古典中国語形態素解析のための品詞体系再構築

山崎 直樹　　守岡 知彦　　安岡 孝一
関西大学外国語学部　　京都大学人文科学研究所

本研究は、古典中国語に対して形態素解析を行うことを目的とし、古典中国語の品詞体系の再構築を試みた過程と現時点での形態素解析の結果報告である。本研究で用いる形態素解析エンジンは、オープンソースの MeCab である。MeCab は最低限、辞書があれば動作するが、形態素情報をもたせた学習用コーパスがあればより精度を高められる。しかし、古典中国語の解析用に作られた辞書も学習用コーパスもこれまでには存在しないため、本研究では、専用辞書の設計を行い、同時にそれに合わせて古典中国語の品詞体系の再構築を行った。この品詞体系の特徴は、(1) 形態的な特徴を手がかりとせず、個々の語彙の意味範疇を細分化した素性を用いること、(2) 作業者に複雑な判断を強いないように、コンテクストに依存した分類をできるだけ排除したことである。本論文では、この品詞体系構築の過程の報告と、それを用いた現段階での試行的な解析結果とを報告する。

Refactoring of Wordclasses for Morphological Analysis of Classical Chinese

YAMAZAKI, Naoki　　MORIOKA, Tomohiko　　Koichi YASUOKA
Faculty of Foreign Language studies　　Institute for Research in Humanities
Kansai University　　Kyoto University

This paper explains an overview of a refactoring of prototype morphological analyzer for Classical Chinese based on MeCab, especially it focuses on the redesigning of wordclasses for the morphological analyzer. The redesigned wordclasses are based on an analysis of morphological corpora developed with the prototype morphological analyzer. We are refactoring our dictionary based on the redesigned wordclasses, and we are also developing new corpora. The characteristics of the redesigned wordclasses use features: (1) that come from subcategorized word meaning, and (2) that are uniquely-determinable and context-free. This paper reports the policy and process of refactoring, and reports some results of test run.

1　はじめに

古典中国語の文献は数・質ともに豊富であり、そこに埋蔵されている情報が科学の諸分野にとって極めて有益であることは、言を俟たない。しかし、古典中国語の文献は豊富であるが故に全貌を一覧することは不可能であるし、また、その読解には高度な知識と技術が要求される。機械処理による情報抽出が待たれる所以である。しかし、古典中国語の文献を機械で処理することには、下記に述べる障碍がある。

例えば、英語であれば、テキストはすでに単語単位に区切られているので、語彙頻度調査やキーワード抽出を機械的に行なうことは容易である。一方、古典中国語の文章は、（空白も含めた）区切り符号をいっさい使わず、漢字のみを連続させる。このようなテキストを「白文」と呼ぶが、この白文のままでは、語彙頻度調査やキーワード抽出等の基本的な解析すらままならない。

そこで我々はオープンソースの形態素解析エンジンである MeCab[1] を用いた古典中国語用形態素解析器の開発を試みている。MeCab は少なくとも辞書があれば形態素解析を行うことができるが、形態素に区切られた学習用コーパスを用意し、ここからパラメーター推定を行うことで解析精度を高めることができる。MeCab 用の学習用コーパスの形式は MeCab の（デフォルトでの）出力形式と同一であり、MeCab に白文を入力し結果を修正することで効率的に入力を行うことができる。逆にいえば、古典中国語用形態素解析器がない状態で全て手で入力作業を行うのは作業者の負担が大きすぎるといえる。そこで、日本語用の形態素辞書をベースに古典中国語用形態素解析器のプロトタイプ [3] とそれを用いた形態素コーパス編集システム [5] を開発し、実際にコーパスの入力作業を行った。

しかし、より精度の高い解析を行なうためには、古典中国語専用に設計された斉一な品詞体系に基づいた辞書が必要である。日本語と古典中国語では品詞体系も構文体系も異なるので、日本語用 IPA 辞書の流用では行き詰まるのはもちろん、日本語用 IPA 辞書の品詞分類（＝日本語用品詞分類）では、古典中国語の語彙を分類することができない。

形態素解析器用の辞書と学習用コーパスの構築のためには何らかの品詞体系が必要であり、プロトタイプではアドホックに決めた品詞体系 [4] を用いたが、この品詞体系は必ずしも古典中国語に適したものとはいえず、特に、日本語用辞書にあった語彙を流用する都合から古典中国語においては意味のない区別を残したものとなっていたといえる。そこで、本研究チームは、古典中国語専用辞書の設計とその品詞体系の再構築を行った。

本発表では、ここで採用した系統的な品詞体系の構築の試みとプロトタイプからの移行手順に焦点を当てて報告したい。

2　解析に必要な道具

2.1　形態素解析エンジン

上述したとおり、形態素解析器としては MeCab を使用する。MeCab は、機械可読な辞書と形態素情報をメタ情報として付与したコーパスを必要とする。しかし、辞書とコーパスのフォーマットに依存しない汎言語的な解析器[1]であるところに特徴がある。

2.2　辞書をどうするか

MeCab の動作には上述の辞書とコーパスとを必要とするが、最低限、辞書だけあれば動作はする。しかし、現状では古典中国語解析用の辞書は存在しないので、日本語用 IPA 辞書を流用し、実験的な解析を行ったところ、一定の成果が見られた。[3] これは、日本語が古典中国語に由来する語彙を多く持ち、なおかつそれを漢字で表記していることに拠る結果である（それがこのような成果を挙げたこと自体は注目に値するが）。

しかし、より高精度の解析のためには、専用に設計された破綻のない品詞体系に基づいた辞書が

[1] MeCab と同様評価の高い形態素解析器である『Chasen（茶筅）』とは、この点で異なる。

表 1: 『大品詞』の下の『通常の品詞』

大品詞	通常の品詞
n	名詞、代名詞、数詞
v	動詞、前置詞、助動詞、副詞
p	助詞、感嘆詞

表 2: 下位範疇化の例

n - 名詞 - 人 -《役割》
（例: 公、侯、司職、司空）
v - 動詞 - 行為 -《役割》
（例: 当、立、為、任、覇、封）

必要である。日本語と古典中国語では言語体系が異なるので、日本語用 IPA 辞書の品詞分類では、古典中国語の語彙を分類することができない。そこで、本研究チームは、古典中国語専用辞書の設計とその品詞体系の再構築をおこなった。

3 品詞体系の設計

日本語用の解析器では形態上の特徴（特定の品詞は特定の語形を持つことが多い）を学習用の手がかりとして利用できるが、古典中国語は形態上の変化がほぼ無いので、これは利用できない。そこで、利用できる情報としては、個々の語彙のもつ意味的なカテゴリーを下位範疇化し、それを素性の束として記述した情報が最も妥当であるということになる。

この考えに基づき、この品詞体系は、まず、『大品詞』(n, v, p) があり、その下に名詞や動詞といった『通常の品詞』があり（表 1)、その下は 2 階層の『意味素性』で分類されている（表 2)。[2] この『意味素性』は、閉じた体系で構成されている。また、異なる『通常の品詞』の下に同一の『意味素性』を敢えて設定し、facet な分類に利用することも可能な体系を目指した。品詞体系の概要を表 3 に示す。

[2]『大品詞』,『通常の品詞』,『意味素性 1』,『意味素性 2』という 4 階層自体はプロトタイプのもの [5] と同じであるが、『通常の品詞』および『意味素性』の構成はプロトタイプのものと異なっている。

4 今回提案する品詞体系の新しい点

4.1 用例ベースの品詞体系

品詞体系構築の手順として、まず、プロトタイプ的形態素解析器 [3][4] を使って入力作業を行い、試験的なコーパスを作成した。[5] そしてそこから抽出した用例辞書を分析して、より整合性のある品詞体系の再設計にとりかかった。このように用例をベースに構築された体系であるという点が新しい点である。

4.2 コンテクストフリーな分類への指向

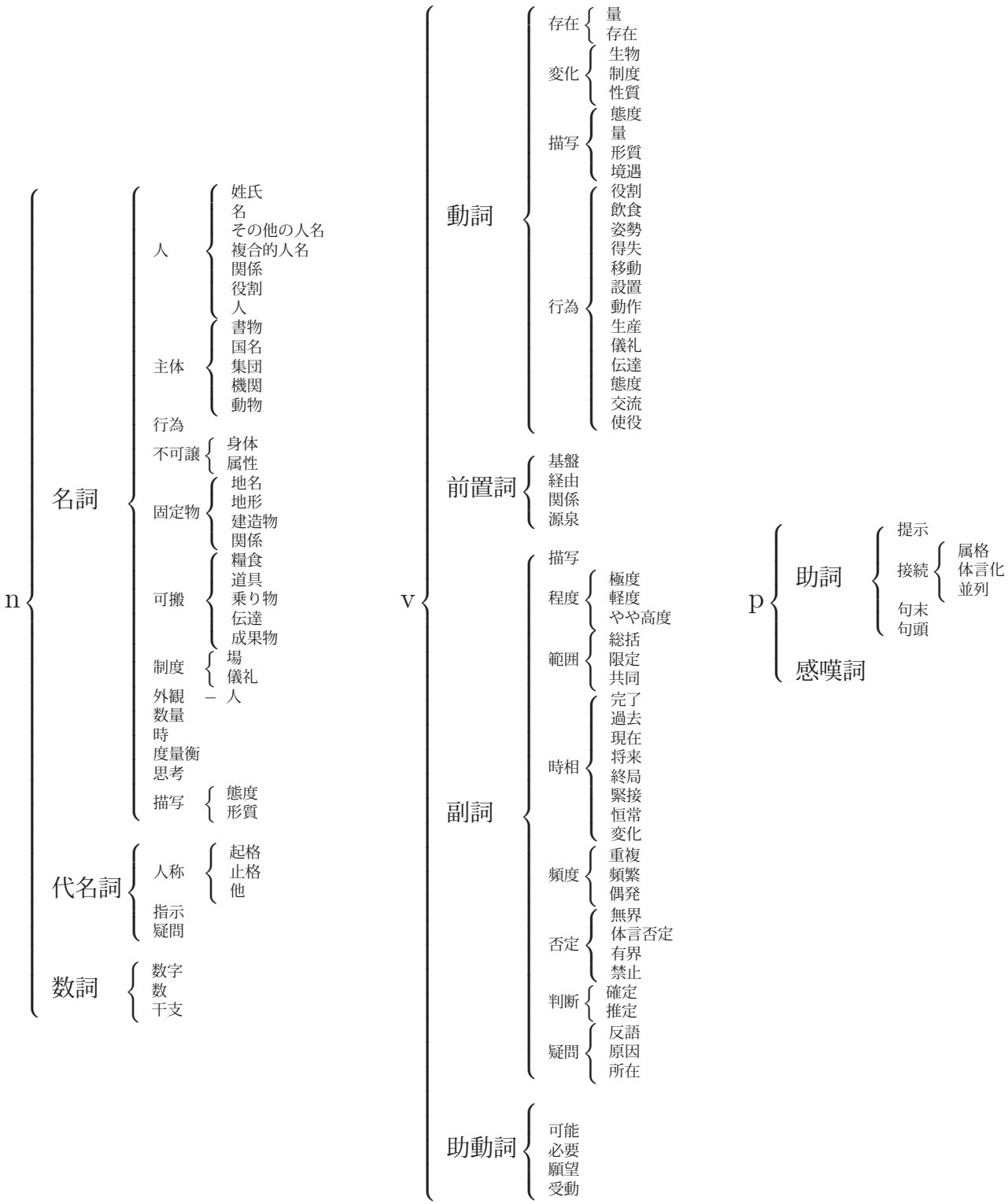
ふつうの品詞体系では、「多機能語」と呼ばれる一群の語がある。同じ形式の語であっても、あるコンテクストでの用法は動詞に分類されたり、別のあるコンテクストでの用法は名詞に分類されたりする語がある。これが「多機能語」である。一般に行なわれている古典中国語の品詞分類では、この「多機能語」が非常に多い、というか、たいていの語彙は多機能語である。[3]

しかし、コーパス入力時の人力によるメタデータ付与の際に、上述の「多機能語」の帰属の判断を、個々の入力作業者に任せることは危険を伴う。それは、作業者に「白文」の意味の正確な把握を要求するのは困難であるという経験的な事実、また、仮に判断できたとしても、その作業結果はばらつきを生じやすいという経験的事実に拠る。

このため、今回の品詞体系では、いかなる作業者であっても、明示的な規則に基づいて、一意的

[3]日本で出版された漢和辞典のうち、品詞名を表示した辞典をご覧いただきたい。

表 3: 品詞体系の概要



に品詞の分類ができる体系を目指した。

4.3 新たな品詞体系の目指すところ

今回、再構築した品詞体系は、将来の古典中国語の統語解析を考慮に入れ、「隣接する要素との共起関係において、似た振る舞いをすると思われる語彙」を1つにまとめる方向で構築した。これには、現代言語学の知見も盛り込まれている。例えば、「意志を持ち、行為の主体になりうる」と「意志を持たない／主体にならない」を区別する、「可搬物」と「固定物」を区別する、「譲渡不可能物」を定義する、などである。

「棒で人を殴り殺した」場合、その「棒」は「道具」なのかあるいは「武器」なのかは、明確な基準がなければ判断が困難である。それゆえ、本研究の品詞体系は派生的・比喩的な分類は廃した。つまり、「棒」は常に「道具」である。

もう1例を挙げると、「呉」という固有名詞は国の名前である。国の名前は地理的な領域を示すのにも使われるし、統治機関も表せば、「呉攻越（呉が越を攻めた）」のように、擬似的な「人」のように、行為をする主体を示すのにも使われる（この問題は上述した）。この研究の体系では、国名は、一貫して「主体」という分類のもとに置かれ、主体とならない地理的領域名や集団名とは区別した。

動詞が名詞的に使われている場合（「笑うことは健康によい」の「笑うこと」がこれにあたる。古典中国語では「〜こと」に当たる形態的特徴はない）、それは動詞なのか名詞なのかを悩むことは、やはり非効率的である。このように「動詞が体言的に使われる」現象、また逆に、「名詞が用言的に使われる現象」などは、古典中国語では極めてふつうの、しかも頻出する現象である。一例を挙げれば、「君君」という2文字からなる古典中国語の文は、「君主が君主らしくある（主語＋動詞）」あるいは「君主を君主として扱う（動詞＋目的語）」のように解釈される。本研究の品詞体系では、これらの派生的用法は取り上げず、すべて本来の用法で分類した。

5 分類の参考

今回、品詞体系を再構築するあたり、『全訳漢辞海』[6]の品詞表示および巻末の古典中国語文法の解説を、参考にした。特に、属する語彙が限られた類である助動詞と副詞は、その下位分類に、すでに詳細な「意義による分類」があり（表4）、本研究の方向性と一致しているので、ほぼそのまま採用した。

表 4: 『漢辞海』の分類例

- 副詞
  - 可能
  - 必要
  - 願望
  - 程度
  - 範囲
  - 時相
    - 完了
    - 過去
    - 現在
    - 将来
    - 終局
    - 緊接
    - 恒常
    - 変化
  - 頻度
  - 否定
  - 判断

逆に、『全訳漢辞海』の分類にある（そして、多くの古典中国語の辞書が採用している）「形容詞」という分類は廃止した。古典中国語において「形容詞」という範疇は動詞の下位分類に過ぎないことはすでに広く知られているし、作業者が「形容詞か動詞か」の判断において、日本語の訓読みに惑わされやすいことが経験にわかっているからである。なお、「形容詞」だったものの多くは、結果的に「動詞 — 描写」に納まった。

今回、語彙の分類体系を再構築するにあたり、現代言語学の分野において、さまざまな言語を対象にさまざまなアプローチで構築された意義素性の集合を、間接的に参考にしている。それらに対

して逐一言及してその影響を記述することはできないが、本研究も、同種の他の研究と同じく、先人の研究成果の恩恵を蒙っている部分も多いことをお断りしておく。

6 『分類語彙表』との関係について

ある言語の語彙を意義素性で包括的に分類したものとして、『分類語彙表』[2] のような語彙体系がある。この語彙体系は、純粋に意味で語彙を分類しているので、理論的には他の言語にも適用可能である。しかし、今回、古典中国語の品詞体系を再構築するにあたり、この分類方法を参考にはしたがそのまま採用することはしなかった。その理由は 3 点ある。

1. 抽象的な階層が多い
   例えば、『分類語彙表』では、「墓（はか）」は、「体, 生産物, 土地利用, 地類（土地利用）」であるが、我々の分類では、「名詞, 固定物, 建造物」である。後者のほうが具体的でわかりやすい。この点は、学習用コーパスを人手で入力する際、作業効率と作業精度において重要になる。

2. 1 つの語彙がコンテクストに依存して複数の類に分類されている
   例えば、「国（くに）」は、『分類語彙表』では、次のように分類されている。

   体, 主体, 公私, 郷里
   体, 主体, 公私, 国
   体, 主体, 公私, 政治区画

   それぞれの分類は、「国」の多義性を表している。しかし、学習用コーパスを人手で入力する場合にこのような分類を行うことを要求すれば、作業者が、いちいち原典のテクストを読解し判断せねばならない。これは作業の速度はもちろん、作業者間での結果の統一性に悪影響を与える。

3. 一意的な分類
   さらに言えば、統治機関としての「国」は意志を持った行為者として、文法的には人間のように振る舞うこともある（例：呉伐越）。このため、本研究の語彙体系では、「魯、宋…」などの古代の国名は、

   名詞, 主体, 国名

   のように分類して、「意志を持った行為者」の用法を持つ他の語とともに、「主体」の下に納め、そのような用法を持たない、

   名詞, 固定物, 地名（※村落名）
   名詞, 固定物, 地形（※山、川など）

   とは、区別をしている。また、本研究では、「国」はすべて一意的に分類されるので、作業者は迷うことがない。本研究においても、必ずしもすべての語彙が一意的に分類されているわけではないが、できる限りこの方針を貫いている。

7 品詞体系の移行

再構築した品詞体系へ移行するためには、形態素辞書と学習用コーパスの品詞・意味素性を適切に（食い違いがないように）書き換える必要があるが、プロトタイプにおける品詞体系と今回再構築した品詞体系は 1 対 1 対応していないために機械的に変換することができない。前述のように形態素コーパスを効率的に入力・編集するためには形態素解析器が必要であるが、形態素辞書や学習用コーパスの移行作業を行わなければ形態素解析器が利用できず、結果的に学習用コーパスの移行作業が進まないというデッドロックに陥ってしまいがちである。ただ、MeCab の場合、形態素辞書が存在すれば学習用コーパスがなくても形態素解析が行えるので、まず形態素辞書の移行作業を行えば良い。そして、ある程度辞書項目が揃ってきた段階で、形態素解析器の辞書のプロトタイプのものから新しいものへと置き換えることで再構築した品詞体系に基づく形態素解析器を構成することができる。

そこで、我々は実際にこのような手順に基づき、プロトタイプを用いて入力したコーパスから形態素を抽出し、それに再構築した品詞体系に基づく品詞・意味素性を付与する作業を行っている。また、こうしてできた形態素辞書に置き換えた形態素解析器を実際に構成した。

ただし現実には、再構築された品詞体系に「ある日突然」移行できるわけではなかった。辞書やコーパスの移行そのものは、かなりの部分が機械的に処理できるのだが、それをおこなう各作業者の「頭」は、必ずしも新たな品詞体系に即座に移行できるわけではない。この結果、作業者に対しては、いわば「クールダウン」の期間を必要とするために、コーパスの移行作業がはかどらず、現状では少数の学習用コーパスを準備するにとどまっている。

8 実験結果

新しい品詞体系を評価するために、従来のアドホックな品詞体系に基づく辞書および学習用コーパスを用いた場合 [4] の認識精度と、新しい品詞体系に基づく辞書および学習用コーパスを用いた場合の認識精度の比較実験を行った（表 5〜7）。この実験には、M（69 語）、K（68 語）、R（320 語）というコーパスを用いた。なお、M は雑多な文例、K は典型的な構文例、R は三国志呉書列伝よりの抜粋である。

新しい品詞体系での学習用コーパスが、現時点では、かなり少数であるにもかかわらず、従来のアドホックな体系に比べて、遜色のない結果である。

しかしながら、R を学習データとして、K を入力データとした場合の比較結果は、かなり悪くなっていると言わざるを得ない。R と K は、語彙の上ではかなり乖離している (R は口語的な表現が多く、一方 K は規範的な古典中国語) ので、それがそのまま反映された形になってしまった。原因としては、共起関係の分離が効きにくくなっている可能性があり、今後コーパスを増やしていく際に、さらなる検討が必要になると考えられる。

表 5: 大品詞と品詞の F 値

	入力データ		
学習データ	M	K	R
M（旧）	100	90/82	90/88
M（新）	100	97/90	97/87
K（旧）	91/90	100	92/89
K（新）	89/85	100	95/88
R（旧）	100/99	90/85	100
R（新）	93/86	85/73	100

表 6: 意味素性の F 値

	入力データ		
学習データ	M	K	R
M（旧）	100	79/78	88/84
M（新）	100	88/80	85/82
K（旧）	89/87	100	89/84
K（新）	82/73	100	83/79
R（旧）	99/96	85/79	100
R（新）	83/80	72/64	100

表 7: 全体の F 値

	入力データ		
学習データ	M	K	R
M（旧）	100	78	83
M（新）	100	75	79
K（旧）	84	100	84
K（新）	70	100	78
R（旧）	93	76	100
R（新）	76	63	100

9 おわりに

実際に入力した形態素コーパスから抽出した用例辞書を分析することで、品詞体系の再構築を行うとともに、用例に基づく形態素辞書の作成を行い、日本語用形態素辞書から抽出して作ったプロトタイプ辞書からの置き換えを試みた。

再構築した品詞体系へ移行するためには、形態素辞書と学習用コーパスの品詞・意味素性を適切に（食い違いがないように）書き換える必要があるが、プロトタイプにおける品詞体系と今回再構築した品詞体系は 1 対 1 対応していないために機械的に変換することができない。形態素解析器なしに手作業で形態素コーパスを入力･編集するのは作業者の負担が大きいため、まず、新しい品詞体系に基づく形態素辞書への移行作業を行っている。また、こうして作られた形態素辞書に置き換えた形態素解析器を実際に構成した。現状ではコーパスの移行作業が不十分であるため、極めて少数の学習用コーパスしかなく、認識精度は必ずしも良くないが、実際の用例を反映した形態素解析器を構成することができた。

今後は形態素辞書の量・質の向上を計るとともに、新しい品詞体系に基づく形態素コーパスの入力作業を進めることで、認識精度の向上を計りたいと考えている。

参考文献

[1] Taku Kudo, et al. MeCab (和布蕪): Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforge.net/.

[2] 国立国語研究所. 分類語彙表. 国立国語研究所資料集, No. 14. 大日本図書, 増補改訂版, 2004 年 1 月.

[3] 守岡知彦. MeCab を用いた古典中国語の形態素解析の試み. 情処研報, Vol. 2008, No. 73, pp. 17–22, 2008 年 7 月. 2008-CH-79.

[4] 守岡知彦. MeCab を用いた古典中国語形態素解析器の改良. 情処研報, Vol. 2009-CH-84, No. 3, pp. 1–5, 2009 年 10 月.

[5] 守岡知彦. 古典中国語形態素コーパス編集システムの開発. 東洋学へのコンピューター利用 第 23 回研究セミナー, pp. 75–83, 2012 年 3 月.

[6] 戸川芳郎（監修）, 佐藤進, 濱口富士雄（編）. 全訳 漢辞海 第二版. 三省堂, 2006 年 1 月.