IIIFとTEIを用いたオンライン翻刻支援システムの開発

一、研究費種目　若手研究

二、課題番号　　一九K二〇六二六

三、研究期間　　二〇一九年度～二〇二二年度

四、研究経費　　直接経費二百万円、間接経費六十万円

五、研究組織

　　研究代表者　中村　覚

六、研究の概要

2020年度より、人間と機械の両者によるテキストデータの作成支援に取り組んでいる。人間によるテキストデータ作成支援に関する研究の成果として、OSSのソースコードエディタであるVSCodeの拡張機能として翻刻支援システムの開発と公開を行なった。具体的には、TEIによるルビ（ruby）や校異情報（app）、および割注などのマークアップを支援する機能をVSCodeの拡張機能として開発した。またXMLファイルの編集に合わせて、表示結果例をリアルタイムにプレビュー可能な機能も提供する。これらの機能を用いることにより、TEIを用いたテキストデータの作成を支援する。機械によるテキストデータ作成支援に関する研究の成果として、「デジタル源氏物語（AI画像検索版）」というウェブサイトを構築・公開した。本ウェブサイトの特徴として、くずし字OCRと編集距離を利用して、テキストデータが類似する写本・版本の画像を自動的に推薦する機能を提供する。本ウェブサイトを通じて、一部認識誤りを含むOCRテキストデータの利用方法の一例を示すことができた。また、本ウェブサイトの構築に合わせて、CODH（人文学オープンデータ共同利用センター）が公開する「くずし字データセット」を用いて独自のくずし字OCRモデルを開発した。深層学習を用いた文字検出と文字認識、および読み順の自動推定を行う3つのステージから構成される。今後、本モデルの公開を行い、第三者が利用可能な環境を整備する。