DHSJRにおける漢字字体の正規化の試み

大島 英之（国立国語研究所非常勤研究員）

2024年11月30日＠国立国語研究所

hid-ohshima@ninjal.ac.jp / a4.ohshima@gmail.com

これまで行ったこと

- DHSJRは、主に、個人個人の研究者が作成・所持していた漢字音データを、一定のフォーマットに従う形で統合することによって構築されている。
  - 入力者によって、入力字体の方針が異なる。
- 2024年度版では、以下の変更を行った。
  - 従来の「単字」「漢語」列を「単字_出現形」「漢語_出現形」に変更。
  - 出現形列におけるデータ間の字体揺れを統合し、その代表字を格納する列として、「単字_見出し」「漢語_見出し」列を新たに作成。

これまで行ったこと

- 「常用漢字表」（2010年告示）において（　）内に示される旧字体のうち、JIS第2水準の範囲に限って、新字から旧字への置換を行った（249字）。
  - 出現形列から、旧字に寄せた結果を、見出し字列に格納。
  - 新字と旧字の区別が字種に関わる11のペア（灯／燈、弁／辨・瓣・辯、糸／絲、欠／缺、缶／罐、台／臺、芸／藝、虫／蟲、証／證、予／豫、余／餘）は置換せず。
  - 詳細は、「見出し」列の正規化に関する注意事項 [https://dhsjr.w.waseda.jp/?page_id=111] を参照。

これまで行ったこと

	A	B	C	D	E	F	G	H	I
1	ID	資料番号	資料名	資料内漢字	資料内漢字	**単字_見出し**	**単字_出現形**	**漢語_見出し**	**漢語_出現形**
13904	20-002-01	20-002-01	医心方_半	1	1	經	経	經方	経方
13905	20-002-01	20-002-01	医心方_半	2	1	方	方	經方	経方
13906	20-002-01	20-002-01	医心方_半	3	2	由	由	由來	由来
13907	20-002-01	20-002-01	医心方_半	4	2	來	来	由來	由来
13908	20-002-01	20-002-01	医心方_半	5	3	五	五	五藏	五蔵
13909	20-002-01	20-002-01	医心方_半	6	4	營	営	營衛	営衛
13910	20-002-01	20-002-01	医心方_半	7	4	衛	衛	營衛	営衛
13911	20-002-01	20-002-01	医心方_半	8	5	盈	盈	盈虛	盈虛
13912	20-002-01	20-002-01	医心方_半	9	5	虛	虛	盈虛	盈虛

- 「盈虛」の「虛」の旧字体「虛」は、JIS第3水準なので置換せず。

現状の問題

- 常用漢字外の字種の異体字
- 新旧以外の関係にある常用漢字の異体字

} 置換未着手。

➢ 字種レベルでの計量的調査は、すぐには実行しにくいのが現状（各自の判断で異体字の統合が必要）

現状の問題

- 教（U+6559）と敎（U+654E）が別々になったまま。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	ID	資料番	資料名	資料内	資料内	**単字_見出し**	単字_出	漢語_見	漢語_出	漢語_al	語種	位置	単字長	声点	声点型	仮名注	仮名型	反切
18276	20-044-0	20-044-0	九条本法	293	293	教	教	教	教			1		平	平	ケウ	ケウ	去詔反
21138	30-006-0	30-006-0	世俗諺文_	704	421	教	教	教	教			1		去	去			
21863	30-006-0	30-006-0	世俗諺文_	1429	859	教	教	教命	教命			1		去	去＊			
22948	30-006-0	30-006-0	世俗諺文_	2514	1535	教	教	傳教	傳教			2			去＊	カウ	フカウ	
25298	30-010-0	30-010-0	阿弥陀経_	648	253	教	教	惡友之教	惡友之教			4	2	平	入緩上上平			
25707	30-010-0	30-010-0	阿弥陀経_	1061	400	教	教	教我觀於	教我觀於			1	2	平	平＊＊上			
25758	30-010-0	30-010-0	阿弥陀経_	1113	423	教	教	教我思惟	教我思惟			1	2	平	平＊去上			
27437	30-010-0	30-010-0	阿弥陀経_	2801	1233	教	教	教令念佛	教令念佛			1	2	平	平去＊＊			
28426	30-012-0	30-012-0	群書治要_	351	308	教	教	教	教			1	2	去	去			
28517	30-012-0	30-012-0	群書治要_	442	387	教	教	五教	五教を			2	2	去	＊去			
28756	30-012-0	30-012-0	群書治要_	681	600	教	教	教	教									
28866	30-012-0	30-012-0	群書治要_	791	702	教	教	教也	教也									
30624	30-012-0	30-012-0	群書治要_	2549	2276	教	教	教	教に									
30636	30-012-0	30-012-0	群書治要_	2561	2287	教	教	教	教									
31707	30-012-0	30-012-0	群書治要_	3632	3236	教	教	教	教									
31864	30-012-0	30-012-0	群書治要_	3789	3373	教	教	教	教の									
32750	30-013-0	30-013-0	尾張国郡	346	221	教	教	教	教									
35737	30-017-0	30-017-0	色葉字類	1062	714	教	教	八教	八教									
38625	30-017-0	30-017-0	色葉字類	3950	2411	教	教	教業坊	教業									
39475	30-017-0	30-017-0	色葉字類	4800	2971	教	教	高教	高教									

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	ID	資料番	資料名	資料内	資料内	**単字_見出し**	単字_出	漢語_見	漢語_出	漢語_al	語種	位置	単字長	声点	声点型	仮名注	仮名型	反切
60312	30-021-0	30-021-0	西方指南	607	288	教	教	教化	教化シ			1	2			ケウ	ケウクヱ	
61024	30-021-0	30-021-0	西方指南	1319	602	教	教	教	教			1	2			ケウ	ケウ	
61340	30-021-0	30-021-0	西方指南	1635	734	教	教	教	教			1	2			ケウ	ケウ	
61343	30-021-0	30-021-0	西方指南	1638	736	教	教	教	教			1	2			ケウ	ケウ	
61345	30-021-0	30-021-0	西方指南	1640	738	教	教	教	教			1	2			ケウ	ケウ	
61353	30-021-0	30-021-0	西方指南	1648	745	教	教	教	教			1	2			ケウ	ケウ	
61364	30-021-0	30-021-0	西方指南	1659	751	教	教	教	教			1	2			ケウ	ケウ	
61531	30-021-0	30-021-0	西方指南	1826	831	教	教	教主釋尊	教主釋尊			1	2			ケウ	ケウシユシヤクソン	
61666	30-021-0	30-021-0	西方指南	1961	905	教	教	傳教大師	傳教大師			2	2			ケウ	＊ケウ＊＊	
61683	30-021-0	30-021-0	西方指南	1978	913	教	教	言教	言教			2	2			ケウ	コンケウ	
61869	30-021-0	30-021-0	西方指南	2164	1012	教	教	教主	教主			1	2			ケウ	ケウシユ	
62161	30-021-0	30-021-0	西方指南	2456	1142	教	教	教主釋尊	教主釋尊			1	2			ケウ	ケウシユシヤクソン	
62825	30-021-0	30-021-0	西方指南	3120	1460	教	教	教主釋尊	教主釋尊			1	2			ケウ	ケウシユシヤクソン	

代表字の基準

- 「単字_見出し」の代表字をどう決めるか？
  - 中古音（≒『広韻』）データとの接続のしやすさを意識したい。
    - 旧字体の側に寄せた理由
  - どの異体字テーブル（あやしくないもの）が適切か？
    - 「史料編纂所データベース異体字同定一覧」？
    - 「漢字データベース」内「異体字データベース」？
    - 特定の漢和辞典に基づいて新たにテーブルを作成？
      - 漢和辞典によって、異体字とするか別字とするか分かれることも多い。

史料編纂所データベース異体字同定一覧

史料編纂所データベース異体字同定一覧

確認日：2024年11月28日

No.		異体字
1	亜	亞
2	唖	啞　瘂
3	悪	惡
4	芦	蘆
5	鯵	鰺
6	圧	壓
7	庵	菴　葺
8	案	桉
9	囲	圍
10	為	爲
11	医	醫
12	育	毓
13	一	弐
14	壱	壹
15	稲	稻
16	因	囙
17	姻	婣
18	飲	飲
19	淫	婬　淫
20	隠	隱
21	陰	陰　陰
22	卯	夘
23	丑	刄
24	嘘	噓
25	欝	鬱
26	厩	廐　廏　廠
27	叡	睿
28	営	營
29	映	暎
30	曳	曳
31	栄	榮
32	穎	潁
33	英	俠
34	衛	衞
35	詠	咏

史料編纂所データベース異体字同定一覧

- 字音の観点からすれば、次のようなペアは異体字とは認めがたい。

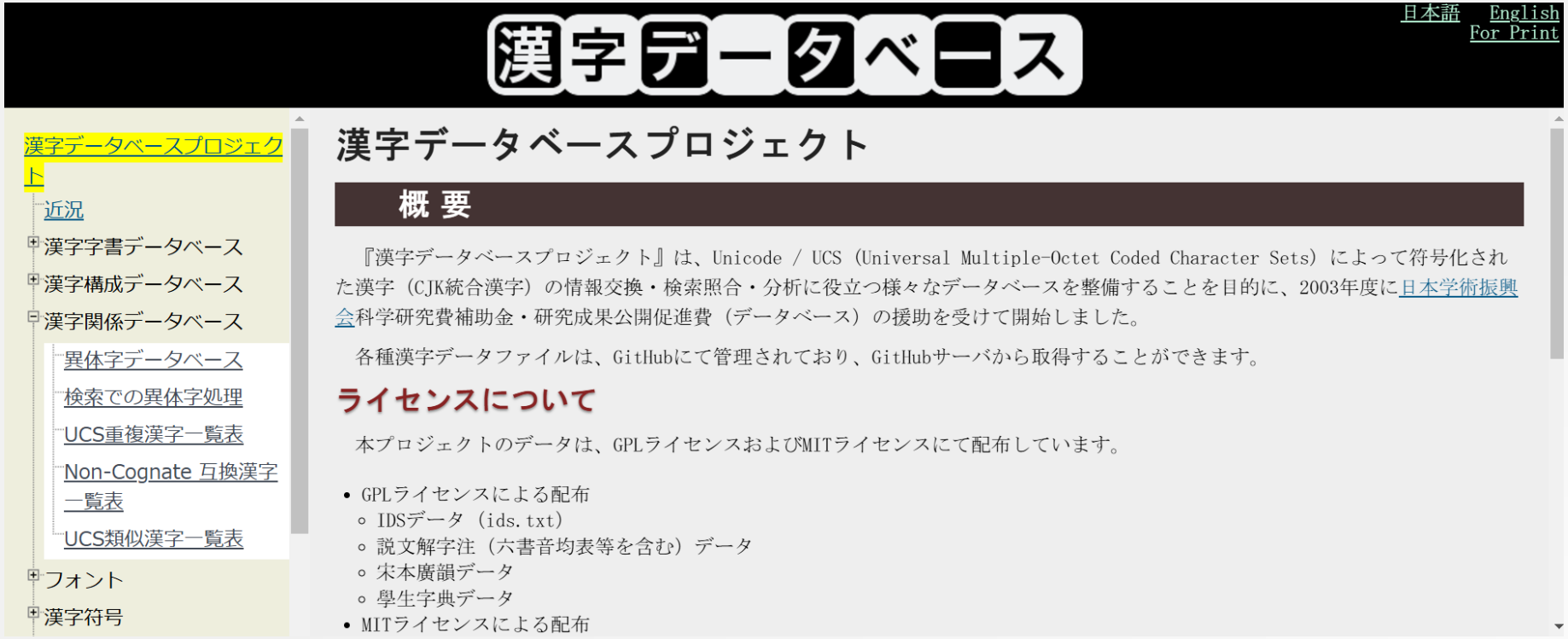
1068	閏	閏　壬

1020	難	亘

- 古文書・古記録における用法から帰納的に判定？

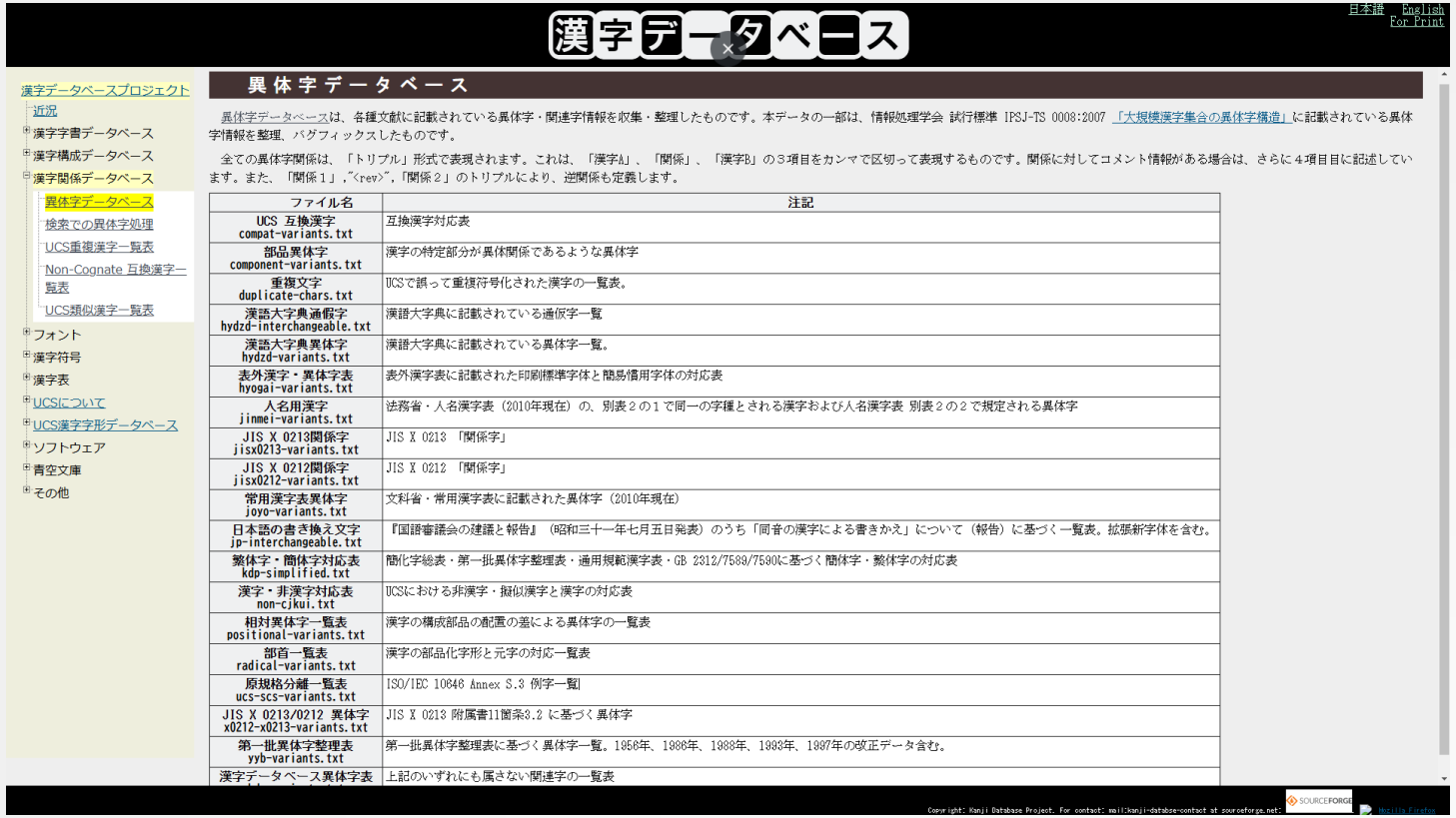
漢字データベース

- https://kanji-database.sourceforge.net/



異体字データベース

- MITライセンスで提供（ライセンス明記のみ必要）



漢字データベース

日本語 English For Print

- 漢字データベースプロジェクト
  - 近況
- 漢字字書データベース
- 漢字構成データベース
- 漢字関係データベース
  - 異体字データベース
  - 検索での異体字処理
  - UCS重複漢字一覧表
  - Non-Cognate 互換漢字一覧表
  - UCS類似漢字一覧表
- フォント
- 漢字符号
- 漢字表
- UCSについて
- UCS漢字字形データベース
- ソフトウェア
- 青空文庫
- その他

異体字データベース

異体字データベースは、各種文献に記載されている異体字・関連字情報を収集・整理したものです。本データの一部は、情報処理学会 試行標準 IPSJ-TS 0008:2007 「大規模漢字集合の異体字構造」に記載されている異体字情報を整理、バグフィックスしたものです。

全ての異体字関係は、「トリプル」形式で表現されます。これは、「漢字A」、「関係」、「漢字B」の３項目をカンマで区切って表現するものです。関係に対してコメント情報がある場合は、さらに４項目目に記述しています。また、「関係１」,"<rev>",「関係２」のトリプルにより、逆関係も定義します。

ファイル名	注記
UCS 互換漢字 compat-variants.txt	互換漢字対応表
部品異体字 component-variants.txt	漢字の特定部分が異体関係であるような異体字
重複文字 duplicate-chars.txt	UCSで誤って重複符号化された漢字の一覧表。
漢語大字典通假字 hydzd-interchangeable.txt	漢語大字典に記載されている通仮字一覧
漢語大字典異体字 hydzd-variants.txt	漢語大字典に記載されている異体字一覧。
表外漢字・異体字表 hyogai-variants.txt	表外漢字表に記載された印刷標準字体と簡易慣用字体の対応表
人名用漢字 jinmei-variants.txt	法務省・人名漢字表（2010年現在）の、別表２の１で同一の字種とされる漢字および人名漢字表 別表２の２で規定される異体字
JIS X 0213関係字 jisx0213-variants.txt	JIS X 0213 「関係字」
JIS X 0212関係字 jisx0212-variants.txt	JIS X 0212 「関係字」
常用漢字表異体字 joyo-variants.txt	文科省・常用漢字表に記載された異体字（2010年現在）
日本語の書き換え文字 jp-interchangeable.txt	『国語審議会の建議と報告』（昭和三十一年七月五日発表）のうち「同音の漢字による書きかえ」について（報告）に基づく一覧表。拡張新字体を含む。
繁体字・簡体字対応表 kdp-simplified.txt	簡化字総表・第一批異体字整理表・通用規範漢字表・GB 2312/7589/7590に基づく簡体字・繁体字の対応表
漢字・非漢字対応表 non-cjkui.txt	UCSにおける非漢字・擬似漢字と漢字の対応表
相対異体字一覧表 positional-variants.txt	漢字の構成部品の配置の差による異体字の一覧表
部首一覧表 radical-variants.txt	漢字の部品化字形と元字の対応一覧表
原規格分離一覧表 ucs-scs-variants.txt	ISO/IEC 10646 Annex S.3 例字一覧
JIS X 0213/0212 異体字 x0212-x0213-variants.txt	JIS X 0213 附属書11箇条3.2 に基づく異体字
第一批異体字整理表 yyb-variants.txt	第一批異体字整理表に基づく異体字一覧。1956年、1986年、1988年、1993年、1997年の改正データ含む。
漢字データベース異体字表	上記のいずれにも属さない関連字の一覧表

Copyright: Kanji Database Project. For contact: mailto:kanji-database-contact at sourceforge.net.

異体字データベース

- 次の説明あり。
  - 異体字データベースは、各種文献に記載されている異体字・関連字情報を収集・整理したものです。本データの一部は、情報処理学会 試行標準 IPSJ-TS 0008:2007 「大規模漢字集合の異体字構造」に記載されている異体字情報を整理、バグフィックスしたものです。

➢ 「JIS X 0213関係字」が使えそう。

  - https://github.com/cjkvi/cjkvi-variants/blob/master/jisx0213-variants.txt

JIS X 0213 関係字の概要

- JIS第1～4水準に含まれる漢字の、2字1組の異体字ペアが、2127字について1320ペア示される。

関係字数	関係字の例	関係字グループ数	グループ内ペア数
2字	笉籅／篇籠／垈墻／幮幮／紵纑／鼆鼀／一弐／万萬／丈丞／与與／丑刄など	882	1（$_2C_2$）
3字	厩廏廐／栄榮荣／箇个ヶ／界畍堺／概槩概／蓋葢盖／学學孝／潟瀉泻など	90	3（$_3C_2$）
4字	焔焰焰焔／秋龝穐烋／畳疊疊疊／体體躰體／鉄鐵銕鐵／島嶌嶋島など	16	6（$_4C_2$）
5字	弁辨辯辨瓣／熙熙熙凞凞／畝畝畝畝畝	3	10（$_5C_2$）
7字	剣劍劒釼劒劒劒／崎嵜埼碕嵜崎碕	2	21（$_7C_2$）

JIS X 0213 関係字の概要

- IPSJ-TS 0008:2007 「大規模漢字集合の異体字構造」のうち、「JIS国内規格 参照異体字」に該当するものと思われる。
  - 選定資料は「JIS X 0208:1997及びJIS漢字字典」「JIS X 0212:1995」「JIS X 0213:2004」 http://www.itscj.ipsj.or.jp/ipsj-ts/ts0008/main.htm（現在はリンク切れ、internet archiveで閲覧可能）

表4.3 異体字関係一覧

番号	名称	参考資料（表4.2における番号）
[1]	JIS X 0208規格 1983年改正に伴う新旧字	1
[2]	JIS国内規格 参照異体字	1，2，3

表4.2 異体字関係選定資料

番号	資料名
[1]	JIS X 0208:1997及びJIS漢字字典（日本規格協会，1997年）
[2]	JIS X 0212:1995
[3]	JIS X 0213:2004

補足：JISコードの歴史

1978	JIS X 0208制定（78JIS）	：第一第二水準が定まる。当時の規格名はJIS C 6226。
1983	JIS X 0208改正（83JIS）	：字体変更・区点位置の字体の入れ替え等。
1990	JIS X 0208改正（90JIS）	：凜・熙の追加。
1990	JIS X 0212制定	：JIS補助漢字が定まる。
1997	JIS X 0208改正（97JIS）	：包摂基準の明確化など。
2000	JIS X 0213制定	：第三・第四水準が定まる。
2004	JIS X 0213改正	：表外漢字字体表制定に伴う補訂。UCS互換10字追加。
2012	JIS X 0208, 0213改正	：常用漢字表の改訂に伴う補訂。

JIS 参照異体字とは

- JIS X 0208:1997（97JIS）の附属書6「漢字の分類及び配列」に、次の項目がある。

n) **参照区点 (参考)**　他に参照すべき区点位置があるとき，それを示す。

16 区 17 点～

16-17	葦	3031 88af b0b1 8466	140 艸 [9]10	[S6883]	[M31437]82,120	あし，イ 地
16-18	芦	3032 88b0 b0b2 82A6	140 艸 4	[S7065]	[M30716]5,82	あし，ロ 地 →73-35 蘆 -78/4 芦 78/4- 芦
16-19	鯵	3033 88b1 b0b3 9BF5	195 魚 8			あじ，ソウ 地 →82-45 鯵 (78 82-45)
16-20	梓	3034 88b2 b0b4 6893	75 木 7	S3637	M14845	あずさ，シ 地
16-21	圧	3035 88b3 b0b5 5727	32 土 2	S1339	M4879’	［常］アツ →52-58 壓
16-22	赦	3036 88b4				

JIS 参照異体字とは

- JIS X 0213:2000も同様の説明。

p) **参照面区点 (参考)**　他に参照すべき面区点位置があるとき，それを示す。この欄は，参考であって規定の一部ではない。

1-14-2 2e22 aea2 87a0 (AAA2)	丈	1（一）3	26 19	26 11	○**セイ** 洒落本・傾城買指南所「丈も大ていじゃ」・辞事典 **人名** **エ** 八丈野 (ヤエノ・姓) **ジョウ** 権丈(ケンジョウ・姓) **タケ** 丈夫 (タケオ) **タケシ トモ** 丈浩 (トモヒロ) **ヒロ** 千丈(チヒロ) **マス** 丈夫 (マスオ) **マスラ** 丈夫 (マスラオ) ◇丈夫・教科書 (人名) **地名** **ジョウ** 岩手県千丈(センジョウ) →丈(1-30-70)	丈
1-14-3 2e23 aea3 87a1 3402	𡈼	1（一）5	37	45	○「「喜」の字の草体「𡈼」が「七十七」と分解できるところから」・辞事典 **人名** **キ** 𡈼代子 (キヨコ) **ヨシ** 𡈼則 (ヨシノリ) ◇𡈼代寿・新聞 (人名) →喜(1-20-78), 憙(1-56-58)	𡈼
1-14-4	丨	2（丨）0	67	75	○コン「沒麻和尚一人集」	

JIS 参照異体字とは

- JIS C 6226-1978（JIS X 0208の78JIS）では「異体字」と明記。

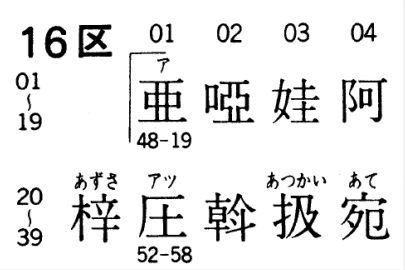
第１水準 又は 第２水準の中で，異体字が近接して配置される場合には，親字と認めるものの位置を定め，その直後にその異体字を配列する。

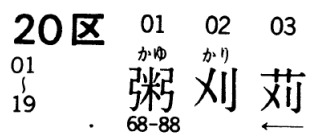
参 考 漢字表では,わく外の区点番号とわく内の漢字だけが規格の一部であり，それ以外の数字，記号，文字等は，規格の一部でない。

第１水準については，それぞれの漢字に目印として代表音訓を付した。音は片仮名，訓は平仮名である。ただし，同音 又は 同訓のものは，先頭のものにだけ付した。また，代表音訓以外に思いつきやすい別音訓の漢字とその区点番号を，ページの下端に示した。

漢字の左上に付したかぎ“「”は,五十音の頭文字 又は 見出し形の 214 部の変わり目を示す。この場合，その行の右欄外に，その頭文字 又は 見出し形を示した。

親字の直後にその異体字がある場合には，異体字の下に矢印を付した。異体字が他の箇所にある場合には，相互参照の区点番号を付した。



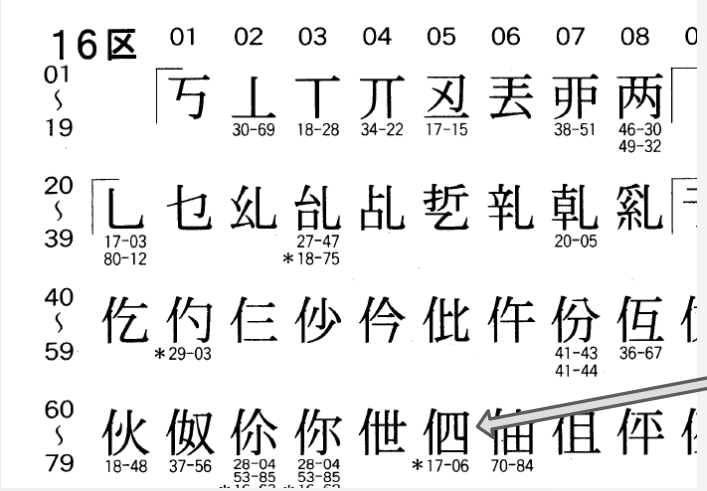


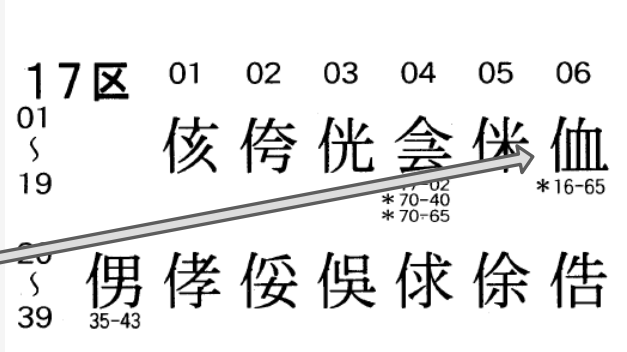
- ※78JISと97JISとの間には異同もある。
  - 例）「澄」と「澂」は、78JISでは参照関係にないが、97JISでは参照関係に。

JIS 参照異体字とは

- JIS X 0212:1990 では「同義漢字」

JIS X 0208及びJIS X 0212に同義漢字がある場合，参照の区点番号を付けた。区点番号だけ示したものは，JIS X 0208の区点番号を示し，＊を記した区点番号は，JIS X 0212の区点番号を示す。同義漢字が四つ以上ある漢字については，欄外にその漢字を示し，漢字の上に当該漢字の区点番号，漢字の下にその漢字の四つ目以降の同義漢字の区点番号を示した。

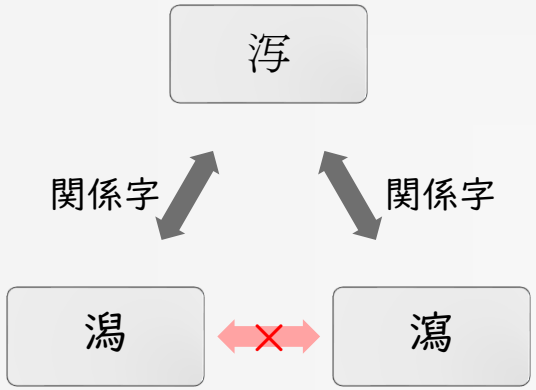




JIS X 0213 関係字とは

- JIS X 0213関係字は、JIS X 0208:1997とJIS X 0213:2004に示された参照面区点に拠って整理したものと思われる。
  - 『増補改訂 JIS漢字字典』に示される「参照」字と、恐らく同じ。
- 全ての組み合わせを示すため、妥当でないペアも若干ある。

```
1796    潟,jisx0213/variant,瀉
1797    潟,jisx0213/variant,泻
1798    瀉,jisx0213/variant,泻
```



今回行ってみたこと

- 「JIS X 0213 関係字」を使って、DHSJRの正規化を試みる。
  - ただし、97JISならびにJIS X 0213が、どの資料に基づいて参照関係を判断したのかは不明。
  - ざっとみた限りでは、一般的な漢和辞典に示される異体字欄と、あまり変わらない印象。

DHSJR見出し字とJIS関係字の対照

- DHSJR_data_all_20240529.txtの「単字_見出し」の異なり字を抽出（6625字）。

字	用例数
生	1467
大	1225
子	1151
法	1135
人	1118
・・・	・・・
・・・	・・・
鮑	1
鰈	1
齷	1
鯀	1
沿	1
計6625字	計164913例

- JIS X 0213関係字と対照

DHSJR見出し字とJIS関係字の対照

- 揺れが出てくる。

	A	B	C	D	E
1	jisx0213/variant_L	jisx0213/variant_R	pair_pattern	L_dhsjr	R_dhsjr
72	冪	冪	1		
73	冽	冽	1		1
74	凄	凄	1	3	1
75	凛	凜	1	3	2
76	凡	凡	1	122	3
77	処	處	1		116
78	函	函	1	24	1
79	刃	刃	1	8	1
80	刈	苅	1		1
81	剋	尅	1	10	3

セル内の値はDHJSRにおける用例数

- どのように正規化するか
  - そもそも一方に寄せていいのか

『広韻』データとの対照

- 「漢字データベース」の「宋本広韻データ」から、見出し字を抜き出し重複字種を削除したデータと対照。

➢ 『広韻』の側に合わせれば、中古音データに接続しやすくなりそう。

- ただし両方ヒットする字（冽冽／凄凄）あり。

	A	B	C	D	E	F	G
1	jisx0213/variant_L	jisx0213/variant_R	pair_pattern	L_dhsjr	R_dhsjr	L_guangyun	R_guangyun
72	冪	冪	1				18555
73	冽	洌	1		1	17537	13427
74	凄	淒	1	3	1	2699	2698
75	凛	凜	1	3	2		8091
76	凡	凡	1	122	3	8590	
77	処	處	1		116		9248
78	函	函	1	24	1	8201	
79	刃	刃	1	8	1	13936	
80	刈	苅	1		1	13917	
81	剋	尅	1	10	3	18843	

宋本広韻データ

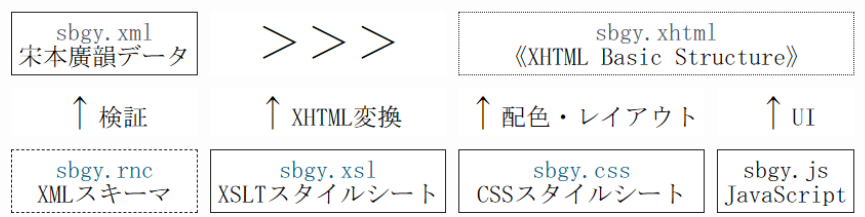
- https://github.com/cjkvi/cjkvi-dict/blob/master/sbgy.xml

宋本廣韻データ

宋本**広韻**データは、周祖謨「校正宋本廣韻」（澤存堂複製※）をUCSで符号化することを目指します。本データは、日本学術振興会科学研究費・基盤研究C「次世代古典文献データベース構築の基礎的研究」（平成14～16年度、課題番号：14510494、研究代表者：村越貴代美）による成果の一部を利用させてもらっています。

宋本廣韻データは、指定された XML スキーマに基いて入力され、XML処理系でデータ処理ができることを目指します。反切部分はマークアップで示され、そこから推定される中古音の例をIPAで示します。親字ＩＤは工藤祐嗣氏の「巻数(1-5)・丁数(2桁数)・表裏面(a/b)・行数・番号」を利用させて頂いています（→参考論文）。

XMLデータのHTML化にあたっては、下図のような構成をとることを考慮しています。



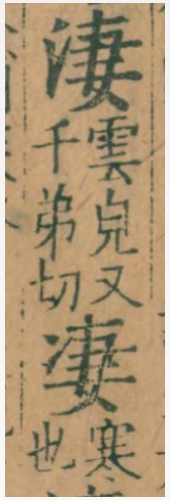
※…清の時代に宋風に彫り直したもの。なお、本データファイルには参考として推定日本語音が付されている場合がありますが、これらは不正確です。日本字音の詳細については、三省堂「言語学大辞典・術語編」の「漢音・呉音・慣用音」の項目や、沼本克明「日本漢字音の研究」(1997、東京堂)などを参照ください。また、宋本廣韻自体が、底本は鉅宋本と大宋本に大別され、それぞれで親字が異なる場合があります（閏と閏など）。本データはそれらの文字を全てカバーしているわけではありませんのでご注意ください。

Copyright: Kanji Database Project. For contact: mail:kanji-databse-contact at sourceforge.net:

「凄」と「淒」

	A	B	C	D	E	F	G
1	jisx0213/variant_L	jisx0213/variant_R	pair_pattern	L_dhsjr	R_dhsjr	L_guangyun	R_guangyun
74	凄	淒	1	3	1	2699	2698

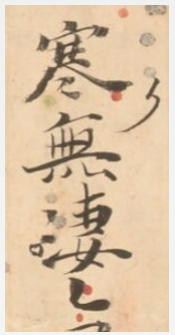
- 中古音は同じだが、意味が異なる。
  - 斉韻清母平声・妻小韻
    - 凄：「雲皃」〈雲のさま〉
    - 淒：「寒也」〈寒い〉
    - ※「淒」に又切（千弟切）があるが、該当する小韻（薺韻清母去声）には見えない。
  - ➢ 『漢辞海』は異体字扱い。『改訂新版 新字源』は別字扱い。

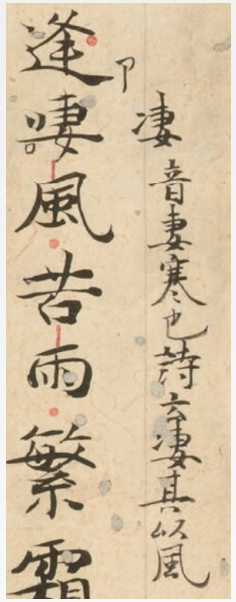


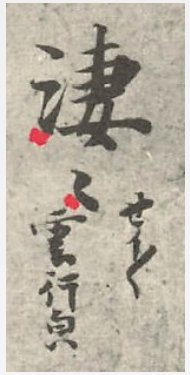
『広韻』

「凄」と「淒」

- DHSJRでは「凄」3例「淒」1例
  - 半井本医心方：巻27「凄々」・巻14「凄風」
    - いずれも〈寒い〉の意味（⇒広韻的には「凄」）
  - 前田本字類抄：「凄々」（実際は「淒」）
    - 〈雲の行くさま〉の意味（⇒広韻的には「淒」）
- 区別するなら字類抄は訂正が必要。
- 区別しなくても、中古音との接続という点では大きな問題はないか。
  - 「又切」を反映したデータに接続する場合は、問題なくもない。







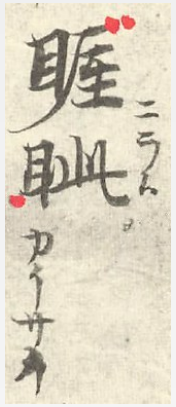
「眥」と「眦」

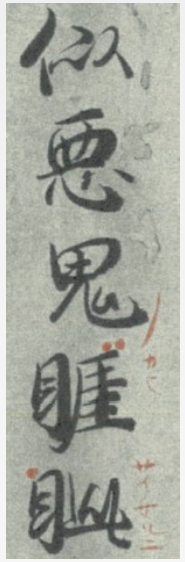
	A	B	C	D	E	F	G
1	jisx0213/variant_L	jisx0213/variant_R	pair_pattern	L_dhsjr	R_dhsjr	L_guangyun	R_guangyun
514	眥	眦	1	1	5	12286	13626

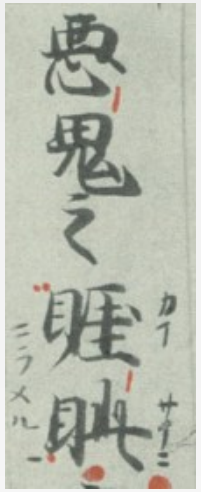
- 異体字関係と認める辞典が多い。
  - 『漢辞海』『改訂新版 新字源』は、いずれも「眥」を親字とする。
- しかし、『広韻』では、中古音が異なる。
  - 眥：寘韻従母・漬小韻「目眥」／霽韻従母・嚌小韻「目際」 →字音シ・セイが期待
  - 眦：卦韻崇母・瘵小韻「睚眦」 →字音サイが期待

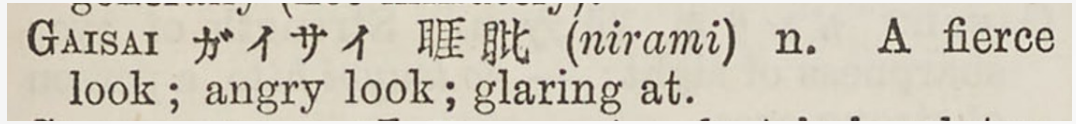
「眥」と「眦」

- DHSJRでは「眥」1例「眦」5例
  - 高山寺本荘子：巻26「眥□女戚」（子斯反・子智反）
  - 前田本字類抄：「睚眦」2例（内1例サイ）
  - 新猿楽記：古抄本「睚眦」康永本「睚眦」（共にサイ）
  - 和英語林集成：「ガイサイ 睚眦」
- ➢「単字_見出し」では、《眥》と《眦》を区別しておきたい。
  - 簡易検索においては、グループ化してよいか。



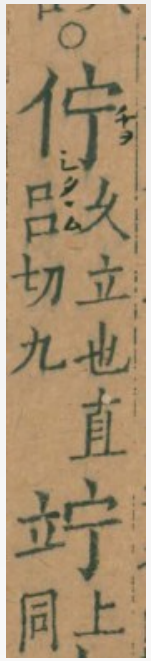





GAISAI ガイサイ 睚眦 (*nirami*) n. A fierce look; angry look; glaring at.

類例

- 『広韻』に共に見えるペア･･･199例
  1. 中古音同音（異体字関係）　79例　例）佇竚
  2. 中古音同音（意味の違いあり）　28例　例）凄淒
  3. 中古音に包含関係あり　42例　例）坂阪
  4. 中古音が異なる　45例　例）眥眦
  5. 広韻データが不審　5例
- 2～4は区別するのも一案。



※本作業にはWeb韻図 [https://suzukish.sakura.ne.jp/search/inkyo/index.php]を使用した。

『広韻』に無くDHJSRに有るケース

- 『広韻』無／DHSJR有：94ペアあり。

	A	B	C	D	E	F	G	H
1	jisx0213/variant_L	jisx0213/variant_R	pair_pattern	L_dhsjr	R_dhsjr	L_guangyun	R_guangyun	dhsjr_or
66	円	圓	1		67			○
67	冊	册	1	1				○
69	取	取	1		1			○
102	卑	卑	1	45	1			○
141	噛	嚙	1	1				○
208	厓	涯	1	3	4			○
250	彝	彝	1		1			○
258	恒	恆	1	1	96			○

➢ 『広韻』非掲出の場合もある（例：噛）が、『広韻』の掲出字体が、JIS第1～4水準にない場合の方が多い。

『広韻』に無くDHJSRに有るケース

関係字ペア	広韻掲出字体
円・圓	圓
冊・册	冊
卑・卑	卑
厏・𢍰	厏
彝・彝	彝
恒・恆	恒

- 「単字_見出し」を広韻掲出字体に改める事はしない方がいいだろう。
  - 今回は保留。中古音データの方を改編して接続するのがよいか。

異体字テーブル（仮）の作成

- ※今回は、「新旧対立のある常用漢字」以外で、かつ、関係字が2字1グループ1ペアのものに限定
- 関係字ペアの両方が『広韻』と一致する場合
  - 字義や字音に差異がみられる場合 ⇒ 今回はテーブルから除外
  - 一方の字が字体注（例：「上同」「古文」）のみの場合
    - 字体注のみの字が、新旧対立のない常用漢字の場合(A) ⇒ 常用漢字体に置換
    - それ以外(B) ⇒ 他方の掲出字体に置換
- 関係字ペアの一方が『広韻』と一致する場合
  - 掲出されない字が、新旧対立のない常用漢字の場合(C) ⇒ 常用漢字体に置換
  - それ以外(D) ⇒ 『広韻』掲出字体に置換
- 関係字ペアの両方が『広韻』と一致しない場合 ⇒ 今回は保留

異体字テーブル（仮）の作成

- 関係字ペアの両方が『広韻』と一致する場合
  - 一方の字が字体注（例：「上同」「古文」）のみの場合
    - 字体注のみの字が、新旧対立のない常用漢字の場合(A) ⇒ 常用漢字体に置換
    - それ以外(B) ⇒ 他方の掲出字体に置換

A : 23例	B : 28例
弌→一, 弍→二, 桒→卓, 卽→即, 譱→善, 妒→妬, 崗→岡, 恠→怪, 桒→桑, 棃→梨, 弃→棄, 况→況, 窰→窯, 黏→粘, 粮→糧, 帋→紙, 緜→綿, 舩→船, 咏→詠, 閈→閉, 隸→隷, 仝→同, 欝→鬱	嘗→嘗, 噪→譟, 墻→牆, 慚→慙, 浣→澣, 狸→貍, 睹→覩, 砧→碪, 竪→豎, 肬→疣, 菱→蔆, 貔→貎, 迩→邇, 竚→佇, 睿→叡, 舵→柁, 槨→椁, 槨→槳, 簷→檐, 燻→熏, 貎→猊, 皋→皐, 笋→筍, 譌→訛, 隽→雋, 鶵→雛, 駞→駝, 鵄→鴟,

異体字テーブル（仮）の作成

- 関係字ペアの一方が『広韻』と一致する場合
  - 掲出されない字が、新旧対立のない常用漢字の場合(C) ⇒ 常用漢字体に置換：190例

丈→丈, 丗→世, 亊→事, 亰→京, 做→作, 侮→侮, 仐→傘, 仂→働, 僧→僧, 兔→免, 冐→冒, 凡→凡, 刄→刃, 苅→刈, 刋→功, 勗→勇, 勉→勉, 勢→勢, 勤→勤, 卆→卒, 咒→呪, 哲→哲, 喝→喝, 嘆→嘆, 囘→回, 埀→垂, 塚→塚, 増→増, 墨→墨, 夛→多, 梦→夢, 姙→妊, 宇→宇, 宜→宜, 冨→富, 寛→寛, 屍→尻, 層→層, 峯→峰, 崕→崖, 巣→巣, 帆→帆, 幣→幣, 廊→廊, 往→往, 徵→徵, 耻→恥, 悔→悔, 懼→惧, 慨→慨, 憎→憎, 憇→憩, 懲→懲, 拘→拘, 掲→掲, 撃→撃, 敏→敏, 旨→旨, 舜→昇, 朗→明, 暑→暑, 暦→暦, 曹→曹, 杁→杉, 棊→棋, 横→横, 橋→橋, 欄→欄, 歩→歩, 歴→歴, 殺→殺, 毎→毎, 冰→氷, 决→決, 冲→沖, 沒→没, 恭→泰, 海→海, 添→添, 渇→渇, 渉→渉, 溫→温, 準→準, 漢→漢, 瀨→瀬, 燈→灯, 爵→爵, 狀→状, 犹→猶, 珎→珍, 瑠→瑠, 瓶→瓶, 甼→町, 畧→略, 踈→疎, 痩→痩, 瞭→瞭, 碑→碑, 社→社, 祈→祈, 祉→祉, 祖→祖, 祝→祝, 神→神, 祥→祥, 禍→禍, 福→福, 穉→稚, 穀→穀, 突→突, 窗→窓, 節→節, 篭→籠, 糺→糾, 縲→累, 緑→緑, 緒→緒, 練→練, 縁→縁, 繁→繁, 罰→罰, 署→署, 羣→群, 者→者, 畊→耕, 脉→脈, 腸→腸, 臭→臭, 舍→舎, 著→著, 薫→薫, 乕→虎, 虜→虜, 蚕→蚊, 裡→裏, 褐→褐, 規→規, 視→視, 觧→解, 諸→諸, 謁→謁, 謹→謹, 戝→財, 贊→賢, 貭→質, 贈→贈, 赱→走, 踴→踊, 軌→軌, 辛→辛, 返→返, 廹→迫, 迯→逃, 逸→逸, 達→達, 郎→郎, 部→部, 鄉→郷, 都→都, 釜→釜, 錬→錬, 錄→録, 镸→長, 陰→陰, 隆→隆, 鄰→隣, 雄→雄, 難→難, 韵→韻, 響→響, 頻→頻, 賴→頼, 顔→顔, 類→類, 凧→風, 飲→飲, 餝→飾, 舘→館, 麻→鹿, 黄→黄, 黑→黒, 皷→鼓, 剥→剝, 叱→叱, 填→塡, 宍→肉, 涌→湧, 瘉→癒, 蹤→踪, 頬→頬,

異体字テーブル（仮）の作成

- それ以外(D) ⇒ 『広韻』掲出字体に置換：304例

刄→丑, 亙→亘, 烝→亟, 仭→仞, 侫→佞, 俎→俎, 傚→傚, 儕→僊, 儔→儔, 儕→儕, 儼→儼, 兇→兇, 寃→冤, 函→函, 尅→剋, 刼→劫, 勻→勻, 滙→匯, 巵→卮, 庬→厖, 咤→咤, 呕→嘔, 嘯→嘯, 埒→埒, 尭→堯, 塹→墩, 垰→墟, 罎→壘, 弉→奘, 嫩→嫩, 嫺→嫺, 嫐→嬲, 屏→屏, 屮→屮, 嵇→嵇, 岷→嶧, 嶮→嶮, 厶→幺, 廏→廠, 慾→慾, 憚→憚, 瀟→瀟, 戞→戛, 扨→扠, 拿→拏, 掎→掎, 撈→撈, 掩→撿, 擣→擣, 斷→斷, 昺→昞, 晄→晃, 曦→曦, 曳→曳, 槓→杠, 桍→栲, 櫁→樒, 橈→橈, 槖→槖, 蘖→糵, 栾→欒, 欄→欄, 氈→氈, 涅→涅, 泮→淬, 渚→渚, 澑→溜, 滔→滔, 滲→滲, 潛→潸, 瀦→瀦, 灞→灞, 灵→灵, 烱→炯, 煢→煢, 燮→燮, 烬→燼, 牖→牖, 犈→犢, 琢→琢, 瑤→瑶, 瓚→瓚, 瓯→甌, 磚→甎, 瘘→瘻, 皹→皸, 蘊→盪, 睨→睨, 瞞→瞞, 瞼→瞼, 矚→矚, 砿→礦, 祐→祐, 祿→禄, 禎→禎, 粃→秕, 頴→穎, 潼→窪, 筐→筐, 筹→籌, 粼→粼, 紘→紘, 縉→縉, 纘→纘, 纏→纏, 綵→纓, 纜→纜, 羌→羌, 菱→菱, 翠→翠, 耋→耋, 胫→脛, 臍→臍, 卧→臥, 舐→舐, 芠→艾, 蒭→芻, 萌→萌, 蕚→萼, 蔥→葱, 蕭→蕭, 蕁→薺, 蘓→蘇, 蜚→蛾, 蝼→螻, 蟒→蟒, 蠏→蟹, 衂→衄, 裵→裴, 襄→襄, 覓→覓, 覷→覷, 誑→誑, 讚→讃, 蹈→蹈, 踌→躊, 躪→躪, 輙→輒, 隶→轟, 轢→轢, 軻→轤, 逈→迥, 遙→遥, 邃→邃, 鋏→鋏, 鋤→鋤, 鈬→鐸, 鑽→鑽, 陋→陀, 隀→隴, 頤→頤, 飃→飄, 飧→飧, 餞→餞, 饒→饒, 馹→驢, 髯→髯, 鮒→鱸, 鶇→鶇, 鷆→鷆, 麼→麼, 齔→齔, 齟→齟, 齬→齬, 徧→徧, 櫥→櫥, 紵→紵, 蠅→蠅, 乍→倅, 俠→俠, 倏→倏, 俱→俱, 兌→兌, 兎→兎, 冦→寇, 冪→冪, 凛→凜, 厠→廁, 厦→廈, 厨→廚, 厮→廝, 呑→呑, 咯→喀, 唖→啞, 嘩→譁, 嚏→嚏, 嚢→囊, 壟→壠, 壷→壺, 奩→匳, 妍→妍, 孽→孽, 寮→寮, 尤→尤, 屡→屢, 弗→岪, 岑→岭, 峨→峩, 崑→崐, 崙→崙, 廸→迪, 廼→迺, 弯→彎, 忰→悴, 懴→懺, 抬→擡, 掴→摑, 掻→搔, 摠→摠, 撹→攪, 旙→旛, 昿→曠, 晋→晉, 栢→柏, 档→檔, 桧→檜, 梼→檮, 棕→椶, 楕→橢, 槙→槇, 槞→櫳, 檃→檃, 檠→檠, 歹→歺, 歿→歿, 殲→殲, 毘→毗, 涛→濤, 涜→瀆, 潅→灌, 灩→灩, 熔→鎔, 犁→犂, 犲→豺, 猯→貒, 獏→貘, 珱→瓔, 瑙→碯, 畴→疇, 疏→疏, 砺→礪, 碍→礙, 祢→禰, 祷→禱, 禀→稟, 穣→穰, 笻→筇, 筝→箏, 箆→篦, 箪→簞, 嵌→嵌, 籤→籤, 紵→紵, 紘→紘, 繋→繫, 繍→繡, 羹→羹, 聊→聊, 胚→胚, 腮→顋, 舮→艫, 芦→蘆, 茘→荔, 莱→萊, 莵→菟, 萸→萸, 蒂→蒂, 蒋→蔣, 夔→夔, 虱→蝨, 蛎→蠣, 蝉→蟬, 蝋→蠟, 蝿→蠅, 蟇→蟆, 蠹→蠹, 禅→禪, 觸→觸, 諌→諫, 譖→譖, 讐→讎, 賍→贓, 賎→賤, 躯→軀, 酆→酆, 醤→醤, 鈎→鉤, 鈩→鑪, 鍬→鍪, 閙→鬧, 隲→騭, 韮→韭, 韲→齏, 頚→頸, 顛→顚, 騨→驒, 鳬→鳧, 鴎→鷗, 鴬→鶯, 鵞→鵝, 鵁→鶒, 鹸→鹼, 麸→麩, 麹→麴,

異体字テーブル（仮）による置換

- 合計545行。
- DHSJRの異なり見出し字数 6625→6442。
- さらなる改善が必要。

今回触れなかったもの

- JIS第1～4水準外字
  - 最初に取り上げた「教」もJISに無いため、今回は置換できていない。
  - ➢個別に判断し、異体字テーブルに追加していく必要あり。
- 複数字種を中和する字体
  - 《潟》《瀉》に対する「泻」など。

複数字種を中和する字体の問題

	A	B	C	D	E
1	jisx0213/variant_L	jisx0213/variant_R	pair_pattern	L_dhsjr	R_dhsjr
79	刃	刄	1	8	1

- DHSJRにおける「刄」1例は、《叉》に相当
  - 正保版魚山私抄九方便「縛曰羅薬叉」
- 刄：《叉》《刃》双方の異体字として出現（山本2016）
  - ➢「単字_見出し字」で特例的に「叉」に校訂すべきか？

複数字種を中和する字体の問題

- 全データから同様の例を洗い出すのは困難。
- 検索システム側へ実装するのも一案か。
  - (刃|刄|叉)・(瀉|泻|潟)などを通用グループとして、検索可能にする。
  - 高橋久子「室町時代の文献に見られる漢字の通用現象に就いて」(1)(2)「南北朝時代の文献に見られる漢字の通用現象に就いて」(1)(2)「鎌倉時代の文献に見られる漢字の通用現象に就いて」(1)(2)「平安時代の文献に見られる漢字の通用現象について」(1)(2)（『東京学芸大学紀要2 人文科学』46～55）などが参考になるか。

参考文献

- 芝野耕司（2002）『増補改訂 JIS漢字字典』日本規格協会
- 矢野啓介（2019）『改訂新版 プログラマのための文字コード技術入門』技術評論社
- 山本秀人（2016）「「靭」「靱」／「靱」／「靱」—「強靭」の「靭」—」『国語文字史の研究十五』和泉書院
- 文化庁文化部国語課（1996）『漢字字体関係参考資料集 JIS情報交換用漢字符号（第一分冊）』
- JIS規格票（JIS X 0208:1997, JIS X 0213:2000）

ご清聴ありがとうございました。