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－

須永　哲矢　　堤　智昭　　高田　智和
国立国語研究所　東京農工大学工学府　　国立国語研究所

言語研究資料としての電子化テキストを作成するという立場から，明治前期雑誌の異体漢字処理の在り方を検討した．現行の国内規格である JIS X 0213 の文字集合および包摂規準が，近代の活字の電子化に対してはどの程度有効かを，明治初期の雑誌『明六雑誌』の異体漢字を例に検証した．JIS X 0213 文字集合によって『明六雑誌』の漢字の 98%以上が表現できるが，言語資料として電子化テキストを使う場合には，2%近くが外字処理に回るのは望ましくない．そこで外字処理をさらに減らす方法として，包摂規準の拡張や別字での代用を提案し，それらを用いて処理した場合の効果も検証した．

Kanji variants and their character codes of the early part of the Meiji period

Tetsuya Sunaga*　　Tomoaki Tsutsumi †　　Tomokazu Takada*
* National Institute for Japanese Language and Linguistics
† Tokyo University of Agriculture and Technology

Sorting out the variants of kanji is an essential task upon digitizing texts of the early Meiji period. The current domestic standard for kanji character codes, JIS X 0213, is aimed mainly at contemporary Japanese. Consequently, not a few kanji in publications of the Meiji period are classified as external codes under JIS X 0213, which is quite inconvenient for researchers who are not interested in characters. As solutions, we propose a modification plan of JIS X 0213, which involves extension of the subsumption standard and substitution of characters. We also examine its capability in the paper.

１．はじめに

紙媒体の文書を電子テキストへ写し取る際には，規格として標準化された符号化文字集合に準拠し，それを運用することが，学術分野・実業分野を問わず，広く行われている．言語資料の電子化では，国内規格 JIS X 0208（第 1 次規格 1978 年）が長く用いられてきたが，電子化に際してその都度，資料に出現した文字を文字集合のどの符号位置に対応させるべきかという問題（文字包摂の問題，粒度の問題）や，文字集合にない文字をどう扱うかという問題（規格外字の問題，文字セットの規模の問題）が指摘されている．

特に，後者の外字問題を背景に，JIS X 0208 を拡張する形で開発された国内規格が JIS X 0213（2000 年）である．例えば，国立国語研究所で開発された「現代日本語書き言葉均衡コーパス」は，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．現代日本語の一般的な文書の電子化に際して，JIS X 0213 を用いることで，外字問題はほぼ解消されたと見てよかろう．

しかし，時代をさかのぼって，近代以前の日本語を対象として，JIS X 0213 文字集合が電子化にあたって有効であるか否かは未だ検証されていない．

本研究は，国立国語研究所で検討されている近代語コーパスの基礎研究として，明治前期雑誌の異体漢字を例に，JIS X 0213 の有効性と限界を見極めようとするものである．

２．「言語研究用コーパス作成」という目的から求められる異体漢字処理方針

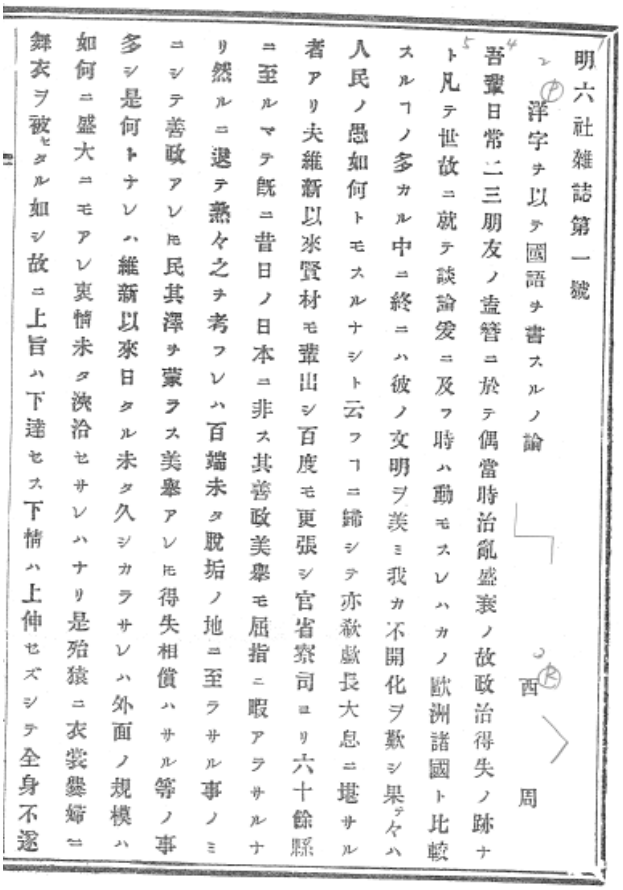
文書を電子化すると言っても，その使用目的によって，もとの文書のどの要素をどこまで再現し，どの要素は再現できなくても良しとするかという方針はさまざまに分かれうる．漢字の字体字形の問題一つをとっても，各字形差を可能な限り正確に表現した方が望ましいとは限らない．「言語研究用のコーパス作成」という場面においては，電子テキスト作成はゴールではなく，あくまで研究の手段としての環境整備，という位置づけとなる．言語研究の素材として使用される電子テキストは，言語資料として「読める」こと，語彙等のサンプルが採集できることが重要となる．そのため，外字処理とされたものが多く，「=」表示ばかりで「読めない」テキストや，動作環境によっては適切に表示されない等，検索しにく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，明治前期雑誌の異体漢字に対する，JIS X 0213 の有効性と限界を検証するとともに，言

語研究の実用に適した字体処理の在り方を模索することとした.

３．『明六雑誌』

『明六雑誌』は，明治７(1874)～８(1875)年の 2 年間にわたって発行された啓蒙雑誌で，近代日本における総合学術誌，学会誌の先駆けと位置づけられる．体裁は 30 字×13 行の活字本，1 号あたり 12~24 ページで全 43 号．総文字数は約 27 万字．広範な読者を獲得し，当時の社会への影響が大きかった点，また，記事の内容が幅広い分野にわたり，その分さまざまな語彙が取り出せる点などから，明治初期の日本語の様相を知るうえで欠かせない資料となっており，これを電子テキストとしたコーパスの作成が求められている．ただし，近代の活字字形は現在のものと異なるものも多く，異体漢字も多様に存在する．そのため，近代語資料を電子テキストで再現するには，字体字形の処理方針を確定することが必要不可欠となる．

そこで今回は，『明六雑誌』電子テキスト化にむけた作業として，字体字形の調査を行った．『明六雑誌』全 43 巻を通じ，実際にどのような字体字形が出現するかを明らかにし，そしてそれらを現行の文字コードの枠内でどのように処理するかを検討した．


明六社雑誌第一號

洋字ヲ以テ國語ヲ書スルノ論

吾輩日常二三朋友ノ盍簪ニ於テ偶當時治亂盛衰ノ故政治得失ノ跡ナト凡テ世故ニ就テ談論爰ニ及フ時ハ動モスレハカノ歐洲諸國ト比較スルフノ多カル中ニ終ニハ彼ノ文明ヲ羨ミ我カ不開化ヲ歎シ果テハ人民ノ愚如何トモスルナシト云フフニ歸シテ亦欷歔長大息ニ堪サル者アリ夫維新以來賢材モ輩出シ百度モ更張シ官省寮司ヨリ六十餘縣ニ至ルマテ既ニ昔日ノ日本ニ非ス其善政美舉モ屈指ニ暇アラサルナリ然ルニ遡テ熹々之ヲ考フレハ百端未タ瑕垢ノ地ニ至ラサル事ノミニシテ善政アレハ民其澤ヲ蒙ラス美舉アレハ得失相償ハサル等ノ事多シ是何トナレハ維新以來日タル未タ久シカラサレハ外面ノ規模ハ如何ニ盛大ニモアレ裏情未タ洽浹セサレハナリ是殆猿ニ衣裳襄婦ニ舞衣ヲ被セタル如シ故ニ上旨ハ下達セス下情ハ上伸セズシテ全身不遂

西周

図１　『明六雑誌』

```
<記事 題名="洋字ヲ以テ国語ヲ書スルノ論" 著者="西周" 文体="文語">
<s>明六社雑誌第一號<l 位置="P001A001"/><br/></s>
<s>洋字ヲ以テ國語ヲ書スルノ論<l 位置="P001A002"/><br/></s>
<s>西周<l 位置="P001A003"/><br/></s>
<s>吾輩日常二三朋友ノ盍簪ニ於テ偶當時治亂盛衰ノ故政治得失ノ跡ナ<l 位置="P001A004"/>ド凡テ世故ニ就テ談論爰ニ及ブ時ハ</s>
<s>動モスレバカノ歐洲諸國ト比較<l 位置="P001A005"/>スル•ノ多カル中ニ終ニハ彼ノ文明ヲ<外字 unicode="7FA1" 代用="1" >羨</外字>ミ我ガ不開化ヲ歎ジ</s>
<s>果<小書>テ</小書><踊字 踊字値="果テ">々</踊字>ハ<l 位置="P001A006"/>人民ノ愚如何トモスルナシト云フ•ニ歸シテ亦欷歔長大息ニ堪ザ>ル<l 位置="P001A007"/>者アリ</s>
<s>夫維新以來賢材モ輩出シ百度モ更張シ官省寮司ヨリ六十餘縣<l 位置="P001A008"/>ニ至ルマデ既ニ昔日ノ日本ニ非ズ</s>
```

図２　『明六雑誌』の電子テキスト（ＸＭＬ形式）化

４．『明六雑誌』における漢字字形

『明六雑誌』を JIS X 0213 文字集合に準拠して電子化する場合の字形処理で，重要な位置を占めるのは包摂規準である．JIS 規格では，漢字字体の包摂規準を定めており，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．

図３　JIS 包摂規準の一例

しかし，近代の活字においては，図４「序」の字形のように，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．

図４　『明六雑誌』にみられる「序」の字形

これらの字形まで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．後述のとおり，『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，約 98.5％が表現可能だが，言語研究資料としてみた場合，100 文字のうち 1.5 文字が読めない電子テキス

トは実用に供さない．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である．そのため，既定の包摂規準を近代の活字資料向けにある程度拡張し，現在の字形に包摂する，あるいは，包摂するのは厳しくとも，本来は外字であるという情報を残したうえでテキスト上は別字で代用する，というような方法を用い，外字を減らしていくことが，コーパス作成という目的にとっては現実的であると考えた．

５．『明六雑誌』漢字字形処理方針

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】
×德(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（下図５参照）．これらに関しては使用しないこととした．

図５ 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】
×𡈼 (1-15-44，U+2131B) → 外字扱い
×𡚒 (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，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．包摂，外字処理に関する方針
５．２．１．包摂規準拡張に関して

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，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字形が多い．そこで，近代語資料用に包摂規準の拡張案を作成し，字形処理に対応することにした．

包摂規準を拡張する場合，結局のところ，どこまで拡張し，どこからを外字とするかが最後まで問題となる．実際の作業においては，明確な線引きをするのはやはり難しいが，

（１）既存の基準の拡大解釈で包摂可能なもの
（２）既存の基準に類例が見いだせるもの

までを規準拡張の範囲とし，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，通用字形のどの字に当たるかが明らかであっても包摂しない．具体例をいくつか以下に示す．

（１）個別字形に対する，既存の基準の明確化
（１－１）常用漢字表「デザイン差」で処理しうるものの明確化

（現行字形） （明六雑誌）

図６ 『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」（図７参照）のひとつとして処理するという方法が考えられるが，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８参照）．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７　既存の包摂規準

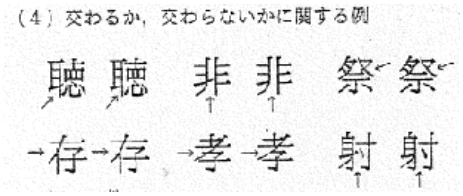


図８　常用漢字表での「デザイン差」字形例

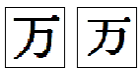


図９　新設した包摂規準

（１－２）『JIS 漢字字典』個別字形例をもとに，包摂規準に格上げ

また，以下のような場合もある.

（現行字形）　　（明六雑誌）

(1-73-22)

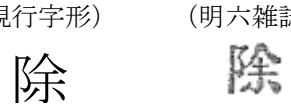
図 10　『明六雑誌』にみられる「藏」の字形

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多く見られる．図 10 にみられる字形の差異も，「万」同様，包摂規準ｂ，デザイン差（４）に照らして包摂すべきと考えられるが，『JIS 漢字字典』の個別字形例には，次図 11 の左から２つめまでの字形が掲げられているのに対し，『明六雑誌』に現れる，最後の字形例は掲げられていない．このような場合，処理上作業者が迷うことも想定されるため，これも規準として明確化した.

図 11　新設した包摂規準

（２）類例を参考に新設

（現行字形）　　（明六雑誌）



(1-29-92)

図 12　『明六雑誌』にみられる「除」の字形

このような字形差に関しても，現行の包摂規準には明示されていないが，「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７参照）の類例に照らし，図 13 のような包摂規準を新設した.

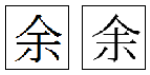


図 13　新設した包摂規準

（３）以下のようなものは包摂しない.

（場合によっては前後の文脈上の使用実態も含め，）通用字のどの字に当たるかはほぼ確定できるが，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，包摂しない.

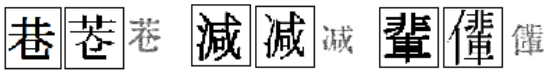


図 14　包摂“しない”字形例

５．２．２．外字処理に関して―「代用」

図 14 のように，包摂規準の拡張はしないものは外字となるが，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「＝」表示となり，読めなくなる字は可能な限り少ないことが望ましい．そこで図 14 のような事例に対しては，「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，拡張包摂規準とは別に，代用字一覧を作成した．図 14 の各字は，それぞれ「巷」(1-25-11)「減」(1-24-26)「輩」(1-39-58)で代用する．

言語研究用という使用目的からは，外字「＝」表示は可能な限り少ないことが望ましい．そこで，包摂規準の拡大解釈で包摂できる字形は包摂し，包摂規準の拡張が厳しい場合は包摂ではなく，別字で代用，という二段構えで「＝」表示を減らしていこうという試みである．

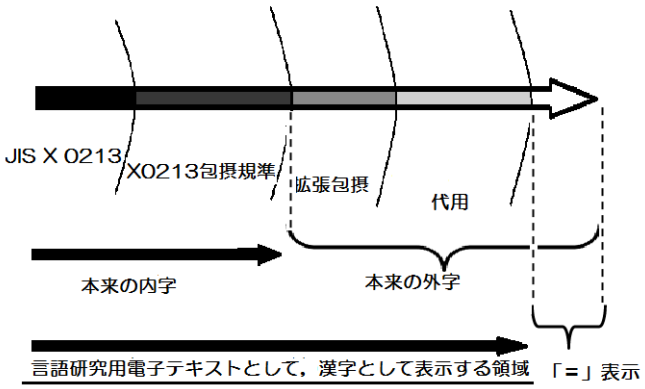


図 15　言語研究用電子化テキストにおける文字処理の理念

電子テキストでは，拡張した包摂規準で包摂したもの，本来外字であるが別字を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】
・拡張した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〉＝〈/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，以下の方針を立てた．なお，ここでは Unicode5.0 を参照している．

①Unicode 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．

【例】
図 14 の「減」（にすい），Unicode:51CF
→〈外字　unicode="51CF" 代用="1" 〉減〈/外字〉

②包摂規準の拡張で処理できそうな場合でも，Unicode で表現可能なものは外字・代用とする．

【例】

(A) 跋 (B) 跋

JIS X 0213 文字集合にあるのは（A）(1-76-77)のみだが，『明六雑誌』では（B）が出現する．この場合，拡張包摂規準を設けてもよいが，Unicode では（A）：U+8DCB，（B）も U+47E6 で表現できる．このような場合，（B）は，外字・代用として処理する．
→〈外字　unicode="47E6" 代用="1" 〉跋〈/外字〉

５．２．３．包摂・代用する字体に関して

『明六雑誌』中の活字のある字形を JIS X 0213 文字集合中のいずれかの文字で包摂・代用する場合，使用する文字の候補が複数存在する場合がある．たとえば図 16 左は，「収」にあたる異体字だが，「収」(1-28-93)，「收」(1-58-32)いずれを使用すべきか．また，右は，「験」で代用すべきだが，「験」(1-24-19)，「驗」(1-81-68)を使用すべきか．

図 16　『明六雑誌』に見られる「収」「験」

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．
②決めかねる場合は正字を使用．

①により「収」，②により「驗」が選択される．

また，拡張した包摂規準の適用にあたっては，単に字形のみを見て判断している場合，意図しないものを包摂してしまうことは十分起こりうる．これは拡張するしないに関わらず，文字包摂では一般的に起こりうることである．ただ，今回の方針は，言語研究用の電子テキスト化というところから出発しており，形のみでなく，読みながら，文脈も含めて判断することが前提となっているため，作業上，意図しないものを包摂することは起こらない．逆に，使用実態上この字に当たる，ということが明らかであれば，かなり思い切った代用を行った部分もある．
【例】

(A) 靭 (B) 靭

（A）（タン）と（B）（ソ）は本来別字であり，JIS X 0213 文字集合にあるのは（A）のみ（Unicode では（A）（B）ともに表現可能）．『明六雑誌』では（B）が出現するが，それは「鞣靼」の「靼」に当たる部分としてである．この使用実態を鑑みて，（B）を（A）「靼」で代用．

【例】 

動詞「すう」に上掲のような字が用いられるが，Unicode では表現できるものの，JIS X 0213 では表現できない．最も一般的な「吸」(1-21-59)とは大きく異なるのは明らかだが，言語研究資料として読めることを優先し，これも「吸」(1-21-59)で代用して処理．

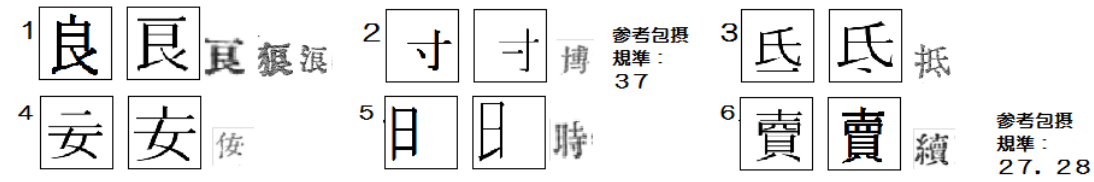
６．X 0213 文字集合／拡張包摂規準／外字代用の検証

以上，包摂規準の拡張，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，処理できる文字数にどのような変化があるかを検証する．

まずは拡張包摂規準，外字代用の一覧を掲げる．

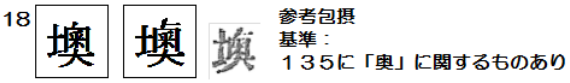
■拡張包摂規準

a 方向・曲直など点画の性質による違い

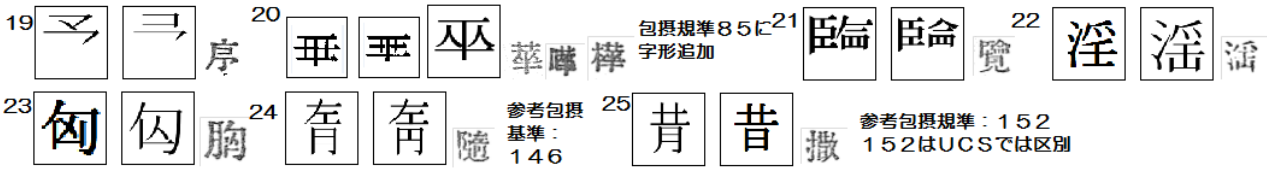


b 2点画の接触交差関係の違い

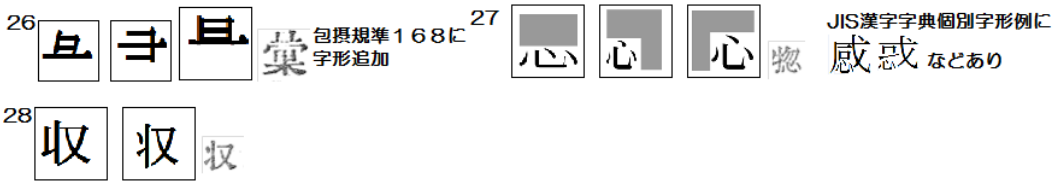
d 1点画の増減の違い



e 類型の統合



f 筆法の簡化の違い



■外字代用一覧
※Unicode では表現可能，X0213 では外字

※Unicode でも表現不可

『明六雑誌』の漢字に対し，JIS X 0213 の文字集合と包摂規準のみに従って文字処理を行った場合と，拡張包摂を適用した場合，さらに外字代用を行った場合の結果を比較してみた．

字形調査には専用ツールを用い，電子テキスト化した『明六雑誌』から確認すべき文字を洗い出し，原資料での該当箇所の字形を確認した．

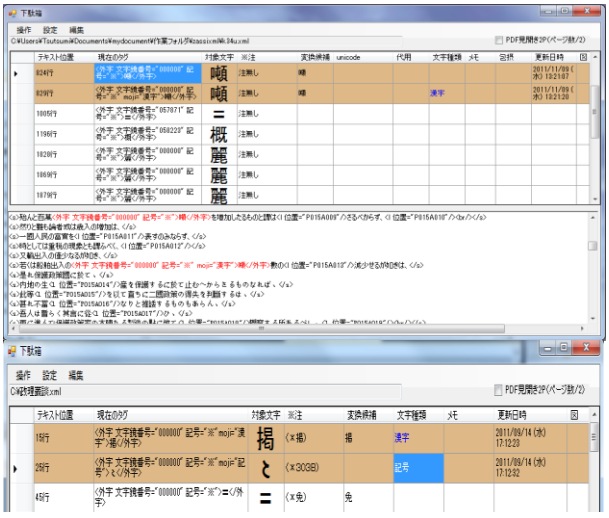


図 17　字形確認用ツール

『明六雑誌』に現れる漢字の総数は 137，897，うちX0213 のみで表現できるものは 135,797，98.5%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厳しい量である．

表１　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

拡張包摂を適用すると，外字となる 2,100 字のうち 1,774 字，さらに代用を適用すると 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，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２　各方針の適用で処理可能な文字

	X0213 包摂	拡張包摂	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

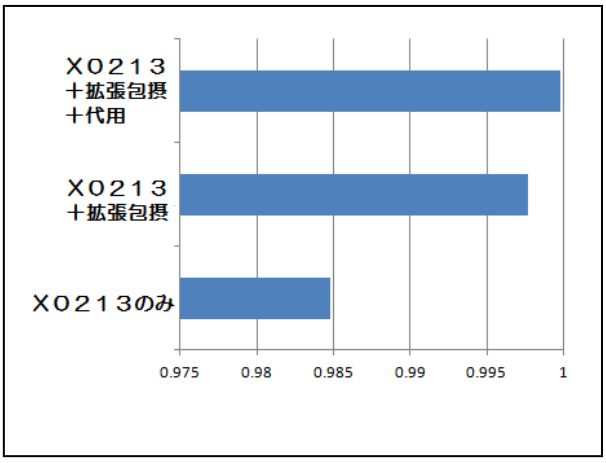


図 18　『明六雑誌』漢字カバー率

以上の処理を経て残る外字処理，「＝」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．

■最終的に「＝」処理となる外字一覧

※Unicode では表現可能

丟　ライトハ本〳〵チユートニツク〴〵＝度尼ノ語ニシテ拉丁ノジュスト云フ語ト同樣ニ(43 号)

踨　故ニ漢婦概スルニ皆＝跛獨歩スベカラズ(8 号)

攙　假冒ノ理ヲシテ其間ニ＝在スルヲ得サラシム(14 号)

蘭　國ノ元氣萎＝シテ振ハズ國威ノ振ハザル所以ナリ(12 号)

醙　又此三ツヨリ＝ナル莫シ(38 号)

軏　所謂輗＝ナキ車何ヲ以テ之ヲ行ンヤ(12 号)

爝　固ヨリ區々ノ心螢＝ノ信汚下ノ甚シキ悖戻ノ尤ナル(36 号)

睊　民心主ナク＝々上ヲ信ゼス上タル者モ亦議論熟セズ(13 号)

睖　其職務ニ從事スレハ天下ノ安キ諸ヲ掌ニ＝ルカ如シ(14 号)

暯　陰＝アルカ如シ(16 号)

懻　是上ニ在ル者一々執拗忮＝唯我ニ同キヲ好ンテ開化ヲ梗塞スル(19 号)

譋　何ソ＝張幻ヲ爲シ狐狸タラザルヲ得ンヤ(20 号)

誑　騙＝盜殺ノ變英雄豪傑ノ爭戰廉耻ヲ忘レ身命ヲ抛チ晝夜計畫措ズシテ(20 号)

髮(A)　之ヲ醫ニ聞ケバ笄釵櫛篦＝(A) 髦髻〈外字文＝(B) 時好ヲ追ヒ新樣ヲ競ヒ壓窄緊束シテ(21 号)

鬢(B)

註　上ニアル者ハ違式＝誤破廉耻不應爲ノ罪アルモ皆避ケテ咎メズ(25 号)

榍(A)　細民ハ＝(A) 子釘＝(B) ノ如シ(35 号)

鉰(B)

阮　闔國＝隍舟ヲ風波ニ泛ブル如ク朝夕變換相保タザルモ亦宜ナリ(35 号)

戇　但是＝愚ノ性愛國過慮ノ痴情止ム能ハス(38 号)

璿　天文學ノ舜ノ＝璣玉衡ニ創マリ地理政誌ノ學禹貢之カ源ヲ開キ(20 号)

夔　刑法ノ皐陶ヲ推シ教育ノ道＝ヲ始メトシ政事ノ伊尹傅説ヲ尊フカ如キ(20 号)

遝　肩摩轂擊倫敦巴黎ト雖往來此ノ如ク雜＝スルヲ見ズ(22 号)

噯　〈〃ガルハニ〃〉＝喇法尼ハ二種ノ金類能ク死蛙ノ腿ヲ動スヲ見テ(40 号)

※Unicode でも表現不可

衹　猶深宮ニ長シタル長袖＝袴ノ兒ノ如シ(32 号)

7．今後の課題

『明六雑誌』に出現する漢字は，JIS X0213 文字集合・包摂規準で 98.5%が表現可能であることが明らかになった．さらに包摂規準の拡張・別字代用の処理を施すことで 99.9%が表現可能となる．

今後の課題は，今回提示した近代語コーパス向け包摂規準拡張案のさらなる検証と洗練である．『明六雑誌』以外の雑誌・書籍・新聞等，ほかの近代語資料でも，この拡張案が有効かどうかについて，詳細な検討を加えていく予定である．

参考文献

[1] 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

[2] 下田正弘・師茂樹(1999)「大正新脩大蔵経データベース（SAT）における外字問題」，『人文学と情報処理』25，pp.35-43

[3] 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

[4] 田中牧郎(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－（国立国語研究所報告 122）』，博文館新社，pp.271-292

[5] 安永尚志(1998)『国文学研究とコンピュータ』，勉誠社