情報処理学会論文誌 Vol.59 No.2 323–331 (Feb. 2018)

古典中国語（漢文）の形態素解析とその応用

安岡 孝一[1,a)] ウィッテルン クリスティアン[1,b)] 守岡 知彦[1,c)] 池田 巧[1,d)] 山崎 直樹[2,e)]
二階堂 善弘[2,f)] 鈴木 慎吾[3,g)] 師 茂樹[4,h)]

受付日 2017年5月9日, 採録日 2017年11月7日

概要：古典中国語（漢文）の解析手法として，MeCab を用いた形態素解析手法を提案する．本手法では，漢文の動賓構造を表現すべく，4 階層の「品詞」からなる新たな品詞体系を構築し，それに基づく MeCab 漢文コーパスを設計した．合わせて，MeCab 漢文コーパスを入力するための専用ツールとして，XEmacs CHISE をベースとしたコーパス入力ツールを開発した．また，MeCab 漢文コーパスを効果的に管理し，さらには品詞体系のリファクタリングを行うべく，MeCab 漢文コーパスの Linked Data 化を行い，WWW 上で公開した．さらに，MeCab を用いた漢文形態素解析の応用として，漢文における固有表現の自動抽出に挑戦した．結果として，地名の自動抽出は高精度に行うことができたが，官職・人名の自動抽出はそれぞれに課題が残った．

キーワード：漢文コーパス，リンクトデータ，固有表現抽出

Morphological Analysis of Classical Chinese Texts and Its Application

Koichi Yasuoka[1,a)] Christian Wittern[1,b)] Tomohiko Morioka[1,c)] Takumi Ikeda[1,d)]
Naoki Yamazaki[2,e)] Yoshihiro Nikaido[2,f)] Shingo Suzuki[3,g)] Shigeki Moro[4,h)]

Received: May 9, 2017, Accepted: November 7, 2017

Abstract: A method to analyze classical Chinese texts is proposed. In the method, we use our original morphological analyzer based on MeCab. We propose a new four-level word-class system to represent the predicate-object structure of classical Chinese. In order to make a corpus for classical Chinese on MeCab, we have constructed a MeCab-corpus editor based on XEmacs CHISE. In order to control the corpus effectively, and to refactor our four-level word-class system, we have converted it into Linked Data on WWW. As an applied study for our morpholgical analysis of classical Chinese texts, we have tried to extract named entities: names of places, job titles, and names of people. As a result we are able to extract names of places from classical Chinese texts almost perfectly. But we have found some difficulties to extract job titles or names of people.

Keywords: classical Chinese corpus, linked data, named entity extraction

[1] 京都大学
Kyoto University, Kyoto 606–8501, Japan
[2] 関西大学
Kansai University, Suita, Osaka 564–8680, Japan
[3] 大阪大学
Osaka University, Minoh, Osaka 562–8558, Japan
[4] 花園大学
Hanazono University, Kyoto 604–8456, Japan
[a)] yasuoka@kanji.zinbun.kyoto-u.ac.jp
[b)] wittern@zinbun.kyoto-u.ac.jp
[c)] tomo@kanji.zinbun.kyoto-u.ac.jp
[d)] ikeda@zinbun.kyoto-u.ac.jp
[e)] ymzknk@kansai-u.ac.jp
[f)] nikaido@kansai-u.ac.jp
[g)] suzukish@lang.osaka-u.ac.jp
[h)] s-moro@hanazono.ac.jp

1. はじめに

古典中国語（漢文）テキストをコンピュータ処理するためには，白文（単なる漢字の列）のままではどうにもならず，テキストを自然言語解析する必要がある．たとえば現代の欧米諸語であれば，テキストは単語単位に区切られており，文末を表す記号も付加されていることが多いので，語彙抽出などを行うのは比較的容易である．一方，古典漢文においては，単語の間にも文の間にも区切りを持たない白文のままでは，せいぜい文字列検索しか行えない．古典漢文のような書写言語の解析においては，まず，単語を認

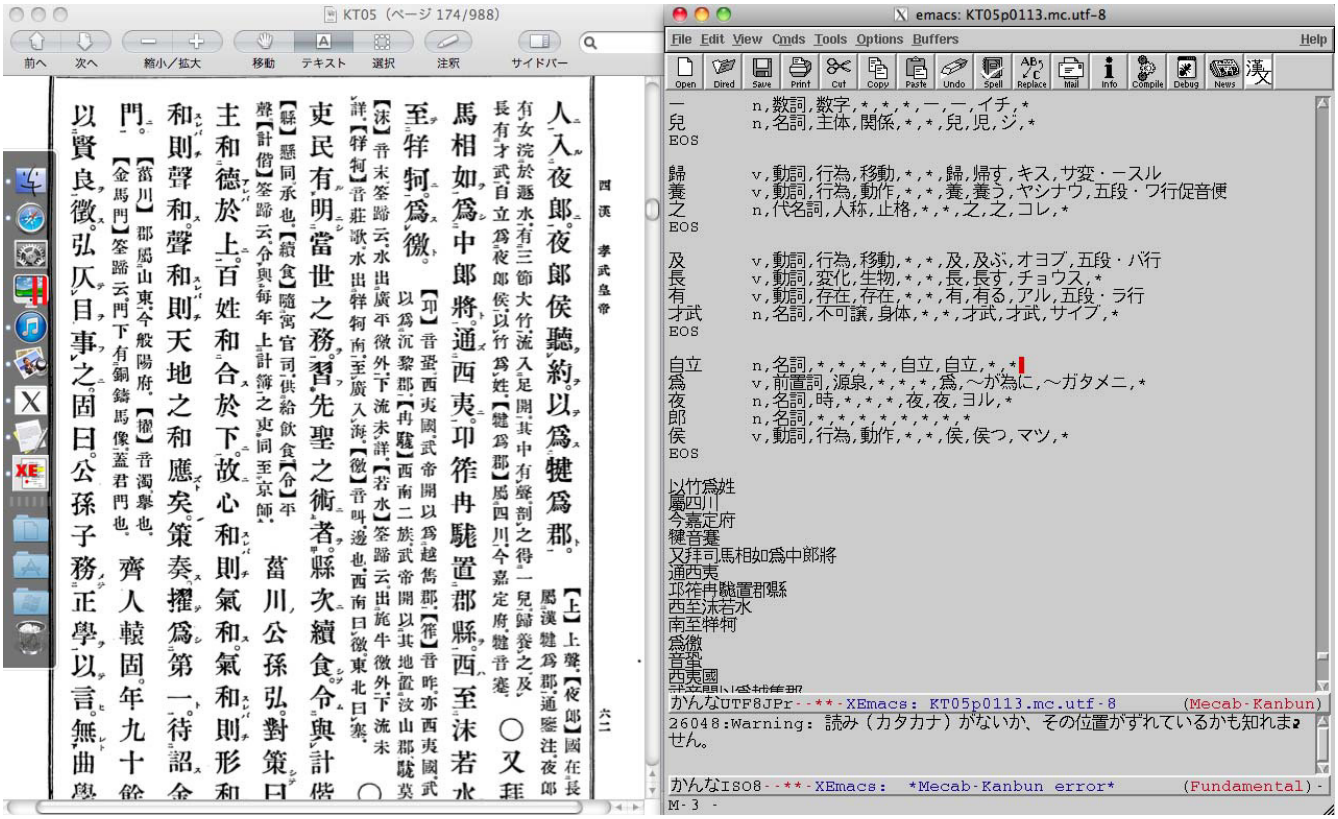

図 1 漢文コーパス入力ツール
Fig. 1 MeCab-corpus editor for classical Chinese.

識することが必須であり，そのためには形態素解析を行わねばならない．

この問題に対し，我々は，2008 年 4 月より京都大学人文科学研究所共同研究班「東アジア古典文献コーパスの研究」を組織し，さらに 2013 年 4 月より京都大学人文科学研究所共同研究班「東アジア古典文献コーパスの応用研究」を組織して，古典漢文に対する形態素解析の研究を行ってきた．現代中国語の形態素解析については，それまでにも数多くの研究がなされていた [1] し，その後も研究は進み続けている [2], [3]．しかしながら，古典漢文の形態素解析には，現代中国語の形態素解析をそのまま応用することができない．古典漢文と現代中国語は，文法的にも語彙的にも，違いがあまりに大きいからである [4]．結果として，古典漢文の形態素解析は，研究分野としてほぼ手つかずの状態だった．

本稿では，古典漢文の形態素解析における，我々の研究成果を俯瞰する．さらにその応用例として，漢文における固有表現抽出について述べる．なお，本稿は [5], [6] を合わせて，加筆，改稿したものである．

2. 漢文の形態素解析

漢文の形態素解析において，我々は，MeCab というソフトウェアを用いることにした [7]．MeCab はオープンソースの形態素解析エンジンで，言語，辞書，コーパスに依存しない汎用的な設計がなされており，辞書とコーパスを準備すればいかなる言語にも対応できる．ならば，漢文（の散文）にも MeCab を使用できるはずだ，というのが，我々の直感だったが，我々以前には誰もそれを試したことがなかった．

MeCab の辞書には「品詞」（複数の階層が可能）が必要なことから，我々は，日本語と漢文をつなぐ「構造」の一種である訓読に着目し，返り点を「品詞」に反映させることを考えた．すなわち，訓読における返り点を，漢文の動賓構造を表しているものとみなし，動詞類に「v」という「品詞」を，賓語に「n」という「品詞」を，そのほかの語に「p」という「品詞」を，それぞれ，MeCab 漢文辞書の「第 1 階層の品詞」（以下「大品詞」と呼ぶ）として定めることにしたのである．次に「第 2 階層の品詞」（以下「品詞」と呼ぶ）だが，これは IPA の日本語辞書から抽出した品詞を，そのまま漢文の品詞として，実験的に使用してみることにした [8], [9]．

この MeCab 漢文辞書（IPA 由来版）と，それに基づいて作った小規模な MeCab 漢文コーパスを用いて，高校教科書の漢文例や，三国志呉書列伝などの白文を，MeCab で形態素解析してみた．そうしたところ，白文を単語に区切るという点に関しては，かなり良好な結果が得られた．そこで我々は，例文入力グループ・デジタル処理グループ・コーパス校訂グループ・品詞分類グループの 4 グループか

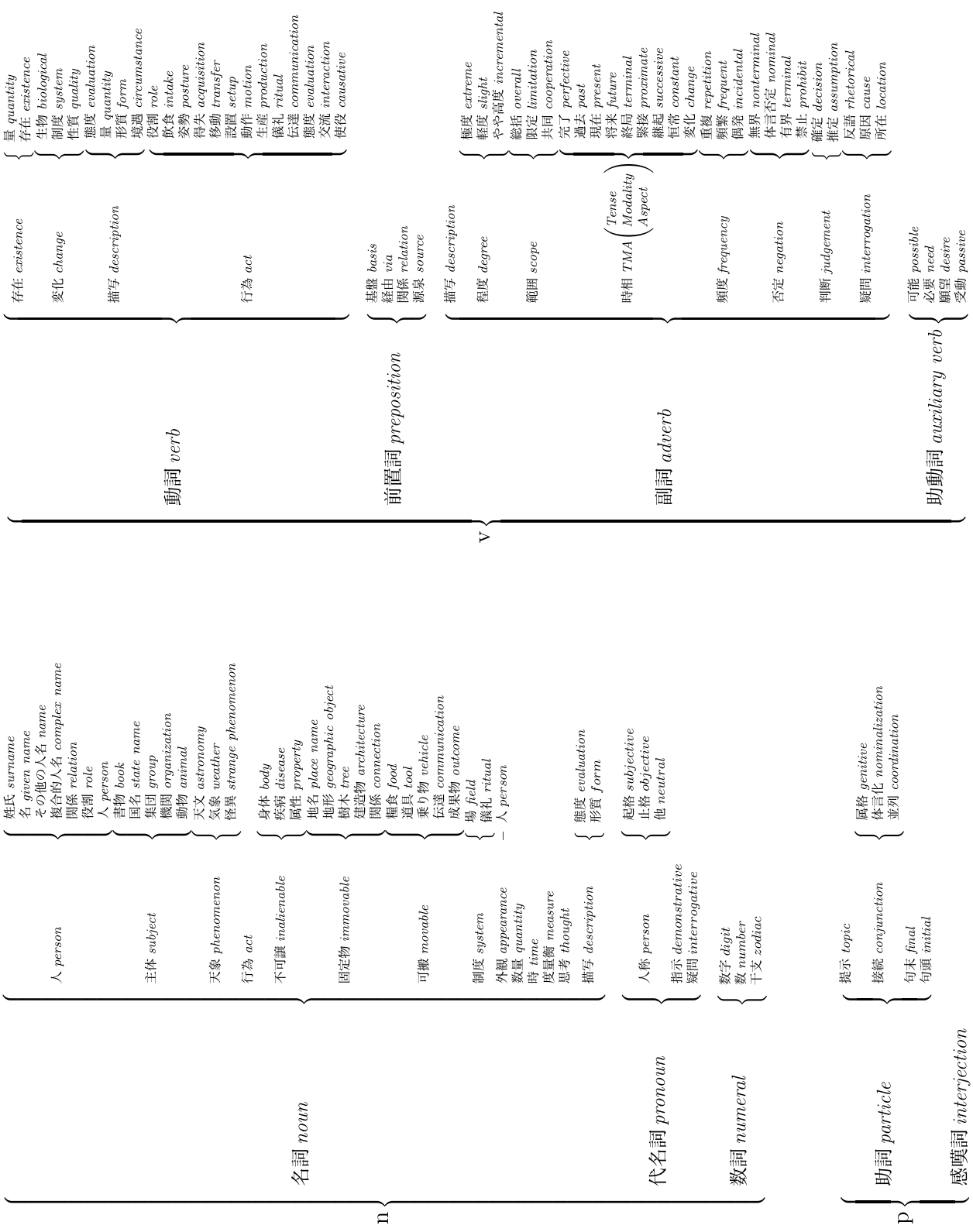

図 2 形態素解析に特化した古典中国語品詞体系
Fig. 2 A new four-level word-class system for classical Chinese.

らなる組織を構成し，MeCab 漢文コーパスの構築を行うこととした．具体的には，安岡を研究代表者とし，共同研究班の班員全員を研究分担者として，2010 年 4 月から 3 年間，科学研究費補助金基盤研究（B）22300087『形態素解析のための品詞情報つき古典漢文コーパスの構築』の研究助成を受けた．

例文入力グループが MeCab 漢文コーパスを直接入力するのは，かなりの困難が予想されたことから，デジタル処理グループは，専用ツールとして，XEmacs CHISE をベースにしたコーパス入力ツールを開発した [10]．このツール

表 1 各学習用コーパスに対する各テストデータの F 値（大品詞/品詞/意味素性/小素性）
Table 1 F-measures on MeCab-corpola for classical Chinese.

	テストデータ M	テストデータ K	テストデータ R
学習用コーパス M	100	97/90/88/80	97/87/85/82
学習用コーパス K	89/85/82/83	100	95/88/83/79
学習用コーパス R	93/86/83/80	85/73/72/64	100

は，白文を入力すると，MeCab を用いた処理をその場で行って，その時点での形態素解析の結果を出力する．結果に問題がなければ，そのまま漢文コーパスに反映し，もし，結果に問題があれば，入力者が手作業で訂正を行って，やはり漢文コーパスに反映する．たとえば，**図 1** の「自立爲夜郎侯」であれば，これを正しく

```
自      v,副詞,範囲,限定,*,*,自,自ら,ミズカラ,*
立      v,動詞,行為,役割,*,*,立,立つ,タツ,五段
爲      v,動詞,行為,役割,*,*,爲,為る,ナル,五段
夜郎    n,名詞,主体,国名,*,*,夜郎,夜郎,ヤロウ,*
侯      n,名詞,人,役割,*,*,侯,侯,コウ,*
```

に訂正してから，漢文コーパスに反映する．このようなやり方で，MeCab 漢文コーパスを効率的に構築できる環境を整えた．

品詞分類グループは，コーパス校訂グループと共同で，MeCab による漢文形態素解析のために，新たな 4 階層の品詞体系を構築した（**図 2**）．（第 1 階層の）大品詞と（第 2 階層の）品詞という構造に加え，「第 3 階層の品詞」（以下「意味素性」と呼ぶ）を構成し，意味素性まででは分類が不十分なものに対しては，「第 4 階層の品詞」（以下「小素性」と呼ぶ）を用いることができるよう，構築している．この新しい品詞体系では，大品詞を「n」「v」「p」の 3 種類とし，品詞を「名詞」「代名詞」「数詞」「動詞」「前置詞」「副詞」「助動詞」「助詞」「感嘆詞」の 9 種類として，従来の漢文文法などで見られた「形容詞」を，[11] に基づき「動詞」と統合しているのが特徴である．これらに加え，44 種類の意味素性と，88 種類の小素性を定義し，形態素解析の結果として得られる各単語を，意味の面からもとらえやすいよう工夫した．また，この新しい品詞体系による MeCab 漢文辞書を作成すると同時に，MeCab 漢文コーパスにもフィードバックし，全体として新しい品詞体系で，MeCab による漢文の自動形態素解析が行えるようにした．

この形態素解析システムを用いて，複数の小さな学習用コーパスに対し，認識精度の比較実験を行った [12]．準備したコーパスは，種々の雑多な漢文文例 M（69 語），高校教科書の漢文用例 K（68 語），三国志呉書列伝の抜粋 R（320 語）であり，これらを互いにテストデータとしても用いた．なお，MeCab 漢文辞書は約 5,500 語で固定とし，実験に用いた MeCab のバージョンは 0.994 である．結果の F 値（大品詞/品詞/意味素性/小素性）を**表 1** に示す．全体としては，R を学習用コーパスとした時の認識精度が，あまり良くなかった．R は口語的な表現が多く，それを学習コーパスとすると，規範的な表現が多い K の認識がうまくいかなくなる，ということである．この結果を見る限り，学習用コーパスは規範的な表現の方が良い，ということだった．

この結果をもとに，我々は，規範的な表現を中心として漢文コーパスを作成することにした．具体的には，漢文コーパスの例文に用いるテキストとして，規範的な表現が多いと考えられる『十八史略』と『孟子』を中心にし，約 46,000 文（複数の入力者による重複を許す）の漢文コーパスを作成した．なお，1 文あたりの平均語数は 3.9 語となった．

3. 漢文コーパスの Linked Data 化

我々の古典中国語品詞体系と MeCab 漢文コーパスを効果的に管理し，さらには品詞体系のリファクタリングを行うべく，我々は，MeCab 漢文コーパスの Linked Data 化を行った [13]．

具体的には，品詞体系の大品詞・品詞・意味素性・小素性のすべてを品詞オブジェクトとし，MeCab 漢文コーパスに対しては，見出しオブジェクト（語），形態素オブジェクト，文オブジェクトの 3 つを準備した．見出しオブジェクトと形態素オブジェクトの間は，対応する品詞オブジェクト（小素性）によってリンクする．形態素オブジェクトは，それを含む文オブジェクトに「用例」としてリンクする．さらに，見出しオブジェクトが 1 文字から構成される場合は，文字オブジェクト（CHISE 文字オントロジー）とリンクする．例として，「自立爲夜郎侯」に関するオブジェクトとリンクを，**図 3** に視覚的に示す．ただし，図 3 は，オブジェクトとリンクを概念的に示したものであり，あくまで全体のごく一部であることに注意されたい．

さらに，これらの Linked Data を WWW 上に実装し，各オブジェクトとリンクを一望できるシステムを実現した [14]．このシステムによって，ある品詞オブジェクトに関する形態素オブジェクトがすべて一望できるようになり，品詞体系の効率的なリファクタリングが可能となった．また，ある見出しオブジェクトに関する品詞オブジェクトも一望できるようになった．たとえば「左右」という見出し語に対しては，「n，名詞，固定物，関係」すなわち位置関係を表す場合と，「n，名詞，人，役割」すなわち官職を表す場合があり，それぞれの用例を簡単にたどれるようになった．このシステムを実現したことで，我々は，漢文の固有表現抽出への手がかりを掴むことができたのである．

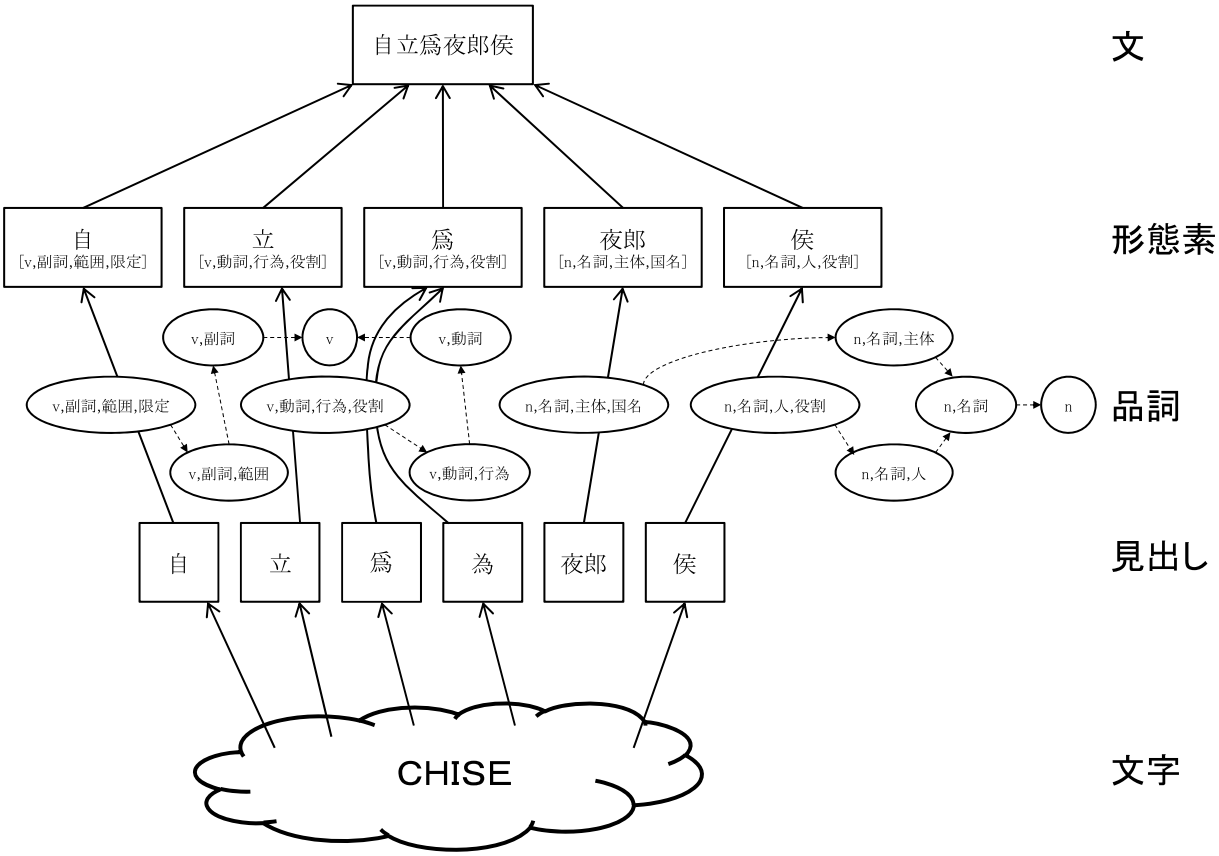

図 3 「自立爲夜郎侯」の Linked Data 概念図
Fig. 3 Linked Data around “自立爲夜郎侯”.

4. 漢文の固有表現抽出

MeCab 漢文コーパスを用いたさらなる応用として，我々は，漢文における固有表現の自動抽出に挑戦した．具体的には，安岡を研究代表者とし，共同研究班の班員全員を研究分担者として，2013 年 4 月から 3 年間，科学研究費補助金基盤研究（B）25280122『品詞素性情報つき古典漢文コーパスの発展的応用』の研究助成を受け，漢文における地名・官職・人名の自動抽出に挑戦した．

4.1 地名の自動抽出

漢文での地名を自動抽出する，という目標に向け，我々は，それまでに作成してきた MeCab 漢文コーパスを洗い直してみた．特に，我々の新しい品詞体系において「n，名詞，固定物，地名」あるいは「n，名詞，主体，国名」に分類されている形態素オブジェクトと，その形態素オブジェクトを含む文例を見直してみた．この結果，我々がたどり着いたのが，「2 文字の地名には地名以外の用例はない」という仮説だった．

この仮説に基づき，我々は「2 文字の地名」の地名以外の用例を，MeCab 漢文コーパスに対して，サンプリング調査してみた．そうしたところ，そのような地名以外の用例は，どの「2 文字の地名」においても 10%未満だった．しかも，それら 10%足らずの用例も「n，名詞，固定物，地形」など，山や川の名前を例文入力グループが地形だとみなしたものが大多数で，これらを仮に地名だとみなしても大した問題は起こらない．「2 文字の地名には地名以外の用例はない」という仮説は，少なくとも 90%の確率で当たっており，地名の自動抽出という観点からは，採用するに値する．

この結論に基づき，我々は，MeCab 漢文コーパスから抽出した「2 文字の地名」を，そのまま MeCab 漢文辞書に追加した．また，3 文字以上の地名は，その多くが「○○府」や「○○縣」の形をとるものだったが，同様に MeCab 漢文辞書に追加した．

では，「1 文字の地名」は，どうなのか．たとえば「渭」のように，地名用例しかないような「1 文字の地名」に関しては，そのまま MeCab 漢文辞書に追加すればよい．しかし，たとえば「夏」という形態素は，王朝名としての「夏」かもしれないし，季節としての「夏」かもしれない．あるいは「莫」という形態素は，地名用例はむしろ少数で，大多数の用例が「v，副詞，否定，禁止」である．もし，「莫」を無理矢理に地名だとみなすような処理を行うと，「v，副詞，否定，禁止」であるべき「莫」を，誤って「n，名詞，固定物，地名」だと処理してしまう危険性がある．その場合，後続の動詞にも悪影響が及ぶので，文法上のミスとしては致命的である．そのようなミスは，絶対に避けなければな

表 2 各辞書に対する各テストデータの F 値（大品詞/品詞/意味素性/小素性）
Table 2 F-measures on MeCab-dictionaries for classical Chinese.

	テストデータ P	テストデータ M	テストデータ R
辞書 α	96/86/85/76	93/90/90/77	96/83/81/71
辞書 β	96/89/88/84	93/90/90/76	96/83/81/71
辞書 γ	96/86/84/73	93/90/90/77	94/81/79/69

らない．

この問題に対し，我々は，たとえ「1 文字の地名」をすべて MeCab 漢文辞書に追加したとしても，MeCab 漢文コーパスを十分に準備すれば，そのようなミスは形態素解析において発生しないだろう，と予想した．「2 文字の地名」という巨大な用例による接続確率（裏を返せば非接続確率）が効いてくるはずで，それによって「1 文字の地名」も正しく認識されるはずだ，という予想を立てたわけである．

もちろん，この予想がうまくいくためには，他の地名用例コーパスも含め，できるだけ多くの地名用例コーパスが必要なうえに，対抗用例コーパスも十全に収録しておかねばならない．たとえば「莫」であれば，「n，名詞，固定物，地名」の「莫」も，「v，副詞，否定，禁止」の「莫」も，いずれも MeCab 漢文辞書に含まれている必要があるし，「莫」の副詞用例コーパスも十全に収録しておかねばならない．また，地名用例コーパスや対抗用例コーパスに加え，それら以外のコーパスも，バランスよく収録しておく必要がある．この目標のために，我々は，約 46,000 文の MeCab 漢文コーパスから，複数の入力者による分析結果が品詞レベルで完全に一致した用例（約 2,000 文，約 6,300 語，うち地名を約 400 語収録）を，本手法の学習用コーパスとして用いることにした．

この手法により，我々の形態素解析システムは，たとえば「莫滅莫」という（かなり人工的な）漢文を

```
莫　v，副詞，否定，禁止，*，*，莫，莫し，ナシ，*
滅　v，動詞，変化，制度，*，*，滅，滅す，ホロボス，五段
莫　n，名詞，固定物，地名，*，*，莫，莫，バク，*
```

「莫を滅すなかれ」と正しく処理できるようになった．

この手法の有効性と，この手法によって引き起こされている悪影響とを，定量的に評価すべく，我々は，以下の 3 種類の MeCab 漢文辞書を準備した [15]．

α：従来，我々が使用してきた MeCab 漢文辞書．
β：辞書 α に，「1 文字の地名」も含め，知りうる限りの古典中国語地名を追加した辞書．
γ：辞書 α から，地名を取り除いた辞書．

辞書 α に収録されていた地名の単語数は 111 語，辞書 β に収録されている地名の単語数は 1,240 語，辞書 γ は 0 である．

さらに，「1 文字の地名」文例およびその対抗用例を，地名テストデータ P（88 語）として準備した．また，地名テストデータ P との比較検討のために，文献 [12] で用いた M（69 語）と R（320 語）も，テストデータとして用いた．なお，比較を容易にするために，辞書 α，β，γ ともに，学習用コーパスは約 2,000 文で固定とした．実験に用いた MeCab のバージョンは 0.996 である．

実験結果として，各辞書に対する各テストデータの F 値（大品詞/品詞/意味素性/小素性）を**表 2** に示す．地名テストデータ P に関しては，辞書 α より辞書 β の方が F 値が上がっている．また，辞書 α より辞書 γ の方が F 値が低いことから，少なくとも地名テストデータ P に関しては，地名は追加すればするほど良い，という結論になると思われる．実際，地名テストデータ P の中で，F 値の良悪を決定づけていたのは，以下のような例文であった．

```
晉　n，名詞，主体，国名，*，*，晉，晉，シン，*
克　v，動詞，行為，交流，*，*，克，克つ，カツ，五段
衞　n，名詞，固定物，地名，*，*，衞，衛，エイ，*
磁　n，名詞，固定物，地名，*，*，磁，磁，ジ，*
洛　n，名詞，固定物，地名，*，*，洛，洛，ラク，*
州　n，名詞，制度，場，*，*，州，州，シュウ，*
```

「晉は衞，磁，洛州に克つ」である．このような「1 文字の地名」が連続している例文において，辞書 α や γ は，「衞」や「磁」や「洛」を，地名以外の名詞だと誤検出してしまうのである．

一方，テストデータ M については，辞書 β で小素性の F 値がわずかに下がっているものの，全体としてほとんど変化が見られない．テストデータ M には地名用例が含まれていないことから，辞書 β における地名の「過剰な追加」は，一般的な漢文の形態素解析に対して，ほとんど悪影響を及ぼさない，と結論づけることができる．

テストデータ R については，辞書 α と辞書 β で F 値に変化がなく，辞書 γ で大幅に F 値が下がっている．これは，テストデータ R に地名が含まれており，辞書 γ においてそれらの地名が取り除かれてしまったために，F 値が下がったと考えられる．一方，辞書 β で追加した地名は，テストデータ R の形態素解析に，良い影響も悪い影響も及ぼしていない．

以上，我々のテストデータに関しては，古典中国語地名を知りうる限り追加した辞書 β が，最も良好な結果を得られたといえる．少なくとも地名テストデータ P に関しては，辞書 β が最も良い結果となっているし，M と R に関し

ては，辞書 β で追加した地名はほとんど悪影響がなかった．

これらの結果から，できる限り多くの地名を MeCab 漢文辞書に追加する手法は，地名を含む漢文の認識精度を高めると同時に，地名を含まない漢文には悪影響がない，ということが，我々の知見として得られた．

4.2 官職の自動抽出

漢文における官職を自動抽出する際も，文字数の短い官職であれば，地名と同様の手法が効果的だった．実際，MeCab 漢文辞書と MeCab 漢文コーパスを十全に準備することで，たとえば「上下左右」の「左右」と，「引置左右」の「左右」を

```
上下　n, 名詞, 固定物, 関係,*,*, 上下, 上下, ジョウゲ,*
左右　n, 名詞, 固定物, 関係,*,*, 左右, 左右, サユウ,*

引　　v, 動詞, 行為, 動作,*,*, 引, 引く, ヒク, 五段
置　　v, 動詞, 行為, 設置,*,*, 置, 置く, オク, 五段
左右　n, 名詞, 人, 役割,*,*, 左右, 左右, サユウ,*
```

という形で正しく見分けることは，我々の形態素解析システムではすでに可能となっている[*1]．

その一方，複数の形態素から構成される（ように見える）官職もあり，これが我々を悩ませた．以下に，いくつかの典型例を示す．

- 丞
  「○○丞」の形を取る名詞は，ほぼすべて官職とみなせる．しかしながら，その形態素解析処理は問題を孕んでいる．たとえば「御史中丞」を 1 つの形態素だとみなしてしまうと，「右御史中丞」や「知御史中丞」をうまく処理できない．「右御史臺中丞」となると，もうどうしていいか分からない．また，「右」は必ずしも最初に付加されるとは限らず，「尚書右丞」「尚書左丞」のような例もある．これらに加え，「湖州長城丞」や「長沙縣丞」のように地名との複合が起こる場合もあって，混沌をきわめる．
- 郎中
  「○○郎中」の形を取る名詞は，まず間違いなく官職である．これらのうち，「兵部郎中」や「司勳郎中」のように，部署名や他の官職との単純な複合は，まだ何とか処理できる．しかしこれが，「兵部左司郎中」や「尚書司勳郎中」という形で複合すると，もはや形態素解析の手に負えない．
- 判～事，知～事，～従事
  「判○○事」「知○○事」「○○従事」の形を取る名詞は，かなりの確率で官職である．しかしながら，形態素解析の立場からすると，「判」「知」「従」はいずれも動詞とみなすべき形態素であり，これが問題を複雑にしている．たとえば「知民事」は，通常は「民事を知る」という文であって，官職ではない．一方「知政事」は官職である．あるいは「知吏部尚書事」は官職だが，内部に他の官職である「吏部尚書」を含んでしまっている．

すなわち，このような（いわば複合的な）官職は，形態素解析で自動抽出するのは，そもそも無理がある．漢文に現れる官職というものの複雑さに対し，我々の認識が甘かったことを示す失敗例だが，あえて本稿に含めておくことにする．

4.3 人名の自動抽出

漢文における人名の自動抽出に向けて，我々は，MeCab 漢文コーパスを洗い直し，「n, 名詞, 人, 姓氏」「n, 名詞, 人, 名」に分類されている形態素オブジェクトと，その形態素オブジェクトを含む文例を見直してみた．その結果，「n, 名詞, 人, 姓氏」については，地名抽出と同様の手法が有効だ，との感触が得られた．しかし，「n, 名詞, 人, 名」については，他の用例とのバッティングが，奇妙な方法で回避されていることが判明した．具体例として，『十八史略』巻之二に現れる「李斯」という人名に関して，我々が得た知見を，以下に述べる．

『十八史略』巻之二には，「斯」という文字が，全部で 16 例，出現する．これらのうち 6 例は，「李斯」という形で出現することから

```
李　n, 名詞, 人, 姓氏,*,*, 李, 李, リ,*
斯　n, 名詞, 人, 名,*,*, 斯, 斯, シ,*
```

であることは確実であり，実際の形態素解析においても，そう処理できる．問題は残る 10 例である．これら 10 例は「斯」が単独で出現するのだが，我々の判断では，最初の 9 例はすべて「n, 名詞, 人, 名」であり，最後の 1 例だけが「n, 代名詞, 指示,*」なのである．具体的には，「李斯」の話が続いている間は，ずっと「斯」は特定の人名である「李斯」を指しており，その後「李斯」が出てこなくなって，かなり文章が進んでから，やっと代名詞の「斯」がたった 1 例だけ出現する．つまり，「李斯」の話が続いている間は，話がややこしくならないよう，代名詞の「斯」の使用をあえて避けているわけである．

このような形で，「ストーリー」全体における用字の分布に異常が見られる場合，それが人名を指している可能性が高い，ということは推定できた．「斯」の例でいえば，曖昧語となりうる「斯」の曖昧性を下げるために，代名詞としての「斯」を使わない，という形で『十八史略』巻之二における用字分布が変わってしまっている．しかしながら，

[*1] 「引」と「置」は，本来は「v, 動詞, 行為, 役割」であるべきだ．しかし，現状の我々のシステムは，この例文において，「引」を「v, 動詞, 行為, 動作」，「置」を「v, 動詞, 行為, 設置」だと読んでしまう．動詞類も，さらに鍛える必要がある，ということだろう．

この推定を，人名の自動抽出にまで結び付けるような手法は，我々には開発できなかった．というのも，「ここで斯が出てきたなら，それは李斯であって，代名詞ではないだろう」ということを理解するには，本質的には「ストーリー」の理解が必要だからである．現状の我々の力不足を，痛感する限りである．

5. おわりに

本稿では，古典漢文に対する形態素解析について，我々が行ってきた研究を概観した．これらの研究の結果，特に名詞まわりの処理については，かなり良い成果が得られたと信じる．

ただし，それはあくまで定性的な側面であり，かならずしも定量的な評価が得られたわけではない．もちろん，F値による評価は，数次にわたって行った [8], [12], [15] のだが，どうも納得がいかないのだ．我々としては，我々が構築している形態素解析システムにおいて，そのインテリジェンス（というか，かしこさ）を評価したいのだが，F値は単純に間違いをチェックするだけである．具体的には，漢文コーパスを増やしていってもF値はあまり変化しないのだが，その結果を我々が読む限りでは，F値は変わらなくても，やはり少しずつ「かしこく」なっているのだ．漢文の文法上，スジのいい間違いとスジの悪い間違いが現実には存在するのだが，それらの差異をうまく引き受けてくれるような評価尺度を，我々はいまだ見つけ得ていないのである．

今後の研究の発展に期待されたい．

参考文献

[1] 黄　昌寧，趙　海：中文分詞十年回顧，中文信息学報，Vol.21, No.3, pp.8–18 (2007).

[2] Jiang, W., Huang, L., Liu, Q. and Lü, Y.: A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging, *Proc. ACL-08*, pp.897–904 (2008).

[3] Shen, M., Liu, H., Kawahara, D. and Kurohashi, S.: Chinese Morphological Analysis with Character-level POS Tagging, *Proc. ACL-2014*, pp.253–258 (2014).

[4] Huang, L., Peng, Y., Wang, H. and Wu, Z.: Statistical Part-of-Speech Tagging for Classical Chinese, *Proc. TSD 2002*, pp.115–122 (2002).

[5] Yasuoka, K., Yamazaki, N., Wittern, C., Nikaido, Y. and Morioka, T.: A Morphological Analysis of Classical Chinese Texts, *Proc. Digital Humanities 2014*, pp.410–412 (2014).

[6] 安岡孝一，Wittern, C., 守岡知彦，池田　巧，山崎直樹，二階堂善弘，鈴木慎吾，師　茂樹：古典中国語（漢文）の形態素解析，東洋学へのコンピュータ利用，第27回研究セミナー，pp.3–14 (2016).

[7] 守岡知彦：MeCabを用いた古典中国語の形態素解析の試み，情報処理学会研究報告，Vol.2008-CH-79, pp.17–22 (2008).

[8] 守岡知彦：MeCabを用いた古典中国語形態素解析器の改良，情報処理学会研究報告，Vol.2009-CH-84, No.3, pp.1–5 (2009).

[9] Morioka, T.: A Prototype of a Classical Chinese Morphological Analyzer based on MeCab, *Proc. Osaka Symposium on Digital Humanities 2011*, p.36 (2011).

[10] 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利用，第23回研究セミナー，pp.75–83 (2012).

[11] Pulleyblank, E.G.: Outline of Classical Chinese Grammar, UBC Press (1995).

[12] 山崎直樹，守岡知彦，安岡孝一：古典中国語形態素解析のための品詞体系再構築，人文科学とコンピュータシンポジウム「じんもんこん2012」論文集，pp.39–46 (2012).

[13] 守岡知彦：古典中国語形態素コーパスのLinked Data化の試み，人文科学とコンピュータシンポジウム「じんもんこん2013」論文集，pp.187–194 (2013).

[14] 守岡知彦：比較的最近のCHISE，東洋学へのコンピュータ利用，第25回研究セミナー，pp.33–46 (2014).

[15] 安岡孝一，守岡知彦，Wittern, C.，山崎直樹，二階堂善弘，鈴木慎吾：古典中国語形態素解析による地名の自動抽出，人文科学とコンピュータシンポジウム「じんもんこん2014」論文集，pp.63–68 (2014).

安岡 孝一

1965年生．1990年京都大学大学院修士課程修了．京都大学博士（工学）．1990年京都大学大型計算機センター助手．1997年同助教授．2000年京都大学人文科学研究所附属漢字情報研究センター助教授．2009年同所附属東アジア人文情報学研究センター准教授．2015年同教授．人文科学と情報科学の橋渡しをすべく，人文情報学の研究に従事．電子情報通信学会，電気学会各会員．

ウィッテルン クリスティアン

1962年生．1991年ハンブルク大学修士（漢学），1998年ゲッティンゲン大学博士（哲学）．1998年中華佛學研究所（台北）副教授，中華電子佛典協會の研究・開発担当．2001年京都大学人文科学研究所附属漢字情報研究センター助教授．2009年同所附属東アジア人文情報学研究センター准教授．2012年同教授．文献学的な手法によって漢籍のデジタル・テキストのあるべきすがたを探る．日本デジタル・ヒューマニティーズ学会，国際仏教学学会会員．

守岡 知彦 （正会員）

1969 年生．1999 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了，博士（情報科学）．1999 年電子技術総合研究所 COE 特別研究員．2000 年京都大学人文科学研究所附属漢字情報研究センター助手．2009 年同所附属東アジア人文情報学研究センター助教．漢字文献を中心とした人文情報学の研究に従事．

池田 巧

1962 年生．1990 年東京大学修士（中国語学），1993 年東京大学大学院博士課程単位取得．山梨県立女子短期大学専任講師，立教大学助教授を経て，1999 年京都大学人文科学研究所助教授，2013 年同教授．専門は漢藏語方言史研究で，語彙の体系および文構造の記述分析を行っている．

山崎 直樹

1962 年生．早稲田大学大学院文学研究科博士前期課程修了．同博士後期課程退学．早稲田大学助手，広島大学専任講師，大阪外国語大学助教授を経て，関西大学外国語学部教授．専門は言語構造の可視化とインストラクション設計．

二階堂 善弘 （正会員）

1962 年生．1985 年東洋大学文学部卒業，1997 年早稲田大学大学院文学研究科博士課程退学，博士（文学）・博士（文化交渉学）．1997 年東北大学大学院国際文化研究科助手，1998 年茨城大学人文学部助教授，2004 年関西大学文学部助教授，2005 年に同教授．大型電算機プログラマの経験あり．専門は中国の民間信仰であるが，人文情報学においても『電脳中国学入門』（好文出版）の著作がある．日本道教学会会員．

鈴木 慎吾

1973 年生．2007 年大阪外国語大学博士（言語文化学）．2008 年京都産業大学外国語学部助教．2011 年大阪大学世界言語研究センター講師．2012 年同言語文化研究科講師．中国語音の歴史的変遷に関する研究のかたわら中国語広東方言の教育に従事．日本中国語学会，日本中国学会，中国語教育学会各会員．

師 茂樹 （正会員）

1972 年生．1995 年早稲田大学第一文学部卒業，2001 年東洋大学大学大学院文学研究科博士後期課程退学，博士（文化交渉学・関西大学）．現在，花園大学教授．仏教学（唯識思想・仏教論理学）とともに，文字符号化，漢字文献情報処理，文化遺産の 3DCG 復元等の研究にも取り組む．