

資料横断的な漢字音・漢語音 データベースの語彙資源化

高田智和（国立国語研究所）

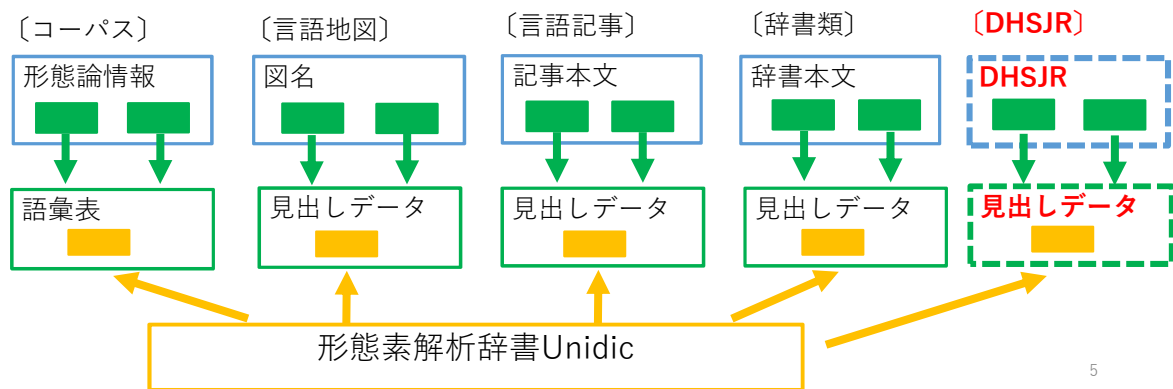
語彙資源

- 語彙表
 - 度数表 例：現代雑誌90種調査
 - 基本語彙 例：教育基本語彙、日本語教育基本語彙
 - 単語親密度 例：NTTデータベースシリーズ「日本語の語彙特性」
- シソーラス 例：分類語彙表
- 辞書 例：人間用辞書、形態素解析用辞書

語彙資源同士をつなぐ試み

Unidicの語彙素IDでつなぐ

- 各資源の個々の見出しに語彙素IDを振る。
- 見出しがなければ、見出しを設定する。



5

「漢字音・漢語音データベース」の項目

- ほかのデータベース・データセットとの接続が可能な項目を持っている。
- 漢字音・アクセントに関する項目
 - 仮名音注、声点、反切、類音注、節博士・・・
- 出典項目（出典資料内の所在情報）
- 単字項目（漢字字種に関する情報）
- 漢語項目

出典項目

単字_見出し	漢語_見出し	声点	所在
大	大恩	平濁	001_006_b_02
恩	大恩	去	001_006_b_02
教	教主	平	001_006_b_02
主	教主	平	001_006_b_02

30-026-03「四座講式」宝暦版

国書データベース

<https://kokusho.nijl.ac.jp/biblio/100309444/10?ln=ja>

【書誌詳細】



単字項目

資料内漢字番号	資料内漢語番号	単字_見出し	単字_出現形	漢語_見出し
21	11	淵	淵	顔淵
80	41	淵	淵	顔淵
92	49	淵	淵	顔淵

50-035-01「論語」京都大学清家文庫蔵、宣賢点

漢字字体規範史データセット

HNG 単字検索

U+6DF5 淵

見出し字: 淵 大字典番号: 06310 大漢和番号: M17692

中国写本

漢書揚雄



中国版本



Unicodeでつなぐ

漢字音・漢語音データベース 50-035-01「論語」京都大学清家文庫蔵、宣賢点

資料内漢字番号	資料内漢語番号	単字_見出し	単字_出現形	漢語_見出し	仮名注
21	11	淵	淵	顔淵	〔墨〕エン
80	41	淵	淵	顔淵	〔墨〕エン
92	49	淵	淵	顔淵	〔墨〕エン

U+6DF5 淵

見出し文字	異体字	統合ID	大字典	UCS	kar_代表字形ID	kar_字体数	kar_用例数
淵	洊淵	06310	06310	6DF5	1579a	5	2

漢字字体規範史データセット kar開成石経論語

漢語項目

• 仮名音注、反切、類音注により語形がわかる歴史的な漢語語彙



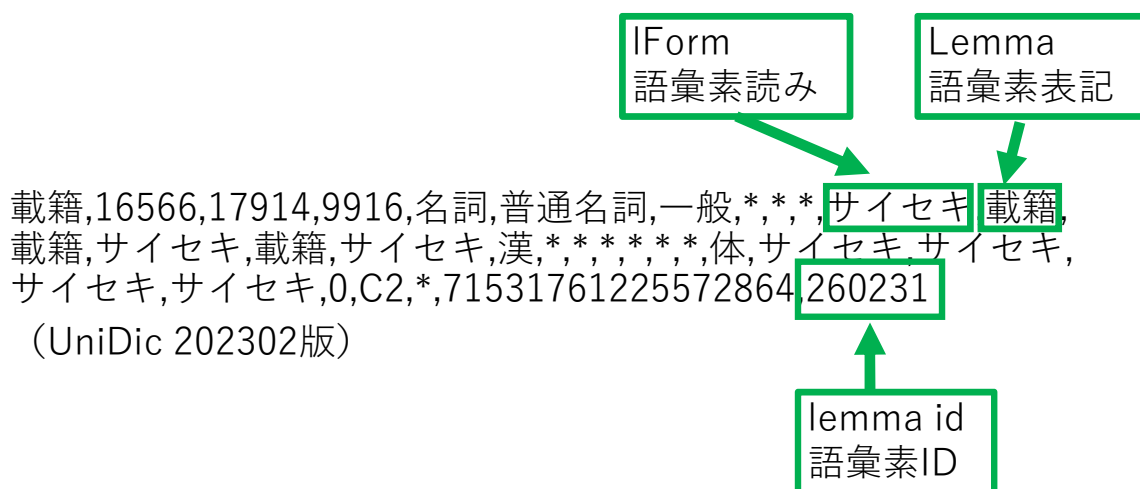
資料番号	単字_見出し	単字_出現形	漢語_見出し	漢語_出現形	声点	仮名注
30-012-01	載	載	載籍	載籍ノ	去	
30-012-01	籍	籍	載籍	載籍ノ	入	
30-012-01	薰	薰	薰風	薰風	平	
30-012-01	炯	炯	炯戒	炯戒	上	ケイ
30-012-01	戒	戒	炯戒	炯戒	去	
30-012-01	祀	祀	千祀	千祀に	上	

30-012-01「群書治要経部」金沢文庫本

「単語」のコード

- 文字（漢字）の場合
 - 文字コード Unicode, JIS
 - 辞書コードブック 新字源、大漢和辞典
→ 文字番号のある辞書
- 単語（日本語語彙）の場合は？
 - 形態素解析用辞書 UniDic
 - 辞書コードブック 熟語番号のある辞書

UniDicの語彙素ID



『日本国語大辞典第二版』（ジャパンナレッジ）のURL



<https://japanknowledge.com/psnl/display/?lid=200201b09c79VpP4ELkv>

コード付け対象

- 20-002-01_医心方_半井家旧蔵本（東博蔵本）
- 30-018-01_三帖和讃

20-002-01_医心方_半井家旧蔵本（東博蔵本）の語彙量

- 漢語数（延べ） 2124語 ※資料内漢語番号から算定
- 漢語数（異なり） 1408語 ※表記の異なり
 - 1字語：296語
 - 2字語：1010語
 - 3字語：87語
 - 4字語：12語 5字語：2語 6字語：1語
 -
 - 仮名注全語形 163語

30-018-01_三帖和讃の語彙量

- 漢語数（延べ） 1664語 ※資料内漢語番号から算定
- 漢語数（異なり） 983語 ※表記の異なり
 - 1字語：62語
 - 2字語：522語
 - 3字語：128語
 - 4字語：219語 5字語：24語 6字語：16語 7-9字語：12語
 -
 - 仮名注全語形 732語

医心方とUnidic

- 漢語数（異なり） 1408語
- 語彙素ID付与できた 390語（27.7%）
 - 中世文語Unidic 313語
 - 近世文語Unidic 285語
 - 近代文語Unidic 353語
 - 現代語Unidic 331語

三帖和讃とUnidic

- 漢語数（異なり） 1664語
- 語彙素ID付与できた 409語（25.6%）
 - 中世文語Unidic 374語
 - 近世文語Unidic 322語
 - 近代文語Unidic 361語
 - 現代語Unidic 335語

医心方と日本国語大辞典第二版

- 漢語数（異なり） 1408語
- “ID”を付与できた 700語（49.7%） cf. Unidic 27.7%
- 仮名注全語形 163語
- “ID”を付与できた 100語（61.3%）
- cf. Unidic 57語（35.0%） → 語誌情報ポータルに接続可

三帖和讃と日本国語大辞典第二版

- 漢語数（異なり） 1664語
- “ID”を付与できた 649語（39.0%） cf. 409語（25.6%）
- 仮名注全語形 732語
- “ID”を付与できた 512語（69.9%）
- cf. Unidic 328語（44.8%） → 語誌情報ポータルに接続可

単位の問題

例 “5字語” 称讃浄土経

- 短単位 (Unidic) では1単位とならない
- 3短単位の接続で表現できる

称讃 ショウサン 16684

浄土 ジョウド 18599

経 キョウ 9144

おわりに

- 形態素解析用辞書、『日本国語大辞典』のどちらを使うにしても、未収録語の取扱いが課題。（“外語”）
- 「漢字音・漢語音データベース」をコードブックにするのが良さそう。（“セルフコードブック”）