様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １５ 日現在

機関番号：１４３０１
研究種目：基盤研究(B)（一般）
研究期間：2017～2019
課題番号：１７Ｈ０１８３５

研究課題名（和文）古典漢文形態素コーパスにもとづく動詞の作用域の自動抽出

研究課題名（英文）Dependency Parsing in Classical Chinese along Morphological Corpora

研究代表者
安岡　孝一（Yasuoka, Koichi）

京都大学・人文科学研究所・教授

研究者番号：２０２３０２１１

交付決定額（研究期間全体）：（直接経費）　　11,200,000円

研究成果の概要（和文）：古典漢文における動詞の作用域、すなわち「動詞の後に置かれる項」のまとまりを、自動抽出する手法の開発をおこなった。具体的には、Universal Dependenciesと呼ばれる文法記述手法を用いて、いわゆる四書（『孟子』『論語』『大學』『中庸』）の係り受けコーパスを制作し、これを用いて、古典漢文の形態素解析と依存文法解析（係り受け解析）をおこなうツールUD-Kanbunを作成した。さらに、このツールを発展させて、動詞の作用域を元に返り点の自動生成をおこない、日本語の活用語尾と助詞を自動で付加することで、自動的に訓読をおこなうツールUD-Kundokuを試作した。

研究成果の学術的意義や社会的意義
古典漢文における単語間の係り受けを自動で抽出する手法を実現したことで、これまで文法的な構造化がおこなわれず白文（単なる漢字の列）のままで放置されてきた大量の古典漢文テキストに対し、その構造化すなわち文法解析をおこなうための端緒を付けることができた。また、高等学校などで教授されている漢文訓読という手法を、現代的なコンピュータ処理によって自動化することにより、漢文訓読が言語学的に（特に依存文法における解析として）何をおこなっているのかを、コンピュータプログラムの形で示すことに成功した。

研究成果の概要（英文）：We have developed a parsing method of dependencies between words in Classical Chinese. Our method is under the syntax scheme of Universal Dependencies, that are developed by natural language researchers all over the world. In a practical point of view, we have developed Universal Dependencies Treebank of the Four Books (孟子, 論語, 大學, and 中庸) in Classical Chinese, then developed a python-module named UD-Kanbun, which is tokenizer, POS-tagger, and dependency-parser for Classical Chinese. We also developed a python-module named UD-Kundoku, which is a transcriptive converter from Classical Chinese into Modern Japanese, along with an encode-reorder-decode scheme.

研究分野：人文情報学

キーワード：文法解析　古典中国語

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 130,000 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）との共同研究で、古典漢籍の全文テキストデータベース化をおこなってきた。

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（単なる漢字の列）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、形態素解析を十全におこなった上で、その結果を元に構文解析を進めていく、という手法を取らざるを得ない。ただし、現代中国語と違って、単語の間にも文の間にも区切りを持たない古典漢文に対しては、現代中国語の解析手法が無力であり、新たな手法を開発しなければならないという問題があった。

この問題に対し、研究代表者は、2008 年度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、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。また、この共同研究班および後身の共同研究班「東アジア古典文献コーパスの応用研究」を母体として、2010～2012 年度に科学研究費基盤研究(B)『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』、2013～2015 年度に科学研究費基盤研究(B)『品詞素性情報つき古典漢文コーパスの発展的応用』により、古典漢文コーパスの構築と形態素解析の研究をおこなった。これらの研究成果は、古典漢文コーパスや辞書ファイルも含めて WWW 上で全て公開しており、古典漢文を形態素解析する環境は、われわれが研究を開始した時点に比べれば、格段に向上していると言える。特に名詞類（名詞・代名詞・数詞）については、『品詞素性情報つき古典漢文コーパスの発展的応用』において、官職・地名・人名を集中的にあつかったことで、研究にかなりの進展が見られた。

これらの研究をさらに進めるべく、われわれは、2016 年度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実証研究」を組織し、古典漢文の構文解析手法を構築すべく、新たな研究に着手した。

古典漢文は、基本的に SVO の語順を有する書写言語であり、VO がまず動詞句を構成し、それが S と結びつく動詞句階層型の言語だと考えられる。ただし、古典漢文は、単語の間にも文の間にも区切りを持たない。このような言語において、文という単位を切り出すためには、個々の動詞が必要とする項のうち、動詞の後に置かれる項を抽出することができれば、文の切れ目を決めることができるはずである。この「動詞の後に置かれる項」のまとまりを、一般に、動詞の作用域という。共同研究班「東アジア古典文献コーパスの実証研究」における検討の結果、古典漢文における動詞の作用域こそが、われわれの新たな研究課題であるとの認識に至った。

２．研究の目的

本研究では、漢から清にかけて大量に蓄積された古典漢文テキストに対し、品詞情報を付加した形態素解析をおこなった上で、動詞の作用域、すなわち、その動詞が後続のどこまでの部分にかかっているか、を自動抽出する手法を構築する。本研究は、古典漢文における構文解析の足がかりとなるものであり、文法的な構造化がおこなわれず白文（単なる漢字の列）のままで放置されている大量の古典漢文テキストに対し、その構造化すなわち文法解析をおこなうための一歩となるものである。

３．研究の方法

古典漢文に対し、形態素解析をおこなった上で、複数の文法解析手法による文法記述と、Treebank を用いた構文解析とを併行しておこない、動詞の作用域の自動抽出手法を構築する。最初に、動詞の作用域の中に他の動詞を含まないような「単文」に対して、手法の構築をおこなう。次に、動詞の作用域の中に他の動詞を含むような「複文」に対し、手法を拡張して、作用域の自動抽出をおこなう。その上で、白文（単語の間にも文の間にも区切りを持たない漢字の列）に対し、さらに手法を拡張して、全ての動詞の作用域の自動抽出をおこなう。

４．研究成果

古典漢文における動詞の作用域、すなわち「動詞の後に置かれる項」のまとまりを、自動抽出すべく、まずは単文（動詞の作用域の中に他の動詞を含まない文）に対して、文法記述手法の検討をおこなった。当初は、Chomsky 流の生成文法およびその変形による手法を試みたが、これらは動詞の作用域を抽出するという目的には、やや不適切であることが明らかとなった。端的に言うと、古典漢文は SVO 型の言語ではなく、S を必ずしも必要としない VO 型の言語であり、しかも末尾に終助詞を伴う predicate-object-final 型の言語として扱うべきだという現実が、浮かび上がってきたのである。このため、初年度（2017 年度）は Treebank の作成に着手できず、研究計画の遅れが生じた。

この遅れを挽回すべく、われわれは様々な文法記述手法を検討し、その中で、Мельчук 流の依存文法による記述手法の一つである Universal Dependencies に辿りついた。Universal Dependencies は、品詞・形態素属性・依存構造情報（単語間の係り受け関係）を言語に依存せず記述する手法である。句構造を考慮せずに係り受け関係を記述できるよう、全ての構文構造を単語間のリンクで記述するのが特徴である。

Universal Dependencies を用いて、Pulleyblank『Outline of Classical Chinese Grammar』の 597 の例文を検討してみたところ、どの例文もほぼ問題なく記述できる上に、記述それ自体が動詞の作用域を示す形となった。そこで、これら 597 例文の Universal Dependencies 記述を、そのまま古典漢文 Treebank の形に仕上げると同時に、東アジア人文情報学研究センターのセンター研究年報 2018『Universal Dependencies にもとづく古典中国語（漢文）の依存文法解析』として出版・公表した。さらに、Universal Dependencies による記述をおこなうための各種ツールを、SVG (Scalable Vector Graphics)と JavaScript によって実装し、公表した。これにより、計画の遅れを補って余りあるスピードで、Treebank の開発を進められるようになった。

このような形で、Universal Dependencies による Treebank 開発の準備が整ったことから、われわれは、いわゆる四書（『孟子』『論語』『大學』『中庸』）の全文記述をおこなった。これらの全文記述を、そのまま古典漢文 Treebank の形に仕上げ、うち『孟子』については、東アジア人文情報学研究センターのセンター研究年報 2018 別冊「古典中国語 Universal Dependencies で読む『孟子』」として出版・公表した。さらに、これらの古典漢文 Treebank を、東アジア人文情報学研究センターの GitLab サーバから WWW 公開すると同時に、プラハ・カレル大学との国際連携により、Universal Dependencies 2.4 および Universal Dependencies 2.5 の一部として WWW 公開した。これら 3 つの WWW 公開ページの URL は、以下のとおり。

https://corpus.kanji.zinbun.kyoto-u.ac.jp/gitlab/Kanbun/ud-kanbun
http://hdl.handle.net/11234/1-2988
http://hdl.handle.net/11234/1-3105

この Treebank をもとに、形態素解析エンジン MeCab と依存文法解析エンジン UDPipe を組み合わせて、古典漢文における動詞の作用域および係り受け関係を自動抽出するツールを作成し、UD-Kanbun という python モジュールとして WWW 公開した。このツールを公開するにあたり、本手法の応用の一つとして、漢文の白文に自動で返り点を打つページを試作した。さらに、デモンストレーションとして、2014～2019 年の大学入試センター試験「国語」の第 4 問（漢文）の本文から返り点を除去し、あらためて本研究の手法で返り点を打ってみせたところ、かなりの確度で返り点を打てることが実証された。すなわち、動詞の作用域が自動で抽出できていることが、実証されたわけである。UD-Kanbun の WWW 公開ページの URL は、以下のとおり。

https://github.com/KoichiYasuoka/UD-Kanbun

この UD-Kanbun のさらなる応用として、動詞の作用域にもとづいて漢文の返り点を自動抽出し、抽出した動詞に日本語の活用語尾を自動で付加し、さらには抽出した名詞に日本語の助詞を自動で付加することで、漢文の自動訓読をおこなうツールを試作し、UD-Kundoku という python モジュールとして WWW 公開した。さらに、デモンストレーションとして、2020 年 1 月の大学入試センター試験「国語」の第 4 問（漢文）の本文から返り点と送り仮名を除去し、あらためて UD-Kundoku で処理したところ、かなり高い精度で自動訓読できることが実証された。UD-Kundoku の WWW 公開ページの URL は、以下のとおり。

https://github.com/KoichiYasuoka/UD-Kundoku

なお、本研究の基礎技術である古典漢文形態素コーパスに関し、情報処理学会論文誌 2018 年 2 月号に『古典中国語（漢文）の形態素解析とその応用』を発表したところ、当該論文が情報処理学会誌特選論文に選定された。これに加え、2018 年 1 月に情報処理学会「人文科学とコンピュータ」研究会で発表した『古典中国語 Universal Dependencies への挑戦』に対し、2019 年度の情報処理学会第 81 回全国大会において、山下研究記念賞を受賞し表彰を受けた。

５．主な発表論文等

〔雑誌論文〕　計16件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　15件）

１．著者名 安岡孝一	４．巻 32
２．論文標題 漢文自動訓読ツールUD-Kundokuの開発	５．発行年 2020年
３．雑誌名 東洋学へのコンピュータ利用	６．最初と最後の頁 3-25
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 守岡知彦	４．巻 61(2)
２．論文標題 内容アドレッシングを用いた多粒度漢字構造情報表現の試み	５．発行年 2020年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 171-178
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安岡孝一	４．巻 94
２．論文標題 漢文の形態素解析・依存文法解析・直接構成鎖解析	５．発行年 2019年
３．雑誌名 東方學報	６．最初と最後の頁 330-322
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安岡孝一	４．巻 2019(1)
２．論文標題 漢日英Universal Dependencies平行コーパスとその差異	５．発行年 2019年
３．雑誌名 人文科学とコンピュータシンポジウム論文集	６．最初と最後の頁 43-50
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Koichi Yasuoka	4．巻 DADH2019
2．論文標題 Universal Dependencies Treebank of the Four Books in Classical Chinese	5．発行年 2019年
3．雑誌名 Proceedings of International Conference of Digital Archives and Digital Humanities	6．最初と最後の頁 20-28
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 山崎直樹	4．巻 37
2．論文標題 古典中国語のテクストをいかに切り分けるか	5．発行年 2019年
3．雑誌名 開篇	6．最初と最後の頁 111-119
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 安岡孝一	4．巻 2019-CH-120(1)
2．論文標題 Universal Dependenciesの拡張にもとづく古典中国語(漢文)の直接構成鎖解析の試み	5．発行年 2019年
3．雑誌名 情報処理学会研究報告	6．最初と最後の頁 1-8
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 安岡孝一	4．巻 2018別冊
2．論文標題 古典中国語Universal Dependenciesで読む『孟子』	5．発行年 2019年
3．雑誌名 センター研究年報(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)	6．最初と最後の頁 1-519
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 守岡知彦	４．巻 2018-CH-118(6)
２．論文標題 古典中国語UDコーパスのIPFSを用いた表現の試み	５．発行年 2018年
３．雑誌名 情報処理学会研究報告	６．最初と最後の頁 1-7
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安岡孝一	４．巻 18
２．論文標題 古典中国語(漢文)の依存文法解析と直接構成素解析	５．発行年 2018年
３．雑誌名 漢字文献情報処理研究	６．最初と最後の頁 56-62
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 師茂樹	４．巻 18
２．論文標題 形態素解析とは何か	５．発行年 2018年
３．雑誌名 漢字文献情報処理研究	６．最初と最後の頁 42-45
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安岡孝一	４．巻 1
２．論文標題 漢文の依存文法解析と返り点の関係について	５．発行年 2018年
３．雑誌名 日本漢字学会研究大会予稿集	６．最初と最後の頁 33-48
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安岡孝一	４．巻 30
２．論文標題 四書を学んだMeCab＋UDPipeはセンター試験の漢文を読めるのか	５．発行年 2019年
３．雑誌名 東洋学へのコンピュータ利用	６．最初と最後の頁 3-110
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安岡孝一、ウィッテルン クリスティアン、守岡知彦、池田巧、山崎直樹、二階堂善弘、鈴木慎吾、師茂樹	４．巻 59(2)
２．論文標題 古典中国語(漢文)の形態素解析とその応用	５．発行年 2018年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 323-331
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安岡孝一	４．巻 2018
２．論文標題 Universal Dependenciesにもとづく古典中国語(漢文)の依存文法解析	５．発行年 2018年
３．雑誌名 センター研究年報(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)	６．最初と最後の頁 1-104
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安岡孝一、ウィッテルン クリスティアン、守岡知彦、池田巧、山崎直樹、二階堂善弘、鈴木慎吾、師茂樹	４．巻 2018-CH-116(20)
２．論文標題 古典中国語Universal Dependenciesへの挑戦	５．発行年 2018年
３．雑誌名 情報処理学会研究報告	６．最初と最後の頁 1-8
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計11件（うち招待講演　1件／うち国際学会　1件）

1．発表者名 守岡知彦
2．発表標題 古典中国語UDコーパスのIPFSを用いた表現の試み
3．学会等名 情報処理学会『人文科学とコンピュータ』研究会
4．発表年 2018年

1．発表者名 安岡孝一
2．発表標題 漢文の依存文法解析と返り点の関係について
3．学会等名 日本漢字学会
4．発表年 2018年

1．発表者名 安岡孝一
2．発表標題 古典中国語Universal Dependenciesへの挑戦
3．学会等名 情報処理学会『人文科学とコンピュータ』研究会
4．発表年 2018年

1．発表者名 安岡孝一
2．発表標題 四書を学んだMeCab＋UDPipeはセンター試験の漢文を読めるのか
3．学会等名 東洋学へのコンピュータ利用第30回研究セミナー
4．発表年 2019年

１．発表者名
安岡孝一

２．発表標題
AIを用いた漢文の文法解析

３．学会等名
日本中国学会・KU-ORCASシンポジウム（招待講演）

４．発表年
2019年

１．発表者名
Koichi Yasuoka

２．発表標題
Universal Dependencies Treebank of the Four Books in Classical Chinese

３．学会等名
DADH 2019: 10th International Conference of Digital Archives and Digital Humanities（国際学会）

４．発表年
2019年

１．発表者名
安岡孝一

２．発表標題
Universal Dependenciesの拡張にもとづく古典中国語(漢文)の直接構成鎖解析の試み

３．学会等名
情報処理学会『人文科学とコンピュータ』研究会

４．発表年
2019年

１．発表者名
二階堂善弘

２．発表標題
漢籍研究環境の変容と今後の課題

３．学会等名
情報化時代の東洋学研究―デジタルアーカイブスの現状と課題―

４．発表年
2019年

１．発表者名
安岡孝一

２．発表標題
漢日英Universal Dependencies平行コーパスとその差異

３．学会等名
人文科学とコンピュータシンポジウム「じんもんこん2019」

４．発表年
2019年

１．発表者名
師茂樹

２．発表標題
漢文古典文献を分析するためのツールの普及に向けた取り組み

３．学会等名
漢字文献情報処理研究会

４．発表年
2020年

１．発表者名
安岡孝一

２．発表標題
漢文自動訓読ツールUD-Kundokuの開発

３．学会等名
東洋学へのコンピュータ利用第32回研究セミナー

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
「東アジア古典文献コーパスの実証研究」共同研究班ログ
http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/kyodokenkyu/archive2020.html

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	山崎　直樹 (Yamazaki Naoki) (30230402)	関西大学・外国語学部・教授 (34416)	
研究分担者	二階堂　善弘 (Nikaido Yoshihiro) (70292258)	関西大学・文学部・教授 (34416)	
研究分担者	師　茂樹 (Moro Shigeki) (70351294)	花園大学・文学部・教授 (34313)	
研究分担者	Ｗｉｔｔｅｒｎ　Ｃ. (Wittern Christian) (20333560)	京都大学・人文科学研究所・教授 (14301)	
研究分担者	池田　巧 (Ikeda Takumi) (90259250)	京都大学・人文科学研究所・教授 (14301)	
研究分担者	守岡　知彦 (Morioka Tomohiko) (40324701)	京都大学・人文科学研究所・助教 (14301)	
研究分担者	鈴木　慎吾 (Suzuki Shingo) (20513360)	大阪大学・言語文化研究科（言語社会専攻、日本語・日本文化専攻）・講師 (14401)	