAM\ C\_O01 : cType_R$, $cForm_R$ / $orthBase_L$

$BIGRAM\ O\_C01 : orthBase_R$ / $cType_L$, $cForm_L$

$BIGRAM\ O\_CO01 : orthBase_R$ / $cType_L$, $cForm_L$, $orth_L$

$BIGRAM\ GC\_O01 : pos1_R$, $cType_R$, $cForm_R$ / $orthBase_L$
$BIGRAM\ GC\_O02 : pos1_R$, $pos2_R$, $cType_R$, $cForm_R$ / $orthBase_L$

$BIGRAM\ O\_GC01 : orthBase_R$ / $pos1_L$, $cType_L$, $cForm_L$
$BIGRAM\ O\_GC02 : orthBase_R$ / $pos1_L$, $pos2_L$, $cType_L$, $cForm_L$

$BIGRAM\ G\_GO01 : pos1_R$ / $pos1_L$, $orthBase_L$
$BIGRAM\ G\_GO02 : pos1_R$ / $pos1_L$, $pos2_L$, $orthBase_L$
$BIGRAM\ G\_GO03 : pos1_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $orthBase_L$
$BIGRAM\ G\_GO04 : pos1_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $pos4_L$, $orthBase_L$
$BIGRAM\ G\_GO05 : pos1_R$, $pos2_R$ / $pos1_L$, $orthBase_L$
$BIGRAM\ G\_GO06 : pos1_R$, $pos2_R$ / $pos1_L$, $pos2_L$, $orthBase_L$
$BIGRAM\ G\_GO07 : pos1_R$, $pos2_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $orthBase_L$
$BIGRAM\ G\_GO08 : pos1_R$, $pos2_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $pos4_L$, $orthBase_L$
$BIGRAM\ G\_GO09 : pos1_R$, $pos2_R$, $pos3_R$ / $pos1_L$, $orthBase_L$
$BIGRAM\ G\_GO10 : pos1_R$, $pos2_R$, $pos3_R$ / $pos1_L$, $pos2_L$, $orthBase_L$
$BIGRAM\ G\_GO11 : pos1_R$, $pos2_R$, $pos3_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $orthBase_L$
$BIGRAM\ G\_GO12 : pos1_R$, $pos2_R$, $pos3_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $pos4_L$, $orthBase_L$
$BIGRAM\ G\_GO13 : pos1_R$, $pos2_R$, $pos3_R$, $pos4_R$ / $pos1_L$, $orthBase_L$
$BIGRAM\ G\_GO14 : pos1_R$, $pos2_R$, $pos3_R$, $pos4_R$ / $pos1_L$, $pos2_L$, $orthBase_L$
$BIGRAM\ G\_GO15 : pos1_R$, $pos2_R$, $pos3_R$, $pos4_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $orthBase_L$
$BIGRAM\ G\_GO16 : pos1_R$, $pos2_R$, $pos3_R$, $pos4_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $pos4_L$, $orthBase_L$

$BIGRAM\ GO\_G01 : pos1_R$, $orthBase_R$ / $pos1_L$
$BIGRAM\ GO\_G02 : pos1_R$, $orthBase_R$ / $pos1_L$, $pos2_L$
$BIGRAM\ GO\_G03 : pos1_R$, $orthBase_R$ / $pos1_L$, $pos2_L$, $pos3_L$
$BIGRAM\ GO\_G04 : pos1_R$, $orthBase_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $pos4_L$
$BIGRAM\ GO\_G05 : pos1_R$, $pos2_R$, $orthBase_R$ / $pos1_L$
$BIGRAM\ GO\_G06 : pos1_R$, $pos2_R$, $orthBase_R$ / $pos1_L$, $pos2_L$
$BIGRAM\ GO\_G07 : pos1_R$, $pos2_R$, $orthBase_R$ / $pos1_L$, $pos2_L$, $pos3_L$
$BIGRAM\ GO\_G08 : pos1_R$, $pos2_R$, $orthBase_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $pos4_L$
$BIGRAM\ GO\_G09 : pos1_R$, $pos2_R$, $pos3_R$, $orthBase_R$ / $pos1_L$
$BIGRAM\ GO\_G10 : pos1_R$, $pos2_R$, $pos3_R$, $orthBase_R$ / $pos1_L$, $pos2_L$
$BIGRAM\ GO\_G11 : pos1_R$, $pos2_R$, $pos3_R$, $orthBase_R$ / $pos1_L$, $pos2_L$, $pos3_L$
$BIGRAM\ GO\_G12 : pos1_R$, $pos2_R$, $pos3_R$, $orthBase_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $pos4_L$
$BIGRAM\ GO\_G13 : pos1_R$, $pos2_R$, $pos3_R$, $pos4_R$, $orthBase_R$ / $pos1_L$
$BIGRAM\ GO\_G14 : pos1_R$, $pos2_R$, $pos3_R$, $pos4_R$, $orthBase_R$ / $pos1_L$, $pos2_L$
$BIGRAM\ GO\_G15 : pos1_R$, $pos2_R$, $pos3_R$, $pos4_R$, $orthBase_R$ / $pos1_L$, $pos2_L$, $pos3_L$
$BIGRAM\ GO\_G16 : pos1_R$, $pos2_R$, $pos3_R$, $pos4_R$, $orthBase_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $pos4_L$

$BIGRAM\ GO\_GO01 : pos1_R$, $orthBase_R$ / $pos1_L$, $orthBase_L$
$BIGRAM\ GO\_GO02 : pos1_R$, $orthBase_R$ / $pos1_L$, $pos2_L$, $orthBase_L$
$BIGRAM\ GO\_GO03 : pos1_R$, $orthBase_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $orthBase_L$
$BIGRAM\ GO\_GO04 : pos1_R$, $orthBase_R$ / $pos1_L$, $pos2_L$, $pos3_L$, $pos4_L$, $orthBase_L$

$BIGRAM\ GO\_GO05 : pos1_R,\ pos2_R,\ orthBase_R\ /\ pos1_L,\ orthBase_L$
$BIGRAM\ GO\_GO06 : pos1_R,\ pos2_R,\ orthBase_R\ /\ pos1_L,\ pos2_L,\ orthBase_L$
$BIGRAM\ GO\_GO07 : pos1_R,\ pos2_R,\ orthBase_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ orthBase_L$
$BIGRAM\ GO\_GO08 : pos1_R,\ pos2_R,\ orthBase_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ orthBase_L$
$BIGRAM\ GO\_GO09 : pos1_R,\ pos2_R,\ pos3_R,\ orthBase_R\ /\ pos1_L,\ orthBase_L$
$BIGRAM\ GO\_GO10 : pos1_R,\ pos2_R,\ pos3_R,\ orthBase_R\ /\ pos1_L,\ pos2_L,\ orthBase_L$
$BIGRAM\ GO\_GO11 : pos1_R,\ pos2_R,\ pos3_R,\ orthBase_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ orthBase_L$
$BIGRAM\ GO\_GO12 : pos1_R,\ pos2_R,\ pos3_R,\ orthBase_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ orthBase_L$
$BIGRAM\ GO\_GO13 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ orthBase_R\ /\ pos1_L,\ orthBase_L$
$BIGRAM\ GO\_GO14 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ orthBase_R\ /\ pos1_L,\ pos2_L,\ orthBase_L$
$BIGRAM\ GO\_GO15 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ orthBase_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ orthBase_L$
$BIGRAM\ GO\_GO16 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ orthBase_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ orthBase_L$

$BIGRAM\ C\_CO01 : cType_R,\ cForm_R\ /\ cType_L,\ cForm_L,\ orth_L$

$BIGRAM\ CO\_C01 : cType_R,\ cForm_R,\ orth_R\ /\ cType_L,\ cForm_L$

$BIGRAM\ CO\_CO01 : cType_R,\ cForm_R,\ orth_R\ /\ cType_L,\ cForm_L,\ orth_L$

$BIGRAM\ G\_CO01 : pos1_R\ /\ cType_L,\ cForm_L,\ orth_L$
$BIGRAM\ G\_CO02 : pos1_R,\ pos2_R\ /\ cType_L,\ cForm_L,\ orth_L$
$BIGRAM\ G\_CO03 : pos1_R,\ pos2_R,\ pos3_R\ /\ cType_L,\ cForm_L,\ orth_L$
$BIGRAM\ G\_CO04 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ cType_L,\ cForm_L,\ orth_L$

$BIGRAM\ GO\_C01 : pos1_R,\ orthBase_R\ /\ cType_L,\ cForm_L$
$BIGRAM\ GO\_C02 : pos1_R,\ pos2_R,\ orthBase_R\ /\ cType_L,\ cForm_L$
$BIGRAM\ GO\_C03 : pos1_R,\ pos2_R,\ pos3_R,\ orthBase_R\ /\ cType_L,\ cForm_L$
$BIGRAM\ GO\_C04 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ orthBase_R\ /\ cType_L,\ cForm_L$

$BIGRAM\ GO\_CO01 : pos1_R,\ orthBase_R\ /\ cType_L,\ cForm_L,\ orth_L$
$BIGRAM\ GO\_CO02 : pos1_R,\ pos2_R,\ orthBase_R\ /\ cType_L,\ cForm_L,\ orth_L$
$BIGRAM\ GO\_CO03 : pos1_R,\ pos2_R,\ pos3_R,\ orthBase_R\ /\ cType_L,\ cForm_L,\ orth_L$
$BIGRAM\ GO\_CO04 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ orthBase_R\ /\ cType_L,\ cForm_L,\ orth_L$

$BIGRAM\ C\_GO01 : cType_R,\ cForm_R\ /\ pos1_L,\ orthBase_L$
$BIGRAM\ C\_GO02 : cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L,\ orthBase_L$
$BIGRAM\ C\_GO03 : cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ orthBase_L$
$BIGRAM\ C\_GO04 : cType_R,\ cForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ orthBase_L$

$BIGRAM\ CO\_G01 : cType_R,\ cForm_R,\ orth_R\ /\ pos1_L$
$BIGRAM\ CO\_G02 : cType_R,\ cForm_R,\ orth_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ CO\_G03 : cType_R,\ cForm_R,\ orth_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ CO\_G04 : cType_R,\ cForm_R,\ orth_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$

$BIGRAM\ CO\_GO01 : cType_R,\ cForm_R,\ orth_R\ /\ pos1_L,\ orthBase_L$
$BIGRAM\ CO\_GO02 : cType_R,\ cForm_R,\ orth_R\ /\ pos1_L,\ pos2_L,\ orthBase_L$
$BIGRAM\ CO\_GO03 : cType_R,\ cForm_R,\ orth_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ orthBase_L$
$BIGRAM\ CO\_GO04 : cType_R,\ cForm_R,\ orth_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ orthBase_L$

$BIGRAM\ GCO\_GCO01 : pos1_R,\ cType_R,\ cForm_R,\ orth_R\ /\ pos1_L,\ cType_L,\ cForm_L,\ orth_L$
$BIGRAM\ GCO\_GCO02 : pos1_R,\ cType_R,\ cForm_R,\ orth_R\ /\ pos1_L,\ pos2_L,\ cType_L,\ cForm_L,\ orth_L$
$BIGRAM\ GCO\_GCO05 : pos1_R,\ pos2_R,\ cType_R,\ cForm_R,\ orth_R\ /\ pos1_L,\ cType_L,\ cForm_L,\ orth_L$
$BIGRAM\ GCO\_GCO06 : pos1_R,\ pos2_R,\ cType_R,\ cForm_R,\ orth_R\ /\ pos1_L,\ pos2_L,\ cType_L,\ cForm_L,\ orth_L$

$BIGRAM\ W\_W01 : goshu_R\ /\ goshu_L$

$BIGRAM\ G\_W01 : pos1_R\ /\ goshu_L$
$BIGRAM\ G\_W02 : pos1_R,\ pos2_R\ /\ goshu_L$
$BIGRAM\ G\_W03 : pos1_R,\ pos2_R,\ pos3_R\ /\ goshu_L$
$BIGRAM\ G\_W04 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ goshu_L$

$BIGRAM\ W\_G01 : goshu_R\ /\ pos1_L$
$BIGRAM\ W\_G02 : goshu_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ W\_G03 : goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ W\_G04 : goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$

$BIGRAM\ GW\_GW01 : pos1_R,\ goshu_R\ /\ pos1_L,\ goshu_L$
$BIGRAM\ GW\_GW02 : pos1_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ goshu_L$
$BIGRAM\ GW\_GW03 : pos1_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ goshu_L$
$BIGRAM\ GW\_GW04 : pos1_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ goshu_L$
$BIGRAM\ GW\_GW05 : pos1_R,\ pos2_R,\ goshu_R\ /\ pos1_L,\ goshu_L$
$BIGRAM\ GW\_GW06 : pos1_R,\ pos2_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ goshu_L$
$BIGRAM\ GW\_GW07 : pos1_R,\ pos2_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ goshu_L$
$BIGRAM\ GW\_GW08 : pos1_R,\ pos2_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ goshu_L$
$BIGRAM\ GW\_GW09 : pos1_R,\ pos2_R,\ pos3_R,\ goshu_R\ /\ pos1_L,\ goshu_L$
$BIGRAM\ GW\_GW10 : pos1_R,\ pos2_R,\ pos3_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ goshu_L$
$BIGRAM\ GW\_GW11 : pos1_R,\ pos2_R,\ pos3_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ goshu_L$
$BIGRAM\ GW\_GW12 : pos1_R,\ pos2_R,\ pos3_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ goshu_L$
$BIGRAM\ GW\_GW13 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ goshu_R\ /\ pos1_L,\ goshu_L$
$BIGRAM\ GW\_GW14 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ goshu_L$
$BIGRAM\ GW\_GW15 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ goshu_L$
$BIGRAM\ GW\_GW16 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ goshu_L$

$BIGRAM\ GCW\_GCW01 : pos1_R,\ cType_R,\ cForm_R,\ goshu_R\ /\ pos1_L,\ cType_L,\ cForm_L,\ goshu_L$
$BIGRAM\ GCW\_GCW02 : pos1_R,\ cType_R,\ cForm_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ cType_L,\ cForm_L,\ goshu_L$
$BIGRAM\ GCW\_GCW05 : pos1_R,\ pos2_R,\ cType_R,\ cForm_R,\ goshu_R\ /\ pos1_L,\ cType_L,\ cForm_L,\ goshu_L$
$BIGRAM\ GCW\_GCW06 : pos1_R,\ pos2_R,\ cType_R,\ cForm_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ cType_L,\ cForm_L,\ goshu_L$

$BIGRAM\ OW\_OW01 : orthBase_R,\ goshu_R\ /\ orthBase_L,\ goshu_L$

$BIGRAM\ GOW\_GOW01 : pos1_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW02 : pos1_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW03 : pos1_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW04 : pos1_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW05 : pos1_R,\ pos2_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW06 : pos1_R,\ pos2_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW07 : pos1_R,\ pos2_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW08 : pos1_R,\ pos2_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW09 : pos1_R,\ pos2_R,\ pos3_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW10 : pos1_R,\ pos2_R,\ pos3_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW11 : pos1_R,\ pos2_R,\ pos3_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW12 : pos1_R,\ pos2_R,\ pos3_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW13 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW14 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW15 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ orthBase_L,\ goshu_L$
$BIGRAM\ GOW\_GOW16 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ orthBase_R,\ goshu_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ orthBase_L,\ goshu_L$

$BIGRAM\ CO\_COW01 : cType_R,\ cForm_R,\ orth_R\ /\ cType_L,\ cForm_L,\ orth_L,\ goshu_L$

$BIGRAM\ COW\_C01 : cType_R,\ cForm_R,\ orth_R,\ goshu_R\ /\ cType_L,\ cForm_L$

$BIGRAM\ COW\_COW01 : cType_R,\ cForm_R,\ orth_R,\ goshu_R\ /\ cType_L,\ cForm_L,\ orth_L,\ goshu_L$

$BIGRAM\ F\_F01 : fConType_R\ /\ fType_L,\ fForm_L$

$BIGRAM\ FO\_F01 : fConType_R,\ orthBase_R\ /\ fType_L,\ fForm_L$

$BIGRAM\ F\_FO01 : fConType_R\ /\ fType_L,\ fForm_L,\ orthBase_L$

$BIGRAM\ FO\_FO01 : fConType_R,\ orthBase_R\ /\ fType_L,\ fForm_L,\ orthBase_L$

$BIGRAM\ I\_I01 : iType_R,\ iForm_R\ /\ iConType_L$

$BIGRAM\ I\_IO01 : iType_R,\ iForm_R\ /\ iConType_L,\ orthBase_L$

$BIGRAM\ IO\_I01 : iType_R,\ iForm_R,\ orthBase_R\ /\ iConType_L$

$BIGRAM\ IO\_IO01 : iType_R,\ iForm_R,\ orthBase_R\ /\ iConType_L,\ orthBase_L$

$BIGRAM\ G\_GF01 : pos1_R\ /\ pos1_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF02 : pos1_R\ /\ pos1_L,\ pos2_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF03 : pos1_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF04 : pos1_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF05 : pos1_R,\ pos2_R\ /\ pos1_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF06 : pos1_R,\ pos2_R\ /\ pos1_L,\ pos2_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF07 : pos1_R,\ pos2_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF08 : pos1_R,\ pos2_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF09 : pos1_R,\ pos2_R,\ pos3_R\ /\ pos1_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF10 : pos1_R,\ pos2_R,\ pos3_R\ /\ pos1_L,\ pos2_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF11 : pos1_R,\ pos2_R,\ pos3_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF12 : pos1_R,\ pos2_R,\ pos3_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF13 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ pos1_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF14 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ pos1_L,\ pos2_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF15 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ fType_L,\ fForm_L$
$BIGRAM\ G\_GF16 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L,\ fType_L,\ fForm_L$

$BIGRAM\ GI\_G01 : pos1_R,\ iType_R,\ iForm_R\ /\ pos1_L$
$BIGRAM\ GI\_G02 : pos1_R,\ iType_R,\ iForm_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ GI\_G03 : pos1_R,\ iType_R,\ iForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ GI\_G04 : pos1_R,\ iType_R,\ iForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$
$BIGRAM\ GI\_G05 : pos1_R,\ pos2_R,\ iType_R,\ iForm_R\ /\ pos1_L$
$BIGRAM\ GI\_G06 : pos1_R,\ pos2_R,\ iType_R,\ iForm_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ GI\_G07 : pos1_R,\ pos2_R,\ iType_R,\ iForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ GI\_G08 : pos1_R,\ pos2_R,\ iType_R,\ iForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$
$BIGRAM\ GI\_G09 : pos1_R,\ pos2_R,\ pos3_R,\ iType_R,\ iForm_R\ /\ pos1_L$
$BIGRAM\ GI\_G10 : pos1_R,\ pos2_R,\ pos3_R,\ iType_R,\ iForm_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ GI\_G11 : pos1_R,\ pos2_R,\ pos3_R,\ iType_R,\ iForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ GI\_G12 : pos1_R,\ pos2_R,\ pos3_R,\ iType_R,\ iForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$
$BIGRAM\ GI\_G13 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ iType_R,\ iForm_R\ /\ pos1_L$
$BIGRAM\ GI\_G14 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ iType_R,\ iForm_R\ /\ pos1_L,\ pos2_L$
$BIGRAM\ GI\_G15 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ iType_R,\ iForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L$
$BIGRAM\ GI\_G16 : pos1_R,\ pos2_R,\ pos3_R,\ pos4_R,\ iType_R,\ iForm_R\ /\ pos1_L,\ pos2_L,\ pos3_L,\ pos4_L$