Vol.2019-CH-120 No.3
2019/5/11

移点ツールの仮名点・語順点への拡張

田島　孝治[†1]　堤　智昭[†2]　高田　智和[†3]　小助川　貞次[†4]

概要：これまでに我々は、訓点資料に付与されたヲコト点に注目したデータ記述方式を検討し、移点を支援するツールを開発してきた。また、ツールを用いて作成したデータを用いた書き下し文の機械的な生成を試みてきた。ヲコト点だけでは語順の入れ替えに対応することが難しく、語順点であるレ点、一二点の電子化は急務であった。また、仮名点を電子化しておくことで、ヲコト点だけでは区別できなかった、読み仮名と助詞・助動詞の識別も可能であると考えられる。そこで、今回これまでの入力ツールを改良し、語順点と仮名点の入力を可能としたのでこれを報告する。またこれに合わせてデータ構造も多少変更し、具体的な資料として、尚書の語順点、仮名点を入力したので、この際の注意事項と、集計結果も合わせて報告する。

キーワード：訓点資料，訓点の電子化，移点，電子化ツール

The Improve of the Tool for Transferring Gloss - Applicate to the inversion gloss and the phonogram gloss -

KOJI TAJIMA[†1]　TOMOAKI TSUTSUMI[†2]　TOMOKAZU TAKADA[†3]
TEIJI KOSUKEGAWA[†4]

Abstract: This paper describes the newly developed software to digitize the glosses in the classical Chinese texts. This software aimed at digitization of (1) main texts and inline notes and (2) wokototen marks at first. The improved software supports to digitize (3) inversion gloss and (4) phonogram gloss. This improvement is an urgent need to create the vernacular reading text from the digitized data. "Shangshu" (old type print version) digitized with developed software for an example as before.

Keywords: Classical Chinese texts, Vernacular reading, Transferring gloss, Digitization tool

1. はじめに

漢文は東アジアの共通書記言語として広範囲で利用されてきた。中国国内で書かれた文献は、日本だけでなく韓国、ベトナムなどに伝えられ、前近代において歴史、政治、経済、宗教などあらゆる分野で利用されてきた。このため、歴史学・文学・仏教学・文化学など広く史的研究を行う上での根幹資料となっている。

漢文訓読は、この漢文を自国語で理解するための技術である。訓点により、漢文の解釈が記入されるため、読者は内容を正確に理解することが容易になる。訓点の種類には、句読点だけでなく、朱や墨で、本文に対して線を引く、文字の周囲に点を打つなどの方法があり、国名や人名などの名詞の種類を表したり、助詞や助動詞、読み仮名や送り仮名を示したりするものがある。これらの解釈方法は主要点図資料（主要ヲコト点図）としてまとめられてきた[1][2]。この結果、資料に付与された加点内容を解釈する方法は確立されている。一方で、ヲコト点などの解釈方法を細かく分析することで、文字、音韻、語彙、語法の変遷を解明することもできると期待されている。

しかしながら、訓点資料研究は、訓点資料の現代語訳を作る、訓点と教学流派の関係を明らかにする研究が多く行われている一方で、定量的な分析が行われておらず、日本語史の研究に活用できているとは言い難い。これは基礎計量を行うための環境が無いことが原因である。訓点資料の加点情報を再構成し、的確に表現して電子化することで、語種の比率や、品詞の比率、加点密度などを正確に計量することが可能になる。これにより、文献どうしの定量的な比較も可能になる。

我々はこれまでに、ヲコト点に対する座標表現を検討してきた。日本古来のヲコト点図には、点の位置と点の解釈がまとめられているが、数値的には扱ってこなかった。そこで、韓国の口訣資料研究[3]を参考に、日本のヲコト点を表す手法を確立し、主要ヲコト点図の電子化を行った[4][5]。さらに、このデータ構造を用いて国語研蔵「尚書（古活字版）」に付与されたヲコト点をすべて電子化する試みを行ってきた[6][7]。この結果、ヲコト点の電子化は完了し、これを用いた書き下し文の自動生成を試みることが出来た[8]。しかし、ヲコト点のみの電子化では語順が考慮でないため、デザインを釈文風にすることで、日本語としては不自然な順に並んでいても、読みやすくする必要があった。

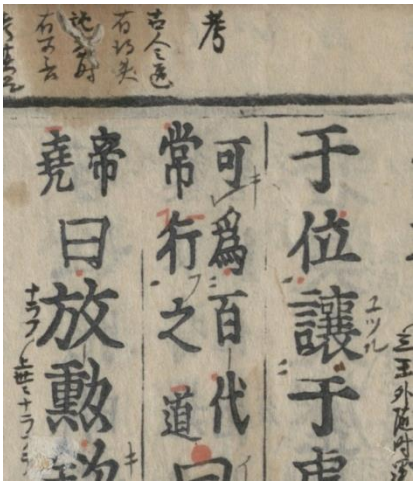
†1 岐阜工業高等専門学校
National Institute of Technology, Gifu College
†2 筑波大学
University of Tsukuba
†3 国立国語研究所
National Institute for Japanese Language and Linguistics
†3 富山大学
University of Toyama

2. 訓点の定量的分析を目的とした電子化

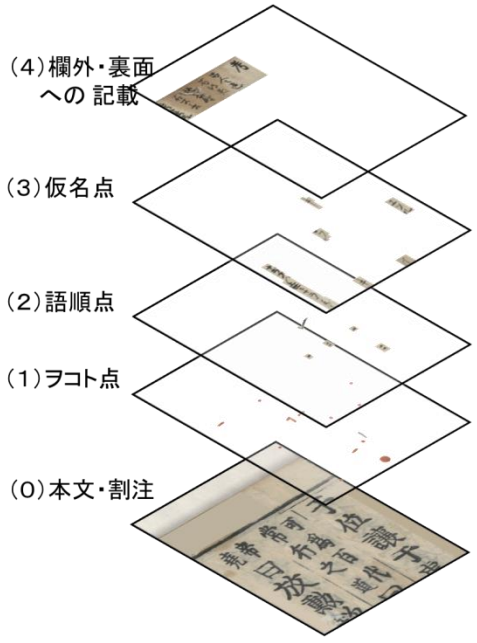
2.1 対象とする加点情報

今回の移点ツールは、訓点資料に付与されている(1)ヲコト点、(2)語順点、(3)仮名点を対象とする。図 1 に、原資料と層ごとに分離した情報を示す。(1)ヲコト点は朱または墨により文字周辺に記号として現れる。音合、訓合だけは、文字と文字にまたがる記号であり、複数の文字をまとめる役割を持つ。(2)語順点は特定の文字の左下に明記され、レ点、一点、二点などがある。(3)仮名点は文字の読み仮名や送り仮名などを表しており、文字の左右に現れる。

訓点資料には、欄外や裏面へ本文の意味を補足するために書き込みが行われている部分もあるが、今回の電子化ではこの部分は対象としない。



(a) 原資料



(b) 要素別に分解したもの

図 1 訓点資料に付与された加点情報

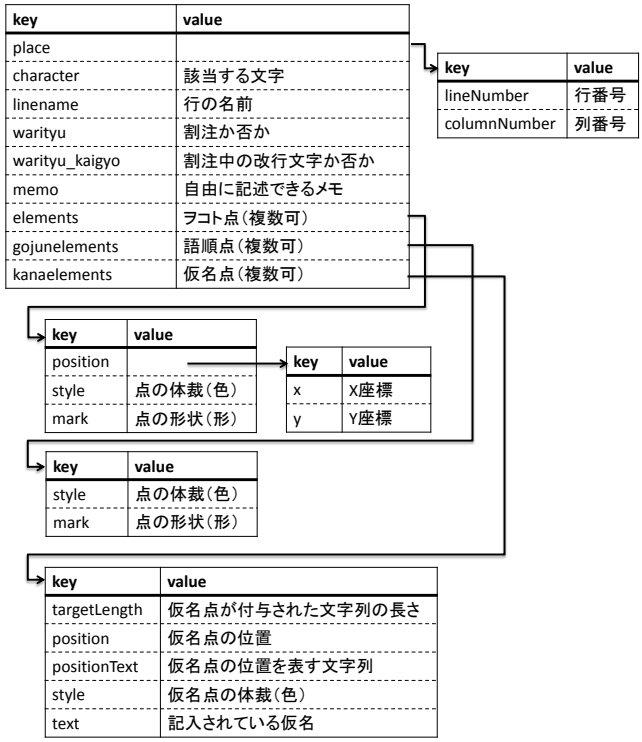
Figure 1 The glosses on the classical Chinese texts.

2.2 加点を表すデータ構造

電子化するヲコト点は RFC8259 に準拠した軽量化オブジェクトである JSON 形式で記述する．これまでに実装してきたソフトウェアで扱ってきたファイルと互換性を持たせるために、ファイル形式を踏襲した。JSON 形式での記述は、多くのプログラミング環境で利用が容易なだけでなく､キーを追加することで新たなデータを自由に記述でき、要素ごとの異なりにも容易に対応できるため、今回の改良においても有用であった。

電子化データの構造を図 2 に示す。データは key-value 型のテーブルであり、文字の場所を示す place のように、value をテーブルとして入れ子にすることもできる。今回、ヲコト点、語順点、仮名点の value はテーブルの配列とし、複数の点を記述できるようにした。

前回のデータ構造からの変更点は(1)語順点の追加、(2)仮名点の追加、(3)割注を表すフラグの追加、(4)メモの追加である。(1)語順点には、記号の色を示す style，記号の形状を示す mark をヲコト点と同様に記述する。ただし、位置を表すキーは入れないことにした。この理由は一、二点であれば文字の左下、レ点であれば文字の下など登場位置が固定されていることが多いことに加え、文字に対応付けられていることが分かれば、書き下し文の生成や統計処理においては十分であり、入力時に位置を記入する必要を無くして効率性を高めたためである。(2)仮名点は、特定の 1 文字だけでなく、音合、訓合で結びつけられた複数の文字に対応付けられることがある。複数の文字に渡り仮名点が付



key	value
place	
character	該当する文字
linename	行の名前
warityu	割注か否か
warityu_kaigyo	割注中の改行文字か否か
memo	自由に記述できるメモ
elements	ヲコト点（複数可）
gojunelements	語順点（複数可）
kanaelements	仮名点（複数可）

key	value
lineNumber	行番号
columnNumber	列番号

key	value
position	
style	点の体裁（色）
mark	点の形状（形）

key	value
x	X座標
y	Y座標

key	value
style	点の体裁（色）
mark	点の形状（形）

key	value
targetLength	仮名点が付与された文字列の長さ
position	仮名点の位置
positionText	仮名点の位置を表す文字列
style	仮名点の体裁（色）
text	記入されている仮名

図 2 JSON 訓点を表す JSON によるデータ構造

Figure 2 The JSON format for glosses.

けられている場合には、仮名点を先頭の文字に記述し、targetLength に対応している文字数を入れることにした。一文字のみに仮名点が付けられている場合はこの値は 1 とする。また、仮名点の位置はヲコト点と違い文字列の長さによって影響を受けるため、座標表現は使わずに「右」「左」「右の右」「左の左」など自由に記述できるようにした。しかし、自由記述だけでは集計が難しく、入力も手間であるため、頻出する上記の 4 種類に関しては 0〜3 の番号を与え、position の中に数値として記録する。style は訓点記号の色を示す記述と同じとし、text は仮名点の記述を現行の片仮名で記述する。(3)割注を表すフラグについては、仮名点の追加に伴い追加した要素である。本文が割注であるか否かによって、仮名点の位置は影響を受けると考えられる。これまでは、ヲコト点の座標のみに注目し、白文データとは独立して処理していたため、割注か否かを考慮してこなかった。今回の改良で白文データを読み込む際に、割注を表すマークを処理するようにしたため、白文のデータを使わずとも JSON データのみで割注かの区別ができるようになった。この結果 JSON データ内に白文のデータから抽出された情報はすべて記述できるようになり、ファイルの一本化が図れた。これまでは、本文の間違いを発見したときに、JSON データと白文データを両方修正し、整合性を取っていたため、この一本化によりデータのバージョン管理も行いやすくなった。(4)メモについては、入力中に気が付いたことを自由に記録していける機能である。ヲコト点の入力作業を行った際に、判断に窮する場合などにメモを残したいという要望があったため追加した。

3. 訓点情報電子化のための移点ツール

3.1 ツールの仕様

移点用のツールは Java 言語を利用して作成した。Java の Runtime が用意できる環境であれば、Windows、Mac OS、Linux（Ubuntu により動作確認済み）での作業が可能である。Java のバージョンは 7 以降を想定している。また、Unicode で記述された本文を確実に表示できるようにするために、IPA ゴシック体および IPA 明朝体をツールは自動的に読み込むように改良した。このためツールの実行時には、同一フォルダ内にフォントを ttf ファイルとして保存しておく必要がある。また、JSON ファイルを扱うためにライブラリである Jackson を利用している。

3.2 ツールの基本画面

移点用のツールのメイン画面を図 3 に示す。従来までのツールでは本文の割注を区別していなかったが、今回の改良でこれに対応したため、メイン画面においても小さい文字で表示し、違いを表現できるようにした。割注内の改行に関しては、実装上の理由から斜線で示し、本文と同一には表示できなかった。しかし、文字数を数える上で、この表示方法は都合がよく、入力作業においても大きな不具合はなかった。

複数の文字にまたがる仮名点はメイン画面から入力する必要がある。CTRL キーを押しながら本文の文字をクリックすると、複数の文字の選択が行える。該当する範囲を選んだ後に、CTRL キーを離してから再度選択中の文字をクリックすると、仮名情報を入力するダイアログが表示され、仮名点の情報を入力できる。また、右クリックを使うことで「レ」点→「一」点→「二」点→「レ」点の順番で語順点を付与していくことが出来る。これらの処理を行うと、図 3(B)に示したように、テキストの背景色や語順点の表示などが変化し、データの入力が行われているかを判断できるようになる。

メイン画面で、本文の文字を左クリックすると図 4 に示す訓点情報の入力画面が表示される。この画面は、これまでヲコト点のみを入力してきたものであった。今回は、ヲコト点選択部分の下部に語順点入力用の領域を設けた。選択式であり、チェックを入れると、右側の一覧表にもその結果が反映される。また、仮名点は文字の左右に頻出するので、文字画像の左右に用意した空白部分をクリックすることで入力できるようにし、直感的な作業を可能とした。それ以外の位置に関しては、文字下部のボタンを押し、専用のダイアログを使って追加する必要がある。また、メモに関してはこの画面から登録し、文字対応づけて記録する。

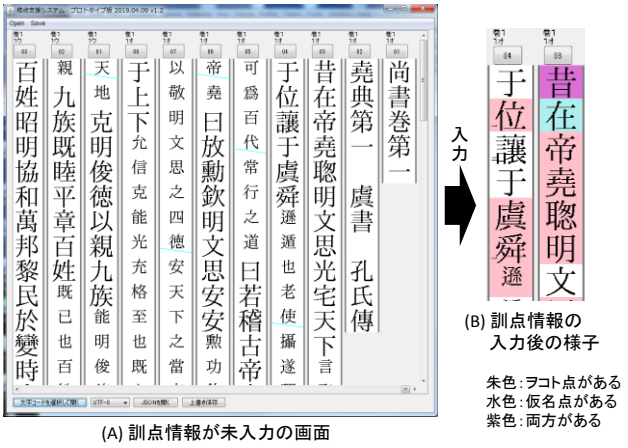


図 3　移点ツールのメイン画面
Figure 3　The main interface of developed tool.

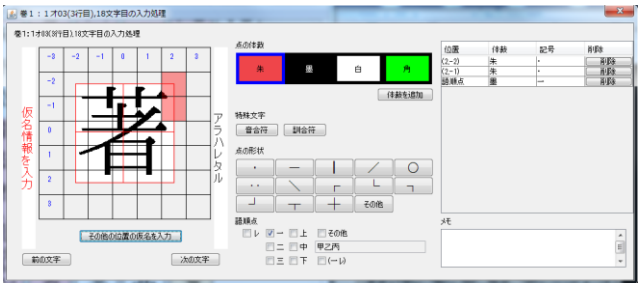


図 4　各文字の用の入力ダイアログ
Figure 4　The gloss input dialog of the developed tool.

4. 対象とする資料と入力結果

4.1 対象とする資料

今回、電子化および分析対象として、国立国語研究所蔵「古文尚書」を用いる[9]。この資料は 1596[慶長元]-1615[慶長 20]年刊のものであり，冬本と奥書が欠けているが、清原宣條の書入本とされている[10]。全編に渡りヲコト点、語順点、仮名点が朱と墨により記入されている。資料は専用の viewer に加え、ページごとの画像データが jpeg ファイルで公開されているため、細部まで拡大して確認することができる。今回の移点は画像データを見ながら行った。冊子本であるため，画像データは 1 丁に対し表裏が存在し，半丁あたり 8 行構成である。

4.2 仮名点の形状

仮名点の電子化を行う前に、仮名の形状がある程度統一されているかを確認した。これまでの電子化におけるヲコト点の入力作業は、人文学分野を専門としない学生に実施してもらい、相互にチェック作業を行った上で、専門家が不明な点やおかしな点を確認するという手順で行ってきた。ヲコト点に関しては、形状と座標を入力するだけなので、資料を見ればそのまま形状や場所を選択できる。一方で、仮名点は記入されている文字が読めない場合には入力が難しい。そこで、巻 1 の冒頭から 1 丁分の仮名点をすべて画像で切り出し、片仮名ごとにまとめて表示した。これを見ることで、同じ片仮名における字形の異なりを調べることができる。また、仮名点のサンプルとしても活用できる。データを見てみると、現行の片仮名とは字体が異なるものもあるが、同じ仮名であれば形状の違いはなく安定しており、専門知識のない学生でも入力が可能であることが分かった。全体の結果は巨大であるため付録 A.1 とし、ここでは現行の片仮名との違いがあり、入力に注意を要したカ、ク、ト、マ、ヲについてのみ表 1 に示す。これらの文字は事前に入力者に示し、注意を促した。

4.3 入力されたデータの総数

語順点、仮名点の基本計量データを表 2 に示す。語順点ではレ点と一、二点はどの巻においても、ほぼ同数であるが、わずかに一、二点の方が多く使われている。また一点と二点の数は一致しておらず、二点のほうがわずかに多い。本文を確認してみると、二点に対応した一点が見当たらない場合があり、今回の作業者の入力ミスによるものだけではないことが分かった。このデータは専門家によるチェックはまだ終わっていないため、今後、再チェックなどで間違いを見つけた場合には微修正する可能性がある。語順点のその他には、四、五点や、甲点、乙点がわずかに見られるものの、20 個は一点とレ点が結合した、一レ点であった。

仮名点に関しては、位置のみを集計した。この結果、文字の右に書かれる仮名点が圧倒的に多く、左側はわずか 14%にとどまった。また、その他は右の右が 63 個、左の左が 6 個であった。

表 1　国語研蔵「尚書（古活字版）」の仮名点（一部）
Table 1　An example of the phonogram gloss in the document.

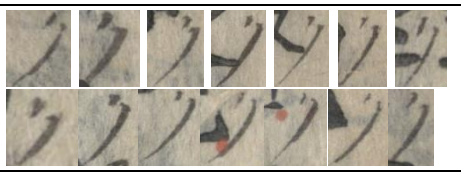
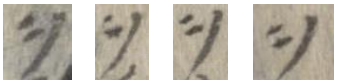
現行仮名(頻度)	仮名画像
カ(18)	
ク(14)	
ト(5)	
マ(1)	
ヲ(4)	

表 2　国語研蔵「尚書（古活字版）」の基本計量データ
Table 2　A key statistics of the inversion gloss and the phonogram gloss in the document.

項目		頻度（割合）	春	夏	冬
文字数		54,504	16,605	16,881	21,018
語順点		16,016(100.0%)	4,317	5,265	6,434
内訳	レ	4,829(30.2%)	1,276	1,662	1,891
	一	5,079(31.7%)	1,427	1,658	1,994
	二	5,121(32.0%)	1,447	1,664	2,010
	三	359(2.2%)	70	103	186
	上	239(1.5%)	41	64	134
	中	99(0.6%)	10	33	56
	下	255(1.6%)	46	72	137
	その他	35(0.2%)	0	9	26
仮名点		16,511(100.0%)	4,811	5,146	6,484
位置	右	14,075(85.2%)	4,175	4,365	5,535
	左	2,367(14.3%)	688	758	921
	その他	69(0.4%)	18	23	28

語順点に関して、より深く考察するために二点が付与された文字に注目して集計してみた。日本語の特性として動詞を最後に読むことを考えると、動詞を表す文字に、二点が付きやすいと推測できる。結果を見てみると、最も多い文字は「有」で頻度は 201(二点中の 3.9%)、次に多いものは「爲」であり頻度は 189(3.7%)あった。以後、頻度が 100 以上あるものを並べると「以」156(3.0%)、「在」123(2.4%)、「不」117(2.3%)、「至」100(2.0%)、「作」100(2.0%)と続く。「有る」、「為す」、「以て」「在り」などはすべて動詞であり予想どおりの結果となった。

5. まとめ

本稿では、移点ツールの語順点、仮名点への対応についてデータ構造、ソフトウェアの改良についてまとめた。また、国語研蔵「尚書（古活字版）」の電子化を行い、得られ結果についてまとめた。語順点に関しては、詳細な分析をこれから行っていくと共に、書き下し文を機械的に生成するツールにおいても対応させていく予定である。また、仮名点に関しては、まだ集計方法を検討中であり、分析の段階に至っていない。仮名点を使ってヲコト点だけでは区別が難しい、助詞、助動詞と読み仮名、送り仮名を区別する方法を確立してきたいと考えている。

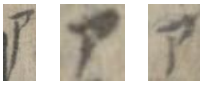
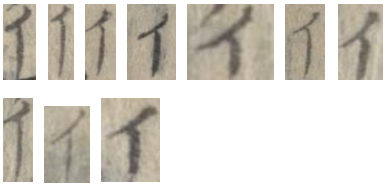
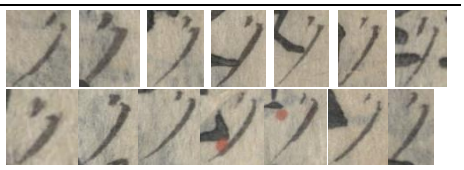
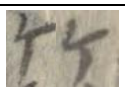
謝辞 本研究はJSPS科研費17K18506の助成を受けたものである．また，本研究は，人間文化研究機構広領域連携基幹研究プロジェクト「異分野融合による総合書物学」の国語研ユニット「表記情報と書誌形態情報を加えた日本語歴史コーパスの精緻化」による成果の一部である．

参考文献

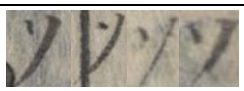
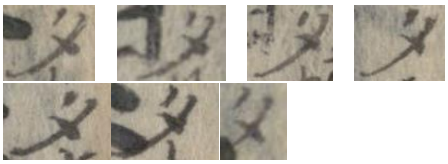
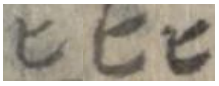
[1] 中田祝夫, 古点本の国語学的研究, 講談社, 1954.
[2] 築島裕, 訓点語彙集成, 汲古書院, 2001.
[3] 朴鎭浩, 文字生活史の観点から見た口訣, 文学第12巻第3号, pp.169-181, 2011.
[4] 堤智昭, 田島孝治, 高田智和, 点図情報入力支援ツールによるヲコト点図の電子化, 人文科学とコンピュータシンポジウム「じんもんこん2015」論文集, pp.185-190, 2015.
[5] 堤智昭, 田島孝治, 小助川貞次, 高田智和, 訓点資料の構造化記述方式と計算機を用いた基礎計量, 情報処理学会論文誌 vol.59 No.2, pp.278-287, 2018
[6] 林昌也, 田島孝治, 堤智昭, 高田智和, 小助川貞次, 訓点資料の加点情報計量のためのデータ構造, 人文科学とコンピュータシンポジウム「じんもんこん2017」論文集, pp.45-52, 2017.
[7] 田島 孝治, 林 昌哉, 訓点資料の移点ツールとデータ校正への活用, 人文科学とコンピュータシンポジウム「じんもんこん2018」論文集,pp. 205-210, 2018.
[8] 林 昌哉, 田島 孝治, 堤 智昭, 小助川 貞次, 電子化した加点情報を用いた書き下し文生成ツールの試作, 人文科学とコンピュータシンポジウム「じんもんこん2018」論文集, pp.21-26, 2018.
[9] http://dglb01.ninjal.ac.jp/ninjaldl/bunken.php?title=syousyo, (参照2019-04-10).
[10] 川瀬一馬, 増補 古活字版之研究, 日本古書籍商協会, 1967.

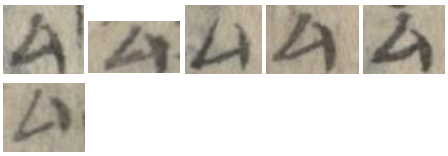
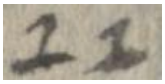
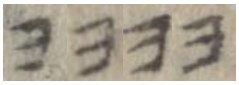
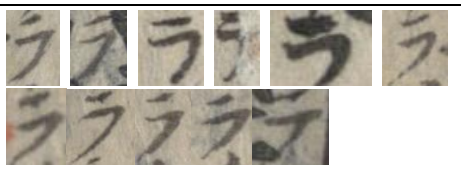
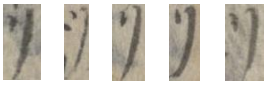
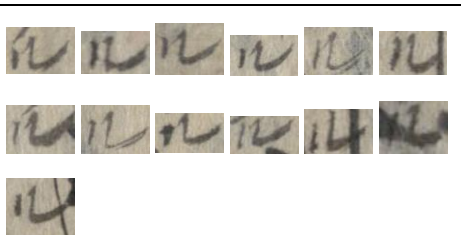
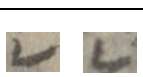
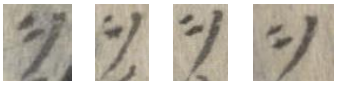
付録

付録 A.1 国語研蔵「尚書（古活字版）」の仮名点一覧

現行仮名(頻度)	仮名画像
ア(3)	
イ(10)	
ウ(8)	
カ(18)	
キ(6)	
ク(14)	
ケ(2)	
コ(5)	
シ(25)	
ス(2)	

付録 A.1「尚書（古活字版）」の仮名点一覧（続き）

現行仮名(頻度)	仮名画像
ソ(4)	
タ(7)	
チ(2)	
ツ(7)	
テ(1)	
ト(5)	
ナ(7)	
ニ(1)	
ヌ(1)	
ノ(2)	
ハ(7)	
ヒ(3)	
フ(8)	
ヘ(7)	
マ(1)	

現行仮名(頻度)	仮名画像
ミ(7)	
ム(6)	
メ(1)	
ヤ(8)	
ユ(2)	
ヨ(4)	
ラ(11)	
リ(5)	
ル(13)	
レ(2)	
ワ(1)	
ヲ(4)	
ン(9)