

京都大学人文科学研究所共同研究最終報告書

1. 研究課題

東アジア古典文献コーパスの実証研究

Empirical Research on Digital Analysis of Classical Chinese Texts

2. 研究代表者氏名

安岡孝一

Koichi Yasuoka

3. 研究期間

2016年04月 - 2020年03月 (4年度目)

4. 研究目的

2010年以来、我々が構築を続けてきた漢文コーパスは、MeCabを用いた形態素解析手法を、漢文処理に適用するものである。この漢文コーパスでは4階層の品詞体系を採用しており、その第2層は「名詞」「代名詞」「数詞」「動詞」「前置詞」「副詞」「助動詞」「助詞」「感嘆詞」の9種類の品詞で構成される。すなわち我々は、従来の漢文文法等で見られた「形容詞」を廃止しているのだが、これが動詞類全体にどのような影響を及ぼしているのかは、必ずしも十分に検討できていない。本共同研究では、漢文コーパスにおける動詞類の実証研究をおこなう。すなわち、実際のコーパスにおいて「動詞」「前置詞」「副詞」「助動詞」の4つのふるまいを研究し、さらに下層の意味素性と小素性についても、現在の品詞体系の妥当性を検証する。

In our recent research we have developed a method to analyze classical Chinese texts. In our method, we use our original morphological analyzer based on MeCab. We proposed our original four-level word-class system for classical Chinese on the MeCab-based analyzer. We designed the top level of the word-class system to represent the predicate-object structure of classical Chinese. The second level is the ordinary word-class of classical Chinese, consisting: “noun,” “pronoun,” “numeral,” “verb,” “preposition,” “adverb,” “auxiliary verb,” “particle,” and “interjection.” The third and fourth levels are word-subclasses to describe detailed behavior of the words in classical Chinese texts. In other words, we excluded “adjective” from the second level of our word-class system. But we did not precisely examine the effect of lack of “adjective” for our morphological analyzer. In this research we will examine the effectiveness of our four-level word-class system,

especially examine “verb,” “preposition,” “adverb,” and “auxiliary verb” in the second level. We will also examine the validity of the third and fourth levels of our word-class system.

5. 研究成果の概要

古典中国語(漢文)における形態素解析手法を発展させて、さらに文法解析へと展開すべく、数々の手法を検討した。具体的には、Chomsky 流の文法解析手法およびその亜種は古典中国語への適用が難しく、М е л ь ч у к 流の依存文法(Dependency Grammar)による解析手法が、古典中国語においては非常に有用であることが明らかとなった。この知見にもとづき、現代的な依存文法記法である Universal Dependencies を用いて、『孟子』『論語』『大學』『中庸』の文法記述をおこなって、デジタル・コーパスの形で WWW で公表した。また、『孟子』『論語』『大學』『中庸』等を機械学習した依存文法解析エンジンを制作した上で、大学入試センター試験『国語』の問題のうち、漢文の本文部分に対して、どの程度の自動解析がおこなえるかを検証した。この成果にもとづき、漢文の自動解析ツール UD-Kanbun と、自動訓読ツール UD-Kundoku を、python モジュールとして公開した。

6. 共同研究会に関連した公表実績

山崎直樹「古典中国語のテキストをいかに切り分けるか」(開篇、Vol.37、2019年4月) 安岡孝一「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」(情報処理学会研究報告、Vol.2019-CH-120、No.1、2019年5月) Koichi Yasuoka 「Universal Dependencies Treebank of the Four Books in Classical Chinese」(DADH2019: 10th International Conference of Digital Archives and Digital Humanities、2019年12月) 安岡孝一「漢日英 Universal Dependencies 平行コーパスとその差異」(人文科学とコンピュータシンポジウム「じんもんこん 2019」、2019年12月)

7. 研究成果公表計画および今後の展開等

『孟子』『論語』『大學』『中庸』の古典中国語デジタル・コーパスは、カレル大学の Universal Dependencies 国際プロジェクトの一部として、公開をおこなった。この成果にもとづき、漢文の自動解析ツール UD-Kanbun と、自動訓読ツール UD-Kundoku を、python モジュールとして公開した。今後は、新規共同研究班「古典中国語のコーパスの研究」において、さらなる応用と発展をはかる予定である。