市民参加型のオンライン翻刻プロジェクト「みんなで翻刻」の資料に対する計量テキスト分析

Quantitative Content Analysis on Historical Earthquake Documents Transcribed by Online Transcription Project “Minna De Honkoku”

*加納 靖之[1]、橋本 雄太[2]
*Yasuyuki Kano[1], Yuta Hashimoto[2]

1. 京都大学防災研究所附属地震予知研究センター、2. 国立歴史民俗博物館
1. Research Institute for Earthquake Prediction, Disaster Prevention Research Institute, Kyoto University, 2. National Museum of Japanese History

京都大学古地震研究会では，2017年1月に「みんなで翻刻【地震史料】」を公開した（https://honkoku.org/）．「みんなで翻刻」は，Web上で歴史史料を翻刻するためのアプリケーションであり，これを利用した翻刻プロジェクトである．ここで，「みんなで」は，Webでつながる人々（研究者だけでなく一般の方をふくむ）をさしており，「翻刻」は，くずし字等で書かれている史料（古文書等）を，一字ずつ活字（テキスト）に起こしていく作業のことである．「みんなで翻刻」では，正式公開から約1年で，東京大学地震研究所図書室が所蔵する資料のうち「古文書」に分類されデジタル画像化されている421点のうち386点のの翻刻がひととおり完了している．総入力文字数は約356万文字である．古地震（歴史地震）の研究においては，伝来している史料を翻刻し，地震学的な情報（地震発生の日時や場所，規模など）を抽出するための基礎データとする．過去の人々が残した膨大な文字記録のうち，活字（テキスト）になってデータとして活用しやすい状態になっている史料は，割合としてはそれほど大きくはない．「みんなで翻刻」によって大量のテキストデータを生成することができた．このテキストデータに対して，計量テキスト分析を行なった．分析には，計量テキスト分析（テキストマイニング）のために開発されたソフトウェアであるKH Coder（http://khc.sourceforge.net/）を利用した．まず，頻出語の計数を行った．頻出語の上位には「地震」「崩」「水」「人」「山」「火」「町」「寺」「宿」「川」「破損」などが挙がった．これらは，地震とその被害に関する語であり，既刊の地震史料集（たとえば，『大日本地震史料』，『新収日本地震史料』など）による翻刻からの印象とほぼ同じである．この印象を定量的に評価できたことになる．また，共起関係についても分析した．「地震」という語には，方角や地名に関する語だけでなく，被害に関する語が伴なうことが多いことがわかった．それぞれの資料で対象となっている地震によって，被害のあらわれ方が違うことから，資料ごとにより詳細に分析することによって，テキスト分析から地震の様相を抽出できる可能性がある．これらのテキスト分析には適切な辞書が必要である．資料の年代や地震記事であることに対応した辞書を作成する必要がある．既存の辞書を利用しつつ，ここでの分析の結果を再帰的に反映させることによって，よりよい辞書を作成できるだろう．謝辞：「みんなで翻刻【地震史料】」は京都大学古地震研究会によって公開・運営されている．「みんなで翻刻【地震史料】」では，東京大学地震研究所所蔵の資料の画像データを利用した．「みんなで翻刻【地震史料】」の翻刻は，有志の参加者によって実施されている．

キーワード：翻刻、くずし字、オープンサイエンス、計量テキスト分析、テキストマイニング
Keywords: transcription, Japanese Kuzushi-ji, open science, quantitative content analysis, text mining

©2018. Japan Geoscience Union. All Right Reserved.