Vol.2018-CH-118 No.7
2018/8/18

尚書（古活字版）の訓点データの基礎計量

林　昌哉，田島　孝治（岐阜工業高等専門学校）
高田　智和（国立国語研究所）

これまでに，国語研蔵『尚書（古活字版）』 を対象として移点作業を行い，訓点（ヲコト点）の抽出を行ってきた．今回はデータのチェックが完了した巻１から巻３までに対し，定量的な分析を行った．ヲコト点と漢字の関係や，複数のヲコト点が付与された文字に注目し分析を行った．その結果，星点がヲコト点全体の 65%を占めることや，加点が文字の頂点と下部に集中していることが明らかになった．さらに，一文字にヲコト点が複数あるパターンは，全部で 140 パターンであり，「シ」「テ」の組み合わせが最も多かった．

A Basic Statistics of Gloss on Syousyo

Masaya Hayashi, Koji Tajima (NIT, Gifu College)
Tomokazu Takada (National Institute for Japanese Language and Linguistics)

We were transferring gloss on the "Syousyo (old type version)" and digitized its gloss. We analyzed the relation between gloss and letters, and letters with multiple glosses in Volume 1, 2, and 3. As a result, the "star point" occupied 65% of the entire Wakoto points, and the points concentrated on the vertex and the bottom of the character. Furthermore, we sum up some patterns with multiple characters in one letter, the combination of "shi" and "te" was the most frequent pattern, and 140 patterns could be confirmed.

１．まえがき

訓点資料の分析は，記述内容の正確な理解を目指し，加点内容を理解することを中心に行われてきた．これまでに，ヲコト点図や釈文の利用により，資料に付与された加点などによる注釈を解釈する方法は確立されている．この結果，現在では訓点資料の現代語訳を容易に手に入れることができる．

一方で，現代語訳の作成に際しては，訓点資料から，書き下し文を一度作り，解釈する作業が行われることが多い．書き下し文は，主に古典中国語で書かれた原文を日本語で読めるように，語順の調整や助詞・助動詞，読み方などを補って作られた文である．どのように原文を解釈するかは，ヲコト点や注釈により記されている．このため，原文から適切な書き下し文を作成するには，文献に関する知識に加え，ヲコト点と訓読方法に対する深い理解が必要である．

ヲコト点を解釈するためには，ヲコト点図だけでは不十分である．紀伝点や喜多院点など代表的なヲコト点に関しては，ヲコト点図に訓点記号と読みとの対応付けが記されている．それに加え，文献固有の対応付けが存在することも多く，経験や知識などにより意味を解釈しなければならないためである．

本研究では，この書き下し文を機械的に生成するために，資料に付与されたヲコト点などの加点情報を電子化する手法を検討してきた [1)]．電子化を行えば，ヲコト点と文字の関係を統計的に処理する，加点内容どうしを比較するなどの基礎研究が実現できる．また，文献固有の対応付けや，ヲコト点図の記述内容に対して，統計的な裏付けが行えるようになると考えられる．

本研究の最終目標は，電子化した訓点資料から，機械的に書き下し文を生成することである．しかしヲコト点に加え，仮名点や返り点といった，全ての要素の電子化には多大な時間を要する．書き下し文の機械的な生成の第一歩として，まずヲコト点のデータのみを電子化する．対象を限定することで，データの構造的な問題や，記述に必要な要素などを発見することができる．

ヲコト点を電子化し，書き下し文の機械的な生成をする上で，複数の点が付与されている文字の解釈は機械には難しい．同じ二点でも，付与されている文字によって訓読の順番が異なる可能性もある．複数の点に優先順位があり，常にそれが守られているかは，統計的に処理しなければ判断できない．ヲコト点の読み順に関係する点を集計し，細かく比較する必要がある．

本稿では，「尚書(古活字版)」の全体を，ヲコト点に注目した構造化方式で電子データとして記述し，その中で確認作業の完了した巻一から巻三までの統計処理を行った結果について述べる．特に，一文字に対し複数の点が付与されているものに注目した結果について詳しく本稿にまとめる．

2．対象とした資料について

今回は国立国語研究所蔵「尚書(古活字版)」を対象として，電子化と統計処理を行う．尚書は書経とも呼ばれ，政治史・政教を記した中国最古の歴史書で，序文と 58 の通篇で構成される [2]．今回電子データを作成した資料は 1596[慶長元]-1615[慶長 20]年刊のものであり，巻一から巻九までの画像データが公開されている．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．今回の電子化は本文である巻一から巻九の全てを対象とした．全 144 丁，約 2300 行のヲコト点情報を電子化した．

電子化終了後，データ入力者とは異なる人物による確認作業を行っている．現在，この確認作業は巻一から巻三まで完了しており，今回はこの範囲で統計処理を行った結果をまとめる．

3．ヲコト点情報の電子化

3.2 構造化記述を試みる加点情報

著者らは過去に尚書に含まれる加点情報を，それぞれ言語依存性の低いレベル A，高いレベル B に区別したデータ構造を提案した [1]．この A と B には文や段落，句読点など様々な要素が含まれており，同時進行でかつ異なる形式で電子化するのは困難であった．

そこで本研究では，段落要素などの形式が大きく異なるものや，仮名点といった入力時に漢文訓読の知識を必要とするものを除外し，ヲコト点に絞って電子化を行った．ヲコト点の位置は一文字に対して図 1 に示す座標のどこかに属する．仮名点と比べても，位置は明確で多少の文字による歪みはあっても，読みが大きく変わるほど見紛うことは少ない．さらに，移点作業時には尚書の画像ファイルの他に，ヲコト点図も参考にして入力を行う．このヲコト点図には，すべてのヲコト点の形状と座標が記載されており，入力者の間違いを抑制できる．

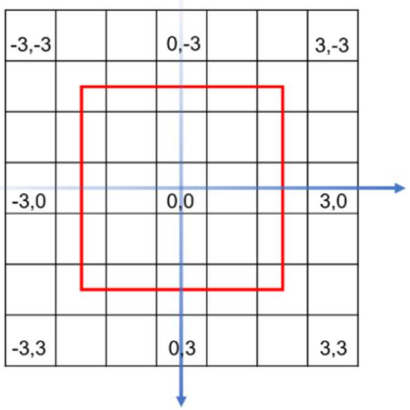


図 1　加点位置の座標系

3.3 データの入力方法

尚書全体の電子化は，3 人の大学と高専生により，巻を分けて入力を行った．漢文訓読の知識に欠ける者も参加したため，このように入力は簡略で単純である必要があった．入力時に不明瞭な点には作業メモとして記録し，確認作業時に確認者が作業メモと照らし合わせて修正を行うことで，より精度の高い電子化を目指した．

データの入力には，図 2 に示す訓点入力ツールを使用した．漢文知識に関係なく，訓点の位置を見たままに入力できるため，手入力に比べ間違いの発生を抑えられる．

3.4 ヲコト点のデータ構造

電子化するヲコト点は RFC7159 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．任意のキーに対して値を関連付ける key-value 型のデータ構造にすることで，必要に応じて要素を追加することが容易である．さらに，統計処理は各種のプログラム言語を利用して行うため，多数のプログラミング言語で標準的に読み込みが可能な点もこの形式を選択した理由である．特に単純な集計処理は Python や JavaScript などのインタプリタで実行可能な言語で処理するため，これらの言語において命令一つでデータ構造を含めて一括して読み込める JSON は利便性が高い．

電子化データの key には文字の場所を示す place，文字の形状を示す character，行の位置を示す linename，付与されている訓点記号を示す elements で構成されている．電子化ツールに尚

図 2　電子化ツール

書の白文のテキストファイルを入力すると，この key に value として各文字の位置や形状が入力された JSON ファイルが生成される．生成時は訓点記号の elements の部分が全て空になっており，入力者は電子化ツールを通じて加点情報を入力していくと図 3 のようになる．elements の value には key が 3 つあり，それぞれ文字に対する訓点記号の位置を示す position，訓点記号の色を示す style，訓点記号の形状を示す mark がある．

過去の構造化で二文字にまたがる訓点記号に対応するために「語要素」を区別して電子化していた．今回の電子化では，この語要素が電子化されていないため，二文字にまたがる訓点記号の訓合符や音合符は，前の文字の下部に付与されているものとして扱った．段落要素も除外したため，段落の頭を示す科段点は文字の上部に付与されているものとして扱った．

3.5 データ入力時の特徴

position は図 1 に示すように，文字に対して中央を(x=0,y=0)とした 7×7 マスの座標で表される．特に，赤い線が文字の輪郭であり，ヲコト点が文字に触れているか否かで意味が大きく異なる．例えば，今回使用したヲコト点図では，文字の右下を表す(2,2)に星点（・）があれば，それは「ハ」と読める．しかし，星点が文字から少しでも離れていれば座標は(3,3)となり，句点の意味を持つ．また，上方向に離れる(0,-3)などの座標にはそもそも点が存在しないため，前の文字に付与された読点と判断できる．尚書では，このような文字に対して訓点記号が触れているか否かは，明確に記載されているものが多く，入力者の間違いは少なかった．

最も入力ミスの多発した訓点記号は，音読みや訓読みを表す墨の「｜」であった．この訓点記号は割注に付与されている場合，左右のどちらのものか判別が難しい．さらに，尚書は行の間に線が引かれており，この線と見間違えることによる入力抜けが多くみられた．このような入力ミスは，確認作業時に作業メモを参考にしながら電子化ツールを通じて修正を行った．

4．統計処理の手法と結果

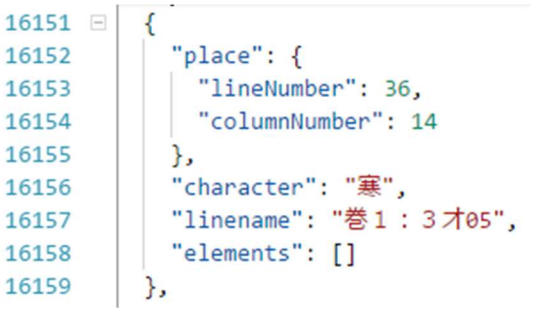
4.1 各要素の個数

電子データの集計用プログラムを Python で実装し，制作したデータの統計処理を行った．電子データは JSON であるため，プログラミング言語が持つ標準の key-value 型のデータ構造を使えば容易に集計できる．

今回集計した範囲の基礎データを表 1 に示す．この範囲の文字は，1620 種類の内，92.5%の文字種が何らかの加点が施されている．加点のある 11889 文字に対し，一文字あたり平均 1.26 個の訓点が付与されていることが分かる．

表 1　基礎データ

	種類数	総数
全ての文字	1,620	16,605
加点のある文字	1,500	11,889
ヲコト点	64	14,977



```
16151  {
16152    "place": {
16153      "lineNumber": 36,
16154      "columnNumber": 14
16155    },
16156    "character": "寒",
16157    "linename": "巻１：３才05",
16158    "elements": []
16159  },
```

(a)　加点情報入力前の JSON

```
16151  {
16152    "place": {
16153      "lineNumber": 36,
16154      "columnNumber": 14
16155    },
16156    "character": "寒",
16157    "linename": "巻１：３才05",
16158    "elements": [
16159      {
16160        "position": {
16161          "x": 2.0,
16162          "y": -2.0
16163        },
16164        "style": "朱",
16165        "mark": "・"
16166      },
16167      {
16168        "position": {
16169          "x": 3.0,
16170          "y": 3.0
16171        },
16172        "style": "朱",
16173        "mark": "・"
16174      }
16175    ]
16176  },
```

(b)　加点情報入力後の JSON

図 3　電子化した JSON データ

4.2 ヲコト点に関する分析結果

ヲコト点の位置別総数をまとめた結果を図 4 に示す．特に文字の頂点部に集中し，文字の下部や中央にも多く見受けられる．特に，全体の 29%を占める，文字の真下，(0,3)は文章中に頻出する読点や，音合が該当するため，重複の多い結果となった．音読みを意味する(3,0)の「｜」は比較的多いが，訓読みを意味する(-3,0)の「｜」は集計内でもかなり少なく，全体の 0.3%しかない．存在する中で一番加点の少ない(-2,-1)は，ヲコト点図の中でも星点の「トキ」の点一種類しかない．文字上部の(0,-3)の点は科段点を示し，巻一から巻三には 9 つの科段点の無い段落があり，それを加えると 115 の段落が存在していることがわかる．

ヲコト点の形状別総数の朱点を表 2 に，墨点を表 3 示す．朱点は，星点が一番多く 9820 個，続いて「―」が 588 個だった．全ヲコト点の中でも圧倒的に星点が多く，ヲコト点の 65%は星点である．墨点は，「｜」が一番多く，3375 個だった．星点にのみ注目して座標別の分布を作ったものを図 5 に示す．大きな傾向は図 4 と変わらないが，文字の上下に現れる点の数が少なくなっている．星点の配置は，文字の四隅と中央，文字下部に偏っていることがより分かりやすい結果となった．また，訓合が 432 なのに対し音合が 2437 と大きな差があることが図 4 と 5 の差分でわかる．

4.3 付与されている文字に関する分析結果

文字の内，加点数が多い順に 10 番目までを表 4 に示す．加点されている 1500 文字の内，最も点の多い文字は「也」の 226 である．「也」は文の終わりに多く，割注の説明で「A は B 也．」という文型が頻出する．そのため「也」には句読点の加点が多く，句点が 52，読点が 174 となった．2 番目に点の多い「言」の位置別点分布を図 6 に示す．(2,2)の点に偏っており，そのほとんどが星点の「ハ」を示している．この「言」に「ハ」が加わった読みは，「言ウ心ハ」となる．割注の文頭に頻出し，意味は「(本文の) この意味は」である．本文の詳しい説明に入る決まり文句であり，「言」の加点の 72%は，この読み方である．

4.4 漢字別の複数の点が付与されている文字

機械的な書き下し文の生成に問題になるのは，複数の点が付与された文字である．ヲコト点には読み順が存在し，適切な読み順で書き下す必要がある．しかし，ヲコト点の読み順に関係しない点も存在する．この点が加点されていても，ヲコト点の読み順の判断から除外できる．ヲコト点の読み順に関係ない点は，墨点，科段点，座標が y=3

表 2　朱点の形状別総数

点形状	総数
・	9,820
―	588
ㄴ	549
ㄱ	117
●	102
／	54
＼	43
｜	43
○	4

表 3　墨点の形状別総数

点形状	総数
｜	3,375
○	202
∞	80

Y＼X	-3	-2	-1	0	1	2	3
-3	0	0	0	**106**	0	0	0
-2	0	**1465**	0	**427**	0	**1626**	0
-1	0	**9**	0	0	0	**213**	0
0	**52**	**129**	0	**966**	0	**57**	**454**
1	0	**73**	0	0	0	**333**	0
2	0	**810**	0	**825**	0	**1262**	0
3	**986**	0	0	**4342**	0	0	**842**

図 4　ヲコト点の座標分布

Y＼X	-3	-2	-1	0	1	2	3
-3	0	0	0	0	0	0	0
-2	0	**1333**	0	**99**	0	**1356**	0
-1	0	**9**	0	0	0	**213**	0
0	0	**117**	0	**966**	0	0	0
1	0	0	0	0	0	**329**	0
2	0	**609**	0	**410**	0	**1078**	0
3	**554**	0	0	**1905**	0	0	**842**

図 5　星点の座標分布

表 4　加点数が多い文字

文字	加点数
也	226
言	196
以	147
天	119
曰	106
五	105
帝	102
水	94
三	94
山	92

の点である．これは，合符，音訓読みを表す点，声点，句読点，返り点兼用の「テ」である．返り点兼用の「テ」の順番は，返り点を読んだ後なので，ヲコト点の読み順には関係がない．

この 3 種のヲコト点の読み順に関係がない点を除き，複数の点が加点されている文字で，その回数が最も多いのは「東」である．「東」の複数加点のある加点位置と形状のパターンの 2 つを図 7 に示す．(a)のパターンは共に星点の位置(-2,0)の「ノ」，位置(0,0)の「カ」があり，ヲコト点の読み順は，「東ノカタ」となる．(b)のパターンは星点の位置(-2,2)の「テ」，形状「 ㄴ」位置(0,2)の「シ」があり，ヲコト点の読み順は仮名点の通り，「東シテ」となる．集計の範囲では(a)は 10 回，(b)は 8 回登場した．しかし，この「東」以外はパターンと文字に相関が無く，特徴的な複数の加点は無い．

4.5 複数の点が付与されているパターン

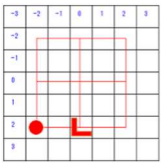
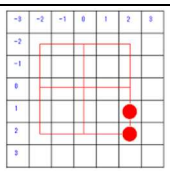
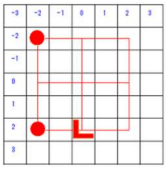
文字に関係なく，ヲコト点の読み順に関係がない点を除いた，複数の点が加点されているパターンのうち 3 番目までを表 5 に示す．最も多かったパターン(A)は(0,2)「シ」と(-2,2)「テ」であり，4.4 の「東」のように～シテと順に読める．2 番目に頻出したパターン(B)の(2,1)「ト」と(2,2)「ハ」の点は，順に～トハとなり，割注などで「Ａトハ Ｂ也」と使われることが多い．3 番目のパターン(C)はそれぞれ(-2,-2)「ニ」，(-2,2)「テ」，(0,2)「シ」であり，順に「～ニシテ」と読める．

このように頻度の多い 3 種のパターンは，付与されている文字に関係なく，ヲコト点の読み順が決定されるものであった．

5 考察とまとめ

集計の結果から，星点の多さや，加点が文字の頂点と下部に集中していることが，定量的に明らかになった．特に，「テ」「ニ」「ヲ」「ハ」といっ

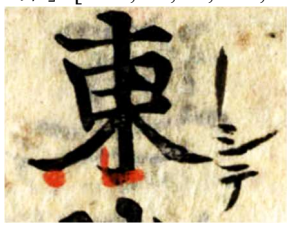
表 5　複数の加点があるパターン

	パターン	総数
(A)		70
(B)		64
(C)		50

Y＼X	-3	-2	-1	0	1	2	3
-3	0	0	0	0	0	0	0
-2	0	**1**	0	0	0	**6**	0
-1	0	**9**	0	0	0	**3**	0
0	**1**	0	0	**2**	0	0	0
1	0	**1**	0	0	0	**1**	0
2	0	**2**	0	**3**	0	**142**	0
3	**12**	0	0	**10**	0	0	**12**

図 6　「言」加点の座標分布

(a)　「東」['-2.0,0.0,・', '0.0,0.0,・']



(b)　「東」['-2.0,2.0,・', '0.0,2.0, ㄴ']

図 7　複数加点のある「東」

た様々な訓点資料に頻出するヲコト点は，加点数が多いことが確認できた．

加点数の多い文字を調べたところ，「也」のような加点の種類が少なく，偏っている文字は少ない．「言」のように，点が１，２か所に集中しながらも，広く加点が分布している文字が大半だった．また，「言ウ心ハ」のような，特有の読み方に変化する点は，集中していると推測できる．

複数の加点のある文字への，読み順の推定を文字に注目すると，「東」のような種類の少ないものは，読み順がわかりやすいことが確認できた．しかし，同じ文字に対して複数点のあるパターンは重複が少ない．文字に焦点を合わせた読み順の推定は，そのパターンの種類の多さや，文字特有の読み順が確認できないため，効率的ではない．

文字に関係なく，複数加点のあるパターンを集計すると，いくつかのパターンに集中していることが確認できた．機械的な書き下し文の生成に向け，ヲコト点の読み順はこのパターンから推測できる．今回確認した 3 パターンを含める，全 140 パターンの読み順を確認することが今後必要になる．

6 あとがき

本稿では国立国語研究所蔵「尚書(古活字版)」を対象として，この加点資料に付与されたヲコト点の電子化と，その電子データの統計処理の結果をまとめた．電子化は全ての巻の入力が終わり，確認作業も巻四から巻九を残すのみである．今後は，確認作業を終わらせるとともに，さらに広範囲での統計処理を行う予定である．

集計したデータから，点の総数や各点の位置分布などの確認を行った．さらに，複数加点のある文字の読み順は，関係のない点を省いたパターンで確認することができた．さらに，点同士の関係性や，読みの優先度を各パターンから導出することができれば，より簡単に書き下し文の機械的な生成ができると考えられる．今後は，集計データの読み順を参考にし，完成した尚書の電子データから，機械的に書き下し文を生成するプログラムを作成し，訓点資料の研究を進める予定である．

謝辞

本研究は JSPS 科研費 17K1850606 の助成を受けたものである．また，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）訓点資料の加点情報計量のためのデータ構造ー国立国語研究所蔵「尚書（古活字版）」を対象としてー，じんもんこん 2017 論文集，Vol.2017，pp.45-52 (2017.12)
2）赤塚忠(翻訳)：中国古典文学大系(1)書経・易経(抄)，(1972.1)