近代語文献を電子化するための異体字処理

近代語文献を電子化するための異体字処理

須永　哲矢（国立国語研究所コーパス開発センター）[1]

１．はじめに

文書の電子化にあたっては、もとの文書のどの要素をどこまで再現し、どの要素は再現できなくても良しとするかという処理方針を定めねばならないが、それはその電子化テキストの使用目的による。漢字の字体字形の問題一つをとっても、各字形差を可能な限り正確に表現した方が望ましいとは限らない。「言語研究用のコーパス作成」という場面においては、電子テキスト化はゴールではなく、あくまで研究の手段としての環境整備、という位置づけとなる。言語研究の素材として使用される電子テキストは、言語資料として「読める」こと、語彙等のサンプルが採集できることが重要となる。そのため、外字として処理された文字が多く、表示上「〓」ばかりで「読めない」テキストや、動作環境によっては適切に表示されない文字が含まれるテキスト等は望ましくない。

そこで、近代語コーパス構築の基礎研究としての本研究では、近代語コーパスの試作となる『明六雑誌』電子テキスト化の作業を通じ、言語研究の実用に適した文字処理の在り方を模索することとした。

２．JIS X0213 文字集合と包摂規準

近代語コーパスの試作としての『明六雑誌コーパス』は、JIS X 0213 文字集合に準拠して電子化することとした。『現代日本語書き言葉均衡コーパス』も JIS X 0213 を依拠する文字集合として文字処理が行われ、およそ 5,800 万字の現代日本語コーパスでは、のべ 99.96%の文字が、JIS X 0213 で表現できることが確認されている（高田ほか 2009）。

JIS 規格では、字体字形の差を処理するために「包摂規準」が定められている。JIS X 0213 では連番で１９９の包摂規準が設定されており、包摂規準の範囲内の差異であれば、同一の符号位置の文字として処理することになる。これにより、明確な規準のもとでの文字処理が可能となる。

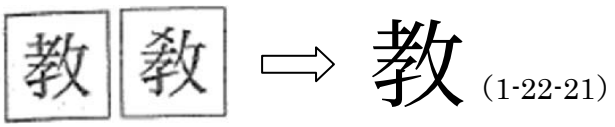


図１　包摂規準の例（連番８）

しかし時代をさかのぼって、近代以前の活字資料を対象とした場合にも JIS X0213 文字集合および包摂規準が有効であるかの検証はいまだなされていない。そこで、明治前期の雑誌である『明六雑誌』のコーパス試作を通じ、JIS X0213 の有効性と限界を見極めたうえで、近代語に適した文字処理方針を構築していくというのが今回の課題である。

３．『明六雑誌』漢字処理上の問題

『明六雑誌』を JIS X0213 に準拠して文字処理していく際、字形処理の実際において問題となるのは大きく分けて次の２つのケースである。

（A）文字集合（ここでは JIS X0213）に含まれない字

図２　『明六雑誌』に出現する JIS 規格外字

[1] tsunaga@ninjal.ac.jp

図２のような文字は、JIS X0213 では用意されておらず、表現することができない。

（B）通用字形とは（僅かな）字形差があるもの

図３『明六雑誌』に出現する「序」「万」「除」の字形（右側）

図３のように、近代の活字では、それが現在の通用字のどの字に当たるかは明らかであるが、字形差があるものが多数見受けられる。JIS 規格では包摂規準が定められているが、図に示した「序」「万」「除」の字形差に関しては、既存の包摂規準の中には明確に適用できるものがない。そのため、既存の包摂規準のみに従って処理していく場合、これらは外字となり、「〓」表示されることになる。

４.『近代語コーパス』のための文字処理方針

近代の活字においては、図３に示した『明六雑誌』での活字のように、既定の包摂規準では包摂してよいのかが明示されていない、わずかな字形の差がある場合が多く見られる。

これらを逐一外字として処理していくと、できあがった電子テキスト内の外字が増え、言語研究資料として実用に供さないものになりかねない。表１に示すとおり、『明六雑誌』の漢字字形に対し、JIS X 0213 の文字集合・包摂規準を適用した場合、その処理だけでのべ約 98.5%が表現可能となる。しかし、言語研究資料としてみた場合、200 文字のうち 3 文字が読めない電子テキストは実用に供さない。

表１　JIS X0213 文字集合・包摂規準を適用して『明六雑誌』の漢字を処理した結果

文字区分	のべ字数	異なり字数
JIS X 0213	135,797	3,218
第 1 水準漢字	117,643	2,066
第 2 水準漢字	17,953	1,061
第 3 水準漢字	118	52
第 4 水準漢字	83	39
外字	2,100	99
計	137,897	3,317
カバー率	98.48%	97.02%

また、『明六雑誌』に出現する字形は、現行の包摂規準だけを拠り所とすると、そのままでは包摂できないものが多く出現するが、その大部分は、現在の通用字体のどれに相当するかは類推でき、字形の差異もわずかなものである。

図４の「〓」表示の内実は、「時」「華」「改」の異体字である。JIS X0213 は、図中の丸囲みのような差異を包摂できる基準を持ち合わせていないため、規格以外字としての扱いとなる。

しかし、このような処理は、JIS X0213 の適用の仕方としては厳密であるが、「〓」表示になった時点で用例としては取り出せなくなってしまうため、用例検索や語彙調査といった、コーパスとしての実用面からは有用性の低い処理になってしまう。むしろ実用面からは、JIS X0213 の適用の仕方が多少ゆるくとも、これらも「時」「華」「改」に包摂し、文字として表示した方が望ましい。

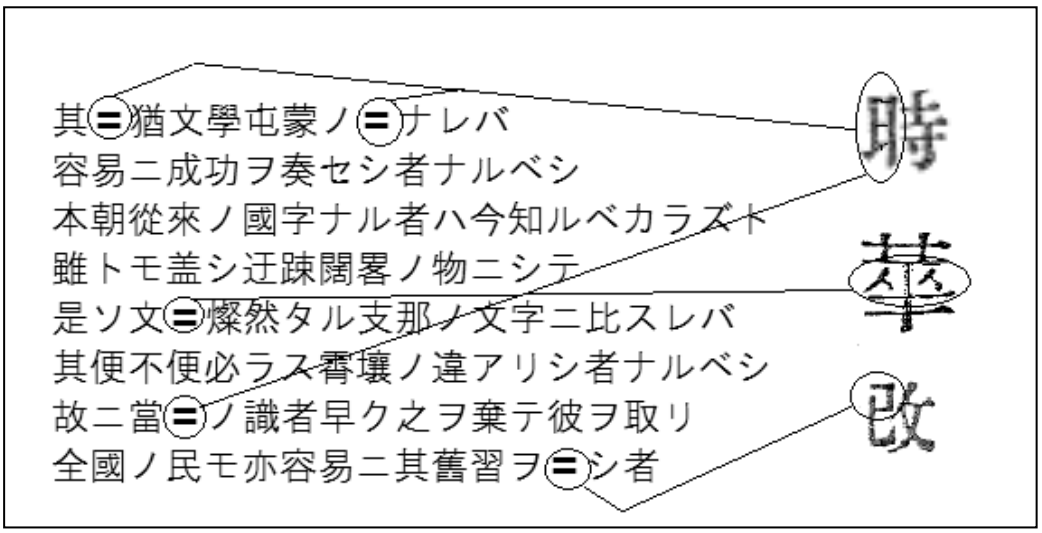


図４　JIS X0213 文字集合・包摂規準を厳密に適用した電子テキスト化の例

「言語研究用コーパス」という目的から求められる漢字処理方針とは、一言で言ってしまえば、可能な限り「〓」表示を減らすこと、つまり、可能な限り読める文字として表現すること、である。しかし、だからといって場当たり的に使える文字を当てていく、というだけでは、作業者によって処理の揺れも生じるうえに、どれが本来の JIS X0213 の範囲で処理したもので、どれが臨時的に処理したものかも後々わからなくなってしまう。そこで、本来の JIS X0213 の範囲を越えた処理をする際には、近代語用の処理基準を設けて、データ上にもタグとして記録を残しておくこととした。

本来 JIS X0213 では外字処理になってしまう文字をもなるべく読める文字として表現する、という目的のもと整備した方針は、大きく以下の２つである。

（１）既存の包摂規準に、近代語用の包摂規準を追加する。

まず、図３、４で示したような近代語特有の差異をカバーするため、既存のＪＩＳ包摂規準に加え、近代語用の追加包摂規準を新設し、その基準に従って字体包摂を行うことで、外字処理を減らす。近代語用に追加した包摂規準によって包摂処理された文字に関しては、タグの形で追加包摂規準により処理されたという情報を埋め込んでおく。

（２）包摂規準の追加では対処にくいものに関しては、別字で代用する。

差異がありすぎる等の理由で、包摂規準の追加では対処しにくい文字に対しても、類似の読みや用法がある文字がある場合、その文字で代用することでコーパス上に表現する。このような代用字に関しても、本来は外字であり別字で代用した、という情報をタグの形で埋め込んでおく。また、どの字をどの字で代用したかの一覧を作成して管理する。

この２つの処理を通して、「〓」表示を極力減らしていくことで、コーパスとしての有用性を高めていけると考える（図５参照）。以上のように追加包摂・別字代用という二つの方策で近代語資料での文字を表現していくという処理は『太陽コーパス』でも採られており、『太陽コーパス』では追加包摂により約３００字、別字代用により約２００字（ともに異なり字数）を処理したという実績がある。ただし、『太陽コーパス』では追加した包摂規準は明示されておらず、実際にどのような字形差を、どのような追加規準で包摂したのかを追跡することはできない。また、別字代用に関しては、情報抽出用アプリケーション『プリズム』を利用して外字一覧を生成することで代用字を閲覧することは可能ではあったが、異なり１０００字を越える「〓」表示の外字とあわせての表示となり、代用情報だけを得るにはやや不便であった。そこで今回の『明六雑誌コーパス』では、追加包摂、

別字代用の処理を行った文字に関してはタグ付けを行い、文字処理の情報を取り出せるようにするとともに、追加包摂規準および別字代用の一実態を一覧として公開することとした。

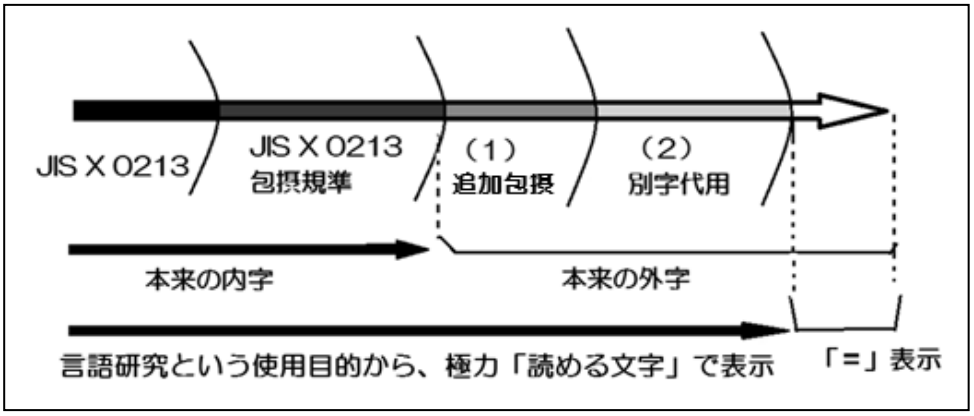


図５　「近代語コーパス」文字処理方針のイメージ

５．『明六雑誌』漢字字形処理方針

近代語文献の文字処理用に追加した包摂規準の詳細、および別字代用の一覧を本節に記す。

５．１ JIS X 0213 文字集合のうち、使用しない領域

今回、『明六雑誌』を JIS X 0213 に準拠して電子化することを試みたが、JIS X 0213 文字集合のうち、使用しない領域を３つ設けたため、ここに記しておく。

①康熙別掲字（104 字）は使用しない。

【例】

×徳(1-84-37) → 徳(1-38-33)を使用

×社(1-89-19) →社(1-28-50)を使用

②UCS 互換字（10 字）は使用しない。

【例】

×叱(1-47-52) → 叱(1-28-24)を使用

×嘘(1-84-07) → 嘘(1-17-19)を使用

康熙別掲字、UCS 互換字は、いわば JIS 包摂規準の例外であり、包摂規準に従うなら、基本的に包摂される字形差である（図６参照）。これらに関しては使用しないこととした。

図６　JIS 包摂規準連番 130、161、78、166

この方針では、本来「德」(1-84-37)で表現できる活字に対しても、包摂規準連番 130 をそのまま適用し、「徳」(1-38-33)として表現することになる。なお、仮に康熙別掲字、UCS 互換字を使用した場合、「德」(1-84-37)と「徳」(1-38-33)がさらに区別されるだけであり、この方針をとらず、康熙別掲字、UCS 互換字まで使用した場合でも、「JIS X0213 で表現される文字の総数」は変わらない。

③CJK 統合漢字拡張Ｂに符号位置が割り当てられる文字（302 字）は使用しない。

【例】

×  (1-15-44、 U+2131B) → 外字扱い

×  (1-15-91、 U+218BD) → 外字扱い

CJK 統合漢字拡張Ｂに関しては、現状では動作環境によっては適切に表示されない等の問題があるため、実用面での判断から使用しない。なお、今回の調査範囲である『明六雑誌』内では、この領域を使えば表現できる漢字は存在しなかったため、この領域を使用した場合でも、『明六雑誌』の範囲内では「JIS X0213 で表現される文字の総数」は変わらない。

５．２ 近代語用包摂基準の設定

JIS X 0213 のうち、上記３領域を除いた文字集合を用いて『明六雑誌』の字形処理を試みることにするが、前述の通り、明治前期の活字字形には、わずかな字形差の活字が多い。それらについては現行の包摂規準には明記されていないものの、感覚的には包摂したいものが多い。そこで、既存の包摂規準を文字処理の規準としたうえで、それに加える形で近代語資料用に包摂規準の拡張案（追加包摂規準）を作成し、字形処理に対応することにした。

近代語での文字処理のため、包摂規準を追加しようという場合、結局のところ、どの程度の字形差までを包摂規準として設定し、どこからを外字とするかが最後まで問題となる。

以下、追加包摂規準の設定のしかた、および追加包摂規準の設定という形では処理しない場合を、具体例と合わせて示す。

５．２．１ 包摂規準を近代語用に追加・修正するケース

（A）既存の基準の明確化

（現行字形）　　（明六雑誌）

(1-43-92)

図７　『明六雑誌』にみられる「万」の字形

このようなパターンについては、漢字字体包摂規準の「ｂ　２点画の接触交差関係の違い」のうち、「抜けるか、抜けないか」のひとつとして処理するという方法が考えられる（図８参照）が、現行の包摂規準内ではこれと完全に一致する字形は示されていない。

このような字形差は、差異の中でも特にわずかな字形差と言いたくなるだろう。漢字の字体字形処理に関しては、JIS 包摂規準以前の前提として、常用漢字表において「デザイン差」とみなされるものは字体の異なりとはしない、という方針があり、そのうち「（４）交わるか、交わらないかに関する例」という例示がなされている（図９参照）。このため、

このような字形差に関しては包摂規準を立てるまでもなく同一字体と処理される、と解釈することもできようが、常用漢字表の「デザイン差」はあくまで例示されるにとどまっており、適用範囲は明確ではない。そこで近代語の電子テキスト化にあたっては、このようなケースに関しても、新たに包摂規準を立て、明確化することとした（図 10）。

図８　包摂規準　連番６０

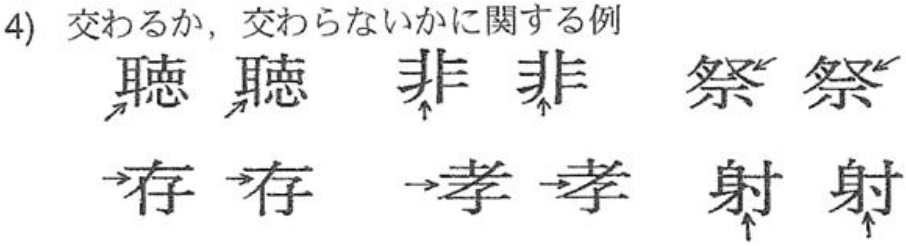


図９　常用漢字表での「デザイン差」字形例

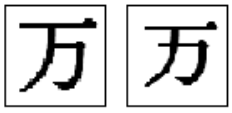


図 10　新設した包摂規準

（B）『JIS 漢字字典』個別字形例を一般化して包摂規準に格上げ

以下のような場合もある。

図 11『JIS 漢字字典』「感」「惑」の個別字形例

字形を包摂するかを判断する手引きとなる『JIS 漢字字典』には、一般規則としての包摂規準のほかに、個別の漢字字体に関して、包摂される複数の字形例が示されている場合がある。図 11 に示した通り、「感」「惑」の「心」の位置の差異に関しては、『JIS 漢字字典』に個別字形例として採られており、この２字に関しては「心」の位置の差異は包摂してよいことになる。近代ではこれ以外の字に関しても類例が見られるため、個別字形例に挙げられている字形差を一般化して包摂規準に格上げした。

図 12　新設した包摂規準

（C）類例を参考に新設

（現行字形）　　（明六雑誌）

(1-29-92)

図 13　『明六雑誌』にみられる「除」の字形

図 13 のような字形差は、現行の包摂規準には明示されていないが、既存の包摂規準「b 2 点画の接触交差関係の違い」のうち、「抜けるか、抜けないか」（図 14 参照）の類例に照らし、図 15 のような包摂規準を新設した。

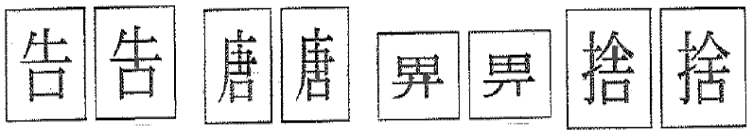


図 14　類例として参考にした既存の包摂規準

図 15　新設した包摂規準

（D）既存の包摂規準に追加

「華」などのの字形差処理のために、包摂規準連番８５が設定されている。（図 16）

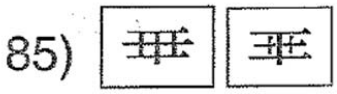


図 16　包摂規準連番８５

『明六雑誌』の「華」「樺」等は、図 17 のような字形で出現する。

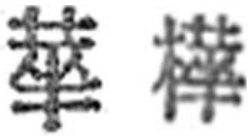


図 17　『明六雑誌』にみられる「華」「樺」の字形

このような字形も包摂するために、包摂規準連番８５の２つの字形に対し新たに図 18 の字形を追加した。

図 18　包摂規準連番８５に追加した字形

（E）既存の包摂規準を統合、より一般化

JIS X0213 では、連番６７、７０、７１のような包摂規準が設定されている。

図 19　包摂規準連番６７、７０、７１

このような字形差に類するとみられる差異で、このままの基準では処理できないものとして『明六雑誌』の「配」「改」「犯」などがある。

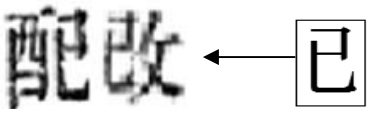


図 20　『明六雑誌』にみられる「配」「改」の字形

図 20 の「配」「改」は「己」の部分が「巳」になっているが、従来の包摂規準では「己」「巳」などの交替を認めても、「巳」は扱われていない。また、図 21 の「犯」は「巳」に交替している例だが、図 22 のとおり、包摂規準連番６７、７０ではこの交替が認められていない。ただし、包摂規準連番７１では、この三者の字形差は認められている。

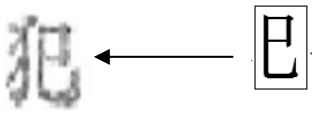


図 21　『明六雑誌』にみられる「犯」の字形

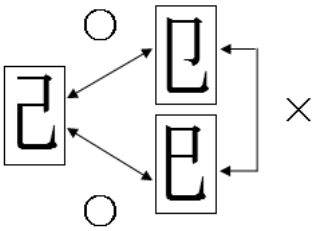


図 22　包摂規準連番６７、７０での交替関係

そこで、このような字形差も処理するために、包摂規準連番６７、７０、７１を統合し、より一般化したうえで、「巳」との交替も認めるという拡張を行った（図 23 参照）。

図 23　連番６７、７０、７１を近代語用に統合・拡張した包摂規準

５．２．２ 追加包摂規準を設定しないケース

（A）部首や部分字形が大きく異なるもの、偏の有無などの差異に関しては、包摂しない。

『明六雑誌』には、図 24、各右側に示すような異体字も出現する。これらのように部首や部分字形が大きく異なるものや偏の有無の違いなどに関しては、包摂規準の新設はせず、包摂しない。これらもコーパス上では、「派」(1-39-41)「脚」(1-21-51)「減」(1-24-26)「輩」(1-39-58)などの通用字で表現するが、包摂（＝同じ字とみなすこと）としてではなく、「別字代用」という形での処理とし、理念上は区別することとする。

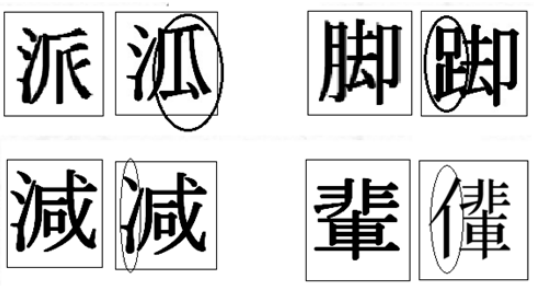


図 24　包摂 ” しない “ 字形例

追加包摂規準を設定して処理する字形差は、もとの JIS 包摂規準に掲げられている、

a）方向・曲直などの点画の性質による違い
b）２点画の接触交差関係の違い
c）２点画の結合分離の違い
d）１点画の増減の違い
e）類型の統合
f）筆法の簡化の違い

という５つに収まる範囲内とする。

（B）Unicode で表現可能な差異は、包摂しない。

将来的な Unicode 対応の可能性を考慮し、理念上は追加包摂規準の設定で処理できそうなものに対しても、Unicode で表現可能なものに対しては包摂はせず、処理上は別字とみなし、「別字代用」として処理する。なお、ここでは Unicode4.0 を参照している。

例えば『明六雑誌』での「跋」は図 25 のようのような字形で出現する。このような部分字形差は、理念上は図 26 のような追加包摂規準の設定を行う方法もありうるが、図 27 のように Unicode4.0 では両者の区別が可能である。このように Unicode では区別可能な字に対しては、将来的な Unicode 対応の可能性を考えると、同一の字にまとめるよりは別字にしておいた方がよいとの判断から、図 26 のような追加包摂規準の設定は行わず、コーパス上は「別字代用」の扱いで処理する。

図 25　『明六雑誌』にみられる「跋」の字形

図 26　追加包摂規準案

(U+8DCB)　(U+47E6)

図 27　Unicode4.0 での表現

５．３ 外字の「別字代用」

異体字のうち、追加包摂規準により同一字とみなされたもの以外は、扱いとしては外字となる。しかし言語研究資料としての使用を考慮した場合、電子テキスト上「〓」表示されていて読めない字というのは可能な限り少ないことが望ましい。そこで、外字認定されたものに対しても、極力「本来は外字であるが、言語研究資料としての使用のため別字で代用する」という手法を取ることにし、追加包摂規準とは別に、別字代用一覧を作成した（後掲）。

言語研究用という使用目的にあわせ「〓」表示は極力減らしたい。そこで、包摂規準の追加・修正で包摂できる字形は包摂し、理念上う、包摂規準レベルでの処理が難しい場合は包摂ではなく、別字で代用、という二段構えで表示上の「〓」表示を減らしていこうという試みである。

５．４ 包摂・代用する字体

本来の JIS X0213 では表現できない文字を、包摂規準の追加・修正または別字代用で表現しようとする場合、図 28 のように使用する文字の候補が複数存在する場合がある。

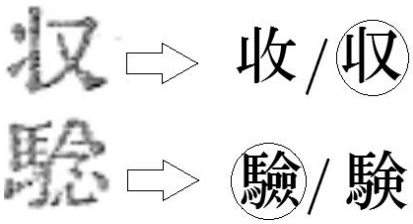


図 28　『明六雑誌』の字形と、包摂・代用候補

このような場合には、以下の方針とした。

①類似点の大きい方を使用する。
②決めかねる場合は正字を使用する。

①により「収」(1-28-93)、②により「驗」(1-81-68)が選択される。

また、別字代用に関しては、言語研究用に「読める」テキストを作成しようというところから出発しており、その目的のためにかなり思い切った代用を行った部分もある。

まず、本来は読み・意味の異なる別字に対しても、コーパス化対象テキストでの使用実態から、代用としての置き換えを認めた場合がある。

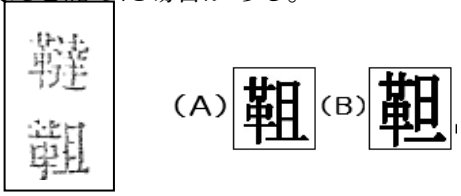


図 29　『明六雑誌』での字形差

図 29 の（A）（ソ）と（B）（タン）は本来読みも異なる別字であり、JIS X 0213 文字集合にあるのは（B）のみ（Unicode では（A）（B）ともに表現可能）である。『明六雑誌』では JIS X0213 外字となる（A）が出現するが、それは上例のように、「韃靼」の「靼」に当たる部分としてである。この使用実態を鑑みて、（A）を（B）「靼」(1-80-58)で代用することとした。

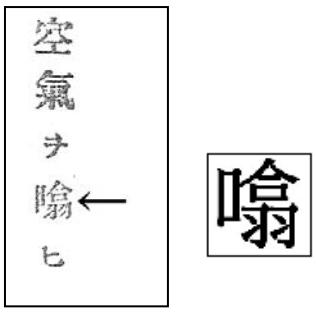


図 30　『明六雑誌』で出現字形

また、字形が大きく異なる場合でも、ほぼ同じ意味・読みの、より一般的な字への置き換えを認めた場合もある。

図 30 のとおり、『明六雑誌』においては動詞「すう」に上掲のような字が用いられる場合があるが、これは Unicode では表現できるものの、JIS X 0213 では表現できない。このような場合にも、言語研究資料として読めることを優先し、「吸」(1-21-59)で代用して処理した。

５．５ 追加包摂・代用した文字の情報管理

追加包摂、別字代用に関しては、まず規則として追加包摂規準、別字代用一覧を整備しておく。さらにコーパスのデータとしては、追加包摂規準で包摂したもの、本来外字であるが別字で代用したものにはそれぞれタグを付け、情報を記述しておく。

【例】
・追加包摂規準により包摂した文字
<包摂>序</包摂>
・代用した文字
<外字　代用=1>減</外字>
・包摂も代用もせず、残った外字
<外字　代用=0>𠮟</外字>

なお、代用としたものの中には、Unicode で表現できる字形も少なくない。そこで将来的な Unicode 対応の可能性も考慮し、Unicode で表現できる代用字に関しては、タグに Unicode 番号を記しておくこととした。

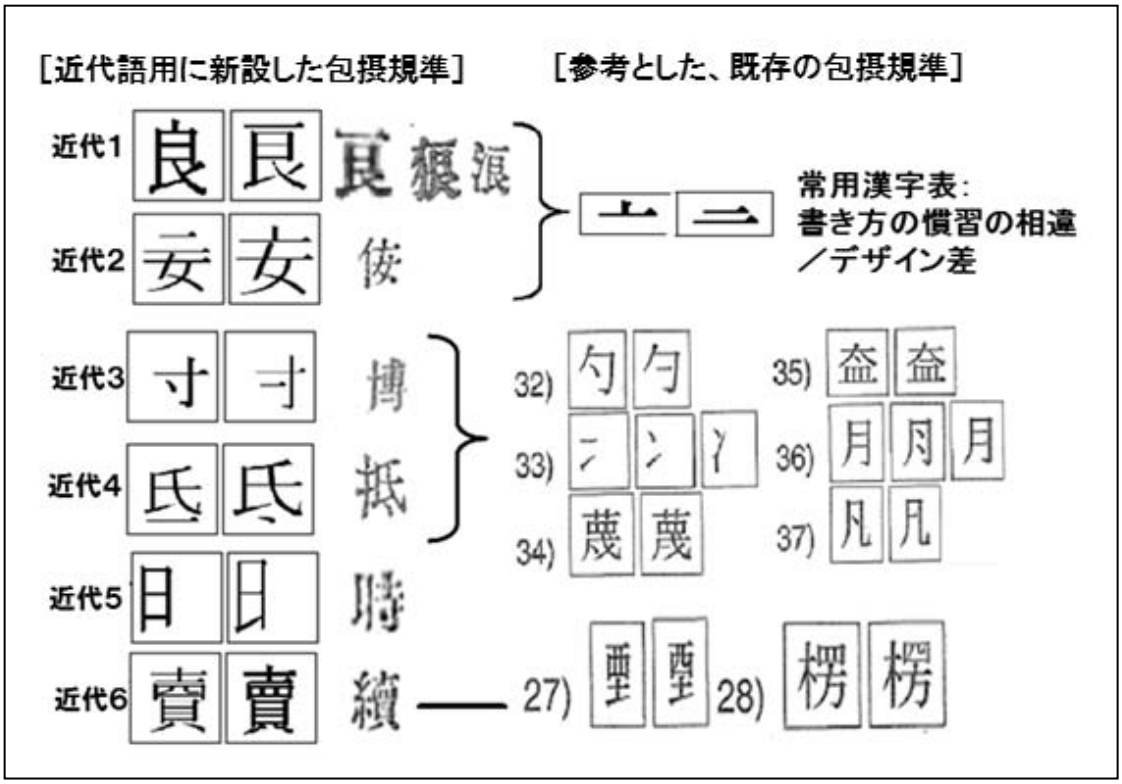
【例】
図 24 の「減」（にすい）、Unicode:51CF
→<外字　unicode="51CF" 代用="1" >減</外字>

６．追加包摂規準・別字代用一覧

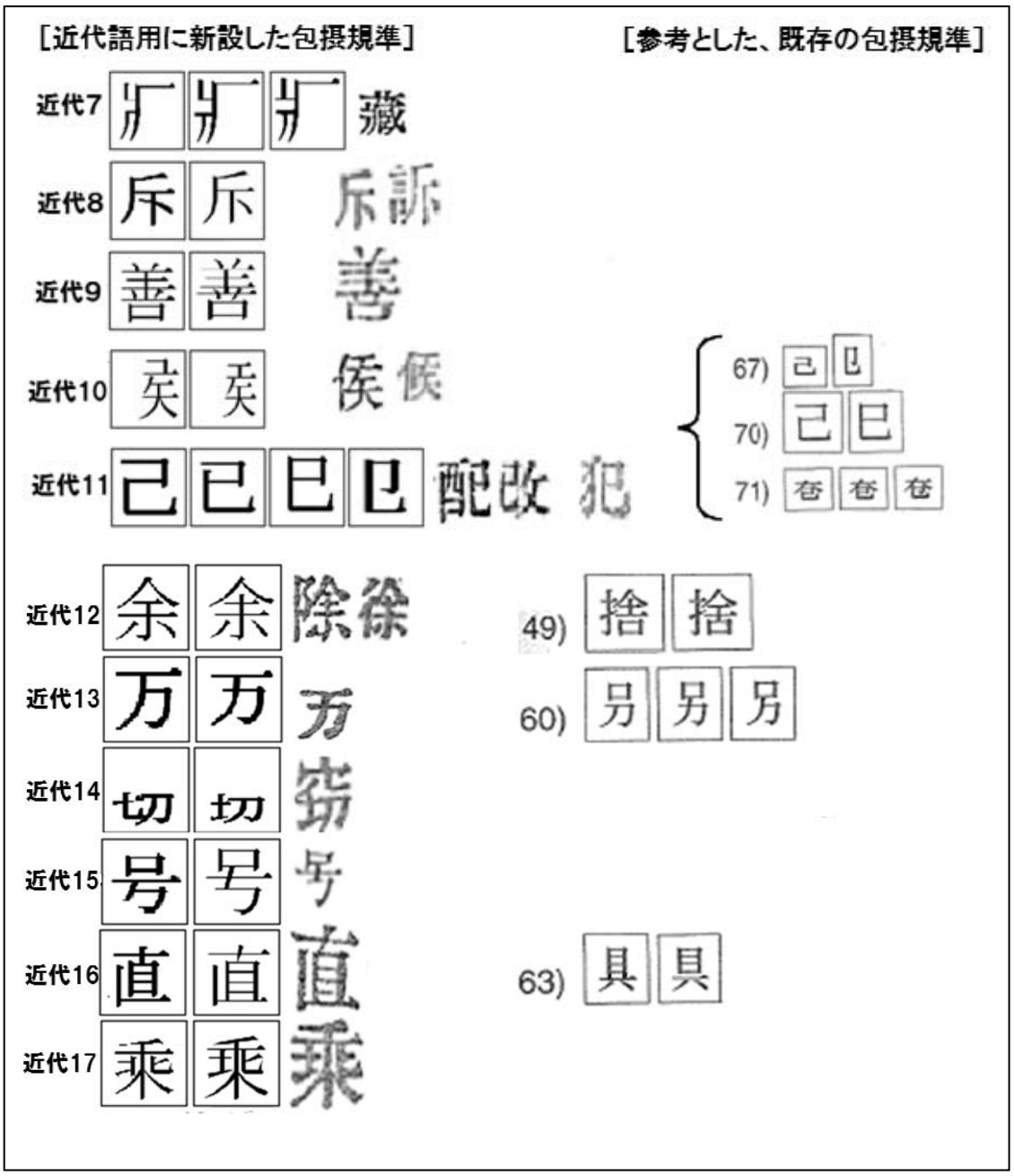
近代語用に新設した追加包摂規準、別字代用の一覧を以下に示す。

６．１　追加包摂規準

ａ）方向・曲直などの点画の性質による違い



b）２点画の接触交差関係の違い

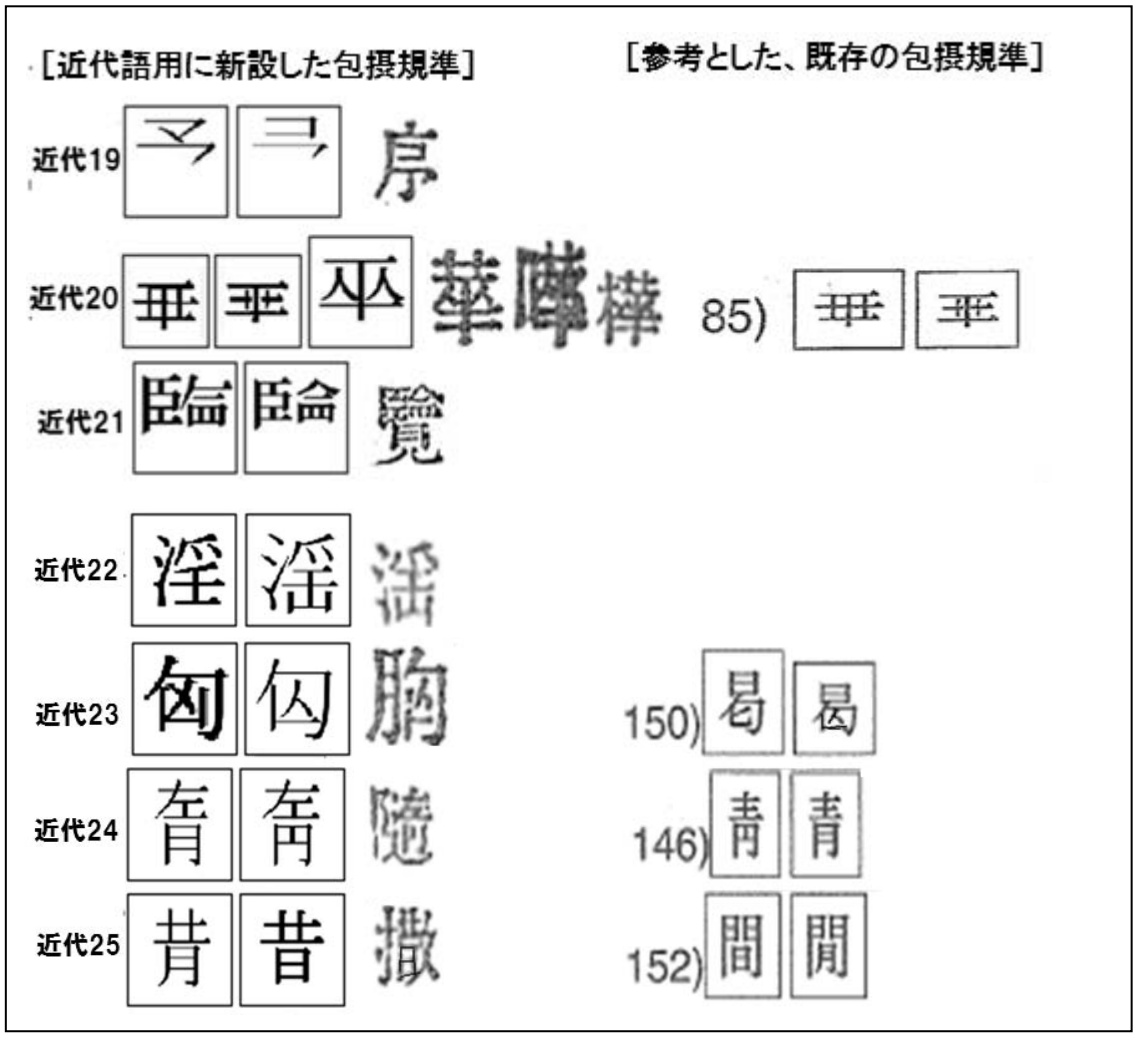


c）２点画の結合分離の違い

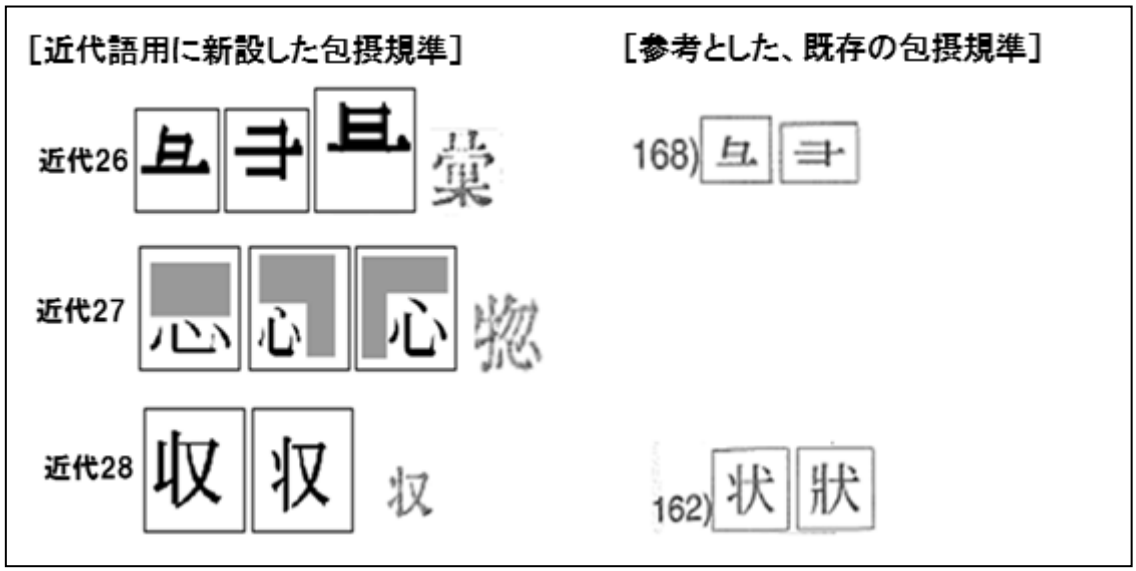
※これに関しては、追加したものはない。

d）１点画の増減の違い

e）類型の統合



f)筆法の簡化の違い



６．２ 別字代用一覧（３８字）

（１）Unicode では表現可能、JIS X0213 では外字（３３字）

（２）Unicode でも表現不可（５字）

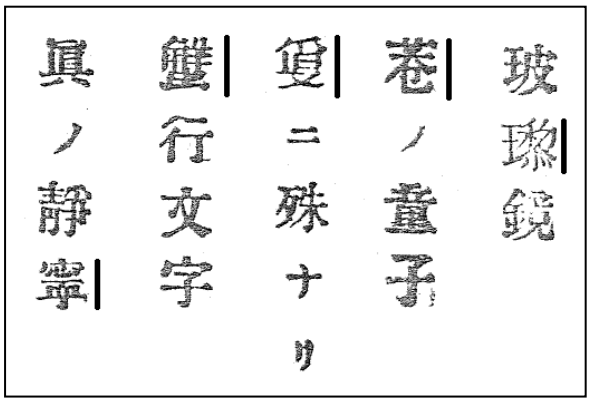


図 31 Unicode でも表現不可の字形

７．X 0213 文字集合／追加包摂／別字代用の検証

以上、包摂規準の追加・修正、外字扱いしたうえで電子テキスト上は代用字を使用する、という方法を考案したうえで、JIS X 0213 規格の包摂規準のみに依拠した場合と、今回提案した処理案を用いた場合とで、処理できる文字数の変化を検証した。

『明六雑誌』に現れる漢字の総数は 137,897、うち JIS X0213 のみで表現できるものは 135,797 字、カバー率にして 98.5%である。残る 1.5%、2,100 字が外字「〓」表示されるテキストとなるが、これは言語研究用の資料としての実用にとっては相当に多い量である。

表２　JIS X 0213 文字集合と『明六雑誌』漢字

文字区分	のべ字数
JIS X 0213	135,797
第 1 水準漢字	117,643
第 2 水準漢字	17,953
第 3 水準漢字	118
第 4 水準漢字	83
外字	2,100
計	137,897

追加包摂規準を適用すると、外字となるのべ 2,100 字のうち 1,774 字、さらに別字代用を適用すると 295 字の処理が可能になり、最終的に「〓」表示となるものは 31 字にまで減少し、99.9%の漢字を表現することができる。これらの処理を通して得られる結果は、割合からすればわずかな差でしかないが、言語研究という要請からは大きな意味を持つとも言える。

表３　各方針の適用で処理可能な文字（のべ）

	X0213 包摂	追加包摂	代用
処理可能文字総数	135,797	137,571	137,866
新たに処理できる文字総数	—	1,774	295
外字総数	2,100	326	31
カバー率	0.98478	0.99764	0.99978

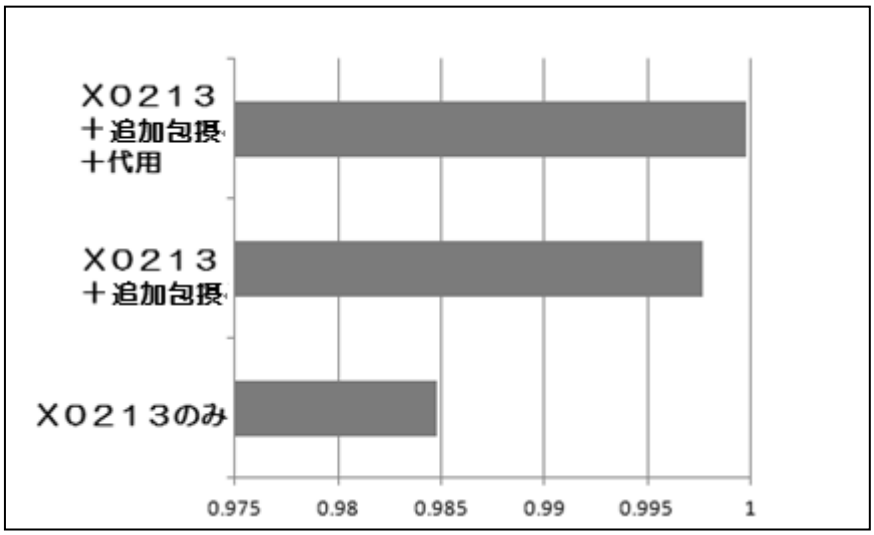


図 32　『明六雑誌』カバー率（のべ）

８．最終的に「〓」表示となる外字一覧

以上の処理を経て残る、最終的な外字、つまり「〓」表示となるものはのべ 31 字、異なりにして 25 字である。なお、これらは１字を除いて Unicode で表現可能である。

（１）Unicode では表現可能なもの（２４字）

（２）Unicode でも表現不可なもの（１字）

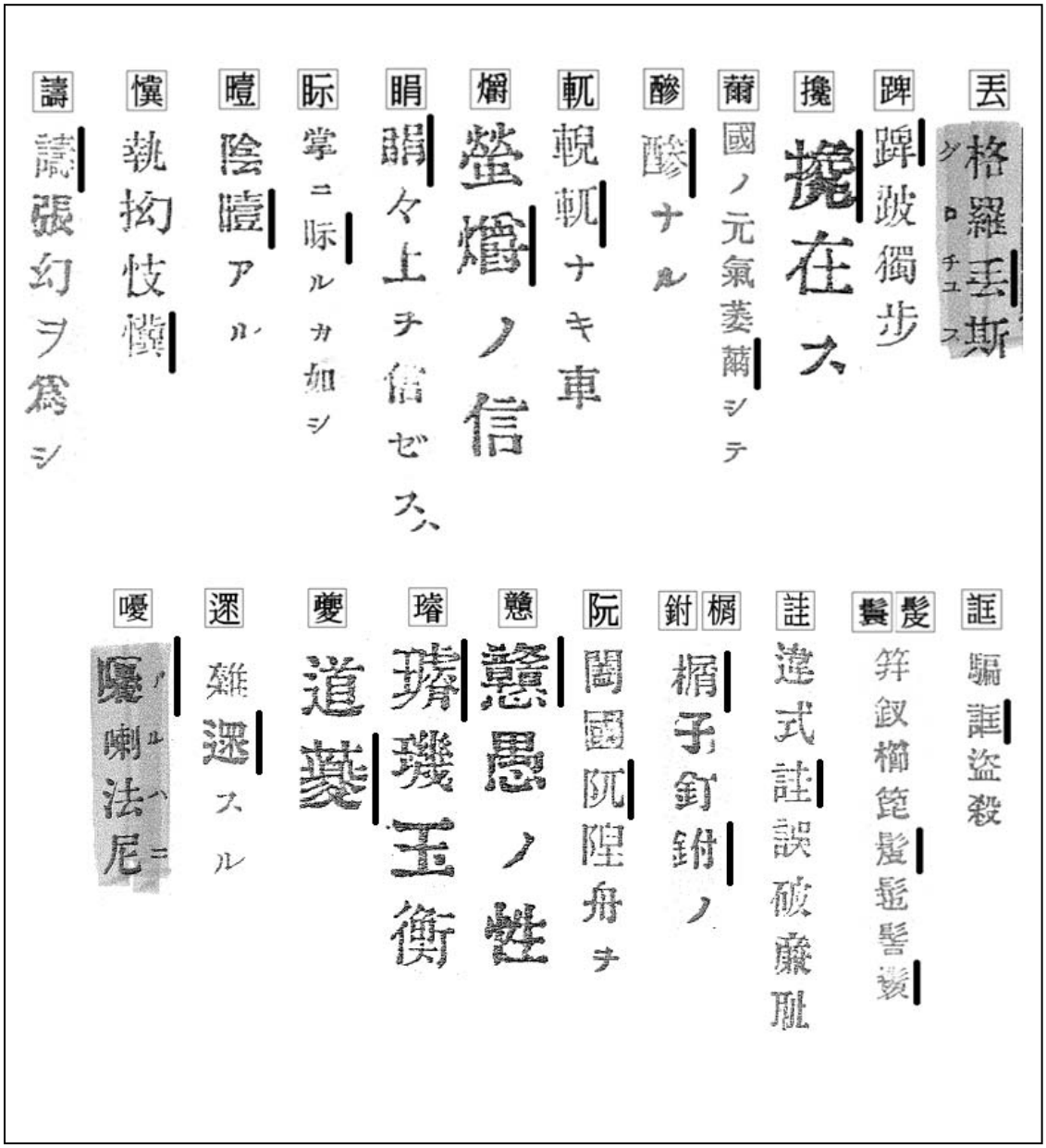


図 33　『明六雑誌』コーパスで「〓」表示となった実字形

９．今後の展望

本稿では『明六雑誌』を対象に、包摂規準の追加・修正と、別字代用の具体的方法を考案したが、『近代語コーパス』全体の文字処理を見据えた場合、今回設定した包摂規準のさらなる検証が必要となる。今後『明六雑誌』以外の活字資料を処理する場合、さらに別の包摂規準を新設する可能性や、本稿での追加包摂規準の修正を行う可能性は十分想定される。例えば『明六雑誌』の異体字処理を通して設定した「近代１９」（図 34）は「序」の異体字に対応するためだけに設けられたものである。

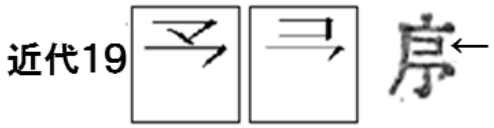


図 34　包摂規準　近代１９

しかし雑誌『太陽』(博文館、1895～1928) での活字を眺めてみると、「序」以外にも「疑」でも同様の字形差がみられ（図 35)、「近代１９」の有用性が確認できる。

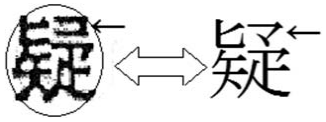


図 35　『太陽』に出現する「疑」の活字字形（左）

さらに『太陽』では図 36 のような類例も見られる。このような字形差も包摂するには、包摂規準「近代１９」を、図 37 のように修正し、さらに一般化していく方向性も考えられる。

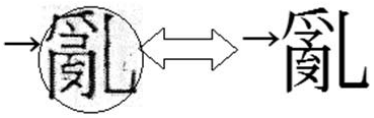


図 36　『太陽』に出現する「亂」(左)

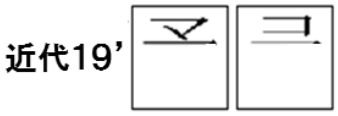


図 37　包摂規準「近代１９」修正案

このような検証を経て、近代活字用の包摂規準の整備を進めていくことが今後の課題となる。

文 献

小池和夫、府川充男、直井靖、永瀬唯(1999)『漢字問題と文字コード』(太田出版)
国立国語研究所(2005)『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』(博文館新社)
柴野耕司編著(2002)『増補改訂ＪＩＳ漢字字典』(日本規格協会)
須永哲矢、堤智昭、高田智和(2011)「明治前期雑誌の異体漢字と文字コード－『明六雑誌』を事例として－」(『人文科学とコンピュータシンポジウム論文集 2011』、pp.381-388)
高田智和、小林正行、間淵洋子、大島一、西部みちる、山口昌也(2009)『JIS X0213:2004 運用の検証（国立国語研究所内部報告書 LR-CCG-09-01）』(国立国語研究所)
田中牧郎(2005)「漢字の実態と処理の方法」(国立国語研究所 2005 所収、pp.271-292)
安永尚志(1998)『国文学研究とコンピュータ』(勉誠社)