資料横断的な漢字音・漢語音データベースと CHISE の統合の試み

守岡 知彦 *

1　はじめに

「資料横断的な漢字音・漢語音データベース」(Database of Historical Sino-Japanese Readings; DHSJR)[4] と CHISE (CHaracter Information Service Environment)[2] の統合の試みについて述べる。

DHSJR は、平安・鎌倉期から現代までの文献資料に現われる漢字音・漢語音を、字音注記（仮名注、声点、反切、類音注、節博士等）に即して検索可能とすることを目指して構築されたデータセットであり、

1. 漢文直読・訓読資料、和化漢文資料、和文資料など文献資料の位相差による漢字音・漢語音の位相的多様性
2. 単漢字単位と漢字連接単位という異なる単位における字音の単位的多様性
3. 中国語原音およびそれを受容した日本語社会における平安・鎌倉期から近世期〜現代までの字音・漢語音韻史の通史的多様性

という３つの多様性を俯瞰し整理することを要件として設計されている。

一方、CHISE (CHaracter Information Service Environment) は著者が中心となって開発している知識処理的アプローチに基づく文字処理環境である。CHISE は文字に関するさまざまな知識を機械可読な知識表現（文字オントロジー）として記述し、その知識に基づいて文字を処理する仕組みを提供している。これはいわば UCS 等の符号化文字集合に対するメタシステムに相当するものといえる。CHISE ではこうした文字知識処理のための枠組だけでなく、その上で記述された大規模な文字知識データセット（CHISE 文字オントロジー）も提供している。

CHISE では、鈴木慎吾氏が開発した「Web 韻圖（廣韻檢索）」[9] へのリンクや、HDIC に収録された宋本玉篇の注文テキスト [3] を自動解析した結果から作成した反切オブジェクト [7] へのリンクを収録している他、網羅的ではないが日本漢字音の音訓データを一部の文字オブジェクト（UCS の抽象文字、互換漢字、もしくは、字体）に付与している。漢和辞典にあるような漢音、呉

* 国文学研究資料館

音、慣用音は古典中国語の中古音と日本漢字音との組織的対応というモデルに基づき分類・再構成したものといえ、必ずしも現実の用例を反映したものとはいえない。また、歴史的仮名遣いに関しても同様である。そこで、DHSJR の持つ情報を統合することにより、HNG が（字書以外のテキストにおける）現実に存在した歴史的字体用例を示すことによって字体標準や字体規範の変遷を示したのと同様に、現実に存在した歴史的な日本 (語) での漢字音やその表記に関する知識を CHISE 文字オントロジーに取り込むことが可能になると期待される。

一方、DHSJR にとっても、CHISE が提供する漢字の部品組合せ方の情報（漢字構造情報）と組み合わせたり、「CHISE IDS 漢字検索」[5] を利用することで従来とは異なる検索や可視化が可能になったり、CHISE が統合している古典中国語コーパスや HDIC 等の情報を組み合わせることにより、DHSJR が収録する漢字連接単位の情報を古典中国語の形態素や語句の情報と比較することが可能になると考えられる。

本稿では XEmacs CHISE ベースの従来型 CHISE / Concord / EsT[6] 実装を用いた DHSJR の統合について概説する。

2　DHSJR の構成

DHSJR は対象とする資料（文献）毎の字音情報データと書誌データから構成され、字音情報データは

1. 資料番号
2. 資料名
3. 資料内漢字番号
4. 資料内漢語番号
5. 単字_見出し
6. 単字_出現形
7. 漢語_見出し
8. 漢語_出現形
9. 漢語_alphabet
10. 語種
11. 漢語内位置
12. 単字長
13. 声点
14. 声点型
15. 仮名注
16. 仮名型
17. 反切
18. 類音

19. 節博士
20. その他
21. 出現位置
22. 備考

という項目からなる。

これらは

- 資料（文献）
- 資料内に出現した漢語（漢字連接単位）
- 資料内に出現した単字

という 3 階層のいずれかを対象としたものといえ、上位のものは下位のものを含み、下位のものは上位のものに含まれるという関係を持つ。

2.1 資料（文献）に関する情報

「資料番号」と「資料名」は対象とする資料（文献）に関する情報であり、それぞれその ID と名前を示す。

2.2 漢語（漢字連接単位）に関する情報

「資料内漢語番号」「漢語_見出し」「漢語_出現形」「漢語_alphabet」「語種」「声点型」「仮名型」「節博士」は漢語（漢字連接単位）に関する情報である。

「資料内漢語番号」はある資料（文献）における漢語（漢字連接単位）の出現順の通し番号である。漢語（漢字連接単位）の ID は資料（文献）の ID である「資料番号」とこの「資料内漢語番号」で構成することができる。

「漢語_見出し」は音注が付された漢字を含む漢語の見出し形を示す。これは異体字の正規化が行われることになっている。

「漢語_出現形」は音注が付された漢字を含む漢語を示す。こちらは異体字の正規化は想定されていないが、データセット全体の統一した方針は存在せず、原則的に入力者の表記方針による。助詞、音合符などが表記される場合もある。

「漢語_alphabet」は欧文による漢語の表記がある場合に用いられる項目である。

「語種」は混種語がある場合に、語種を示す。ただし、入力者によって対応はまちまちである。

「声点型」は漢語に対する声点の組合せを示す。声点がない単字については＊で表す。

「仮名型」は漢語に対する仮名注の組合せを示す。仮名注がない単字については＊で表す。

「節博士」は声明等音楽資料に付される博士譜などを示す。

2.3　単字に関する情報

「資料内漢字番号」「単字_見出し」「単字_出現形」「漢語内位置」「単字長」「声点」「仮名注」「反切」「類音」は単字（漢字の出現；字形[*1]）に関する情報である。

「資料内漢字番号」はある資料（文献）における漢字の出現順の通し番号である。漢字（の出現；字形）の ID は資料（文献）の ID である「資料番号」とこの「資料内漢字番号」で構成することができる。

「単字_見出し」は音注が付された漢字の見出し形を示す。これは異体字の正規化が行われることになっている。

「単字_出現形」は音注が付された漢字を示す。こちらは異体字の正規化は想定されていないが、データセット全体の統一した方針は存在せず、原則的に入力者の表記方針による。助詞、音合符などが表記される場合もある。

「漢語内位置」は単字の漢語内での位置を示す。例えば 1 文字目ならば 1 となる。

「単字長」は単字の拍数を示す。[*2]

「声点」は単字に対する四声（平上去入）、六声（平平軽上去入軽入）及び清濁を示す。

「仮名注」は仮名表記による字音注（仮名反切を含む）を示す。

「反切」は単字に対する反切注を示す。

「類音」は単字に対する類音注を示す。

2.4　その他

「その他」の使い方は資料毎に異なる。例えば、40-045-01「法華経音訓」（東洋文庫）では、「その他」欄に「同」「異本」等の注記と、一括されている掲出字を掲出順にを記している。一方、50-041-01「邦訳日葡辞書」では、複数の表記が推定されている場合に推定表記出現形の情報が記載される。また、60-028-01「平曲譜本」では譜から推定されるアクセント型が記される。このように、「その他」の指示対象や形式は資料毎に異なり統一的に扱うことができない。

「備考」も同様に資料毎に使われ方が異なり、指示対象が単字なのか漢語なのかもまちまちである。

「出現位置」も同様に資料毎に指示対象や形式が異なり統一的な扱いが難しい。

3　設計

DHSJR は単字と漢語（漢字連接単位）という 2 つの字音記述単位に対し、収録対象となる資料（文献）における記述対象となる文字と字音情報に関わるアノテーションの情報を表現しているが、

[*1] そこに書かれた文字という個物を示す。グリフ包摂階層上の単位としての含意はない。

[*2] 現状、この項目はほとんど入力できておらず、今後の検討課題とされている

記述対象となる文字はその出現位置の情報、ないしは、それによって表現されるその場所に書かれた字形（列）に加えて、それを UCS の抽象文字列で表現した出現形とそれを正規化した見出し形が記載されている（表 1)。

記述単位	出現	出現形 / 見出し形
単字	字形オブジェクト	文字オブジェクト (form/lemma)
漢語	漢語出現オブジェクト	漢語見出しオブジェクト (form/lemma)

表 1　記述対象の種別

同様に、字音情報に関わるアノテーションもそれが書かれた場所における出現形（アノテーションの出現、あるいはアノテーションの字形列）とその指示対象となる音の情報の対応関係が考えられるし、同様な出現形を持つ複数のアノテーションの出現の集合を考えることができるだろう（表 2)。

記述単位	出現	注記
単字	字形オブジェクト	音節オブジェクト
漢語	漢語出現オブジェクト	漢語音オブジェクト

表 2　注記情報の種別

漢語（漢字連接単位）に関するアノテーションとしては「声点型」、「仮名型」、「節博士」があり、これらは漢語音オブジェクトで表現するものとする。

また、単字に関するアノテーションとしては「声点」「仮名注」「反切」「類音」がある。

「声点」と「声点型」、「仮名注」と「仮名型」は、それぞれ、単字と漢語に対応し、後者は前者の列であると考えられる。但し、資料によっては「仮名注」は注が書かれた場所や色・種類などを表現するために構造化されており、そうした場合に、「仮名型」は「仮名注」のカナ部分（これを仮名主部と呼ぶことにする）だけを入れていることが多く、単純に「仮名型」を「仮名注」の列と見なすことができない。そのため、「仮名型」を主部、種別部、位置部、備考部で構成される複合オブジェクトとし、「仮名型」は仮名主部の列として扱うことにする。

4　実装

今回は C で実装した従来型 Concord 実装と XEmacs CHISE ベースの従来型 CHISE を用い、この環境において Emacs Lisp ベースの EgT を用いて Web アプリケーション化を行なった。

4.1 単字

4.1.1 character@dhsjr

従来、CHISE では、字形を表現可能な Concord ジャンルとして、character ジャンルにおける字形粒度のオブジェクト、glyph-image ジャンルの字形オブジェクト、image-resource ジャンルのオブジェクト（全文画像中の領域を表現することができる）などを設けていた。

DHSJR における「単字の出現」は意味的には文字の出現に他ならないので字形に相当するものの、包摂粒度上の単位としての含意はないので、character ジャンルで扱うのは適切でないといえる。また、字形画像情報も含意されておらずその用意もないため、image-resource ジャンルでは扱えない（character ジャンルにおける字形オブジェクトも字形粒度のグリフ情報を要求するため不適といえる）。一方、glyph-image ジャンルの字形オブジェクトで表現することは可能といえるが、現状、glyph-image ジャンルのオブジェクトは 1 つ以上の image-resource を持つことを想定しているため、image-resource を持たない glyph-image ジャンルのオブジェクトを設けるとこれを想定していないプログラムに悪影響を与える可能性が考えられる。

そこで、DHSJR における「単字の出現」を表現するために、新たな専用の Concord ジャンルとして character@dhsjr を設けることとした。

character@dhsjr ジャンルに作成する「単字の出現」オブジェクト（以下、単字出現オブジェクトと呼ぶことにする）は次の素性を持つ：

=dhsjr-character 素性　資料番号-資料内漢字番号 で構成される DHSJR データセットに属する単字出現オブジェクトの ID. この素性値 *ID* を使って構成される URL https://dhsjr.kojisho.com/character/*ID* は「横断漢字音簡易検索」の検索結果における単字 ID に対応する詳細ページに対応する。EgT においては、この素性に対応する欄（「= [DHSJR] 単字 ID」）の素性値部分はこの URL で示される詳細ページへのリンクとなる。

=id 素性　dhsjr-資料番号-資料内漢字番号 で構成される単字出現オブジェクトの ID.

=title 素性　単字出現オブジェクトの表題。他のオブジェクトからリンクされる際の人間可読性を考慮して構成される。

character 素性　単字出現オブジェクトに対応する文字オブジェクトのリスト。

->part-of 素性　単字出現オブジェクトが出現した漢語出現オブジェクトを示す。型はオブジェクトのリストであるが、漢語出現オブジェクトは 1 つになるはずである。

->tone-mark-annotation 素性　単字の出現に声点が存在する時、その声点に対応する漢字音節オブジェクトへのリンクを示す（オプショナル）。型は漢字音節オブジェクトのリスト。

->kana-annotation 素性　単字の出現に仮名注（仮名表記による字音注（仮名反切を含む））が存在する時、その仮名注に対応する漢字音節オブジェクトへのリンクを示す（オプショナル）。型は漢字音節オブジェクトのリスト。

->fanqie-annotation 素性　単字の出現に反切注が存在する時、その反切注に対応する漢字音節オブジェクトへのリンクを示す。型は漢字音節オブジェクトのリスト。

->similar-sound-note 素性　単字の出現に類音注が存在する時、その類音注に対応する漢字音節オブジェクトへのリンクを示す。型は漢字音節オブジェクトのリスト。

mora-length 素性　単字の拍数を示す（オプショナル）。型は整数のリスト。

4.1.2　単字に対する声点

DHSJR における単字の出現に対する声点の項目には、

- 四声（平上去入）
- 六声（平平軽上去入軽入）
- 清濁

が記載されるが、実際には「〔墨〕入濁」や「〔朱〕上」や「〔淡朱〕入軽」といった墨の色や、「〔圏〕去」や「〔墨〕去濁（圏点）」のような形状、「〔朱〕平（虫）」や「虫損」や「(破損)」といった虫損に関する情報、あるいは、「〔墨〕去濁（ママ）」のようなコメント情報も記載されている。

また、「去／平」や「〔墨〕上／〔朱〕平／〔朱〕去」のように複数の項目が「／」で区切られて併記されている場合がある。なお、区切り文字として「・」を用いている箇所も存在する。

このように、記載内容や形式は資料毎に異なり統一が取れていない。

ここでは、複数項目の区切り文字として「／」を用い、各項目は

[“〔” *type* “〕”] *mark* [“（” *comment* “）”]

という形式のもののみを扱うことにした。但し、[...] は省略可であることを示す。

ここで、type は「〔墨〕」や「〔圏〕」のように記載された墨の色や形状等の情報である。色と形状といった別種のものが混じっており、また、意味論的には資料毎の用途を考慮すべきであるが、これらは今後の整理に委ねることとし、「〔...〕」内に書かれたものは一律に『声点タイプ情報』として扱うことにした。

また、*mark* は声点（四声（平上去入）、六声（平平軽上去入軽入）及び清濁）といった声点の本体部を示す。

また、*comment* は「(虫)」や「(ママ)」といった「(...)」内に書かれたものを一律にコメントとして扱うものである。意味的に考えれば、「(虫)」は声点の本体部の虫損を示すものといえ、声点の本体部のデータとして記載すべきものと考えられるが、簡単化のため今回はコメントとして扱うことにする。

このように、DHSJR の「声点」欄の情報は、声点の本体データと付加情報付きの複合的声点データが混在しているが、これを整理するために、声点の本体データを示すためのオブジェクト（声点本体オブジェクト）と付加情報付きの複合的声点データを示すためのオブジェクト（声点複

合体オブジェクト）に分けて考え、関係素性 `->head` および逆関係素性 `<-head` を用いて

$$\text{声点本体オブジェクト} \xleftarrow{\text{head}} \text{声点複合体オブジェクト} \tag{1}$$

のようなオブジェクト間の関係で表現することにした。

付加情報のうち、『声点タイプ情報』は声点複合体オブジェクトの `tone-mark-annotation-type` 素性の値として記載する。また、コメントは `*note` 素性の値として記載する。

なお、ある単字出現オブジェクトが `->tone-mark-annotation` 素性を持ち、その素性値に声点本体オブジェクト、もしくは、声点本体オブジェクトを持つ声点複合体オブジェクトが存在する時、単字出現オブジェクトの `character` 素性値で表現される文字オブジェクトの `sound@ja/tone-mark` 素性にその声点本体オブジェクトが追加される。これにより、該当する文字オブジェクトに DHSJR の声点情報が付加されるとともに、EgT の該当ページから DHSJR の声点情報のページへのリンクが構成される。

なお、ここで素性名の言語を表現するドメイン名としては、古典中国語を示す `lzh` や古典日本語を示す `ojp` を使うことも考えられたが、DHSJR における声点欄の内容は資料毎に多岐に渡り、古典中国語の中古音に対応しないものも多々あるため `lzh` の使用は除外した。また、時代も一様ではないことや文法を扱わないことを鑑み、通時的な日本語音を扱うという観点で今回は日本語を示す `ja` を用いることとした。

4.1.3 単字に対する仮名注

DHSJR における単字の出現に対する仮名注の項目には、仮名表記による字音注（仮名反切を含む）が記載されるが、声点の場合と同様に、墨の色や形状、虫損に関する情報、コメント情報等の情報も記載されている。

また、記載形式としては

- 「ア」のように括弧なしで直接カナを記載したもの
  - 「[左] イク」や「[朱] サイ」のように位置情報や墨の色や形状等のメタデータを U+005B「[」と U+005D「]」で囲み、その後にカナ情報を置く形式
  - 「［右］カフ／［下］コウ」のように位置情報を U+FF3B「［」と U+FF3D「］」で囲み、その後にカナ情報を置く形式（複数項目は「／」で区切る）
  - 「〔左〕ア」や「〔左〕〔淡朱〕アウ」のように、位置情報や墨の色や形状等のメタデータを U+3014「〔」と U+3015「〕」で囲み、その後にカナ情報を置く形式
    - 「〔朱〕クン（クシを朱でンに訂正)」や「〔左〕〔朱〕クエン（ママ)」のように、さらに、注記情報を U+FF08「(」と U+FF09 「)」で囲んで置く形式
    - 「〔墨〕（虫）イ」や「〔墨〕（墨抹）ン」のようにカナ情報の前に（あるいはその一部として） U+FF08「(」と U+FF09 「)」で囲まれた虫損情報や編集情報等が現れるケース
  - 「アン〔左〕」や「ア〔左〕〔朱〕」のように、カナ情報の後に、位置情報や墨の色や形状

等のメタデータを U+3014「〔」と U+3015「〕」で囲んで置く形式
  - 「シチ（ママ)」「ケウ（フを訂正)」「ケイ（後筆)」のように、カナ情報の後に、注記情報を U+FF08「(」と U+FF09 「)」で囲んで置く形式
- 「(カ)」や「(アク／ヲ)」のように U+0028「(」と U+0029「)」でカナを囲む形式
  - 「ヽ（コ)」「ゝ（セイ)」のように括弧の前に形状を記載したもの
    - 「ヽ（ス）〔左〕」のようにさらに位置情報を記載したもの
      - 「ヽ（ス）〔左〕〔朱〕」のようにさらに墨の色を記載したもの
  - 「(虫)」
- 「(カゥ)」のように U+FF08「(」と U+FF09「)」でカナを囲む形式
  - 「(仮名なし)」、「(虫)」等
- 「「シウ」(消)」
- 「／＼」でアクセント等を表現したもの?
  - 「／＼（シヤウ)」「／＼（ヤウ)」のように括弧内にカナを記載したもの
  - 「／サム」のように括弧なしでカナを記載したもの
- その他

などがある。

複数項目を並置するための区切り文字としては、声点の場合と同様に、「／」が用いられる場合が多いが、「・」を用いている箇所も存在する。

このように、記載形式や内容は声点の場合と同等かそれ以上に多岐に渡っており統一が取れていない。

ここでは、複数項目の区切り文字として「／」を用い、各項目は ABNF [1] 形式によって式 2 のように定義される形式のもののみを扱うことにした。

$$\text{仮名注} = \text{仮名注}_1 \ / \ \text{仮名注}_2 \tag{2}$$

$$\text{仮名注}_1 = [\ \textit{type-annotation}\ ]\ [\ \textit{location-annotation}\ ]\ \text{カナ}\ [\ \textit{comment-annotation}\ ] \tag{3}$$

$$\text{仮名注}_2 = \text{カナ}\ [\ \textit{location-annotation}\ ]\ [\ \textit{type-annotation}\ ] \tag{4}$$

$$\begin{aligned}
\textit{type-annotation} = (\ &\text{“［”} && \text{; U+005B}\\
&\textit{type}\\
&\text{“］”}\ ) && \text{; U+005D}\\
/\ (\ &\text{“［”} && \text{; U+FF3B}\\
&\textit{type}\\
&\text{“］”}\ ) && \text{; U+FF3D}\\
/\ (\ &\text{“〔”} && \text{; U+3014}\\
&\textit{type}\\
&\text{“〕”}\ ) && \text{; U+3015}
\end{aligned} \tag{5}$$

$$
\begin{aligned}
\textit{location-annotation} = (\ &\text{“[”} && \text{; U+005B}\\
&\textit{location}\\
&\text{“]”}\ ) && \text{; U+005D}\\
/\ (\ &\text{“［”} && \text{; U+FF3B}\\
&\textit{location}\\
&\text{“］”}\ ) && \text{; U+FF3D}\\
/\ (\ &\text{“〔”} && \text{; U+3014}\\
&\textit{location}\\
&\text{“〕”}\ ) && \text{; U+3015}
\end{aligned}
\tag{6}
$$

$$
\textit{type} = \text{“墨”} \ /\ \text{“朱”} \ /\ \text{“濃朱”} \ /\ \text{“淡朱”} \ /\ \text{“スリケシ”} \ /\ \text{“院政”} \ /\ \text{“後筆”} \tag{7}
$$

$$
\begin{aligned}
\textit{location} = \ &\text{“左”} \ /\ \text{“左上”} \ /\ \text{“左下”} \ /\ \text{“右”} \ /\ \text{“右上”} \ /\ \text{“右下”} \ /\ \text{“上”} \ /\ \text{“下”}\\
/\ \ &\text{“踊字”}\\
/\ \ &\text{“朱”}
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
\textit{comment-annotation} = \ &\text{“（”} && \text{; U+FF08}\\
&\textit{comment}\\
&\text{“）”} && \text{; U+FF09}
\end{aligned}
\tag{9}
$$

ここで、式 8 はフォールバックのために追加したものであり、本来、*type* で扱うべきものである。

このように、DHSJR の「仮名注」欄の情報は、声点の場合と同様に、本体データと付加情報付きの複合的仮名注データが混在しているが、これを整理するために、仮名注の本体データを示すためのオブジェクト（仮名注本体オブジェクト）と付加情報付きの複合的仮名注データを示すためのオブジェクト（仮名注複合体オブジェクト）に分けて考え、関係素性 `->head` および逆関係素性 `<-head` を用いて

$$
\text{仮名注本体オブジェクト} \xleftarrow{\text{head}} \text{仮名注複合体オブジェクト} \tag{10}
$$

のようなオブジェクト間の関係で表現することにした。

付加情報は声点複合体オブジェクトの次の素性に記載する：

`kana-annotation-type` 素性　仮名注タイプ情報 (*type*)
`kana-annotation-position` 素性　仮名注位置情報 (*location*)
`*note` 素性　コメント情報

なお、ある単字出現オブジェクトが `->kana-annotation` 素性を持ち、その素性値に仮名注本体オブジェクト、もしくは、仮名注本体オブジェクトを持つ仮名注複合体オブジェクトが存在する時、単字出現オブジェクトの `character` 素性値で表現される文字オブジェクトの `sound@ja/kana` 素性にその仮名注本体オブジェクトが追加される。これにより、該当する文字オブジェクトに DHSJR の仮名注情報が付加されるとともに、EgT の該当ページから DHSJR の仮名注情報のページへのリンクが構成される。

4.1.4　単字に対する反切注

DHSJR における単字の出現に対する反切の項目には、反切注が記載されるが、声点や仮名注の場合と同様に、墨の色や形状、虫損に関する情報、コメント情報等の付加的情報も記載されている。

また、反切注本体は、「苦瓜反」や「強魚切」のように声母字と韻母字による漢字 2 文字の後に「反」または「切」を書くのを基本とするが、最後の「反」「切」は省略されている場合もある。また、U+303B「〻」が書かれているケースがあるがこれは「反」の草書を表現したものであろうか。ただ、これは、本来、重文記号（揺すり点）を示すためのものであるので、別の抽象文字で表現することが望ましいと言える。

いずれにせよ、反切注本体は漢字 2 文字もしくはその後に反切を示すための接尾辞が付いた 3 文字で表現されるはずであるが、DHSJR では「丁 [平] 交 [去]」や「力 [入] 張 [平] 〻」のように声母字や韻母字の後に声点が付けられているケースがある。

複数の反切を並置する場合、「／」や「・」で区切るケース（声点や仮名注の場合と同様、区切り文字は統一されていない）の他に、「羊昌蕪孝二反」のように記載されている場合や「奴但反又他丹反」や「儒隹反又奴禾反又乃四反」のように「又」で繋いでいるケースもある。また、「初佳反／又初宜反」のように「又」の前に「／」を入れているケースもある。

また、「頭莫報反」や「脚女江反」のように位置情報を記載しているケースもあり、さらに「脚説文所祐反」や「頭玉力谷反」や「頭玉篇云…余摂反」のように出典も記載するケースもある。

このように、DHSJR の「反切」欄のデータは、付加的情報に関しては、コメントを除いて、基本的に前置方式で記述されており、仮名注に比べれば比較的整っているといえるが、声点や反切注に比べて反切注本体の構造が複雑であり、また、複数の反切の並置を自然言語（漢語）で記述していて、それによって表現される複数の反切全体に付加情報が付くと考えられるケースが存在する。つまり、

- 声母字・韻母字に付加情報がつく場合
- 声母字＋韻母字（＋接尾字）で構成されるある漢字音節を表現した反切に付加情報がつく場合
- 反切記述全体に付加情報がつく場合

という 3 つのパターンが考えられる訳である。このため、全体の構造としては、声点や反切注の場合に比べて複雑になってしまうといえる。

ここでは、複数項目の区切り文字として「／」と「・」を用いるとともに、「〜二反」「〜反又〜反」「〜反又〜反又〜反」という並置形式を受理し、各項目は ABNF [1] 形式によって式 11 のように定義される形式のもののみを扱うことにした。

$$\text{複合反切注} = [\ \textit{location-annotation}\ ]\ [\ \textit{type-annotation}\ ]\ \text{反切本体部} \tag{11}$$

$$\begin{aligned} \textit{location-annotation} = (\ &\text{“〔”} && ;\ \text{U+3014} \\ &\textit{location} \\ &\text{“〕”}\ ) && ;\ \text{U+3015} \end{aligned} \tag{12}$$

$$\begin{aligned} \textit{type-annotation} = (\ &\text{“〔”} && ;\ \text{U+3014} \\ &\textit{type} \\ &\text{“〕”}\ ) && ;\ \text{U+3015} \end{aligned} \tag{13}$$

$$\begin{aligned} \textit{location} = \ &\text{“左”} \ /\ \text{“右”} \ /\ \text{“欄外”} \ /\ \text{“上欄”} \ /\ \text{“下欄”} \\ /\ &\text{“イ本”} \ /\ \text{“上欄イ本”} \ /\ \text{“下欄イ本”} \end{aligned} \tag{14}$$

$$\textit{type} = \ \text{“濃朱”} \ /\ \text{“淡朱”} \tag{15}$$

$$\text{反切本体部} = \ \text{反切} \ /\ \text{反切}_2 \ /\ \text{反切}_3 \tag{16}$$

$$\text{反切} = \ \text{声母字}\ [\ \text{声母字声点}\ ]\ \text{韻母字}\ [\ \text{韻母字声点}\ ]\ [\ (\ \text{“反”} \ /\ \text{“〻”} \ /\ \text{“切”}\ )\ ] \tag{17}$$

$$\begin{aligned} \text{反切}_2 = (\ &\text{声母字}_1\ [\ \text{声母字声点}_1\ ]\ \ \text{韻母字}_1\ [\ \text{韻母字声点}_1\ ] \\ &\text{声母字}_2\ [\ \text{声母字声点}_2\ ]\ \ \text{韻母字}_2\ [\ \text{韻母字声点}_2\ ]\ \text{“二反”}\ ) \\ /\ (\ &\text{声母字}_1\ [\ \text{声母字声点}_1\ ]\ \ \text{韻母字}_1\ [\ \text{韻母字声点}_1\ ]\ \text{“反又”} \\ &\text{声母字}_2\ [\ \text{声母字声点}_2\ ]\ \ \text{韻母字}_2\ [\ \text{韻母字声点}_2\ ]\ \text{“反”}\ ) \end{aligned} \tag{18}$$

$$\begin{aligned} \text{反切}_3 = \ &\text{声母字}_1\ [\ \text{声母字声点}_1\ ]\ \text{韻母字}_1\ [\ \text{韻母字声点}_1\ ]\ \text{“反又”} \\ &\text{声母字}_2\ [\ \text{声母字声点}_2\ ]\ \text{韻母字}_2\ [\ \text{韻母字声点}_2\ ]\ \text{“反又”} \\ &\text{声母字}_3\ [\ \text{声母字声点}_3\ ]\ \text{韻母字}_3\ [\ \text{韻母字声点}_3\ ]\ \text{“反”} \end{aligned} \tag{19}$$

$$\text{反切内声点} = \text{“[”}\ \text{声点}\ \text{“]”} \tag{20}$$
$$\text{声母字声点} = \text{反切内声点} \tag{21}$$
$$\text{声母字声点}_1 = \text{声母字声点} \tag{22}$$
$$\text{声母字声点}_2 = \text{声母字声点} \tag{23}$$
$$\text{声母字声点}_3 = \text{声母字声点} \tag{24}$$
$$\text{韻母字声点} = \text{反切内声点} \tag{25}$$
$$\text{韻母字声点}_1 = \text{韻母字声点} \tag{26}$$
$$\text{韻母字声点}_2 = \text{韻母字声点} \tag{27}$$
$$\text{韻母字声点}_3 = \text{韻母字声点} \tag{28}$$

このように、DHSJR の「反切」欄の情報は、声点や仮名注の場合と同様に、本体データと付加情報付きの複合的反切データが混在しているが、声点や仮名注の場合と異なり、付加情報は反切本体ないしはその母字と注記全体の 2 階層に付き得るので、これをナイーブに表現すると、トータル

で３階層ないしは４階層となり、階層が深くなってしまう。しかしながら、実際には、声点が併用されるケースにおいて、「又」などを用いた自然言語による並置形式と並置記述全体に対する付加情報を表現している例はないため、実際にはたかだか２階層で済むと考えられる。

そこで、これを整理するために、反切注の本体データを示すためのオブジェクト（反切オブジェクト）と付加情報付きの複合的反切注データを示すためのオブジェクト（反切複合体オブジェクト）に分けて考え、関係素性 `->head` および逆関係素性 `<-head` を用いて

$$\text{反切オブジェクト} \xleftarrow{\text{head}} \text{反切複合体オブジェクト} \tag{29}$$

のようなオブジェクト間の関係で表現することにした。また、反切オブジェクトは HDIC の宋本玉篇データの統合の際に導入したもの [8] を流用した。この反切オブジェクトは声母字オブジェクトと韻母字オブジェクトに対して

$$\text{声母字オブジェクト} \xleftarrow{\text{initial-character}} \text{反切オブジェクト} \xrightarrow{\text{final-character}} \text{韻母字オブジェクト} \tag{30}$$

という関係をもち、声母字オブジェクトと韻母字オブジェクトの双方での集計を可能としている。

付加情報は反切複合体オブジェクトの次の素性に記載する：

`fanqie-annotation-type` **素性**　反切注タイプ情報 (*type*)
`fanqie-annotation-position` **素性**　反切注位置情報 (*location*)
`fanqie-initial-tone` **素性**　声母字の声点
`fanqie-initial-note` **素性**　声母字に対するコメント情報
`fanqie-final-tone` **素性**　韻母字の声点
`fanqie-final-note` **素性**　韻母字に対するコメント情報
`fanqie-suffix` **素性**　反切を示す接尾辞（「反」「切」「ʒ」等）

なお、ある単字出現オブジェクトが `->fanqie-annotation` 素性を持ち、その素性値に反切オブジェクト、もしくは、反切本体オブジェクトを持つ反切複合体オブジェクトが存在する時、単字出現オブジェクトの `character` 素性値で表現される文字オブジェクトの `sound@ja/fanqie` 素性にその反切オブジェクトが追加される。これにより、該当する文字オブジェクトに DHSJR の反切情報が付加されるとともに、EsT の該当ページから DHSJR の反切情報のページへのリンクが構成される。

なお、ここで素性名の言語を表現するドメイン名としては、古典中国語を示す `lzh` や古典日本語を示す `ojp` を使うことも考えられたが、DHSJR における反切欄の内容は資料毎に多岐に渡り、古典中国語の中古音に対応しないものも多々あるため `lzh` の使用は除外した。また、時代も一様ではないことや文法を扱わないことを鑑み、通時的な日本語音を扱うという観点で今回は日本語を示す `ja` を用いることとした。なお、反切オブジェクトに関しては HDIC 宋本玉篇と同じ形式の `hanzi-syllable` ジャンルのオブジェクトで表現するようにしたため、宋本玉篇等の古典中国語における漢字音を示す反切と統合されており、両者を比較することができる。

4.1.5　単字に対する類音注

DHSJR における単字の出現に対する類音の項目には、類音注が記載されるが、声点や仮名注や反切の場合と同様に、墨の色や形状、虫損に関する情報、コメント情報等の付加的情報も記載されている。

類音注本体は「音力」のように「音」の後に音が似た漢字を示すのが典型的な例であるが、DHSJR における類音の項目には「七反」や「七ゝ」、あるいは、「七音」のような形式で書かれているものもある。

複数項目の区切り文字としては

- 「・」を使うケース。例：「音卦・音圭」
- 「／」を使うケース。例：「音接／徐音集」
  - 「／又～」。例：「音洛／又音岳」
- 「、」を使うケース。例：「音接／徐音集」
- 「～又～」。例：「音八又音拜」「駄ゝ又土ゝ」
- 「～二音」。例：「會活二音」

「又」に関しては、「又主音」や「又於ゝ」のように、少なくとも、DHSJR のデータ上、区切り文字ではなく先頭に現れるケースもある。

ここでは、複数項目の区切り文字として「・」「／」「、」を用い、各項目は ABNF [1] 形式によって式 31 のように定義される形式のもののみを扱うことにした。

$$\begin{aligned} \text{類音注} = [\ \textit{location-annotation}\ ]\ [\ \textit{type-annotation}\ ] \\ \text{類音注本体部}\ [\ \textit{comment-annotation}\ ] \end{aligned} \tag{31}$$

$$\begin{aligned} \textit{location-annotation} = (\ &\text{“〔”} && ;\ \text{U+3014} \\ &\textit{location} \\ &\text{“〕”}\ ) && ;\ \text{U+3015} \end{aligned} \tag{32}$$

$$\textit{location} = \text{“左”}\ /\ \text{“右”} \tag{33}$$

$$\begin{aligned} \textit{type-annotation} = (\ &\text{“〔”} && ;\ \text{U+3014} \\ &\textit{type} \\ &\text{“〕”}\ ) && ;\ \text{U+3015} \end{aligned} \tag{34}$$

$$\textit{type} = \text{“濃朱”}\ /\ \text{“淡朱”} \tag{35}$$

$$\text{類音注本体部} = \text{類音注本体}_{1p}\ /\ \text{類音注本体}_{1s}/\ \text{類音注本体}_{2p}\ /\ \text{類音注本体}_{2s} \tag{36}$$

$$\text{類音注本体}_{1p} = \text{類音接頭辞}\ \ \text{類音字} \tag{37}$$

$$\text{類音接頭辞} = [\ (\ \text{“又”} \mid \text{類音典拠指示子}\ )\ ]\ \text{“音”} \tag{38}$$

$$\text{類音注本体}_{1s} = \text{類音字}\ \ \text{類音接尾辞} \tag{39}$$

$$\text{類音接尾辞} = \text{“反”}\ /\ \text{“ゝ”}\ /\ \text{“音”} \tag{40}$$

$$\text{類音注本体}_{2p} = \text{“音”}\ \text{類音字}_1\ \text{“又音”}\ \text{類音字}_2 \tag{41}$$
$$\text{類音注本体}_{2s} = \text{類音字}_1\ \text{類音字}_2\ \ \text{“二音”} \tag{42}$$

このように、DHSJR の「類音」欄の情報は、類音注の本体データと付加情報付きの複合的類音注データが混在しているが、これを整理するために、類音注の本体データを示すためのオブジェクト（類音注本体オブジェクト）と付加情報付きの複合的類音注データを示すためのオブジェクト（類音注複合体オブジェクト）に分けて考え、関係素性 `->head` および逆関係素性 `<-head` を用いて

$$\text{類音注本体オブジェクト} \xleftarrow{\text{head}} \text{類音注複合体オブジェクト} \tag{43}$$

のようなオブジェクト間の関係で表現することにした。

付加情報は類音注複合体オブジェクトの次の素性に記載する：

`similar-sound-note-position` **素性**　類音注位置情報 (*location*)
`similar-sound-note-type` **素性**　類音注タイプ情報 (*type*)
`similar-sound-note-prefix` **素性**　類音接頭辞
`similar-sound-note-suffix` **素性**　類音接尾辞
`*note` **素性**　コメント情報

なお、ある単字出現オブジェクトが `->similar-sound-note` 素性を持ち、その素性値に類音注本体オブジェクト、もしくは、類音注本体オブジェクトを持つ類音注複合体オブジェクトが存在する時、単字出現オブジェクトの `character` 素性値で表現される文字オブジェクトの `sound@ja/similar-sound-note` 素性にその類音注本体オブジェクトが追加される。これにより、該当する文字オブジェクトに DHSJR の類音注の情報が付加されるとともに、EgT の該当ページから DHSJR の類音注の情報のページへのリンクが構成される。

なお、ここで素性名の言語を表現するドメイン名としては、古典中国語を示す `lzh` や古典日本語を示す `ojp` を使うことも考えられたが、DHSJR における類音欄の内容は資料毎に多岐に渡り、古典中国語の中古音に対応しないものも多々あるため `lzh` の使用は除外した。また、時代も一様ではないことや文法を扱わないことを鑑み、通時的な日本語音を扱うという観点で今回は日本語を示す `ja` を用いることとした。

4.2　漢語

4.2.1　word@dhsjr

DHSJR における「漢語の出現」を表現するために、新たな専用の Concord ジャンルとして `word@dhsjr` を設けることとした。

`word@dhsjr` ジャンルに作成する「漢語の出現」オブジェクト（以下、漢語出現オブジェクトと呼ぶことにする）は次の素性を持つ：

`=dhsjr-word` **素性**　資料番号-資料内漢語番号 で構成される DHSJR データセットに属する漢語

出現オブジェクトの ID.

=id 素性　dhsjr-資料番号-資料内漢語番号 で構成される漢語出現オブジェクトの ID.

=title 素性　漢語出現オブジェクトの表題。他のオブジェクトからリンクされる際の人間可読性を考慮して構成される。

sequence 素性　漢語出現オブジェクトを構成する文字（出現）オブジェクトの列（リスト）。DHSJR に単字出現の項目がある場合、単字出現オブジェクト、そうでない場合、CHISE の文字オブジェクトが該当位置の要素となる。

<-part-of 素性　この漢語出現オブジェクトに含まれる単字出現オブジェクトのリスト。

->word 素性　この漢語出現オブジェクトの出現形を表現する古典中国語見出し文字列オブジェクト。

->word@lemma 素性　この漢語出現オブジェクトの見出し形を表現する古典中国語見出し文字列オブジェクト。

->tone-mark-annotation 素性　漢語出現に声点型が存在する時、その声点型に対応する漢語音オブジェクトへのリンクを示す（オプショナル）。型は漢語音オブジェクトのリスト。

->kana-annotation@occurrence 素性　漢語出現に仮名型が存在する時、その仮名型に対応する漢語音オブジェクトへのリンクを示す（オプショナル）。型は漢語音オブジェクトのリスト。

neuma 素性　漢語出現オブジェクトの節博士の項目（オプショナル）。型は文字列。

5　おわりに

「資料横断的な漢字音・漢語音データベース」(DHSJR) の CHISE 文字オントロジーとの統合の試みについて述べた。

DHSJR には多様な情報が収録されていて、現代および歴史的な日本語における漢字音・漢語音の豊富な用例を提供している一方、収録対象となった文献資料や記述法の多様性などのためにその内容は複雑であり、現状、貴方の統一や構造化の点で問題を抱えているといえる。

今回、DHSJR を CHISE 文字オントロジーに統合し、古典中国語コーパスや HDIC 等で導入した既存の形態素や反切等の構造化（オブジェクト化、Linked Data 化）と調和するように、DHSJR における単字と漢語を、特に、声点、仮名注、反切注、類音注に関して、現在の記載内容に基づき、比較的少数の規則でなるべく多くの記述内容をサポート可能な構文規則を定義し、その解析器を実装し、構造化を行なった。

しかしながら、今回は節博士をはじめいくつかの項目について十分な構造化を行うことができなかった。また、DHSJR では資料毎に異なる入力方針がとられている箇所があり、こうした資料依存の記載方針や形式についてサポートすることができなかった。また、サポートできていない記述内容や項目もいくつか残されている。

DHSJR の可用性を高めるためには、資料依存の形式をなくすとともに、資料依存のポリシーを記述するためのメタ形式を適切に定義し機械可読化を行う必要があると考えられる。また、併記形

式や付加項目の順序など資料の性質にあまり関わらないと思われる部分に関しては、今後、形式の統一が進むことが望ましいと考えられる。こうしたことに今回の構文規則や構造化が役立てば幸いである。なお、著者は漢字音史研究者ではないため今回の分析に誤りがあったり構造化が適切でない可能があり、こうした問題に関して識者からのご指摘が頂ければ幸いである。

参考文献

[1] D. Crocker (ed) and P. Overell. *Augmented BNF for Syntax Specifications: ABNF*. The Internet Society, 2008 年 1 月. RFC 5234, STD 68.

[2] Tomohiko Morioka. Multiple-policy character annotation based on CHISE. *Journal of the Japanese Association for Digital Humanities*, Vol. 1, No. 1, pp. 86–106, 2015 年 11 月.

[3] 池田証壽. HDIC Database Project. `https://github.com/shikeda/HDIC`, 2022 年 3 月.

[4] 加藤大鶴, 他. 資料横断的な漢字音・漢語音データベース. `https://dhsjr.w.waseda.jp/`, 2025 年 3 月.

[5] 守岡知彦. CHISE IDS 漢字検索. `https://www.chise.org/ids-find`.

[6] 守岡知彦. Wiki 的手法に基づく構造化データの編集について. 人文科学とコンピュータシンポジウム論文集 —人文工学の可能性 〜異分野融合による「実質化」の方法〜, 情報処理学会シンポジウムシリーズ, 第 2010 巻, pp. 33–40. 情報処理学会, 情報処理学会, 2010 年 12 月.

[7] 守岡知彦. CHISE における HDIC 統合の試み. 情処研報, Vol. 2022-CH-129, No. 12, pp. 1–6, 2022 年 5 月.

[8] 守岡知彦. Json-ld を用いた古字書注文構造化の試み. 情処研報, Vol. 2023-CH-133, No. 6, pp. 1–8, 2023 年 9 月.

[9] 鈴木慎吾. Web 韻圖（廣韻檢索）. `https://suzukish.sakura.ne.jp/search/inkyo/`, 2023 年 3 月.