

## [A2] HTR プログラム Transkribus による日本語キリシタン版『コンテムツス・ムンヂ』のデジタルアーカイブ化

○ノイツラ・ゾフィー<sup>1)2)</sup>, 宮川創<sup>3)</sup>

1)日本学術振興会外国人特別研究員, 三重大学人文学部文化学科, 〒514-8507 三重県津市栗真町屋町 1577

2)ルール大学ボーフム, ドイツ, 44801 ボーフム, ユニヴェルジテーツ通り 150

3) 国立国語研究所, 東京都立川市緑町 10-2

E-mail: so-miyagawa@ninjal.ac.jp

ORCID: 0000-0002-4584-6507, 0000-0002-2950-7193

## Application of Transkribus for a Digitization of the Japanese Christian *Contemptus Mundi*

NEUTZLER Sophie<sup>1)2)</sup>, MIYAGAWA So<sup>3)</sup>

1)JSPS International Research Fellow (Graduate School of Science, Mie University 1577 Kurimamachiyacho, Tsu, Mie, 514-8507 Japan)

2)Faculty of East Asian Studies, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum, Germany

3)National Institute for Japanese Language and Linguistics (NINJAL), National Institutes for the Humanities (NIHU), 10-2 Midori-cho, Tachikawa, Tokyo, Japan

### 【発表概要】

2017年にドイツのヘルツォーク・アウグスト図書館（HAB）で発見された新出キリシタン資料ローマ字本日本語訳『コンテムツス・ムンヂ』の翻刻プロジェクトで用いられた技術と公開方法について論じる。本論文著者両名による日独共同研究プロジェクトにおいて、ルール大学ボーフムのオースタキャンプ・スエン教授の指導のもと、本文献のデジタル化を目標としたデジタル技術活用に関する議論と実践がなされた。ここでは、機械学習による自動翻刻ソフトウェア Transkribus を用いて、その HTR（Handwritten Text Recognition）モデルに学習させ、自動および手動修正で文献のレイアウト・補助記号などを忠実に再現したデジタル翻刻が行われた。本プロジェクトは、このデジタル翻刻に基づいて、本キリシタン版をデジタルアーカイブ化し、可能な研究対象として公開し、国際的な研究に活用できるようにするモデルを提示する。

### 1. はじめに

16～17世紀のキリシタン版の現存する原本は、日本国内でのキリシタン弾圧により非常に少なく、書誌学的に貴重なものである。新たに発見されるケースは近年にも希に見られるが、発見地がドイツというのは珍しい。そのため、キリシタン資料が初めてドイツのヘルツォーク・アウグスト図書館（HAB）で発見されたことは、人々の関心を集めた。

そのキリシタン版は、2017年に Triplett 氏[1]によって発見されたキリシタン版は日本語訳『コンテムツス・ムンヂ』（*Contemptus Mundi jenbu*）であり、1596年にイエズス会による天草にて印刷されたローマ字本である

（図 1）。それまで現存する日本語訳『コンテムツス・ムンヂ』の原本は2冊しか確認されていなかったため、今回発見されたのは3冊目である。判型はオクタボであり、頁数は

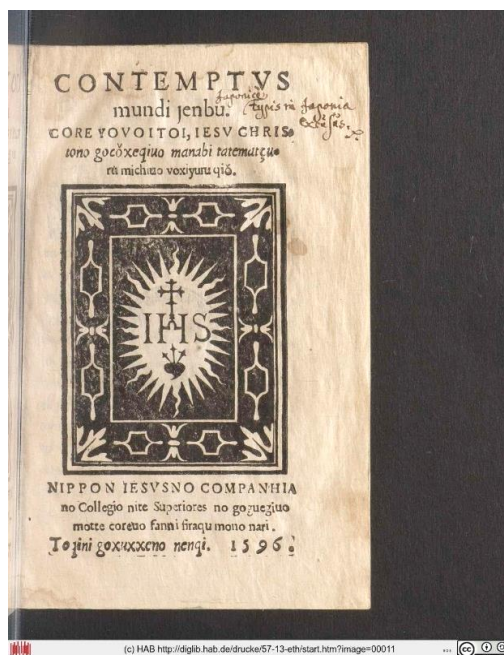


図 1. HAB の『コンテムツス・ムンヂ』の標題紙 [2]

450 頁を越えており、当時の日本語文献としては、本書のテキスト量は非常に多いといえる。さらに、標題紙には歴史的な手書きノートが確認でき、日本でこの本が製作された経緯が明確に示されている。

現在、『コンテムツス・ムンヂ』は、漢字かな混じり表記への字訳[3]が既に刊行されているが、ローマ字本通りの忠実な翻刻 (diplomatic transcription) は公開されていない。

本共同研究は、初めてローマ字本を忠実に翻刻し、作成したデータを公開し、デジタルアーカイブ可の『コンテムツス・ムンヂ』を様々な分野の研究へのアクセスを可能にすることを旨とする。キリシタン版のローマ字本の一つの特徴は借用語 (外来語) の表記である。本研究による翻刻の公開によって、このキリシタン版の特殊な表記のデジタル分析も可能になる。

## 2. Transkribus の作業サイクル

日独共同による本研究は、手書き文字認識 (Handwritten Text Recognition; HTR) プログラムである Transkribus[4],[5]を適用して翻刻を行った。このプログラムは、LSTM (長期短期記憶) モデルをベースにした人工ニューラルネットワークモデルを使用している。

### 2.1 自動レイアウト解析とその修正

最初にアップロードした古文書・文書の画像ファイルに、Transkribus の自動レイアウト解析を適用する。自動レイアウト解析はテキスト領域 (TR: Text Region)、ベースライン (BL: Base Line)、ポリゴンなどを認識する機能を含む (図 2、緑の矢印)。

ただし、本資料は古文書のため、経年による汚れや染みによって文字認識の際に様々



図 3. Transkribus での見出し文字 (左: V のイニシアル) と装飾小模様 (右) の自動レイアウト問題[7]

な問題が生じる。また、見出し文字 (initial) や装飾小模様 (cuts, vignettes) などがプログラムによって数々の文字やベースラインとして誤って認識されることが多い (図 3)。

その上、本文上部の空白には章の表題とページ番号が印刷される「欄外表題」という部分があり、紙の裏から透けて見える文字もよく観察される。従って、自動レイアウト解析の修正は必要である。ベースラインのレイア

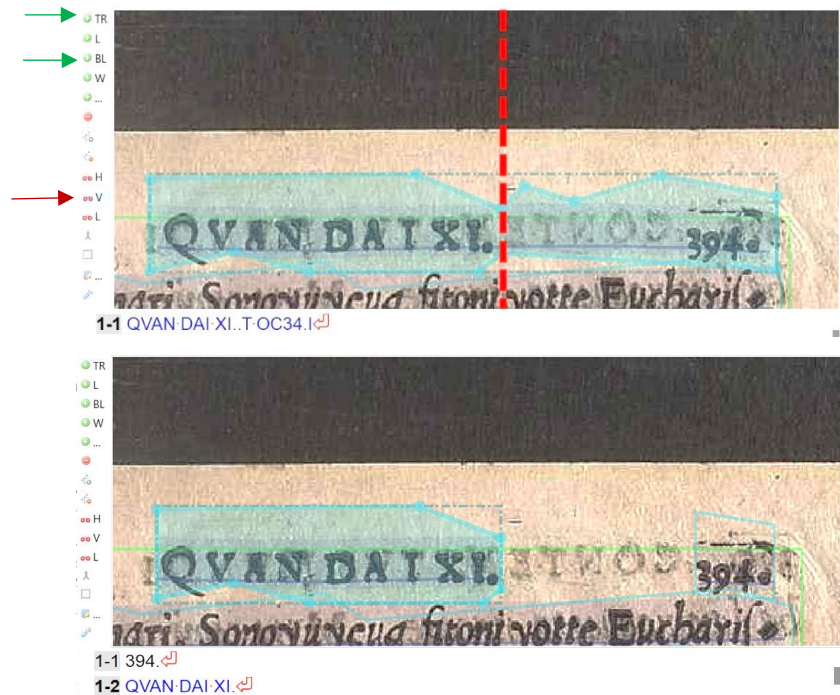


図 2. Transkribus の自動レイアウト解析: 修正前と後[6]

ウトを 2 分割するツールを使って、縦に (V: Vertical) 分割する機能を実行する (図 2、赤い矢印)。その後、裏写りしている文字を、Transkribus が認識する範囲を規定するベースラインから削除することができる。

## 2.2 グラントゥールス (ground truth)

次に、Transkribus が認識したリージョンおよびベースラインにテキストデータを入力する。人手で書き写したデータはグラントゥールス(ground truth)と呼ぶ。そのテキストデータを基に HTR モデルを学習される。

## 2.3 HTR モデル

書き写したグラントゥールスのデータを訓練データにして、学習させ、HTR モデルを作る。学習済モデルは、この文書のデータのみで学習したため、特定のタイプセットに視覚的に適応し、テキストに出現する特定の語彙に内容的に適応していることが一つの特徴である。

より多くの教師データを学習させるほど HTR モデルは明確になる。学習させる場合は、既存のモデルをベースモデルとして、新しいデータで再学習させることができる。モデル

の精度を表す指標は CER (Character Error Rate) と呼ばれ、文字エラー率を意味する。本研究では、Transkribus の作業サイクルを繰り返し適用することで、文字エラー率 1%以下の HTR モデルを実現した。この HTR モデル『R\_Japanese\_print\_1596(ver5)』は、11472 行、58246 語の教師データを学習し、UNICODE 文字セット (Character Set) を 90 以上用いており、CER は 0.98% / 2.09% (Train Set / Validation Set)である (図 4)。

その結果、HTR 技術の使用は、テキスト量の多い本資料を手作業で書き写すのに比べ、作業工程の効率性を非常に上げることができた。

## 3. デジタル翻刻における言語学的発見

デジタル翻刻では、ローマ字本の特殊な外来語はイタリック体でマークアップされ、原文のままに翻刻されている。例えば、原文の省略形「De<sup>o</sup>」(=デウス〔神〕)、「Aīa」(=アニマ〔魂〕)、「Orō」(=オラシヨ〔祈り])などはそのまま明確に書き写してある[9]。また、本キリシタン版では、エスの文字に対して、「s」と「l」(長い s)の両方が使用されていることが明らかになった。例えば、「Christo」と「Chrifto」や、「Eucharistia」と「Euchariftia」も見られる[10]。その上、珍しい「β」が確認され、「Miffa」の他に「Miβa」、「Confiffão」の他に「Confißam」も使われていることが分かった[11]。本研究で作成したデジタルテキストのデータは、他のデジタルソフトウェアを通して計量言語学・計量文献学的な統計分析 (statistical analysis) が可能になる。

## 4. デジタル翻刻の公開

本共同研究は、貴重資料の保存とそのアクセシビリティをオープンデータ基本指針に基づき、データを公開し、様々な研究者に国際的に活用できるようにすることを目標としている。そのため、本デジタル翻刻の成果である TEI XML データは、既に英国図書館蔵・天草版『平家物語』『伊曾保物語』『金句集』を公開している国立国語研究所のデジタルア

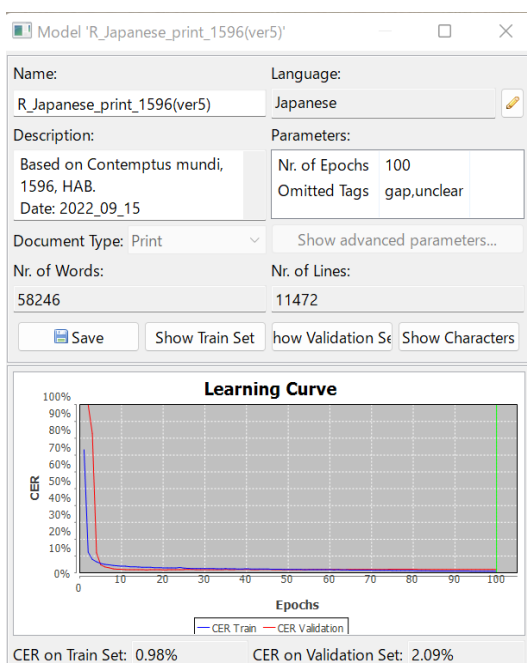



図 4. Transkribus HTR モデル『R\_Japanese\_print\_1596(ver5)』[8]



一カイクに追加することを検討している。それらのキリシタン版も同じようにローマ字本であり、イエズス会によって印刷された17世紀の古文書である。国立国語研究所のデジタルアーカイブに追加する『コンテムツス・ムンヂ』の翻刻もオープンアクセスで利用でき、FAIR (Findable, Accessible, Interoperable, Reusable)のスタンダードを満たすことを目指す。ユーザーインターフェース (user interface) に『コンテムツス・ムンヂ』の画像・複写とテキストデータを同時に提示することを計画している。

## 5. おわりに

以上、本共同研究の成果、Transkribus を用いたデジタル翻刻のテキストデータ、新しいHTRモデルの作成、最初の検証結果と、テキストデータの公開計画について述べた。本文献に使用されているローマ字は、ポルトガル語式ローマ字であるため、日本語音韻史研究において重要な資料であると思われる。このように東西交流史を象徴する本文献のデジタルアーカイブを国際社会の文字文化遺産に加えることが本研究の目的である。

## 参考文献

- [1] Triplett, Katja (2018). The Japanese Jesuit Contemptus Mundi (1596) of the Bibliotheca Augusta: A Brief Remark on a New Discovery. *Journal of Jesuit Studies* 5.1. 2018, p. 123-127.
- [2] Herzog August Bibliothek Wolfenbüttel <http://diglib.hab.de/drucke/57-13-eth/start.htm?image=00011> 

- [3] 小島幸枝. コンテムツス・ムンヂの研究. 資料篇. 研究篇. 東京: 武蔵野書院. 2009.
- [4] Jaillant, Lise: *Archives, Access and Artificial Intelligence. Working with Born-Digital and Digitized Archival Collections*. Bielefeld: transcript Verlag. 2022, p. 185-186.
- [5] Muehlberger, Guenter, et al. Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study. *Journal of Documentation* 75.5. 2019, 954-976.
- [6] Transkribus-1.20.1 のスクリーンショット (参照 2022-09-06).  
資料の画像は Herzog August Bibliothek Wolfenbüttel <http://diglib.hab.de/drucke/57-13-eth/start.htm?image=00407> 
- [7] Transkribus-1.20.1 のスクリーンショット (参照 2022-09-12).  
資料の画像は Herzog August Bibliothek Wolfenbüttel <http://diglib.hab.de/drucke/57-13-eth/start.htm?image=00014>  
<http://diglib.hab.de/drucke/57-13-eth/start.htm?image=00015> 
- [8] Transkribus-1.20.1 のスクリーンショット (参照 2022-09-08).
- [9] 『コンテムツス・ムンヂ』 *Contemptus Mundi jenu*. Japanese College of the Society of Jesus. 1596, p. 366, 381, 388, 391.
- [10] 『コンテムツス・ムンヂ』 *Contemptus Mundi jenu*. Japanese College of the Society of Jesus. 1596, p. 44, 70, 369, 367.
- [11] 『コンテムツス・ムンヂ』 *Contemptus Mundi jenu*. Japanese College of the Society of Jesus. 1596, p. 360, 369, 381, 383.



この記事の著作権は著者に属します。この記事は Creative Commons 4.0 に基づきライセンスされます (<http://creativecommons.org/licenses/by/4.0/>)。出典を表示することを主な条件とし、複製、改変はもちろん、営利目的での二次利用も許可されています。