和歌解析用MeCab辞書の開発
—八代集解析済みコーパスによる学習—

山元 啓史

東京工業大学

本稿では八代集 (905 年頃〜1205 年) 用の品詞解析済みコーパスと品詞タグつき辞書を用いて、CRF (Conditional Random Field) 法による連接コスト計算を実施し、二十一代集 (905 年頃〜1439 年) に対応した和歌用形態素解析辞書の開発について述べる。八代集テキストの CRF による連接コスト学習の結果、90.3%の連接が正しく解析できた。この辞書をもとに今後徐々に二十一代集テキストを増やしながら、辞書を育て、最終的には二十一代集対応の辞書に仕上げていく。
キーワード：和歌、辞書編集、形態素解析、八代集、二十一代集、連接コスト

Development of the MeCab Dictionary for Classical Japanese Poems Based on the *Hachidaishū* Corpus

Hilofumi Yamamoto

Tokyo Institute of Technology

This paper addresses the development of the dictionary of the *Nijūichidaishū* (ca. 905–1439) for MeCab, morphological analysis parser, based on the dictionary of the *Hachidaishū* (ca. 905–1205). The CRF (Conditional Random Fields) method is used to calculate connection rules weights which indicate the coherance of any two words (bigrams) in corpus/sentences. As a result of the compilation of a dictionary, 90.3% of the parsing accuracy is obtained.
Keywords: classical Japanese poetry dictionary compilation morphological parser, the *Hachidaishū*, the *Nijūichidaishū*, connection rules weights

1 はじめに

山元 [7] は和歌用の形態素解析辞書および形態素解析システム kh を開発した。その対象は八代集（905 年頃〜1205 年）に限定されていた。本研究の大きな目的の一つは、その辞書を八代集から二十一代集（905 年〜1439 年）に解析可能対象を拡大することである。ところが、この八代集辞書には連接規則情報がないため、未知のフレーズはうまく解析できず、入出力を逐一確認し、適切な解析が得られるように辞書を育てなければならなく、膨大な作業を要していた。そこで、本研究では八代集辞書を初期基本辞書として用い、二十一代集の処理を通して用語の連接規則情報を計算処理によって学習し、二十一代集用解析辞書に仕上げることを試みる。

1.1 語彙の体系を目で見てみたい

さて、語と語は互いに結びつきあって、どんな意味のまとまりを作っているのだろうか。和歌の場合なら「梅と鶯」、「桜と時鳥」、「吉野と桜」、「龍田と紅葉」のように和歌ならではのコンビネーションが思い浮かぶが、このようなコンビネーションはい

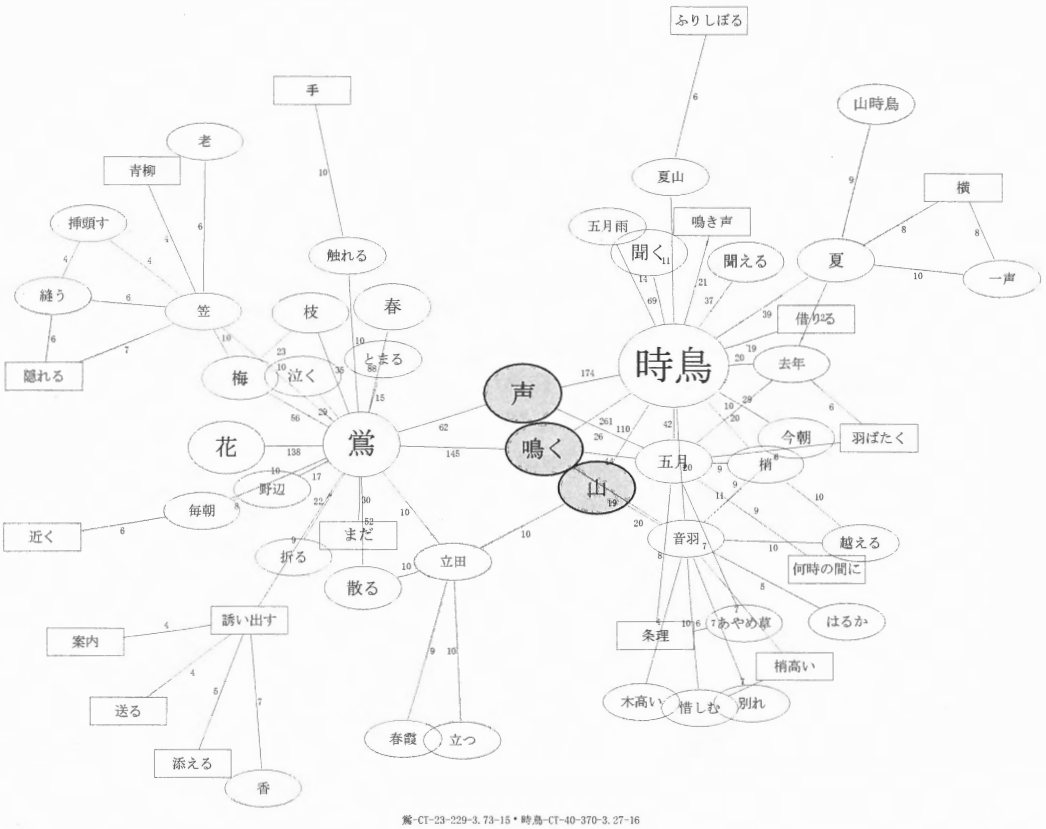

図 1: 鶯と時鳥の合成グラフモデル: 鶯と時鳥の２語について、古今集の和歌とその現代語を比較し、グラフで描いた。網掛けは共有ノード。楕円は和歌の語、矩形は現代語訳にのみ現れた語。エッジの数字は共出現の頻度。山元 [6] より。

くつぐらい、どんな種類が存在し、どんな意味合いで、どの時代から使われはじめ、それらは互いにどれぐらいの強さで結びついているのだろうか。

和歌研究者の直観や経験だけでは即答しにくいコンビネーションを実際に和歌データから獲得する企みとして、図 1 に示す可視化モデルを作成し、和歌用語の体系について論考を重ねてきた。たとえば、地名の例でいうなら「龍田」は紅葉彩る秋の風景、「吉野」は桜をとりまく春の花模様として有名であるが、可視化モデルを通して見るとそれだけでなく「龍田」は「神の地」、「吉野」は「人間世界／世俗の地」というまとまりをも観察することができた [6]。その後、2009 年までに八代集用語について辞書とシソーラスを整備し [7, 8]、八代集限定ではあるが、和歌用語の可視化モデルを完成させた。本研究はこれを基礎にして、八代集（905 年頃〜1205 年、9440 首）の 300 年間だけでなく、二十一代集（905 年頃〜1439 年、25,648 首）、534 年間の大きな古典の知識を蓄積し、体系化を進めるのが究極的な目的ではあるが、いくつかの基本的な問題点があるので、ここではその問題と解決について多少説明を加えたい。

1.2 辞書開発の必要性

八代集から二十一代集に処理対象を拡張すると、分析できる時代が 500 年間に広がるだけでなく、歌の数も 25,000 首以上になる。データが多くなればなるほど、語と語の組合せ頻度もある程度得られ、語相互の連接規則を統計的に推定するには都合がよい。しかし、そのためには単位分析（歌を単語に分割し、品詞名を各単語に付ける作業）が必要であるが、25,000 首のすべてに対し、手作業で行うには限界がある。たとえば、単語の分割は一通りではなく、長く切る場合（例「うらふきかへす」）もあれば、短く切る場合（「うら/ふき/かへす」）もあり、切り方を統一しておかなければ、語彙一覧表に見られる単語の種類や頻度が異なり、結果的に出現頻度計算が無意味になってしまう。さらに、これを人手によって行うとなると切り方の判断に揺れが生じ、

不統一なデータができてしまう。均一な処理を何度も繰り返し実施するには、量の多少に関わらず計算機で行うべきである。

従来にも和歌を計算機処理する試みはいくつか存在する。たとえば、近藤ら [1, 2] によるNグラム統計による方法や竹田ら [4] の LCS（Longest Common Subsequence）法などである。いずれも辞書を用いずに文字列のみを操作して、その目的に応じた研究成果をあげている。しかしながら、古典知識を蓄積するためには、活用語を基本語に変換、表記を1つ（あるいは意味コード）に統一して、文法や意味の構造も柔軟かつ汎用的に取り扱いたい。文字列をそのまま扱う（あるいは一般の古語辞典を転用する）方法では、異なる表記の同語（異形同語：たとえば、京都の地名「音羽川」はその形が「おとは／をとは／音羽」と「かは／がは／川／河」の組合せ数存在する）、同じ表記の異語（同形異語：たとえば、ワ行下二段動詞「植う」の未然形あるいは連用形と一般名詞「上」は共に「うへ」である）の判別、宛て字（たとえば「立覧 [たつらむ]」、「契剣 [ちぎりけん]」、「思ふ蝶 [おもふてふ]」など）における単語の分割と品詞の特定、基本単語（特に目、手のような身体語）、助詞・助動詞に多く見られる一音節単語の特定は難しい。

従来、機械学習によって辞書の連接情報を得るには大量の処理済みコーパスが必要とされてきた。しかしながら、和歌の電子テキストはあるものの、機械学習によって連接規則を得るほどの量の【処理済みコーパス】はなく、ひとつひとつ手作業によって解析済みコーパスを作成せざるを得なかった。

1.3 和歌特有の問題点

和歌を計算処理するためには和歌特有の問題点があり、十分注意して取り扱わなければならない。たとえば、和歌の表記に①二句切れ、三句切れのように歌の途中で意味上終るものがあるが、句点などなく明示的に文の終りを示す手がかりがない。歌の途中で文が終っているのか、次の句を修飾しているのかがわからないため、連体形なのか終止形なのか判別できない。②和歌大系本やデータベースの中には、あらかじめ「／」のような句の切れ目を示す記号を入れたものがある場合とない場合がある。③「ゝ」や「〳〵」などの踊り字がある場合とない場合がある。特に「ゝ」の場合、前の語尾に同じ音が語頭に続く場合、たとえ単語をまたいでいても「ゝ」が当てられている場合があり、分割・集計した後、それがどの語であったのか分からなくなってしまう。④仮名文には清濁の明示はないが、大系本には読者への便宜を図り、清濁や漢字を適当に施したものがある。しかし、清濁を明示しないことで掛詞を示す場合は清濁をつけないこともある。

1.4 八代集辞書の問題点

山元 [7] の八代集辞書には連接規則に関わる情報がない。その代わりに八代集に見られる単語の連鎖パターンをすべて登録してあり、適切な出力が得られるようにしてある。古文には単音節の単語が複数連接して、ひとまとまりの意味を示すことが多い。たとえば、「ながめせしまに」は「ながめ」「せ」「し」「ま」「に」に分解され、「ながめ」以外は一音節のすべて品詞の異なる単語が連接している。山元 [7] は、このような複合した語群を単純に辞書に登録して、見出語と解析済品詞列を入れ換えるだけの、最長一致法による品詞タグづけシステムを開発した[1]。

しかしながら、上記方式では未知のパターンはうまく解析できず、新しい連鎖パターンが出てくるたびに辞書を逐一育てなければならなく、膨大な作業を要していた。現代語の形態素解析では、すでに大量の解析済みデータと現代語辞書、そして標準的な日本語の表記があるため、機械学習による連接規則の学習が可能であり、その規則を有する辞書を用いて新たな現代文をほぼ完全に近い形で解析することができている。一方、古語には大量の解析済みデータもなく、連接規則情報つきの辞書もなく、そして標準的な表記もない。

そこで、本研究では今までに蓄積してきた八代集辞書と八代集解析済みデータを元手にして、徐々にテキスト量を増やし、最終的に二十一代集処理用に仕上げることを計画する。まず計算処理により、八代集までの連接規則を学習させ、その未熟な辞書を用いて二十一代集テキストを少しずつ処理させていく。始めから二十一代集すべてを処理させたのでは未知語が多く、正しい方向性を持った連接コストの計算が保証されない。少しずつ処理させ、未知語や誤解析を修正し、それを辞書に反映させ、徐々にテキスト量を増やしていき、解析精度も高め、最終的に二十一代集対応の辞書に仕上げるのである。

[1] 言い換えれば、八代集（約 9,500 首）に存在する連接パターンを異なる表記も含めてすべて人手で辞書に登録したにすぎない。しかし、この作業がなければ、揺れのない解析結果は得られなかった。

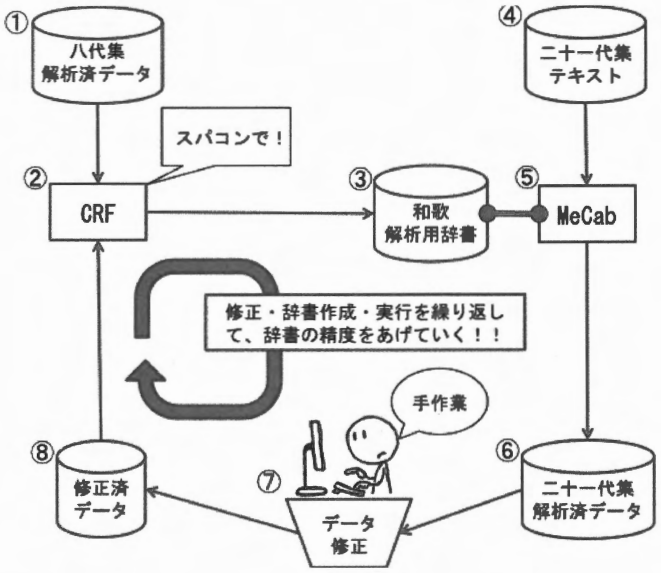

図 2: 辞書開発と連接規則獲得の手順: 山元 [7] 開発の①八代集用辞書を②CRF（連接パラメタ推定プログラム）で処理し③二十一代集用の初期辞書を作成。④二十一代集テキストを準備し⑤MeCab（形態素解析器）と③で⑥二十一代集を解析。誤りや未知語は手作業で⑦修正。⑧修正済データを再び②CRF で処理し、③辞書を作成。②〜⑧を繰り返し、徐々に精度の高い辞書を得る。

2 方法

二十一代集のための辞書開発　研究方法は図 2 に沿って説明する。まず、山元 (2007) [7] 開発の①八代集用の解析済みデータと②CRF (Conditional Random Fields) 法[2]を用いて仮の解析辞書を作成する。CRF は語と語のつながりの程度（コスト）を統計的に推定するプログラム（連接パラメタの推定）で、辞書の主要な部分を生成する。

多種多様な表現形式の収集　つぎに④二十一代集テキストの準備である。これは国文学研究資料館の二十一代集データベースを用いる[3]ほか、古典文学大系本その他をスキャンし、表記情報（漢字仮名混じり、送り仮名などの異なりや揺れを調査したものを追加し、多種多様な表記に対応したテキストデータを作成する。この作業にはドキュメントスキャナを用いて、電子テキスト化し、コンピュータプログラムによって表記の異なりや揺れを一括して収集整理できる状態にしておく。しかしながら、手作業によるところも多い。

MeCab で形態素解析　仮の辞書と二十一代集テキストが準備できれば、⑤MeCab で形態素解析を行う。MeCab を本研究に採用した理由は、これが既成の品詞体系に依存しない設計になっており、現代語のみならず古代語であっても独自の品詞体系で形態素解析器が自作できるからである。たとえば、守岡 [5] は MeCab のこの特徴に注目し、古典中国語を形態素解析するための辞書を開発している。本研究の場合でも、和歌にありがちな独自の品詞体系も十分に設定できるものと判断した。MeCab 辞書の作り方は、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`）の手順に従った。

MeCab で処理した後、⑥二十一代集の解析済みデータを得られるが、このデータには誤りや未知語があるので、それを手作業で⑦修正・追加し、⑧修正済みデータを作成する。

辞書の精度をあげていく　⑧修正済みデータを用いて、再度②CRF で前回よりも精度のよい③辞書を作成する。ただし、この CRF による辞書作成には大量のメモリと計算速度が必要なので、東京工業大学のスパコン TSUBAME を用いて、効率的に行う。二回目以降の修正作業では、単に辞書の追加や修正

[2] `http://mecab.sourceforge.net/`
[3] 既に同館知的財産委員会より利用許諾は得ている。

だけでなく、品詞体系の見直しや新たな連接規則の導入も試みる。これには辞書やテキストを実際に目で追いかけながら、連接規則を分析的に眺める作業が不可欠である。この点が本研究における本質的かつ忍耐力が必要な部分である。以上、②〜⑧までを何度も繰り返し、徐々に精度の高い辞書を作成して行く。

時代別・歌集別の辞書の検討　二十一代集テキストは一度にすべてを処理せず、歌集ごとに辞書に項目を追加しながら辞書を育てて行く方法にしておけば、類似の誤解析を減らすことができるだけでなく、その育てていく過程において、時代別・歌集別に辞書を分割しておいた法が望ましいかどうかを検討していく。時代別・歌集別に分割した方が効率が良ければ、辞書開発を通して、それぞれの特徴が抽出できることも考えられる。

以下では、第一段階として、従来の八代集辞書と八代集処理済みコーパスを用いて、八代集における用語連接コストの学習と、MeCab で八代集テキストを処理してみた結果、どの程度の再現率が得られたのかについて報告する。

3　材料：八代集辞書の収録内容

初期辞書として利用される八代集辞書には、新編国歌大観 CD-ROM 版の二十一代集に相当するデータ [9]、国文学研究資料館編集二十一代集データベース [3]、新日本古典文学大系本二十一代集に相当する書籍その他、新潮日本古典集成の新古今集、ヴァージニア大学日本語テキストイニシアティブ（`http://etext.lib.virginia.edu/japanese/`）監修の二十一代集データから、それらにすべてに見られる用語がそれぞれの表記で登録されている。MeCab による処理実験に用いるテキストは国文学研究資料館二十一代集データベースの中の八代集のすべての和歌を用いる。

4　解析結果

八代集辞書を MeCab 用辞書に変換して、再び八代集テキストを解析し、評価を実施した。また、八代集以外のテキスト（新後撰和歌集）についても解析実験を行った。CRF による連接コスト学習は、所有の自作パソコン、Linux Kernel version 2.6.27.15 (gcc version 4.2.4)、Intel(R)Core(TM)2 Duo CPU E7200 2.53GHz (cache size:3072 KB)、全記憶容量 2GB、スワップサイズ 20GB で計算した。すべての和歌テキストを連接コストの学習に用いたかったが、メモリを使い尽したため、テキストを国文学研究資料館二十一代集データベースを中心にメモリ容量極限の 11,119 行に限定し、連接コストの計算をやり直した。

表 2 は八代集収録以外の歌、新後撰和歌集の 4 番歌を新編国歌大観 CD-ROM と国文学研究資料館データベースの 2 種類のテキストで処理したものである。「まで／まて」のように清濁の有無は問題なく解析できているが、「みよし野の／みよしのゝ」は踊字「ゝ」がうまく解析できていない。EOS の前の「かな」はいずれにおいても解析できている。

どの程度正しく解析できたかを評価するために、MeCab パッケージ標準添付の `mecab-system-eval` プログラムを用いて、MeCab の結果とすでに山元 [7] において処理した結果との差異（`precision / recall`）を計算した。その結果、1 番目の素性（主に品詞名のみ）の特定については、99.7%、すべての素性（活用形などに代表される品詞の下位分類）の特定については、90.3%の解析再現が確認できた。

5　おわりに

本稿では八代集用の辞書を用いて、連接コストを CRF により学習し、二十一代集を処理するための初期基本辞書の試作を行った。現行では手持ちのパソコンの計算速度ならびに記憶容量の制約ですべての八代集テキストを用いた試作実験が行えなかった。今後はスパコン上に処理環境を作り、まず今回行えなかった八代集の全テキストで初期辞書を作成していく。その上で、二十一代集を処理しつつ、徐々に精度をあげて、二十一代集すべての和歌が柔軟に処理できる辞書に仕上げる予定である。さらに、この作業を通して得られる連接情報をもとに、和歌の連接規則の理論化も進めていきたいと考えている。

参考文献

[1] 近藤みゆき：n-gram 統計による語形の抽出と複合語—平安時代語の分析から—, 日本語学, Vol. 20, pp. 79–89 (2001).

[2] 近藤泰弘, 近藤みゆき：平安時代古典語古典文学研究のための N-gram を用いた解析手法, 言語処理学会第 7 回年次大会発表論文集, 第 7 巻, pp. 209–212 言語処理学会 (2001).

[3] 中村康夫, 立川美彦, 杉田まゆ子：国文学研究資料館データベース 古典コレクション『二十一代集』(正保版本) CD-ROM, 岩波書店, 東京 (1999).

表 1: **MeCab** 用に整理しなおした八代集辞書（**Seed** 辞書の例）

```
うつつ,0,0,0, 名詞, 一般,*,*,*,*, 現, うつつ,*,*
うつぶしそめ,0,0,0, 名詞, 一般,*,*,*,*, 空五倍子染め, うつぶしぞめ,*,*
うつぶしぞめ,0,0,0, 名詞, 一般,*,*,*,*, 空五倍子染め, うつぶしぞめ,*,*
うつぶし染め,0,0,0, 名詞, 一般,*,*,*,*, 空五倍子染め, うつぶしぞめ,*,*
うつま,0,0,0, 動詞,*,*,*, 四段・マ行, 未, 埋む, うづむ, 埋ま, うづま
うつまき,0,0,0, 名詞, 一般,*,*,*,*, 渦巻, うづまき,*,*
うつむ,0,0,0, 動詞,*,*,*, 下二段・マ行, 終, 埋む, うづむ, 埋む, うづむ
うつもるゝ,0,0,0, 動詞,*,*,*, 下二段・ラ行, 体, 埋もる, うづもる, 埋もるる, うづもるる
うつもれ,0,0,0, 動詞,*,*,*, 下二段・ラ行, 未, 埋もる, うづもる, 埋もれ, うづもれ
うつもれ,0,0,0, 動詞,*,*,*, 下二段・ラ行, 未用, 埋もる, うづもる, 埋もれ, うづもれ
うつもれ,0,0,0, 動詞,*,*,*, 下二段・ラ行, 用, 埋もる, うづもる, 埋もれ, うづもれ
うつら,0,0,0, 動詞,*,*,*, 四段・ラ行, 未, 移る, うつる, 移ら, うつら
うつら,0,0,0, 動詞,*,*,*, 四段・ラ行, 未, 映る, うつる, 映ら, うつら
うつら,0,0,0, 名詞, 一般,*,*,*,*, 鶉, うづら,*,*
うつり,0,0,0, 動詞,*,*,*, 四段・ラ行, 用, 移る, うつる, 移り, うつり
うつりか,0,0,0, 名詞, 一般,*,*,*,*, 移り香, うつりが,*,*
うつりが,0,0,0, 名詞, 一般,*,*,*,*, 移り香, うつりが,*,*
うつり香,0,0,0, 名詞, 一般,*,*,*,*, 移り香, うつりが,*,*
```

表 2: **MeCab** による解析結果 (例): 国文学研究資料館のデータベースには「／」が含まれているが、取り除いた上で解析した。

	新後撰集（新撰国歌大観版）４番歌
000004	昨日までふる郷ちかくみよし野の山もはるかにかすむ春かな
000004	記号, 一般,*,*,*,*,
昨日	名詞, 一般,*,*,*,*, 昨日, きのふ,*,*
まで	助詞, 一般,*,*,*,*, まで, まで,*,*
ふる郷	名詞, 一般,*,*,*,*, 故郷, ふるさと,*,*
ちかく	形容詞, ク,*,*,*, 用, 近し, ちかし, 近く, ちかく
みよし野	名詞, 地名,*,*,*,*, み吉野, みよしの,*,*
の	助詞, 格助詞,*,*,*,*, の, の,*,*
山	名詞, 一般,*,*,*,*, 山, やま,*,*
も	助詞, 係助詞,*,*,*,*, も, も,*,*
はるかに	形容動詞, ナリ,*,*,*, 用, 遥かなり, はるかなり, 遥かに, はるかに
かすむ	動詞,*,*,*, 四段・マ行, 終体, 霞む, かすむ, 霞む, かすむ
春	名詞, 一般,*,*,*,*, 春, はる,*,*
かな	助詞, 終助詞, 詠嘆,*,*,*, 哉, かな,*,*
EOS	
	新後撰集（国文学研究資料館データベース版）４番歌
000004	昨日まて／ふる里ちかく／みよしのゝ／山もはるかに／かすむ春かな
000004	記号, 一般,*,*,*,*,
昨日	名詞, 一般,*,*,*,*, 昨日, きのふ,*,*
まて	助詞, 一般,*,*,*,*, まで, まで,*,*
ふる里	名詞, 一般,*,*,*,*, 旧里, ふるさと,*,*
ちかく	形容詞, ク,*,*,*, 用, 近し, ちかし, 近く, ちかく
みよしの	名詞, 地名,*,*,*,*, み吉野, みよしの,*,*
ゝ	助詞, 格助詞,*,*,*,*, と, と,*,*
山	名詞, 一般,*,*,*,*, 山, やま,*,*
も	助詞, 係助詞,*,*,*,*, も, も,*,*
はるかに	形容動詞, ナリ,*,*,*, 用, 遥かなり, はるかなり, 遥かに, はるかに
かすむ	動詞,*,*,*, 四段・マ行, 終体, 霞む, かすむ, 霞む, かすむ
春	名詞, 一般,*,*,*,*, 春, はる,*,*
かな	助詞, 終助詞, 詠嘆,*,*,*, 哉, かな,*,*
EOS	

[4] 竹田正幸, 福田智子, 南里一郎：歌集間における表現特徴の自動抽出—部分文字列の生起頻度にみる—, 情報処理学会研究報告 00-CH-47, Vol. 47, pp. 39–46 (2000).

[5] 守岡知彦：MeCab を用いた古典中国語の形態素解析の試み (セッション 1), 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, Vol. 2008, No. 73, pp. 17–22 (2008).

[6] 山元啓史：コンピュータによる歌枕の分析, イタリア日本語教育協会、第３回シンポジウム論文集, pp. 373–382, イタリア日本語・日本語教育学会 (2006).

[7] 山元啓史：和歌のための品詞タグづけシステム, 日本語の研究, Vol. 3, No. 3, pp. 33–39 (2007).

[8] 山元啓史：分類コードつき八代集用語のシソーラス, 日本語の研究, Vol. 5, No. 1 (2009).

[9] 新編国歌大観編集委員会（編）：CDROM 版新編国歌大観, 角川書店 (1996).