TEI Lex-0 を利用した日本古辞書の構造化記述

李　媛[1]

１　はじめに

TEI P5 は、人文科学資料のマークアップに関する最新のガイドラインで、1990 年に主に欧州および北米の大学や研究機関を中心に公表された。元々は欧米の文献資料が主な対象だったが、近年は東アジアでも TEI に基づくプロジェクトが増え、その使用範囲が広がっている。古写本古辞書のような文献資料は、現代の辞書と比較すると、整然とした凡例や項目構造を持つものが少なく、特に欧州言語の辞書項目と比べると、字音、字義、異体字表記などの内容に独特の配置構造を持つことが多い。また、日本の古辞書では、和訓記述や漢字音注記などの要素が加わり、内容および構造がさらに複雑になる。

日本古辞書の構造化記述に関する研究としては、申（2016）、劉ほか（2017）、藤本・韓・高田（2018）、李（2018）、申（2018）、岡田（2020）、李（2022）などが挙げられる。中でも劉ほか（2017）や藤本・韓・高田（2018）は独自システムを用いた構造化記述に関するもので、その手法や内容は参考になる。

李（2018）は『篆隷万象名義』の TEI P5 に基づく構造化記述を試みた例である。申（2018）は、『類聚名義抄』の構造化を検討した研究で、申（2016）を基に改訂された。岡田（2020）は、TEI に基づく日本古辞書の効率的な符号化モデルに関する論考で、大きな参考になる。李（2022）は、岡田（2020）で示されたモデルを基に、『玉篇』系の字書の構造化記述に取り組んだものである。

これらの研究は、いずれも TEI P5 辞書モジュールの範囲で検討した内容である。一方で、TEI P5 辞書モジュールのほか、TEI Lex-0 という辞書、語彙、言語学データなど、言語リソースをエンコードするために特化している規格があり、言語学関係の研究者やプロジェクトに採用されることも多い。

今年 11 月 17 日（土）には、東京ビッグサイトで TEI Lex-0[2]に関するワークショップ「Digitization of Historical Lexicography and TEI-Lex0」が開催された。ここでは TEI Lex-0 の主要メンバーである Toma Tasovac 氏が講演し、日本における辞書デジタル化に取り組む研究者等が発表した。Tasovac 氏は、TEI Lex-0 が欧州言語の辞書を基にして作成されたことから、東アジア、特に日本の古辞書の特徴や構造に関心が寄せられていると指摘した。

一方で、筆者はこれまで特に辞書モジュールに焦点を当てた TEI P5 Guidelines を研

---

[1] 京都大学人文科学研究所附属人文情報学創新センター

[2] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

究し、東アジアでの TEI の普及状況を踏まえながら、『篆隷万象名義』のような比較的構造がシンプルな古辞書のエンコーディングに取り組んできた。しかし、構造の面での差異が大きく、TEI P5 の現行タグの範囲内でのエンコーディングには多くの困難が伴った[3]。

TEI-C 東アジア／日本語分科会[4]の研究会に参加してから、TEI の全体的なシステム、内容、そして細部について、様々な分野の研究者と議論する機会を得たが、同じ分野の研究者が少ないため、古辞書の内容や構造に関する具体的な議論、さらに TEI P5 辞書モジュールの本来の限界性の影響もあり、日本古辞書のエンコーディング作業に困難が伴うことが多かった。

TEI Lex-0 は、辞書や語彙、言語学データなどの言語リソースをエンコードするために特化した規格である。先に述べた TEI Lex-0 のワークショップで、筆者は『篆隷万象名義』を例に、日本古辞書へのデジタルアプローチを紹介した[5]。このワークショップでの講演や内容に触発され、本発表では TEI Lex-0 のスキーマを基に、日本古辞書の特有な部分のエンコーディングを検討する。

２　対象資料

高山寺本『篆隷万象名義』

『篆隷万象名義』は平安初期（830 年頃）に編纂された日本現存最古の字書である。

内容は原本玉篇から抜粋し、六帖の構成で、約 16,000 字が収録される。前の四帖は空海撰述であるが、後の二帖は別人の手による。永久二年（1114）に書写された高山寺本が唯一の古写本である。原本『玉篇』は中国で早く散逸し、残巻と逸文を伝わるのみである。それゆえ、完本である篆隷万象名義は原本玉篇の元の姿を窺う資料として価値が高い。

ただ、字体の誤写や義注の誤りが多いと言われる。さらに、誤写かと思われる例の中には、中国南北朝時代以来の字形の古態を残すものもあって、誤写か古い字体かを判別するのは容易ではない。高山寺本篆隷万象名義の本来の価値を見出すには、精密な校訂が要求される。

天治本『新撰字鏡』

新撰字鏡 12 巻は、序文によると、撰者の昌住が寛平 4 年（892 年）夏に唐の玄応一切経音義 25 巻を基に、「三軸」からなる部首分類の字書を編纂したことに始まる。その

---

[3] 筆者は、既存のタグ（element）の範囲で、日本の古辞書の内容のエンコーディングを試みたが、京都大学人文科学研究所の Wittern Christian 氏から、必要なタグ（element）があれば、TEI 委員会に提案することを勧められた（2019 年の TEI 会議、東京）。

[4] TEI-C 東アジア／日本語分科会：https://tei.dhii.jp

[5] LI (2023)

後、昌泰年中（898-901 年）に切韻、玉篇、私記類を得て増補し、12 巻とする。現存本には完本と抄録本の 2 系統がある。完本は天治元年（1124 年）に書写された宮内庁書陵部蔵の天治本が唯一のものである。抄録本は 12 巻系統の本から和訓のある項目を抜き出したもので、享和本、群書類従本、大東急記念文庫蔵本などが存在する。

天治本は、160 の部に分かれ、単字をその字形により部首に分類した部（天・日・月・肉・雨など）と、語句を意義分類した部（親族・小学篇字・本草・重点・連字・臨時雑要字など）から構成される。部首分類には、天や人、衣、食、住など、意味に基づく分類が取り入れられている。引用される主な資料は一切経音義（玄応）、切韻、玉篇の三書で、それぞれの原典の本文の体裁を反映し、一群となって引用される。

観智院本『類聚名義抄』

『類聚名義抄』は原撰本と改編本の二種類がある。原撰本は法相宗と真言宗を兼学する学僧により 1081 年以降に成立したとされ、図書寮本が唯一の伝本である。改編本『類聚名義抄』は、観智院本、高山寺本、蓮成院本、西念寺本、宝菩提院本が知られる。このうち観智院本が唯一の完本である。改編本『類聚名義抄』の成立は、逸文の伝存状況から 1178 年までには成立していたとされる。撰者は、真言宗の学僧による編纂とされる。原撰本の仏教事典的要素を削り、原撰本の漢文注を省略、万葉仮名を片仮名に改め、片仮名注を増補している。原撰本に比べ、見出しは 2.1 倍、和訓は 4.1 倍に増加し、120 の部首に約 32,000 の掲出語（項目）、約 34,710 の和訓、声点（アクセント）付きは 1 万を有する。

３　対象資料のテキストデータ

筆者の所属していた北海道大学大学院文学研究院言語科学講座池田研究室では「平安時代漢字字書総合データベース（HDIC）」を構築するプロジェクトを推進していた。

平安時代漢字字書総合データベース（Integrated Database of Hanzi Dictionaries in Early Japan, 略称 HDIC, https://hdic.jp）は、7 万字を超える漢字の処理ができる Unicode を用いて、平安時代を代表する部首分類体字書である『篆隷万象名義』『新撰字鏡』『類聚名義抄』を総合した全文テキストデータベースを整備・構築し、日本古辞書の研究基盤の確立を目指す。これらの漢字字書は、語彙・音韻・漢字字体等の日本語史研究の重要な資料である。これらの漢字字書に見える掲出字と注文とを電子テキスト化し、掲出字のリレーションシップを設定することによって、自在に検索できるシステムを構築し、その公開によって学術研究の発展に資することを意図してきた。

HDIC は古辞書翻刻・入力作業の効率化をはかるため、まず古版本である宋本『玉篇』をデータ化し、これを土台に高山寺本『篆隷万象名義』の入力作業も完成した。2016 年 4 月に宋本『玉篇』、9 月に『篆隷万象名義』の全文テキストデータを順次に公開された。

その後、2018 年 6 月に天治本『新撰字鏡』、2022 年 3 月に観智院本『類聚名義抄』の全文テキストも公開された[6]。

4　TEI P5 の辞書モジュールと TEI Lex-0

TEI P5（Text Encoding Initiative）の辞書モジュールと TEI Lex-0 は、テキストエンコーディングのための標準的なガイドラインであり、それぞれ異なる目的を持つが、いくつかの共通点もある。TEI P5 は、テキスト文書をエンコードするための包括的なガイドラインであり、文学テキスト、歴史文書、学術論文など、さまざまな種類のテキストをエンコードするために使用される。TEI P5 は、XML ベースのテキストエンコーディングのための一般的なフレームワークを提供する。辞書モジュールは、TEI P5 の一部で、辞書や語彙リソースをエンコードするための仕様を提供する。このモジュールを使用すると、辞書の構造、見出し語、定義、例文などを XML 形式でエンコードできる。

TEI Lex-0 は、TEI P5 の一部ではなく、独立したプロジェクトで、主に辞書と語彙リソースのために設計された TEI ベースのスキーマである。TEI Lex-0 は、辞書、語彙、言語学的データなど、言語リソースをエンコードするために特化しており、辞書エントリ、見出し語、定義、発音、例文などを包括的に記述するための要素と属性を提供する。

TEI Lex-0 は、特に言語学的な研究や言語リソースのエンコーディングに適しており、TEI P5 とは異なるスキーマであるが、TEI P5 の一部として組み込むこともできる。

5　古辞書における漢字の字形・字音・字義についてのエンコーディングの実際

字形

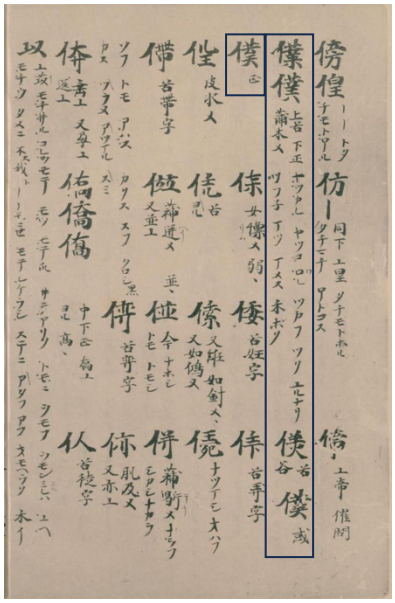

図 1 観智院本『類聚名義抄』の例　「僕」

---

[6] https://github.com/shikeda/HDIC
HDIC について、池田（2014・2018）などを参照されたい。

僕僕　上俗下正　ヤツカレ(LHHL)　ヤツコ　ロレ(HH、ロ-ワ)　ツカフ　ツク ユルナリ　蒲木反　ツフネ　マツ　マメス　和ホク(N@)[7]

僕β　俗 谷

僕γ　或

僕δ　正　　（観智院本類聚名義抄 仏上 人部）

<entry xml:id="K01004211" type="mainEntry" xml:lang="zh">
  <form type="lemma">
    <orth>僕</orth>
  </form>
<form type="variant">
        <orth>僕</orth>
      <graphic url="https://glyphwiki.org/glyph/hdic_hkrm-01004211.100px.png"/>
      </form>
      <form type="variant">
        <orth>僕</orth>
      <graphic url="https://glyphwiki.org/glyph/hdic_hkrm-01004212.100px.png"/>
      </form>
      <usg type="glyph class">上俗下正</usg>
      <sense xml:id= K01004211_sense1 xml:lang="jp">
        <def>ヤツカレ(LHHL) </def>
      </sense>
      <sense xml:id= K01004211_sense2 xml:lang="jp">
        <def>ヤツコ</def>
      </sense>
      <sense xml:id= K01004211_sense3 xml:lang="jp">
        <def>ロレ(HH、ロ-ワ)</def>
      </sense>
      <sense xml:id= K01004211_sense4 xml:lang="jp">
        <def>ツカフ</def>
      </sense>
      <sense xml:id= K01004211_sense5 xml:lang="jp">

[7] 声点は平（L)、上（H)、去（R)、入（S)、平軽（F)、入軽（T）のように括弧内の略称で示す。声点が和訓の一部分のみに施されている場合は、声点が施されていない箇所を半角のアットマーク（@）で表す。池田ほか（2020）および申（2015）を参照されたい。

```
        <def>ツク</def>
      </sense>
      <sense xml:id= K01004211_sense6 xml:lang="jp">
        <def>ユルナリ</def>
      </sense>
      <usg type="phonetic" xml:lang="zh">蒲木反</usg>
      <sense xml:id= K01004211_sense7 xml:lang="jp">
        <def>ツフネ</def>
      </sense>
      <sense xml:id= K01004211_sense8 xml:lang="jp">
        <def>マツ</def>
      </sense>
      <sense xml:id= K01004211_sense9 xml:lang="jp">
        <def>マメス</def>
      </sense>
      <usg type="phonetic" xml:lang="jp">和ホク(N@)</usg>
      </entry>

  <entry xml:id="K01004241" type="subEntry" xml:lang="zh">
    <form type="variant">
      <orth>僕β</orth>
      <graphic url="https://glyphwiki.org/glyph/hdic_hkrm-
01004241.100px.png"/>
    </form>
    <usg type="glyph class">谷</usg>
    <usg type="glyph class">俗</usg>
  </entry>

  <entry xml:id="K01004242" type="subEntry" xml:lang="zh">
    <form type="variant">
      <orth>僕γ</orth>
      <graphic url="https://glyphwiki.org/glyph/hdic_hkrm-
01004242.100px.png"/>
    </form>
    <usg type="character class">或</usg>
  </entry>
```

<entry xml:id="K01004310" type="subEntry " xml:lang="zh">
<form type="variant">
<orth>僕δ</orth>
<graphic url="https://glyphwiki.org/glyph/hdic_hkrm-01004310.100px.png"/>
</form>
</entry>

字音

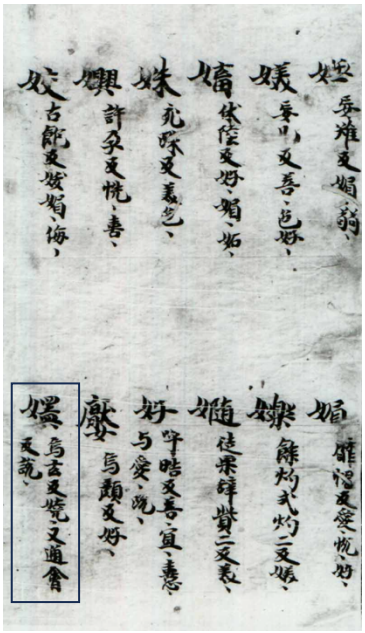

図２ 『篆隷万象名義』の例　「嬽」

嬽 烏玄反。嬈也。又通會反。說也。(篆隷万象名義 第 1 帖 72 丁表)

<entry xml:id="1_072_A62" xml:lang="zh">
<form type="lemma">
<orth>嬽</orth>
</form>
<sense xml:id="1_072_A62_sense1" xml:lang="zh">
<pron>烏玄反</pron>
<def>嬈也</def>
</sense>
<sense xml:id="1_072_A62_sense2" xml:lang="zh">

又<pron>通會反</pron>
<def>說也</def>
</sense>
</entry>

字義

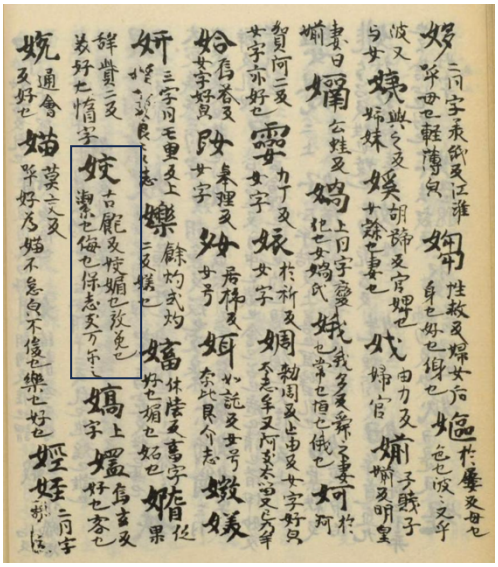

図３ 『新撰字鏡』の例　「姣」

姣　古飽反。姣媚也、放免也、潔也、侮也。保志支万尓之。（新撰字鏡 第３巻 女部）

<entry xml:id="s0319b701" xml:lang="zh">
<form type="lemma">
<orth>姣</orth>
</form>
<pron>古飽反</pron>
<sense xml:id=" s0319b701_sense1" xml:lang="zh">
<def>姣媚也</def>
</sense>
<sense xml:id=" s0319b701_sense2" xml:lang="zh">
<def>放免也</def>
</sense>
<sense xml:id=" s0319b701_sense3" xml:lang="zh">
<def>潔也</def>
</sense>
<sense xml:id=" s0319b701_sense4" xml:lang="zh">

```
        <def>侮也</def>
    </sense>
    <sense xml:id=" s0319b701_sense5" xml:lang="jp">
        <def>保志支万尓之</def>
    </sense>
</entry>
```

６　おわりに

本稿では、TEI Lex-0 のスキーマを基に、日本古辞書の特徴的な項目に着目し、それらの構造化記述を行うエンコーディングを検討した。しかし、この検討は個別の事例に焦点を当てたものであり、古辞書全体の構造を総合的に把握するには不十分である。日本古辞書に特有の内容や項目構造をさらに詳細に整理し、TEI Lex-0 の範囲でどこまで記述可能かを把握することが必要である。また、追加すべきタグ（element）に関する報告も今後の重要な課題となる。

附記：本研究は JSPS 科研費 JP21K18013 の助成を受けたものである。

文献資料画像

図１　『類聚名義抄』[2],貴重図書複製会,昭和 12. 国立国会図書館デジタルコレクション https://dl.ndl.go.jp/pid/2586891 (参照 2023-12-18)

図２　北海道大学大学院言語科学講座所蔵焼付写真（石塚晴通名誉教授）

図３　昌住 著 ほか『新撰字鏡 ： 天治本』巻 3 巻 4,六合館,大正 5. 国立国会図書館デジタルコレクション https://dl.ndl.go.jp/pid/1241741 (参照 2023-12-18)

参考文献

池田証壽 「平安時代漢字字書総合データベースの構築」、『北海道大学文学研究科紀要』 142、北海道大学文学研究科、pp.79-90、2014 年 3 月.

池田証壽「篆隷万象名義全文テキストデータベース構築からＨＤＩＣへの展開」、『平成 29 年度高山寺典籍文書綜合調査団研究報告論集』、pp.34-44、2018 年 3 月.

池田証壽「観智院本類聚名義抄全文テキストデータベース －その構築方法と掲出字項目数等の計量－」、『訓点語と訓点資料』144、pp.105-129、2020 年 3 月.

岡田一祐「日本平安期古辞書の符号化モデル: TEI をもとにした符号化」、『デジタル・ヒューマニティーズ』vol.2、pp.26-54、2020 年 11 月

申雄哲「図書寮本類聚名義抄の本文解読とデータベース作成の問題点」漢デジ 2016、2016 年 8 月.

申雄哲「図書寮本『類聚名義抄』の構造化テキストの設計と実践」国際シンポジウム 「古

辞書研究の射程」、2018 年 8 月.
申雄哲「図書寮本類聚名義抄の基礎的研究」北海道大学大学院博士論文、2015 年 3 月.
藤本灯・韓一・高田智和「古辞書の構造化記述の試み: 和名類聚抄を例に」日本語学会 2017 年度秋季大会、2017 年 10 月.
劉冠偉・李媛・鄭門鎬・張馨方・池田証壽「部首分類体日本古辞書の項目構造の多様 性に対応したマークアップ・ツールの開発」『人文科学とコンピュータシンポジウム 2017 予 稿集』(2017 年 12 月)、pp. 97–102.
李媛「TEI P5 Dictionaries モジュールに基づく古辞書の構造化記述の試み: 篆隷万象名 義を中心に」『情報処理学会研究報告』2018-CH-117, no. 5(2018 年 5 月): pp. 1–8.
李媛「玉篇系字書の構造化記述に関する TEI マークアップについて」、『東洋学へのコンピュータ利用 第 35 回研究セミナー 予稿集』、東アジア人文情報学研究センター、pp.13-27、2022 年 7 月.
TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange, Version 4.4.0, revision ff9cc28b0（April 2022）, https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf
Yuan LI, Digital Humanities approaches to early Japan dictionaries research: A case of Tenreibanshomeigi, Workshop on "Digitization of Historical Lexicography and TEI-Lex0", Tokyo Big Sight, 2023 年 11 月 17 日.