様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年　6 月　7 日現在

機関番号：14301
研究種目：基盤研究(B)
研究期間：2010～2012
課題番号：22300087
研究課題名（和文）　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

研究課題名（英文）　Constructing a Corpus of Classical Chinese with Word-Class for Morphological Analysis

研究代表者
安岡　孝一　（YASUOKA KOICHI）
京都大学・人文科学研究所・准教授
研究者番号：20230211

研究成果の概要（和文）：古典漢文の白文（句読点や区切りや返り点のない単なる漢字の列）に対し、形態素解析をおこない、品詞情報つきの形態素に分解するシステムを構築した。また、形態素解析に必要な古典漢文コーパスと古典漢文辞書を、汎用の形態素解析エンジン MeCab に即した形式で作成し、WWW で公開した。これらと合わせ、古典漢文コーパスを構築するためのツール群も作成し、同じく WWW で公開した。

研究成果の概要（英文）：We constructed a morphological analyzer for plain texts in classical Chinese. Our analyzer segments classical Chinese sentences into a morpheme sequence, and it is based on MeCab, which is a language-independent morphological analyzer. We developed a “dictionary” and a “corpus” for MeCab, and also developed our own tools including a corpus editor. We released the dictionary, the corpus, and the corpus editor via WWW.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	6,200,000	1,860,000	8,060,000
2011 年度	3,700,000	1,110,000	4,810,000
2012 年度	3,300,000	990,000	4,290,000
年度			
年度			
総　計	13,200,000	3,960,000	17,160,000

研究分野：情報学
科研費の分科・細目：図書館情報学・人文社会情報学
キーワード：文学情報

１．研究開始当初の背景

京都大学人文科学研究所附属東アジア人文情報学研究センターは、その前身である附属東洋学文献センター時代から現在に至るまで、約 110,000 タイトルの古典漢籍文献を収集し、その保存と公開につとめてきた。また、1980 年代から、京都大学大型計算機センター（現、京都大学学術情報メディアセンター）とのコラボレーションにより、古典漢籍の全文テキストデータベース化に携わってきた。

これらの膨大な古典漢文テキストをコンピュータで処理するためには、白文そのままではなく、テキストを自然言語解析する必要がある。古典漢文のように、単語の間にも文の間にも区切りを持たない書写言語の解析

では、まず、単語を認識することが必須であり、そのために形態素解析をおこなわなければならない。しかしながら、古典漢文に対しては、このような形態素解析システムは、これまで構築されてこなかった。

研究代表者は、みずからが主催する京都大学人文科学研究所共同研究班「漢字情報学の構築」（平成 16～19 年度）において、古典漢文に対し、漢語文本切分与詞性標注（北京大学計算語言学研究所の現代中国語コーパス）を流用した形態素解析を試しにおこなってみたが、これは全くの失敗に終わった。失敗の原因を分析してみたところ、単語の出現頻度や品詞構造が、古典漢文と現代中国語とではかなり異なっていて、それぞれに異なるコーパスが必要となる、ということが明らかになった。それはつまり、現代中国語コーパスに関連して現在おこなわれている様々の研究は、残念ながら古典漢文には直接適用できない、ということを意味する。

そこで研究代表者は、平成 20 年度より新たに、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、古典漢文に対する形態素解析の研究を開始した。この共同研究班において、われわれは、言語に依存しない形態素解析エンジンとして MeCab を選び、さらに古典漢文を形態素解析するための品詞分類を研究した。これと平行して、IPAL(計算機用日本語基本辞書)から古典漢文にも出現する名詞の抽出作業をおこない、あわせて人名・地名辞典から固有名詞の抽出作業もおこなった。また、パイロットプランとして、品詞情報を含む古典漢文の例文も 200 例ほど入力した。

２．研究の目的

本研究の第一の主眼は、古典漢文の形態素解析に必要な漢文コーパスの作成である。すなわち、約 18,000 例の漢文に対して、本研究で定義する品詞情報および意味素性情報を付加したコーパスを作成する。例として用いる漢文は、『漢文大系』（冨山房）など、基本的に過去の研究によって釈文が確定しているものを用い、読み下し文を適宜参照することによって、例文の品詞の精度をできる限り高める。さらに汎用の形態素解析エンジン MeCab と併用することで、各単語間の接続確率を導出して、古典漢文コーパスを完成する。

本研究の第二の主眼は、作成した古典漢文コーパスをベースにした漢文解析システムの研究である。すなわち、作成した古典漢文コーパスを基に形態素解析システムを構築し、このシステムを用いて、一般的な漢文（白文）の形態素解析がおこなえることを示す。この際に、形態素解析が十分におこなえなかった漢文に対しては、その原因を解析過程に遡ってつきとめ、原因を除去するために必要な品詞の分類や接続確率を、古典漢文コーパスにフィードバックすることで、古典漢文コーパスと漢文解析システムをより良いものに仕上げていく。

３．研究の方法

本研究の中心をなすのは、品詞つき古典漢文コーパスである。品詞つき古典漢文コーパスの各例文は、元になる漢文を単語に分解し、それぞれに大品詞・品詞・意味素性・小素性を付加したものである。たとえば「自立爲夜郎侯」という例文に対しては、以下のようになる。

自	v,副詞,範囲,限定
立	v,動詞,行為,役割
爲	v,動詞,行為,役割
夜郎	n,名詞,主体,国名
侯	n,名詞,人,役割

大品詞は古典漢文の動賓構造に対応しており、「v」（動）、「n」（賓）、「p」（その他）の 3 種類としている。品詞は、最終的には、「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、従来の漢文文法等で見られた「形容詞」を廃止したのが特徴である。さらに、43 種類の意味素性と、80 種類以上の小素性を、最終的には定義した。

本研究計画のおおまかなデータフローは、以下の通り。

①品詞つき古典漢文コーパスのフォーマットにしたがい、例文入力グループ（中国学を専攻する博士課程学生およびオーバードクターで構成）が、例文を入力する。ただし、上記フォーマットをいきなり手で打ち込むのではなく、例文に対する仮の形態素解析をおこない、その出力結果（当然ながら誤りを含む）を手作業で補正する形で入力作業をおこなう。

②入力されたデータは順次、デジタル処理グループ（師・守岡）に渡される。デジタル処理グループは複数人による入力を突き合わせ、矛盾がない場合には、古典漢文コーパスに追加し形態素解析エンジンにも反映させる。矛盾している場

合には、問題のデータをコーパス校訂グループ（安岡・二階堂・ウィッテルン）に渡す。コーパス校訂グループは、データの矛盾が簡単に直せる場合は、直したデータをデジタル処理グループに戻す。直せない場合、すなわち品詞分類の修正等が必要だと考えられる場合は、問題のデータを品詞分類グループ（山崎・池田・鈴木）に渡す。

③品詞分類グループは、問題のデータにより、品詞分類に変更を加える必要があるかどうかを検討する。その結果、新たな品詞・意味素性・小素性を拡充した場合は、それをデジタル処理グループに伝え、形態素解析エンジンに反映させる。また、必要があれば、過去に入力したデータも、デジタル処理グループに自動修正させる。同時に、修正後の新たな品詞分類表を、例文入力グループにも伝え、その後の例文入力に反映させる。

４．研究成果

平成22年度は、古典漢文コーパス構築のための基礎作業として、まず『漢文大系』の全文画像を構築し、さらに全文テキスト化のための目次情報を構築した。これに並行して、品詞分類グループは品詞処理のためのプロトタイプを設計したが、その際に、散文と韻文とでかなり文法構造が異なることが発見された。そこで、散文と韻文とを分離すべく、デジタル処理グループと共同で、韻文の基本構造を自動抽出する手法を考案し、例文ベースでの検証をおこなった。この手法により散文と韻文を分離することが可能となり、それぞれの文法構造に応じた形態素解析エンジンを作りこむことが可能となった。

また、平成 22 年度の研究成果を、2011 年 3 月 28～29 日開催の国際シンポジウム Osaka Symposium on Digital Humanities 2011 において発表すべく、3 件の extended abstract を研究代表者・研究分担者ともども投稿したところ、見事 3 件とも採択された。しかし、2011 年 3 月 11 日に発生した東日本大震災により、シンポジウムが 2011 年 9 月に開催延期となってしまった。

平成 23 年度は、古典漢文コーパス構築のためのシステム設計をおこない、さらに『漢文大系』から、複数の初期コーパスを作成する作業をおこなった。デジタル処理グループとコーパス校訂グループは、この初期コーパスを検討し、品詞分類に多少の改善を施す必要を認めた。これに基づき、品詞分類グループは改訂版の品詞分類を提案し、例文入力グループにフィードバックして、新たな品詞分類による基本コーパスの入力作業を開始した。

また、平成 22～23 年度の研究成果を、2011 年 9 月 12～14 日開催の国際シンポジウム Osaka Symposium on Digital Humanities 2011 に 3 件の発表論文として再投稿した。これらは 3 件とも採択となり、1 セッションまるまるを本研究「古典漢文コーパス」に割り当てていただいた。漢文のコンピュータ処理というのは、人文情報学の中でもかなりマイナーな分野なのだが、われわれの手法が他の古典言語に対しても適用可能だという「熱気」を、これら 3 件の発表に対する他の研究者の質問および意見として、強く感じとることができた。

平成 23 年度までに作成した古典漢文コーパスにより、品詞分類に改善の必要が認められたことから、平成 24 年度は、新たな品詞分類の設計をおこなった。新たな品詞分類では、大品詞を「n」「v」「p」の 3 種類とし、動賓構造を「v」と「n」の組み合わせで表現することにした上で、その下位分類での品詞を「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、従来の漢文文法等で見られた「形容詞」を廃止した。これらに加え、43 種類の意味素性と、80 種類以上の小素性を定義し、形態素解析の結果として得られる各単語を、意味の面からも捉えやすいよう工夫した。

さらに、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、例文入力グループにフィードバックして、MeCab 漢文コーパスの入力をおこなった。また、MeCab 漢文辞書と MeCab 漢文コーパスを元に、MeCab による漢文の自動形態素解析をおこなえるようにした。この形態素解析システムで、高校教科書の漢文例や、『三国志呉書列伝』などの白文を実際に解析してみたところ、大品詞の F 値は平均で 92、品詞の F 値は平均で 84 と、まずまずの高成績が得られ、白文の単位切りはほぼ完璧だった。

本研究の最終仕上げとして、全体の MeCab 漢文辞書（約 5,000 語）および MeCab 漢文コーパスデータ（約 18,000 例）を WWW で公開し、当初研究計画をほぼ予定通り終了した。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者に

は下線）

〔雑誌論文〕（計 2 件）
① 安岡孝一、「東アジア古典文献コーパスの研究」共同研究班報告、東方學報(京都)、査読有、第 88 冊、2013、掲載予定
② Tomohiko Morioka、Christian Wittern、Koichi Yasuoka、Naoki Yamazaki、A Study of Linguistic Analysis for Classical Chinese Texts、International Conference on Culture and Computing、査読有、2013、掲載予定

〔学会発表〕（計 5 件）
① 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」、2012 年 11 月 17 日、札幌
② 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、2012 年 3 月 16 日、京都
③ Tomohiko Morioka、A Prototype of a Classical Chinese Morphological Analyzer based on MeCab 、Osaka Symposium on Digital Humanities 2011、2011 年 9 月 14 日、大阪
④ Naoki Yamazaki 、Toward Syntactic Frame Retrieval of Classical Chinese Rhymes Using Japanese `kun' Readings and Syntactic Parallelism of Couplets、Osaka Symposium on Digital Humanities 2011、2011 年 9 月 14 日、大阪
⑤ Koichi Yasuoka、Toward a Syntactic Analysis of Classical Chinese Texts、Osaka Symposium on Digital Humanities 2011、2011 年 9 月 14 日、大阪

〔その他〕
ホームページ等
http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/kyodokenkyu/archive2013.html

６．研究組織
(1)研究代表者
安岡　孝一（YASUOKA KOICHI）
京都大学・人文科学研究所・准教授
研究者番号：20230211

(2)研究分担者
山崎　直樹（YAMAZAKI NAOKI）
関西大学・外国語学部・教授
研究者番号：30230402

二階堂　善弘（NIKAIDO YOSHIHIRO）
関西大学・文学部・教授
研究者番号：70292258

師　茂樹（MORO SHIGEKI）
花園大学・文学部・准教授
研究者番号：70351294

クリスティアン　ウィッテルン
（WITTERN, CHRISTIAN）
京都大学・人文科学研究所・教授
研究者番号：20333560

池田　巧（IKEDA TAKUMI）
京都大学・人文科学研究所・准教授
研究者番号：90259250

守岡　知彦（MORIOKA TOMOHIKO）
京都大学・人文科学研究所・助教
研究者番号：40324701

鈴木　慎吾（SUZUKI SHINGO）
大阪大学・言語文化研究科・講師
研究者番号：20513360

(3)連携研究者
（　　）

研究者番号：