2008－CH－79 (3)
2008／7／25

MeCab を用いた古典中国語の形態素解析の試み

守 岡 知 彦†

古典中国語（漢文）電子テキストの蓄積が進む中、自然言語処理技術の重要性は高まって来ているが実装は少ない。本論文では、MeCab を用いた古典中国語用形態素解析器のプロトタイプについて概説するとともに、本格的な古典中国語文法コーパス作成のためのワークフローについて考察する。

A Prototyping of Morphological Analyzer for Classical Chinese based on MeCab

MORIOKA TOMOHIKO†

This paper explains an overview of an experimental Morphological Analyzer for Classical Chinese based on MeCab. In addition, the paper considers a workflow to develop grammatical corpus for Classical Chinese.

1. はじめに

現在、古典中国語（漢文）資料の電子化が各方面で行われているが、単純なテキストの電子化に比べて、マークアップやオントロジーといった知識処理のための基盤整備には労力がかかり、結果的に、その進展が遅れたり、バランスを欠いたものになりがちであるといえる。作業の多くを専ら人手に負っているのが現状であろう。こうした現状は、

(1) ツールチェーンの欠如
(2) 電子化資料の再利用性の悪さ

に起因していると思われる。こうした状況を改善する上で現実的な自然言語処理技術の利用は重要であるといえ、特に、形態素解析や構文解析といった自然言語解析のための基盤技術を整備していくことは極めて重要であるといえる。しかしながら、古典語の場合、開発者が少なく、文法コーパスの整備も遅れているのが現状であり、少ない開発リソースでの実現が求められるといえる。また、解析用辞書や文法コーパスの整備においては古典中国語や古典中国学の知見が必要となり、必ずしも自然言語処理に通じていない人でも参加可能なワークフローが望まれる。

こうした条件を鑑みれば、フルスクラッチからツール、辞書、文法コーパスを開発するのはあまり現実的ではないといえ、既存の利用できそうなものはなるべく利用してなんらかのプロトタイプを実現し、それを元に少しずつ辞書や文法コーパスを改良していくような漸進的アプローチを採るのが良いと考えられる。ツールや辞書、コーパスは互いに絡み合っており、また、闇雲にデータを増やせば良いというものではなく、『次元の罠』という言葉に象徴されるように、データを増やしたことによってかえって認識率が下がるということもあり得る。無意味な労力を避けるためにも、十分に処理を意識したデータ整備が望まれるといえ、そのためにも実際に動くプロトタイプを用意することは有効であるといえる。

そこで、著者は MeCab[2] と IPA 辞書をベースに、極めて少ない労力で、古典中国語のための形態素解析器のプロトタイプを実現することにした。本論文では、このプロトタイプの構成法について概説するとともに、プロトタイプのリファクタリングに基づく古典中国語のための辞書および文法コーパス作成のためのワークフローについて考察する。

2. MeCab とは

MeCab[2] は工藤拓氏によって開発されている形態素解析エンジンで☆、オープンソース・ソフトウェアとして公開されている☆☆。MeCab は言語、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、コーパス、品詞体系等を用意することで現代日本語以外の言語でもサポート可能な構造になっている。また、UTF-8 をサポートしており、UCS 統合漢字[1] が利用

† 京都大学人文科学研究所
Institute for Research in Humanities, Kyoto University

☆ 京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクトを通じて開発された。

☆☆ GPL, LGPL, または、BSD ライセンスに従って使用、再配布することができる。

可能である☆。こうした特徴から、古典中国語のための形態素解析エンジンとして有望であるといえる。

MeCab で日本語の形態素解析をするためには、現在、IPA コーパスに基づく「IPA 辞書」と京都コーパスに基づく「Juman 辞書」が公開されており、この内、前者が推奨されている。

3. 必要となるデータ

MeCab を利用するためには少なくとも辞書が必要である。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、学習用コーパスからパラメータ推定を行うことで、連接コストを考慮した解析が可能となっている。

このパラメータ推定を行うには、

(1) Seed 辞書
(2) 学習用コーパス

の 2 つが必要である。

Seed 辞書は利用者が用意する辞書であり、これを学習用コーパスとともに処理することで、連接コストの情報が付与された辞書が作成される。

3.1 辞　　書

MeCab の辞書はコンマ区切り (Comma Separated Value; CSV) 形式になっている。最初の 4 カラム目：

(1) 表層形（単語そのもの）
(2) 左連接状態番号
(3) 右連接状態番号
(4) コスト

は必須項目であり、2〜4 番目は Seed 辞書では使われないので 0 とすることになっている。

5 番目以降の欄は「素性」と呼ばれる項目で、利用者は好きなだけ素性を付与することができるようになっている。品詞、活用、読み、発音といった単語に関する情報はこの素性を用いて記述することができる。☆☆ この特徴により、MeCab は言語、辞書、コーパス独立性を得ているといえる。

3.2 学習用コーパス

学習用コーパスは、タブとコンマで区切られた複数の行と EOS のみの行で一文を表したものを連ねたものである（図 1)。

これは、MeCab のデフォルトでの出力と同一の形式であり、MeCab の出力を使って容易にコーパスを編集できるように設計されている。

```
智者　　n,*, 名詞, 一般,*,*, 智者, 智者, チシャ
不　　v,t, 副詞, 否定, 否定,*, 不, 不, ズ
惑　　v,i, 動詞,*,*,*, 惑, 惑, マドウ
EOS
不　　v,t, 副詞, 否定, 否定,*, 不, 不, ズ
遠　　v,t, 形容詞,*,*,*, 遠, 遠, トオシ
千　　n,*, 数詞, 数,*,*, 千, 千, セン
里　　n,i, 量詞,*,*,*, 里, 里, リ
EOS
無　　v,t, 動詞, 否定, 禁止,*, 無, 無, ナシ
恻隠　　n,*, 名詞, 一般,*,*, 恻隠, 恻隠, ソクイン
之　　p,t,-p-,+adj,*,*, 之, 之, ノ
心　　n,i, 名詞, 一般,*,*, 心, 心, ココロ
非　　v,t, 形容詞, 否定, 否定,*, 非, 非, アラズ
人　　n,i, 名詞, 一般, 人,*, 人, 人, ヒト
也　　p,i,-p., 語気, 陳述,*, 也, 也, ナリ
EOS
夫　　p,*,p-, 語気, 発語,*, 夫, 夫, ソレ
礼　　n,*, 名詞, 一般,*,*, 礼, 礼, レイ
者　　p,*,-p,+n,*,*, 者, 者,*
自　　n,*, 人称代詞, 一人称, 単数,*, 自, 自, ミズカラ
卑　　v,i, 動詞,*,*,*, 卑, 卑, ヒククス
而　　p,*, 接続詞,*,*,*, 而, 而,*
尊　　v,t, 動詞,*,*,*, 尊, 尊, タットブ
人　　n,i, 名詞, 一般, 人,*, 人, 人, ヒト
EOS
```

図 1　学習用コーパスの例

4. 辞書の作成

4.1 IPA 辞書からの漢語の抽出

IPA コーパスは現代日本語のコーパスであり、当然のことながら、IPA 辞書を用いたのでは古典中国語の形態素解析はできない。しかしながら、現代日本語には古典中国語由来の語彙が多数含まれ、そうした語彙を切り出すことにより、古典中国語の辞書をでっちあげることができると考えられる。

さて、どういったものが古典中国語の語彙となり得るかであるが、漢字以外の文字を含む語は明らかに古典中国語の語彙とはなり得ないといえる。また、品詞の規則的対応が存在することも必要であり、助詞や助動詞などは機械的に変換できないといえる。

動詞および（文語）形容詞は、古典中国語においても、それぞれ、動詞および形容詞となるものが多いと考えられ、語幹を取り出すことにより、機械的に変換可能であるといえる。

同様に、一般名詞は古典中国語においても名詞になるものが多いと仮定できる。また、サ変名詞は動詞になるものが多いと仮定できる。ただ、名詞の場合、古典中国語において句や文を構成していたものが日本語において名詞化したものが多数あり（例：非常、不祥)、こうしたものを除去する必要がある。

こうしたことから、次の品詞を持つ語彙を機械的に変換することにした（表 1)。

☆ BMP の範囲は完全に対応しており、Ext-B も（少なくとも入力文字列としては）利用可能なようである。

☆☆ なお、各単語において、各欄が意味する素性の定義は揃っていなければならない。

表 1　機械的に変換可能な品詞

IPA 辞書	古典中国語	変換語彙数
名詞（一般）	名詞（一般）	37552
名詞（副詞可能）	名詞（副詞可能）	606
名詞（数）	名詞（数）	23
名詞（サ変接続、1 文字）	動詞	129
動詞（基本形）	動詞	1683
形容詞（文語基本形）	形容詞	221

4.2　追加する必要のある語彙

IPA 辞書から機械的に変換した語彙だけでは、副詞や助動詞、助詞、前置詞、接続詞等の文法的に必要な語彙を欠き、十分な解析が行えない。そこで、こうした語彙を追加することにした。

4.3　その他利用可能な情報

月、干支、王朝名、元号、地名を追加した。

4.4　学習用コーパスからの変換

学習用コーパスの EOS のみの行を除く各行の形式は、3.1 節で述べた辞書の形式のうち、第 2, 3, 4 欄を削り、その部分をタブに置き換えたものになっている。よって、学習用コーパスの各行のタブを ,0,0,0, に置き換えることで容易に辞書の形式に変換可能である。例えば、UNIX 系の OS であれば、

```
grep -v EOS corpus \
| sed 's/    /,0,0,0,/' \
| sort | uniq > misc.corpus.csv
```

という処理でコーパスから辞書 misc.corpus.csv を作成することができる。

5.　品詞・素性

古典中国語と日本語は文法が全く異なるので、なんらかの品詞体系を用意する必要がある。3.1 節で述べたように、MeCab の辞書は第 5 欄以降に任意の素性を付けることができ、品詞に相当する情報はこれらの素性を使って記述するようになっているので、階層的な品詞体系を用いたり、意味素性を記述したりするようなことも可能である。

どのような品詞・素性を付けるかは、どのような処理をしたいかによる。MeCab は 1 つでも素性が異なれば別のエントリとして扱ってくれ☆、同じ見出し語のエントリが複数あったとしても、コーパスを用いた学習等によって、適切なエントリを推定してくれるので、文脈から読み取れるような情報であれば、素性として付ける価値があり得る。例えば、訓に関わる情報を付けておけば、訓を推定することが可能であり、これを使って自動訓読システムを実現することができるかも知れない。あるいは、自動的に返り点を付けるシステムを作りたいのであれば、返り点を取るかどうかという情報を付ければ良いといえる。あるいは、日付、人名、地名等の情報を抽出するには、簡単な意味素性を付ければ良いかも知れない。

このようにいろいろな品詞・素性を付けることが考えられるが、多数の素性を付けるには辞書やコーパスを作成するための労力が増える上、認識率が悪化するといえ、落しどころをどのへんにするかは問題である。おそらく、どうしても試行錯誤が必要となると思われるので、階層的な品詞・素性体系を用い、必要ならば素性を拡張するという方法が良いかも知れない。

著者は、幾つかの形式を試行錯誤した結果、

(1)　品詞 0（大品詞）
(2)　品詞 1（修飾）
(3)　品詞 2（通常の品詞）
(4)　品詞 3（意味的素性 1）
(5)　品詞 4（意味的素性 2）
(6)　品詞 5（関係的素性）
(7)　原表記
(8)　代表表記
(9)　日本語での読み
(10)　日本語表記
(11)　日本語での活用の種類

という形式を用いることにした。

ここで、「品詞 0」素性（大品詞）と「品詞 1」素性（補語の種類）は 7.2 節で述べる返り点付き漢文コーパスの利用を考慮して設けたものである。「品詞 0」素性（大品詞）は

n　名詞類
v　動詞類
p　助詞類

からなる大雑把な品詞分類である。「品詞 1」素性（修飾）は次に取り得る要素を表すのに用いることにし、

i　次に来る語を修飾しない（目的語、自動詞等）
t　次に来る語を修飾する（他動詞、助動詞、副詞等）

という情報を記載している。ここで、**t** は連用修飾と連体修飾を区別していない。

品詞体系は、「全訳 漢辞海 第二版」[4] の区分を参考に表 2, 3, 4 に挙げたものを定義した。

6.　ワークフロー

3 節で述べたように、MeCab で古典中国語の形態素解析をするには、Seed 辞書と学習用コーパスが必要である。この内、後者は、4.4 節で述べたように、前者から変換可能であるので、ある程度の網羅性を有した辞書を作った後には、学習用コーパスを作ることに注力すれば良いといえる。

6.1　文例の収集

3.2 節で述べたように、MeCab の学習用コーパスの形式は MeCab のデフォルトでの出力と同一の形式である。このため、MeCab の出力を元に作成するのが効率的であると考えられる。よって、まず、必要となるのは MeCab の入力となる白文コーパスであると

☆ 素性のサブセットからなる内部エントリも作成するようである。

表 2　品詞一覧（名詞類）

0	1	2	3	4
n	*	名詞	一般	*
n	*	名詞	副詞可能	*
n	*	名詞	日付	月
n	*	名詞	人	*
n	*	名詞	動物	*
n	*	固有名詞	日付	元号
n	*	固有名詞	地域	一般
n	*	固有名詞	地域	国
n	*	固有名詞	人	*
n	*	数詞	数	*
n	*	数詞	序数	干支
n	*	量詞	*	*
n	*	量詞	長さ	*
n	*	量詞	面積	*
n	*	量詞	人数	*
n	*	接尾辞・一般	*	*
n	*	接尾辞・副詞可能	*	*
n	*	地域接尾辞	*	*
n	*	地域接尾辞	人	*
n	*	地域接尾辞	人	称号
n	*	人名接尾辞	*	*
n	*	人称代詞	一人称	単数
n	*	人称代詞	一人称	複数
n	*	人称代詞	二人称	単数
n	*	人称代詞	二人称	複数
n	*	人称代詞	三人称	単数
n	*	人称代詞	三人称	複数
n	*	指示代詞	近称	*
n	*	指示代詞	遠称	*
n	*	指示代詞	虚称	*
n	*	指示代詞	傍称	*
n	*	指示代詞	無称	*
n	*	疑問代詞	人	*
n	*	疑問代詞	事物	*
n	*	疑問代詞	場所	*
n	*	疑問代詞	理由	*

いえる。

6.2　学習用コーパスの作成

前節で述べた白文を MeCab で解析した結果をチェックし、誤りを修正することによって、学習用コーパスを作成することができる。

この作業には古典中国語の知識が必要であり、古典中国語の知識を持った人がチェック・修正する必要があるといえる。

6.3　学　　習

学習用コーパスを更新したら、これを元に辞書を作成する (4.4 節) とともに、パラメータ推定を行う。これは次のような工程で行う：☆

(1)　Seed 辞書を元に学習用バイナリ辞書を作成 (mecab-dict-index)

(2)　CRF パラメータの学習 (mecab-cost-train)

(3)　配布用辞書の作成 (mecab-dict-gen)

☆ 筆者はシェルスクリプトによって、この工程を一括して行うようにしている。

表 3　品詞一覧（動詞類）

0	1	2	3	4	(Note)
v	*	動詞	*	*	
v	i	動詞	*	*	(自動詞)
v	t	動詞	*	*	(他動詞)
v	t	動詞	否定	禁止	
v	t	助動詞	可能	may	
v	t	助動詞	可能	能力	
v	t	助動詞	可能	実現	
v	t	助動詞	可能	価値	
v	t	助動詞	必要	推奨	
v	t	助動詞	必要	当然	
v	t	助動詞	必要	必須	
v	t	助動詞	願望	是非	
v	t	助動詞	願望	望む	
v	t	助動詞	願望	我慢	
v	t	助動詞	願望	積極	
v	t	助動詞	願望	強引	
v	t	助動詞	受動	受身	
v	*	形容詞	*	*	
v	i	形容詞	*	*	
v	t	形容詞	*	*	
v	t	形容詞	否定	否定	
v	t	副詞	程度	極度	
v	t	副詞	程度	軽度	
v	t	副詞	程度	やや高度	
v	t	副詞	範囲	総括	
v	t	副詞	範囲	限定	
v	t	副詞	範囲	共同	
v	t	副詞	時間	過去	
v	t	副詞	時間	現在	
v	t	副詞	時間	将来	
v	t	副詞	時間	終局	
v	t	副詞	時間	緊接	
v	t	副詞	時間	恒常	
v	t	副詞	時間	変化	
v	t	副詞	数量	重複	
v	t	副詞	数量	頻度	
v	t	副詞	謙敬	謙譲	
v	t	副詞	謙敬	表敬	
v	t	副詞	否定	否定	
v	t	副詞	否定	禁止	
v	t	副詞	語気	確定	
v	t	副詞	語気	推定	
v	t	副詞	語気	反語	
v	t	前置詞	時間/場所	*	
v	t	前置詞	原因	*	
v	t	前置詞	方式	*	
v	t	前置詞	関係	*	
v	t	p-	+n	*	(「所」)

(4)　解析用バイナリ辞書の作成 (mecab-dict-index)

6.4　実験と評価

前節のようにして、新たな学習用コーパスを元にした解析用バイナリ辞書を作成すれば、これを用いて MeCab で古典中国語の形態素解析を行うことができる。これにより、また、6.2 節の作業に戻ることができる。

一方、作成したコーパスや辞書、形態素解析処理系を評価することも重要である。MeCab はこのための

表 4　品詞一覧（助詞類）

0	1	2	3	4	(Note)
p	*	接続詞	*	*	
p	*	-p-	*	*	(助詞「之」)
p	*	-p-	+adj	*	(助詞「之」)
p	t	-p-	+adj	*	(助詞「之」)
p	*	-p	+n	*	(助詞「者」)
p	i	-p.	語気	陳述	(語気助詞・文末)
p	i	-p.	語気	疑問	(語気助詞・文末)
p	i	-p.	語気	反語	(語気助詞・文末)
p	i	-p.	語気	詠嘆	(語気助詞・文末)
p	i	-p.	語気	推測	(語気助詞・文末)
p	*	p-	語気	発語	(語気助詞・文頭)
p	*	p-	語気	ease	(語気助詞・文頭)

ツールを提供しており、テスト用コーパスがあれば、評価過程を自動化することができる。

このテスト用コーパスの形式は、学習用コーパスと同様、MeCab のデフォルトでの出力と同一の形式である。よって、6.2 節で述べた工程により、テスト用コーパスを作成する必要がある。

また、評価のためには、テスト用コーパスと学習用コーパスは別であるべきなので、6.2 節で述べた工程は、学習用コーパスの作成というよりも、テスト用コーパスや学習用コーパスの元となる文法コーパスを作る工程という風に理解できる。そして、この工程では、幾つかの異なるソースに対して文法コーパスを作るのが望ましいと考えられる。

その上で、評価過程では、幾つかのテスト用コーパスのセットに対して、それと異なる幾つかのコーパスをブレンドして作成した学習用コーパスを用いて実験し、適切な学習用コーパスをチューンするのが望ましいと考えられる。

7. 省力化の展望

MeCab で古典中国語の形態素解析をするには、学習用コーパスやテスト用コーパスとして用いられる文法コーパスの蓄積が必要となる。しかしながら、この工程では、古典中国語の知識を持った人によるチェックや修正といった手作業が必要となり、大きなコーパスを蓄積するには多数の労力が必要であると考えられる。

この労力を低減するための手法として、構文解析や返り点付きデータの利用が考えられる。

7.1 構文解析

構文解析により明らかな非文を除去することにより、人間がチェックすべき箇所を減らすことが考えられる。ただ、このためには、古典中国語のための構文解析器が必要であり、構文解析のための文法コーパスもあった方が良い。そのため、シリアスに考えれば、省力化のためにさらなる労力が必要となる結果になりかねないが、ここで必要となるのは省力化のための構文解析であり、MeCab で複数の候補を出した上で、その中から非文を除去することだけを考えた、簡易なものを実現すれば良いのではないかと考えられる。ただ、いずれにせよ、完全な自動化は無理であり、人間による作業を省力化するための適切な UI を考えることが重要であるといえる。

7.2 返り点付きデータの利用

山崎直樹氏は訓点付き漢文の返り点からの統語情報を抽出する手法を提案しており[3]、鈴木慎吾氏は、実際に、訓点付き漢文を構文解析し、その統語情報を XML で出力するツールを試作している（図 2）。これらの

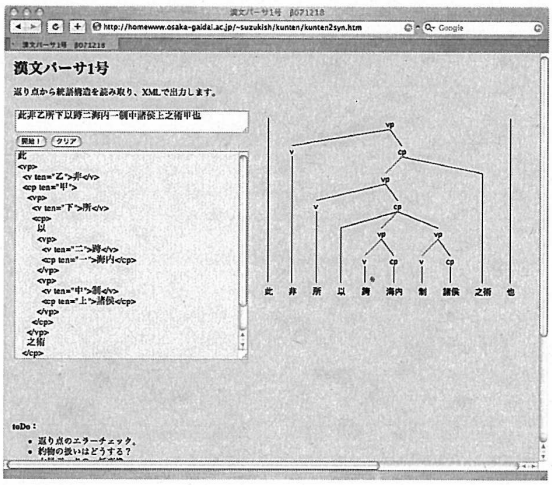

図 2　漢文パーサ 1 号

研究に基づき、現在、京都大学人文科学研究所「東アジア古典文献コーパスの研究」共同研究班では返り点の付いた漢文の情報を機械可読化することを目指している。

訓点付き漢文は電子化されてないものを含めれば多数のテキストが存在し、著作権の切れたテキストも多い。訓点付き漢文を単に入力するだけであれば、古典中国語に関する知識は必要なく、電子化を行う上で有利であると考えられる。

しかしながら、訓点付き漢文の返り点が持つ統語情報は完全な文法情報ではないという問題点がある。そもそも、返り点は日本語と語順が異なる場合にのみ付けられているのであり、語順が一致している場合には返り点は付かず、その部分の統語情報は得られない。また、MeCab で必要とする文法コーパスにおいては、品詞情報が非常に重要であるが、返り点の情報だけでは非常に限定的な品詞情報しか得られない。すなわち、『動詞類』か『目的語（補語）』かという情報しか判らず、『動詞類』においてそれが動詞、形容詞、助動詞、副詞、前置詞等のどれかは判別が付かないし、『目的

語（補語）』内に返り点がなければその中の構文情報も得られない。

こうした制約のもとで、訓点付き漢文から MeCab のための文法コーパスを作ることを考える場合、幾つかの工夫が必要である。

まず、品詞情報の不十分さを鑑みれば、返り点の情報からのみ得られるような品詞情報を品詞階層の大範疇として立てるのが良いと考えられる。即ち、『動詞類』（動詞、助動詞、副詞、助詞（「所」）等）と『名詞類』というような非常にラフな大品詞と、次に『目的語』を取るかどうかという情報を品詞 0, 1 素性として立てる訳である。

また、統語情報が不足しているために、そのままでは文法コーパスにならないが、MeCab の「制約付き解析（部分解析）」機能を利用することが考えられる。例えば、

此 非乙 所下 以 跨二 海内一 制中 諸侯上 之 術甲 也

という訓点付き漢文を

```
此
非      v
所      v
以
跨      v
海内    *
制      v
諸侯    *
之術
也
EOS
```

という MeCab 用の制約情報形式に変換し、`-p` (`--partial`) オプションを付けて MeCab を起動することで、返り点の情報から判る構文情報を形態素の分割位置や品詞 0 に対する制約として利用することができる。

8. おわりに

古典中国語テキストの電子化が進む中、自然言語処理技術の利用は重要な課題であり、そのためには、古典中国語のための文法コーパスの整備は非常に重要である。よって、現実的なワークフローを設計し、効率良く文法コーパスを整備・改良して行く方策を考えることは重要な問題であるといえる。

本論文では、MeCab[2] を用いた古典中国語用形態素解析器のプロトタイプについて述べるとともに、このプロトタイプを用いた漸進的な文法コーパスの開発のためのワークフローについて検討した。現代日本語は古典中国語由来の語彙を多数持っており、名詞・動詞・形容詞を中心にある程度まとまった量の語彙を機械的に変換できるといえる。そして、それに副詞や助動詞、助詞、前置詞、接続詞、代詞等の文法的に必要な語彙を追加することで、形態素解析器のプロトタイプを得ることができる。

しかしながら、ちゃんとした形態素解析器を実現するためには、やはり、ちゃんとした辞書と文法コーパスの蓄積が必要である。このためのツールとして、MeCab を用いた古典中国語用形態素解析器のプロトタイプは有用であると思われる。MeCab は出力結果をそのまま文法コーパスにすることができ、そこから簡単に辞書を作ることもできるので、誤っている部分を修正するだけでコーパスが作成でき、全てを人手で入力するのに比べて省力化が計れるといえる。また、作業の進展に従って、認識精度の向上が期待でき、作業の進展に従って省力化が進んで行くと考えられる。

最後に、本研究を行う上で、京都大学人文科学研究所「東アジア古典文献コーパスの研究」共同研究班のメンバー諸氏、特に、山崎直樹氏、鈴木慎吾氏、池田巧氏からさまざまな示唆を受けたことに感謝する。

参考文献

1) International Organization for Standardization (ISO). *Information technology — Universal Multiple-Octet Coded Character Set (UCS)*, March 2003. ISO/IEC 10646:2003.
2) MeCab: Yet Another Part-of-Speech and Morphological Analyzer. `http://mecab.sourceforge.net/`.
3) 山崎直樹. 訓点付き漢文の返り点から統語情報を導出し XML で構造化する試み. 漢字文献情報処理研究, Vol.8, , October 2007.
4) 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, January 2006.