訓点資料の加点情報に対する階層的なデータ化の試み
—春秋経伝集解を事例として—

田島　孝治（岐阜工業高等専門学校　電気情報工学科）
堤　智昭（東京電機大学）　高田　智和（国立国語研究所）　小助川　貞次（富山大学）

訓点資料の分析は記述内容と加点内容の分析を中心に行われてきており，記述内容の正確な理解や加点内容であるヲコト点などに対する理解は十分な段階に達している．一方で，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研究は進行中で，この実現には，加点内容を適切に記述できる構造化記述手法が必要である．本稿では加点資料を理解するために必要な知識段階に基づいた階層的なデータ記述方法を提案し，鎌倉時代の加点資料である宮内庁書陵部蔵「春秋経伝集解」を対象としてデータ化を行った結果についてまとめる．

A Hierarchically Designed Data Structure for Glossed Material
— Case Study of *Shunjukeidenshikkai* —

Koji Tajima (Dept. of Electrical and Computer Engineering, NIT, Gifu College)
Tomoaki Tsutsumi (Tokyo Denki University)
Tomokazu Takada (National Institute for Japanese Language and Linguistics)
Teiji Kosukegwa (University of Toyama)

This paper describes a hierarchically designed data structure for glossed material. Understanding the contents of the glossed material and the system of gloss is already done by many researchers. In the future, we will start the statistical analysis for gloss or glossed material. By the reason, we proposed a hierarchically designed data structure based on the knowledge necessary to understand the glossed material.　And we performed it by "*Syunjukeidenshikkai*".

1．まえがき

現在の訓点資料の分析は記述内容と加点内容の分析を中心に行われている．現状の訓点資料の分析と今後の展望を図１に示す．書き下し文の作成により記述内容の正確な理解は十分に達成されている．また加点内容であるヲコト点などに対する理解は，現存する文献の移点や過去のヲコト点図の活用により，ほぼ完了したと考えられる．

一方，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研究は進行中で，この実現には，加点内容を適切に記述できる構造化記述手法が必要である．

訓点の定量的分析を目的として，著者らは漢文加点資料の構造化記述手法を検討してきた[1]．この記述手法ではヲコト点を中心に座標系を定め，返点や語などに対する符号を木構造で記述していた．

しかしながら，これまでの手法では複数の文字にまたがって付与された点の扱いが煩雑で，どの要素に点の情報を付与するか一意に定まらないなどの問題があった．この問題の原因は，解釈や用途が多岐に渡る漢文の加点要素に対し，付与された符号のみに着目して階層を作ったことが原因であった．今後の研究で研究者が必要とする加点内容の比較や統計分析など高度な定量的分析のためには，加点者の違いや，点の持つ意味の違いに注目して集約可能なデータ構造が望ましい．現状では，加点者の違いに関しては，厳密に分類

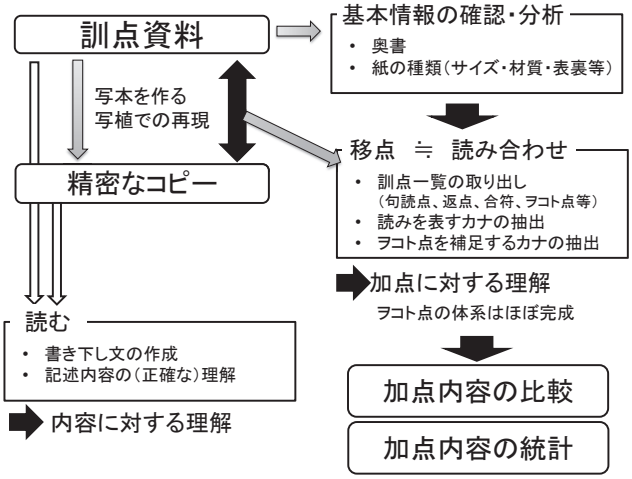

図１　現状の加点資料の分析と今後
Figure 1　The current status and future of analyzing the glossed material

することは困難である．これは，加点に使われた墨や朱などの違いを見分けることが難しいだけでなく，写本の作成時に加点内容も書き写していることが原因である．一方で，点の持つ役割に関しては，内容の理解が完了しているため分類が可能であると考えられる．今回はこの点に注目し，漢文加点資料の特質に即した新たな構造を検討し，記述方法を変更し，実際の文献に適用する．

2．目的

本稿では改良したデータ記述方法を提案し，鎌倉時代の加点資料である宮内庁書陵部蔵「春秋経伝集解」を対象としてデータ化を行う．提案する記述方法は，加点された点の持つ意味や役割などの特性に基づいた階層的なデータ構造とする．これにより，加点資料の特定の点に注目した統計的な分析や，同一文献に異なる加点が行われた場合の文献比較を行えるように設計する．

3．加点内容の階層化

今回提案するデータ構造は，資料を理解するために必要な知識段階に基づいて階層化する．本章では，加点をどのように階層に分けるかを議論する．図２に加点資料における，原文と加点，読み手の関係をまとめる．加点資料は写本という形で継承されてきており，写本の作成においては原本の書き手と，写本の作成者である読み手の文化や知識に違いがあることが多い．書き手と読み手が言語，文化など同じ知識を共有しているのであれば，資料を通じて情報を伝えるためには加点の必要はない．しかし，言語や文化の違いがある場合，読み手が資料を理解するには，辞書や注釈書などを参照する必要がある．この作業結果を資料に重ねて記述したものが加点であり，読み手の理解を補助する役割がある．

また写本は単なる複製品を作成したいわけではなく，本文を理解し利用することを目的としているので本文を書き写した後に，本文を理解するための作業が行われる．これには原本の加点内容をそのまま書き写し，理解のために読み手が注釈

表１　読み手の理解度に基づく加点の階層化

Table 1　The hierarchically design of the glossed material based on the steps of knowledge necessary to understand

大区分	小区分	特徴
本文	文字	オリジナルのテキストで区別できる
	書（書式）	
加点レベルA	文	言語に依らない分割区分である
	語	
	段落	
加点レベルB	助詞・助動詞	仮名書き，ヲコト点で記述される．膠着語固有の文法を表すために使う
	送り仮名	
	読み仮名	

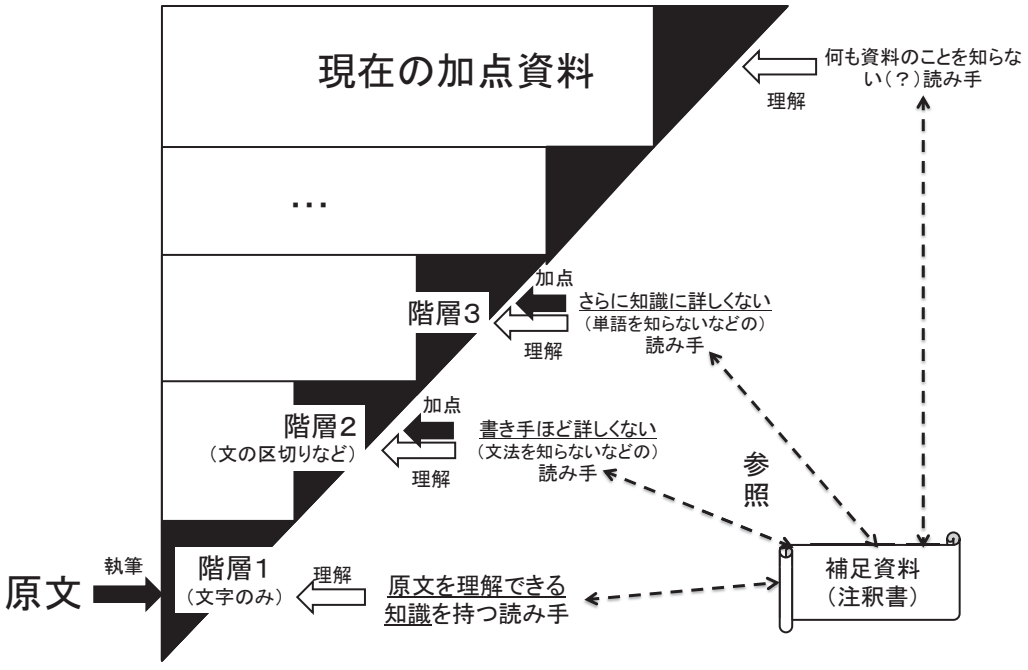

図２　読み手の理解度に基づく加点の階層化

Figure 2　The steps of knowledge necessary to understand the glossed material

書を見ながら，さらに加点を行うことが一般的であろうと推測される．この結果，写本作業を重ねることに加点内容は増えていく．また，付与される点の持つ意味も初めは文の区切りなどの単純な物であるのに対し，だんだんと単語の持つ属性や，その単語の意味などと複雑になっていく．さらに，原本と異なる言語を持つ読み手が内容を解釈する場合には，文法的な異なりを補完するための記号も必要になってくる．返り点や送り仮名などはこれらに該当する．この現象は加点資料にとどまらず，外国人や児童に配慮したルビや注釈入りの日本語のパンフレットなどとも共通する．

今回のデータ化にあたり，表１のように記述内容を大きく３区分に分け，さらにそれぞれの中で細かく分けることにした．大きな区分は，小助川（2015）を参考に，(1)本文，(2)加点レベル A，(3)加点レベル B とした[2]．本文はオリジナルテキストの段階で分割可能な要素であり，文字の文献上の物理的な出現位置，書式を含んでいる．加点レベル A は言語に依らない分割区分で，句読点，句点などで区切られた文，文の集まりである段落，人名など特殊な単語を表す集合である．加点レベル B は膠着語に固有の要素を表す符号である．膠着語の特徴である活用語尾を含む助詞・助動詞，読みを補うための送り仮名，字の読みを記述した読み仮名の三つとした．

このデータ構造は，複数の写本間での点の加点内容の比較や内容の統計を意識したものである．加点レベル A は言語に依らないため，韓国語やベトナム語など他の言語で漢文に加点された資料に関しても共通である可能性が高い．一方で，加点レベル B は言語への依存性が高く，資料間での比較ではなく，加点内容の統計処理などを意識している．この階層化に従って分割した具体例を図３に示す．この例は，春秋教伝集解巻第二の第７紙５行目に記されているものである．加点レベル A の要素は，語を音でつなぐことを表す合符（音合）と読点である．加点レベル B の要素はヲコト点が８つと，仮名で書かれた語の意味（訓）がある．

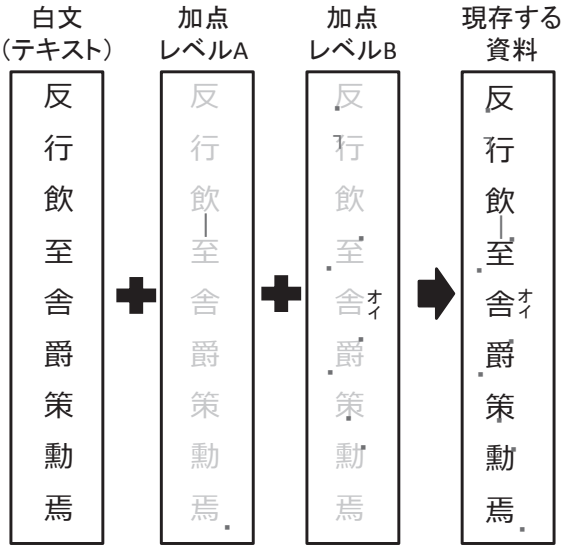

図３　本文の階層化の例
Figure 3　An example of hierarchically designed data structure

表２　原文用の簡易タグ
Table 2　The simple tag for original (no-glossed) material

区分	種別	フォーマット
行	物理行	行番号:本文
符号	割行	〈本文〉
	割行内改行	/
	墨筆補入	+b{本文}
	白筆補入	+w{本文}
	朱筆補入	+r{本文}
	墨筆見消	$b{本文}
	後補部分	#a{本文}
外字	諸橋番号	=[Mxxxxx]
	UCS	=[U+zzzz]

４．各段階における記述方法

4.1　本文の記述方法

本文は UTF-8 でエンコードしたプレーンテキストとして記述する．物理的な行を表すためにコロン（:）を用いて行番号と本文を区切ることにする．また，本文中の割書，補入などは表２に示す<>や+b などテキストで表現可能な簡易タグを使って記述する．これらは一般的なテキストエディタにおいても入力しやすく，環境依存のない文字を利用している．図３に示した白文は，第７紙５行目の途中であり，周辺２行を含めると図４のようなテキストとなる．07-004 は紙面上の位置であり，第７紙４行目を意味している．また，紙面の都合上 07-006 は折り返しているが，データファイル上は処理しやすいように改行文字単位で紙面の１行に対応させることにする．

```
07-004:〈好/也〉冬公至自唐告于廟也凡公行
07-005:告于宗廟反行飲至舎爵策勳焉
07-006:禮也〈爵飲酒器也既飲置爵則書/勳勞於策言速紀有功也〉
```

図４　春秋経伝解巻第二・第７紙４行目～６行目
Figure 4　An example of the simple tagged text

4.2 加点レベル A の記述方法

加点レベル A の要素は 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．加点レベル A に属する要素は文とそれを表す句読点，語，段落であり，各要素が持つパラメータは大きく異なる．そこで，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，要素の異なりに対応することにした．また，今後の統計処理は各種のプログラム言語を利用して行う予定であり，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予定であるため，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．

加点レベル A の key の必須要素は，`type` と `id` である．`type` として指定する値は，`sentence`，`period`，`comma`，`word`，`paragraph` を想定しており，それぞれ文，句点，読点，語，段落を表すために利用する．これ以外の `key` には，他のオブジェクトへの関連付けである `sequence` と `charcount`，付与されている符号の形状を表す `mark`，付与されている符号の位置を表す `x`，`y` がある．位置を表すための座標系は図５のようになっており，点図集[3]の点を全て網羅できるように設計した．この座標系は中心を原点とし，右方向と下方向を正とする７×７の正方形で考える．これは，これまでに作ってきたデータと同じ値であり，他のデータとの整合性が崩れないように考慮している．このデータ構造で図３の部分を具体的に表すと，図６のようになる．

4.3 加点レベル B の記述方法

加点レベル B の要素は XML 形式により記述する．著者らはこれまでに本文を含め１つの XML とする構造を想定しデータを作ってきた[2]．これは XML のもつ木構造のデータ構造が，文，文字，

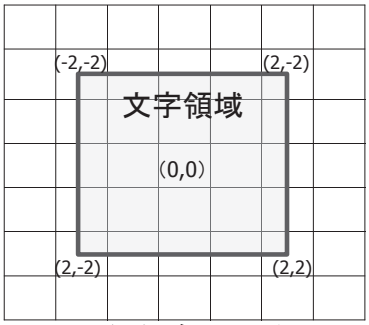

図５　加点データの座標系
Figure 5　The Coordinate system of our data structure

```
{
        "type":"paragraph",
        "id":"r001",
        "sequence":["s001","s002"],
        "mark":"●",
}
```

(a) 段落要素

```
{
        "type":"sentence",
        "id":"s001",
        "charcount":5,
        "sequence ":["07-004-003",
             "07-004-004","07-004-005",
             "07-004-006","07-004-007"]
}
{
        "type":"sentence",
        "id":"s002",
        "charcount":4,
        "sequence ":["07-004-008",
              "07-004-009","07-004-010",
              "07-004-011"]
}
                (以下省略)
```

(b) 文要素

```
{
        "type":"priod",
        "id":"p001",
        "char":"07-004-007",
        "mark":"・",
        "x":3,   "y":3
}
                (以下省略)
```

(c) 句点要素

```
{
        "type":"comma",
        "id":"c001",
        "char":"07-004-012",
        "mark":"・",
        "x":3,   "y":3
}
                (以下省略)
```

(d) 読点要素

```
{
        "type":"word",
        "id":"w001",
        "sequence ":["07-005-003",
                     "07-005-004"],
        "mark":"｜",
        "x":0,   "y":3
}
```

(e) 語要素

図６　加点レベル A の記述方法
Figure 6　An example JSON date for the level A

ヲコト点という構造に向いているからであったが，合符などヲコト点や他の訓点が１文字の子要素にならない場合に対応できていなかった．

今回，階層化することで加点レベルBが表す要素が確定したため，これに合わせ，構造を改めることにする．今回新たに設計したXMLの構造を図7に示す．このデータ構造では1文字または1語を親要素とし，そこに情報を付与していく．図7は親要素を`Letter`として示しているが，これは`Word`でもかまわない．`Letter`の場合は任意の１文字を表し，加点レベル`A`で定義した文字の`ID`を属性に持たせる．`Word`は加点レベル`A`で定義された語であり，こちらも`ID`が付与してあるため，同様に処理できる．また，これらのタグには文字そのものである`Character`を必須の子要素とし，加点内容は`Annotation`として任意の数だけ子要素にする．`Annotation`には`Mark`，`LocationX`，`LocationY`が必須の子要素となっており，形状と位置を表している．また，`Annotation`タグにも複数個の区別のために属性として`ID`を付与できるようにする．

図８に図３冒頭部分のヲコト点をデータ化した結果をまとめる．この形状は一文字に多くの付加情報が付与されてもその文字に関連付けて表記可能な形式であり，XMLの持つ木構造という特性も活かせる．

このデータの処理は，加点レベルAとは異なり，オブジェクトや構造体をしっかりと定義した上で利用することを想定しているため，インタプリタ方ではなく，Javaなどのオブジェクト指向型のプログラミングで分析する予定である．

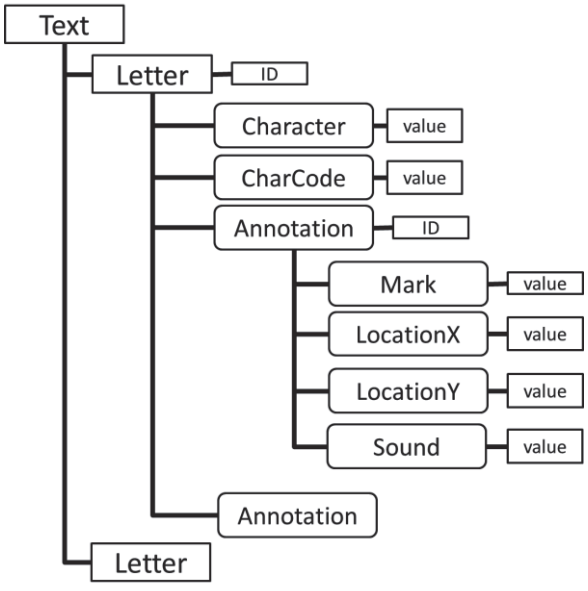

図7　加点レベルBを表すXMLの構造
Figure 7　Data stricture of the XML for Level B

加点レベルAではJSON，加点レベルBではXMLという異なるデータ構造を採用した理由は，これらレベルの異なりに応じてデータの趣旨が異なっているからであり，本文は絶対的な基準であり変化しない基準となるレベルである．また，レベルAは集約の段階であり，複数の文字がどのように集まってグループを構成しているかを表すことを目的としている．一方でレベルBは本文とレベルAで定義したユニット（文字や語など）に対して情報を追加することが目的であり，この段階で新しいユニットが発生することは無い．また，付与される情報に関してはAnnotationタグとして要素の統一を図った．

５．あとがき

本稿では，宮内庁書陵部蔵「春秋経伝集解」を対象とし，この加点資料に付与された情報を階層的にデータ化した結果についてまとめた．現在は，巻２第７紙の一部のみの記述であるため，今後は巻２全体を電子化し，構造上の不具合が無いかを確認する予定である．また，巻２に関しては，藤井斉成会有鄰館の比較的加点が少ない資料と比較することを考えている．少ない加点がレベルAに集中している，さらにはレベルAの加点内容は宮内庁書陵部本と同じであるがレベルBは少

```
<Text>
    <Letter id="07-005-005">
        <Character>反</Character>
        <Annotation id="001">
          <Mark>・</Mark>>
          <LocationX>2</LocationX>
          <LocationY>-2</LocationY>
          <Sound>て</Sound>
        </Annotation>
    </Letter>

    <Letter id="07-005-006">
        <Character>行</Character>
        <Annotation id="002">
          <Mark>┐</Mark>>
          <LocationX>-2</LocationX>
          <LocationY>-2</LocationY>
          <Sound>より</Sound>
        </Annotation>
    </Letter>

      （以後省略）
</Text>
```

図8　加点レベルBの例
Figure 8　An example XML date for the level B

し異なるという結果になるのではないかと想定している．

また，他の巻では巻１０が東洋文庫に所蔵され原色原寸の複製本も出版されている．このため，宮内庁の巻１０についても本構造により各段階のデータを作り，加点内容の異なりを比較できるかを検証する予定である．さらに，京都大学附属図書館は「春秋経伝集解」の電子版も公開しておりこちらとも比較ができると考えている．

謝辞

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構築」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」（代表者：高田智和）の成果の一部である．

参考文献

1）田島孝治，堤智昭，高田智和：ヲコト点電子化のためのデータ構造と入力支援システムの試作，じんもんこん 2012 論文集，Vol.2012，No.7，pp.211-216　(2012.11).
2）小助川貞次：漢文訓読の多面的意義，JSL 漢字学習研究会，Vol.7，pp56-65　(2015.3).
3）築島裕：訓点語彙集成〈第１巻〉，ヲコト點概要，汲古書院　(2007).