様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　８　月　２１　日現在

機関番号：　３２６５７
研究種目：　研究活動スタート支援
研究期間：　2015～2016
課題番号：　１５Ｈ０６８３３

研究課題名（和文）訓点資料電子化のためのデジタル記述法と電子化プロセスの確立

研究課題名（英文）Modeling of Digital Description for Computerization of KUNTEN and Establishment of a Computerization Process

研究代表者

堤　智昭（TSUTSUMI, Tomoaki）

東京電機大学・情報環境学部・助教

研究者番号：８０７５９０３５

交付決定額（研究期間全体）：（直接経費）　　2,000,000円

研究成果の概要（和文）：東アジア諸国の漢文で書かれた古典書籍には、記載される漢文を訓読するための訓点と呼ばれる注釈が書き込まれている。訓点は判別が難しい複雑な書き入れが多く、コンピュータを用いた調査・解析はあまり行われていない。そこで本研究では、訓点情報を損なうことなくコンピュータに電子テキストとして取り込むための、電子的な記述手法と電子化プロセスの確立を目標に研究を行った。主に次の3つの研究成果をあげた。①訓点の一つであるヲコト点の構造化記述方式の提案。②電子化ツールの開発。③電子化したヲコト点情報を用いた基礎計量を行いヲコト点の性質を明らかにした。

研究成果の概要（英文）：The numerous classical material written by the Chinese classics is left for East Asian nations. The guiding marks for rendering Chinese into Japanese to read the Chinese classics in their Japanese pronunciation and a called notation are written in the material. These marks are called WOKOTOTEN. It's difficult to distinguish the meaning of the WOKOTOTEN. Therefore, an investigation using a computer and an analysis aren't performed so much.
In this research, we studied targeting the establishment of the electronic description step and computerization process to take it in a computer as e-text, without damaging information of guiding marks for rendering Chinese into Japanese. And we got 3 of next study results mainly. Ⅰ: Suggestion of a structured description system of WOKOTOTEN. Ⅱ: Development of a computerization application. Ⅲ: The nature of the WOKOTOTEN was made clear by a basic quantitative.

研究分野：　コンピュータネットワーク

キーワード：　日本語史　訓点資料　ヲコト点　デジタルアーカイブ

１．研究開始当初の背景

近年、デジタルアーカイブの普及により、近代の活字書籍や古典籍写本・版本の電子化が進められている。活字書籍の電子化にはデジタル画像として取り込む一次資料化と、電子テキストとして取り込む二次資料化がある。一次資料は貴重な原本資料の保存や、拡大縮小等の画像処理により肉眼では確認しづらい資料状態の把握ができる。二次資料は、コンピュータを用いた自然言語処理やデータ解析を行う場合には必須である。一次、二次どちらの資料もインターネットを通じて広く公開され古典文化に触れる機会を広げるとともに、研究環境の向上にも寄与している。

一次資料の代表例としては、「漢字字体規範データベース」(http://joao-roiz.jp/HNG/)、「拓本文字データベース」(京都大学人文科学研究所、http://coe21.zinbun.kyoto-u.ac.jp/djvuchar)、「電子くずし字字典データベース」（東京大学史料編纂所、http://www.hi.u-tokyo.ac.jp/ships/）など東洋学・考古学・歴史学分野の成果がある。研究代表者と共同研究者が制作した、日本語史研究資料（国立国語研究所所蔵）や米国議会図書館蔵『源氏物語』画像 (桐壺・須磨・柏木)も一次資料のデータベースである。

文字の史的研究においてこれらの画像データベースを利用する場合、画像による用例提示だけでなく、二次資料として利用するための、用例字形そのもののデジタル記述法が求められる。近代語のように現代語と語形差が少ない場合、既存の文字コード（JIS 漢字や Unicode）に包摂し電子化するといった手法がとられている。そのために国内 JISX0208 規格や JISX0213 規格では、包摂のため「漢字の字体の包摂規準」が定められている。

研究代表者と共同研究者はこれまでに、活字資料のコーパス化を目的とし、近代書籍を中心に書籍活字の 2 次資料としての電子化を行ってきた。研究成果として JIS 規格が定める包摂規準の有効性と妥当性の検証や、電子化のための文字処理工程を確立した。その中でも研究代表者は特に、電子化を行うための処理工程の確立と、それを効率的に実行するための処理ツールの作成を担当した。

これらの成果を受け、本研究では近代書籍に次いで、古典に分類される書籍の二次資料としての電子化を試みる。平安時代以降の資料には、漢字のみで記され、後にその読み方を表記するための訓点が付与された漢文訓点資料が多く存在し、研究対象となっている。そこで、訓点の情報を保持したまま二次資料化を行うためのデジタル記述法を提案する。

現在、日本における漢文訓点資料は図 1 に示すように、一次資料化ではカメラ等により撮影し画像ファイルとして電子化する手法が取られている。この方法では、視覚的に訓点自体の情報を残すことができるが、統計的分析は難しい。二次資料化では、原文のみまたは訓読した文としてテキストデータ化されている。この方法は、先に述べた JIS 漢字と包摂規準など既存のコンピュータシステムを用いて実現できる。しかし訓点の情報は記録できないため、訓点自体を分析する資料としては利用できない。また、漢文訓読に関する専門的な知識がなければデータ化が難しい。書き下しをした文では、解読において他の解釈を行うことが難しいといった問題もある。

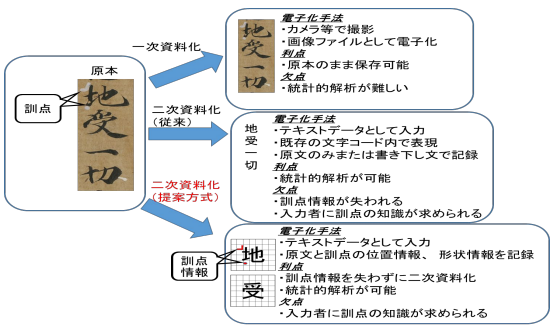

図 1：従来の電子化手法と提案手法

２．研究の目的

本研究の目的は、古典籍資料の中でも文章中に訓点が記されている訓点資料を対象とし、訓点情報を損なうことなくコンピュータに電子テキストとして取り込むための、電子的な記述手法と電子化プロセスを確立することである。本研究によって訓点情報を残したまま電子テキスト化することで、訓点自体を分析する資料として利用することが可能となる。それにより、従来のデジタル画像として電子化した資料や原文のみ、または書き下し文の形式で電子化したテキストデータでは難しかった分析を行うことが可能となる。また、開発した電子化システムや、開発時に電子化したテキストデータを、文字研究者のみに限らず、漢文訓読を学ぶ学生や教育者など様々な人に利用しやすい形式で広く情報公開を行う。また、電子化したデータを用いてコンピュータを用いた計量を行い、訓点の性質を定量的に明らかにする。情報工学の分野から言語学、文献学の分野の研究推進に貢献することができる研究である。

３．研究の方法

（1）ヲコト点電子化のための構造化記述手法の確立と有効性検証

① 文字研究の専門家を交えて、研究活動に有用な構造化記述方式を検討
② 検討結果を受けて、訓点資料を電子化するためのツールを改良（プロトタイプの記述方式と、入力支援ツールは試作済み）
③ 訓点資料の電子化実験を行う
④ 電子化したデータを文字研究者に利用してもらい、フィードバックを得る

（2）電子化システム及び電子化データの公開

① 電子化データを広く研究者に利用してもらうためホームページを通じて公開
② 電子化データを用いて訓点データベース作成の検討を行う。

４．研究成果

研究目的である、訓点情報を損なうことなく訓点資料をコンピュータに電子テキストとして取り込むための、電子的な記述手法を提案し、実際に電子化データを作成するための支援アプリケーションを実装・公開した。また支援アプリケーションを用いてヲコト点図の電子化を行い、電子化プロセスを確立した。また、実際に主要なヲコト点 26 種（以下、主要ヲコト点 26 種）を電子化することで、提案する座標系に全てのヲコト点を配置可能であることを確認した。

研究を遂行する過程で得られた文字研究者からのフィードバックにもとづき、ヲコト点の性質を定量的に明らかにするため、電子化したヲコト点図の情報を用いヲコト点の基礎計量を行った。基礎計量はヲコト点を構成する要素である「位置」「形状」「読み」の 3 要素について行い、以下の性質を明らかにした。

本研究では、ヲコト点を構成する 3 要素のうち 1 つでも異なる場合、それぞれ別のヲコト点であると定義している。主要ヲコト点 26 種中に記載されているヲコト点では,「位置」「形状」「読み」の 3 要素のいずれかが異なるヲコト点は 2,943 個存在した。最も多くヲコト点が記載されていたヲコト点図は宝幢院点であり 263 個であった。最もヲコト点の記載が少なかったヲコト点図は智証大師点であり、32 個であった

（1）ヲコト点の読みの計量と性質

「ヲ」「コト」といったヲコト点の「読み」は、主要ヲコト点 26 種中には、594 種類存在した。594 種の読みの中で主要ヲコト点 26 種に最も多く出現したものは「ス」「ナル」「ナリ」「タリ」であり、出現回数はそれぞれ 37 回であった。次に多く出現したものは「ヨリ」「シ」であり 35 回であった。ここで、登場

回数がヲコト点図の数 26 種よりも多いのは、一つのヲコト点図の中で同一の読みを表すが、形状が異なる点が存在するためである。例えば、仁都波迦点には「ス」と読むヲコト点が「・」と「―」の形状で 2 つ存在する。

（2）ヲコト点の位置の計量と性質

ヲコト点は漢字の四隅に最も多く付与されその中でも右上が最も多く 337 個のヲコト点が存在する。四隅に次いで漢字の上辺下辺左辺右辺のそれぞれ真ん中及び漢字の中心に付与される傾向がある。漢字の外部では、左側よりも右側に多く付与され、上側と下側では下側に多く付与される。

（3）ヲコト点の形状の計量と性質

代表的なヲコト点の「形状」は星点「・」だが、それ以外にも様々な形状が存在する。主要ヲコト点 26 種中には 67 種類の形状が存在した。最も多く登場した形状は「・」であり、その回数は 337 回であった。2 番目に多い形状は「└」で 250 回であった。形状「・」は「└」よりも 87 回多く登場しており、ヲコト点図においては形状「・」が最もよく使われることが定量的に示された。また、登場回数が 200 回を超える形状は「・」「└」「｜」「―」「＼」「┐」「／」の 7 種類であった。登場回数の多いヲコト点は、一筆書きで付与できる形状が多く、簡単に書ける形状ほど、多くのヲコト点に使われる傾向がある。

以上、提案する方式により主要ヲコト点 26 種が電子化可能であったこと、電子化したデータを用いてヲコト点の性質を明らかにするための分析が行えたことから、訓点のデータベース化は可能であり、研究推進のための有用性も認められると考えられる。本研究の研究成果は、人文科学とコンピュータシンポジウム 2015，2016 で学会発表を行った。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　　件）

〔学会発表〕（計　２件）

堤智昭，田島考治，高田智和,小助川貞次「コンピュータを用いた主要ヲコト点の関係性の解析」，人文科学とコンピュータシンポジウム「じんもんこん 2016」，じんもんこん 2016 論文集, Vol.2016, pp.139-146(2016)，2016 年 12 月 10 日，国立国語研究所（東京都、立川市）.

堤智昭，田島考治，高田智和「点図情報入力支援ツールによるヲコト点図の電子化」，人文科学とコンピュータシンポジウム「じんもんこん 2015」，じんもんこん 2015 論文集，Vol.2015，pp.185-190(2015)，2015 年 12 月 19 日，同志社大学京田辺校（京都府京田辺市）.

〔図書〕（計　　件）

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕

ホームページ等

ヲコト点図作成ツールダウンロードページ

http://pba.tsu-lab.sie.dendai.ac.jp/lab/?page_id=67

６．研究組織

(1)研究代表者

堤智昭　（TSUTSUMI Tomoaki）　東京電機大学 情報環境学部・助教

研究者番号：80759035

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：