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
古典中国語のコーパスの研究
Study of Classical Chinese Corpora

2. 研究代表者氏名
安岡孝一
Koichi YASUOKA

3. 研究期間
2020 年 4 月-2023 年 3 月

4. 研究目的
2010 年以来、われわれが構築を続けてきた古典中国語(漢文)コーパスは、MeCab を用いた形態素解析を古典中国語に適用した上で、UDPipe を用いた依存文法解析を適用するものである。これにより、単語の品詞や、単語と単語の係り受け関係を、自動で解析できるようになった。
本共同研究では、古典中国語に対する形態素解析と依存文法解析をさらに押し進め、単語より大きな単位、すなわち句や文について、それらの振る舞いや関係性を解析すべく、さらなる古典中国語解析手法を研究・開発する。
Since 2010, we have developed Classical Chinese Corpora. We first constructed the Corpora using MeCab-Kanbun, a morphological analyzer for Classical Chinese texts. Then we applied UD-Kanbun, a dependency parser based on Universal Dependencies, into the Corpora. Using the Corpora, now we can analyze Classical Chinese texts in word-level: word segmentation (tokenization), Part-Of-Speech tagging, and dependency parsing.
In this study, we will investigate to analyze Classical Chinese texts in phrase- and sentence-levels, enhancing the Classical Chinese Corpora.

5. 研究成果の概要
古典中国語(漢文)Universal Dependencies を検討しつつ、実際にコーパス化をおこなった。具体的には『禮記』『十八史略』『楚辞』『唐詩三百首』『佛説阿彌陀經』『金剛般若波羅蜜經』『維摩詰所説經』『摩訶般若波羅蜜大明呪經』『日本漢詩』『戦国策』『世説新語』を検討対象とし、順次コーパス化をおこなった。また、これらのコーパスのうち、検討が終了したものに関しては、以前に製作した『孟子』『論語』と合わせ、Universal Dependencies 2.11 (2022

年 11 月リリース)として、カレル大学 LINDAT/CLARIN と共同で WWW 公開した。

6. 共同研究会に関連した主な公表実績
なし

7. 研究成果公表計画および今後の展開等
Universal Dependencies 2.12 およびそれ以後のバージョンを通じて、順次、カレル大学と共同で公開する計画である。