明治中期雑誌の異体漢字と JIS 漢字

－『国民之友』を事例として－

須永 哲矢　堤　智昭　近藤　明日子　木川　あづさ　服部　紀子

昭和女子大学　東京農工大学　国立国語研究所　国立国語研究所　国立国語研究所

国内規格 JIS X 0213 文字集合および包摂規準の，明治期活字に対する有効性を，明治中期の雑誌『国民之友』の活字を例に検証た．その結果，約 99.0%の文字が表現可能であることが明らかになった．同時に，JIS X 0213 をベースに，以前『明六雑誌』をもとに作成した近代活字用追加包摂規準の適用を試みたところ，全 28 の追加包摂規準のうち 22 が『国民之友』でも稼働し，これにより現代語資料と遜色ないカバー率(99.7%)を達成した．これにより追加包摂規準は明治期の活字の電子化に際し，ある程度一般的に適用できるものであることが検証された．さらに『国民之友』での活字調査を通じ，近代活字用追加包摂規準の 3 つを修正，32 を追加し，さらなる精密化を目指した．新たな包摂規準の適用で『国民之友』は JIS X 0213 ベースで 99.9%の文字が表現可能となる．

Kanji variants of the middle Meiji period in comparison with the JIS standard kanji

———The case of the magazine *Kokumin no Tomo*

Tetsuya SUNAGA[a]　Tomoaki TSUTSUMI[b]　Asuko KONDO[c]
Azusa KIKAWA[c]　Noriko HATTORI[c]

[a]Showa Wemen's University
[b]Tokyo University of Agriculture and Technology
[c]National Institute for Japanese Language and Linguistics

JIS X 0213, the current domestic standard for character codes, covers approximately 99.0 percent of the printing types in *Kokumin no Tomo*, a typeset magazine published in the middle of the Meiji period. In the paper we introduce the additional unification criteria to JIS X 0213, which accept certain non-standard kanji characters as equivalent to the specific characters of JIS X 0213. Though the criteria were initially configured for *Meiroku Zasshi*, the typeset magazine of the early Meiji, they proved to be effective as well as for the middle Meiji printing types; as many as 22 out of the total 28 criteria were applicable to *Kokumin no Tomo*. With further revision to the unification criteria (3 amendments and 32 criteria newly added), 0.1 percent of the printing types of *Kokumin no Tomo* is successfully covered by JIS X 0213.

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．言語資料の電子化に際しては，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）が指摘されてきた．

国内規格としての JIS 漢字では，前者の問題に対しては「漢字の字体の包摂規準」の設定，後者の問題に対しては，第 3・第 4 水準漢字への拡張という形で対応がなされ，2000 年には国内規格 JIS X 0213 が制定された．現代日本語の一般的な文書の電子化に際しては，JIS X 0213 がある程度有効であることは確認されており，例えば国立国語研究所で開発された約 5,800 万字からなる『現代日本語書き言葉均衡コーパス』では，JIS X 0213 に依拠する文字文字処理を行った結果，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[6]．

しかし，時代をさかのぼって，現代活字とはやや異なる日本語活字資料を電子化するにあたっても，JIS X 0213 文字集合が有効であるか否かは未だ十分な検証はなされていない．

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構築のための基礎研究として，明治期雑誌における異体漢字を例に，JIS X 0213 の有効性と限界を見極め，コーパス構築における文字処理のあり方を検討するものである．

２．研究背景 - 明治初期雑誌と JIS 漢字

２．１　『明六雑誌』電子化の事例

国立国語研究所では「近代語コーパス」の構築が構想されており，その一環として明治初期の学術啓蒙雑誌『明六雑誌』を電子化した『明六雑誌コーパス』が既に公開されている．この『明六雑誌』の電子化にあたっては符号化文字集合 JIS X 0213 が採用され，JIS 包摂規準に従って字体包摂が行われた．その結果，のべ 98.47%の文字が JIS X 0213 で表現できることが確認された．これは現代日本語の活字資料をもとにした『現代日本語書き言葉均衡コーパス』でのカバー率を 1.5 ポイントほど下回る結果である[3]．

２．２　言語研究のための文字処理方針と，追加包摂規準

『明六雑誌』において JIS 外字となった活字の大部分は，図 1 に示すように，現在のどの文字にあたるかは明らかであるが，JIS 包摂規準の範囲内では包摂できない差異が存在するものである．

（現行字形）　（『明六雑誌』）

図１　『明六雑誌』にみられる「序」の活字字形

電子データ化の際，漢字の字体字形の差異をどの程度意識すべきかは，そのデータの使用目的による．「言語研究用のコーパス作成」という場面においては，差異を重視して「＝」表示したテキストを作成するよりも，多少の差異は許容し，語が語として取り出せるよう，規格内字で表現した方が有用である．そのため，図 1 のような差異に対しても，「近代語コーパス」構築の場面においては字体包摂を行うこととした．実際，明治期の活字に対する JIS X 0213 のカバー率の高さ（のべ 98.47%）は，包摂規準適用によるところが大きく，『明六雑誌』全 43 号中，2 号分のみのサンプリング調査ではあるが，包摂規準を適用しなかった場合，カバー率は 85.96%にまで落ちることが確認されている[4]．この事実からも，包摂規準は，文字セットそのものを増やすことなく，活字字形の差異に対応できるという有用性がわかる．そこで，明治期の活字に対応するため，『明六雑誌』の活字調査を経て，近代用追加包摂規準を設定した．規準設定の基本方針は，大まかに以下のとおりである．

（１）現行の包摂規準を参考に増補

現行の包摂規準に類例がある場合，それを参考に包摂規準を増補する．図 2 は「直」の近代活字に対応するために設定した追加包摂規準であり，包摂規準連番 63 を参考としている．

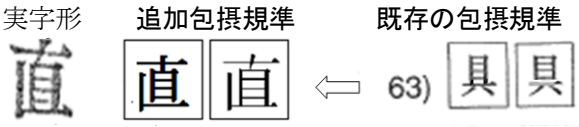


図２　追加包摂規準の設定例（１）

（２）常用漢字表での「デザイン差」等の明確化

実際の活字字形に見られる差異の中には，常用漢字表における「デザイン差」として処理する可能性もある，特に小さな違いも多い．しかし，どこまでをデザイン差とし，どこからを部分字体の差とみなすかの線引きはそもそも難しい上に，常用漢字表が掲げるデザイン差はあくまで例示にとどまっており適用範囲が明確ではない．そのため，実際の作業上迷いそうなものに関しては，追加包摂規準として明確化したものがある．図 3 は「万」の近代活字であり，常用漢字表のデザイン差に掲げられる「交わるか，交わらないか」の事例として処理する可能性もあるが，包摂規準として明確化した．

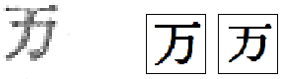


図３　追加包摂規準の設定例（２）

（３）『JIS 漢字字典』の個別字形例をもとに一般化

『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，複数の字形例が示されている場合が多く見られる．図 4 左は『JIS 漢字字典』における「感」の字形例で，「心」の位置の差異が示されている．近代活字では「感」に限らず，他のさまざまな字でも似たような差異が生じるため，包摂規準として設定した．

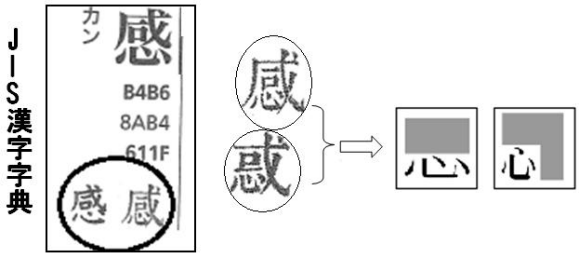


図４　追加包摂規準の設定例（３）

（４）包摂しない差異

偏や旁が別の偏・旁になっている，あるいは偏の有無などの差異は包摂しない．また，JIS X0213 文字集合内で区別されている差異に類するものは包摂しない．

以上のような方針のもと，近代活字用追加包摂規準を 28 種設定した（図 5）．

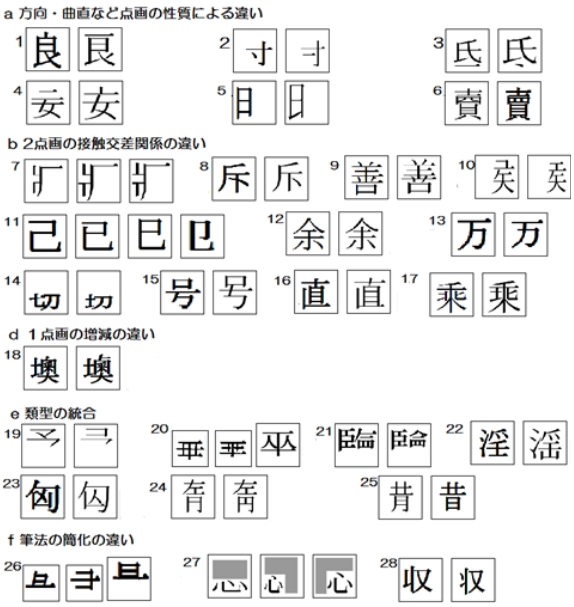


図 5 『明六雑誌』用追加包摂規準（須永ほか 2011[3]）

『明六雑誌』の異体字処理に関しては，この追加包摂規準により，本来 JIS 外字となるのべ 2,100 字のうち，1,774 字，つまり 8 割以上が処理でき，98.48 %であったカバー率も 99.76%，現代語に 0.1 ポイント差にまで迫ることができるという結果を得た[3]．

２．３　課題

以上のように，JIS X 0213 文字集合および包摂規準の検証・追加包摂規準の設定を行ったが，これはあくまで明治初期の『明六雑誌』という 1 種類のテキストのみをもとにしている．やや時代の異なる別資料に対しても JIS X 0213 や追加包摂規準の適用実験を行い，『明六雑誌』用に作成した追加包摂規準の有効性の検証・規準自体の修正・洗練を経て，近代活字に広く適用できるような追加包摂規準を画定していくことが課題となっている．

３．明治中期の雑誌『国民之友』

本稿では明治中期の雑誌『国民之友』（図 6）の異体字調査を行い，JIS X 0213 そのもの，および発表者らが作成した近代活字用追加包摂規準の有効性を検証することとする．『国民之友』は，徳富蘇峰によって 1887（明治 20）年に創刊され，1898（明治 31）年，372 号まで刊行された日本初の総合雑誌である．今回はこのうち創刊から 2 年間の 36 号分，追加包摂規準作成のもととなった『明六雑誌』 1874（明治 7）～1875（明治 8)年から 13 年後の資料である．調査範囲の総文字数は約 164 万字，追加包摂規準作成のもととなった『明六雑誌』約 14 万字のおよそ 12 倍であり，追加包摂規準の検証としても有意義な文字数といえよう．

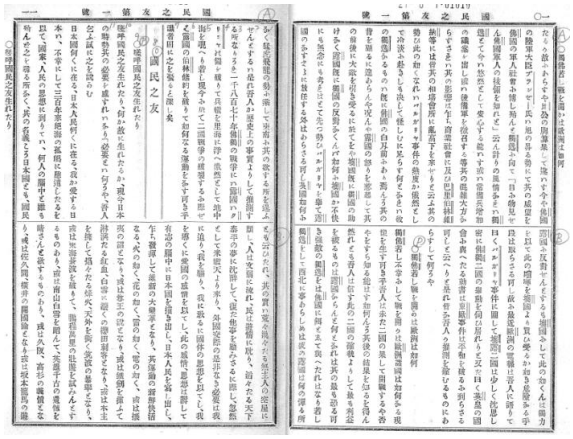


図 6　『国民之友』

４．調査方法

『国民之友』の異体字調査にあたっては，原資料のデータ化（１次入力）時，テキストデータ校正時にそれぞれ目視確認によって異体字を探し出した．さらに字形差がよく見られる要確認字や，見つかった異体字と部分字形を同じくする字等に関しては，文字調査支援ツール（図 7）[8]を用い，３次確認まで行った．

図 7　文字調査支援ツール「＝箱（げたばこ）」

５．『国民之友』活字と JIS X0213

『国民之友』創刊 2 年分の文字数はのべ 1,642,791 字，うち JIS X0213 で表現できるものは 1,627,078 字，カバー率は 99.04%となった（表 1）．このカバー率は，現代語を対象とした『現代日本語書き言葉均衡コーパス』(99.96%)，『国民之友』以前の『明六雑誌』(98.48%)の中間に位置する結果となった（図 8）

表 1　『国民之友』活字と JIS X0213

文字区分	のべ字数	異なり字数
JIS X0213	1,627,078	4,868
第 1 水準	1,525,348	2,788
第 2 水準	100,170	1,818
第 3 水準	1,225	147
第 4 水準	335	115
外字	15,713	85
総計	1,642,791	4,953
カバー率	99.04%	98.28%
『明六雑誌』カバー率	98.48%	97.02%

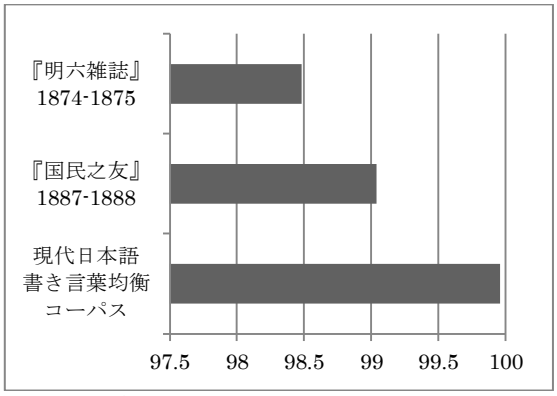


図 8　カバー率の比較（のべ字数）

6．『国民之友』における JIS 外字

『国民之友』を電子テキスト化する場合，JIS 外字となるものは，全体の 1%程度とはいえ，実数としてのべ 1 万字以上が「=」表示などによって読めなくなるというのは，研究利用を考えた際には望ましいことではなく，可能な限り文字の形で「読める」ように電子化を実現したい．そこで，これらをそのまま「=」表示で放置するのではなく，なんからに救い上げる方策を考えたい．

JIS 外字となるものは，大きく三つに分けられる．一つは図 9 に示すように，文字集合にない字，いわば外字らしい外字である．

図 9　『国民之友』にみられる JIS 外字例（１）

続いては，本文中での使用実態から，現行の通用字のどの字にあたるかは明らかではあるものの，偏や旁が異なるもの．図 10 左，『国民之友』に見られる活字は，現代での「障害」にあたると思われるが，「障」の字の偏が「石」になっている．

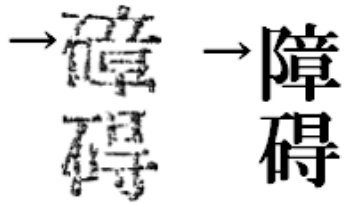


図 10　『国民之友』にみられる JIS 外字例（２）

残る一つは，実際の字形の類似及び本文中での使用実態から，現行の通用字のどの字にあたるかは明らかではあり，外形上差異も偏や旁の交替といった図 10 のようなレベルではなく，より部分的なわずかな差が認められるにすぎないが，その差異は JIS 包摂規準の範囲内では包摂できないものである．図 11 は『国民之友』にみられる「誤」の異体漢字であるが，JIS 包摂規準連番 48 では，この差異を包摂できず，包摂規準に厳密に依拠するならば，外字という扱いになってしまう．

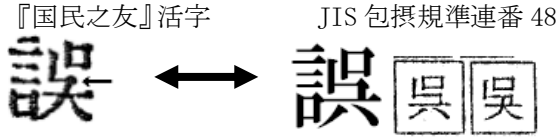


図 11　『国民之友』にみられる JIS 外字例（３）

以上のように，外字といってもその程度によって三つのタイプに分けられる．これらに対処する方法としては，JIS X0213 より大きな文字集合を適用する，あるいは JIS 漢字の範囲内で，近代活字に対応できるよう，包摂規準を拡張・修正するという方法が考えられよう．主に図 9，10 のような外字に対応するためには前者の方法が有効に思えるが，現時点では規模の多い文字集合を適用したところで，近代の活字には対応しきれるものではない．例えば JIS X 0213 に代わって，その約 8.5 倍の規模の文字集合 Unicode4.0 で文字処理を行った場合でも，のべ 15,713 の JIS 外字のうち Unicode で表現できるのはわずか 2%ほど，のべ 364 字にすぎず，15,349 字は結局外字のままとなり，さほど結果は変わらない．

そこで，むしろ効果が期待されるのは後者の方法，包摂規準の拡張である．実際，『国民之友』で JIS 外字となるものの大半は，図 11 のようなわずかな差異を持つものであり，近代活字に対応するための包摂規準の設定により，1 万字単位での JIS 外字を処理できることが期待される．

７．近代活字用包摂規準

７．１　『明六雑誌』版追加包摂規準と『国民之友』

前掲図 5,『明六雑誌』での活字をもとに作成した近代用追加包摂規準を『国民之友』での活字処理に適用してみた．結果は表 2 のとおり，近代用追加包摂規準の適用により，15,713 字あった JIS 外字のうちの約 78%，12,296 字が JIS X 0213 規格内で表現できるようになった．カバー率の面でも『明六雑誌』に適用した場合とほぼ同様の結果となり，近代用追加包摂規準が，『明六雑誌』以外の近代活字資料に対しても有効であることが確認された．

また，追加包摂規準の稼働率という観点からも設定した 28 の規準のうち 22 が『国民之友』の文字処理にも適用された．約 8 割が稼働したということからも，近代用包摂規準の有効性が検証できたといえよう．

表 2『国民之友』活字に近代用追加包摂規準を適用した結果（のべ字数）

	JIS X0213	追加包摂規準適用
	文字数	文字数
処理可能字	1,627,078	1,639,324　(+12,296)
第 1 水準	1,525,348	1,536,711　(+11,363)
第 2 水準	100,170	101,103　(+933)
第 3 水準	1,225	1,225
第 4 水準	335	335
外字	15,713	3,417　(−12,296)
カバー率	99.04%	99.79%
『明六雑誌』カバー率	98.48%	99.76%

さらに，約 14 万字の『明六雑誌』から約 164 万字の『国民之友』に適用規模を広げたところ，『明六雑誌』時点では見られなかった活字字形にも同様の包摂規準が適用できる，という事例も多くみられ，この面においても近代活字用追加包摂規準の一般性が確認できた．例えば図 12 は『明六雑誌』において「候」「侯」の活字に対応するために設定された追加包摂規準「近代 10」だが，『国民之友』では，この 2 字に加え，新たに出現した「喉」の活字に同様の規準が適用された．

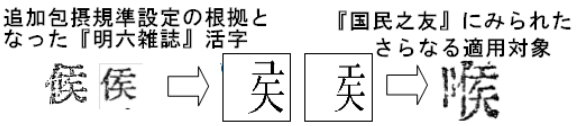
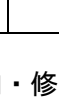
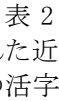


図 12　近代活字用追加包摂規準適用例

一方，近代活字用包摂規準のうち，『国民之友』では稼働しなかったものは表 3 のとおり．これらの活字体に関しては，13 年前の『明六雑誌』と比べて現代の活字体に近づいていることが確認されたことになる．

表 3『国民之友』で稼働しなかった近代用追加包摂規準

近代活字用包摂規準	『明六雑誌』活字	『国民之友』活字
寸 寸		
𠂢 𠂢 巫		
臨 臨		
淫 淫		
收 收		
号 号		（「号」の活字なし）

７．２　近代用追加包摂規準の追加・修正

表 2 より，『明六雑誌』の活字をもとに作成された近代活字用追加包摂規準が，『国民之友』での活字に対しても同等の有効性を持つことは確認されたが，調査対象の文字数を増やせば増やすほど，さまざまな字体字形の差異が目につくようになる．『明六雑誌』の約 12 倍の文字数をもつ『国民之友』での事例収集から，近代用追加包摂規準を追加・修正することで，処理可能な文字をさらに増やせると同時に，追加包摂規準自体の一般性・有用性も高められると考える．

例えば図 13 は『明六雑誌』での活字をもとに設定された近代活字用包摂規準の一つであるが，『国民之友』での活字を調査した結果，図 14 のようにさらに細かい差異のパターンが認められるため，図 14 の矢印部分の差異に関しても明記するよう修正する．

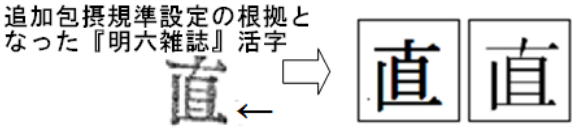


図 13　近代活字用追加包摂規準「近代 16」

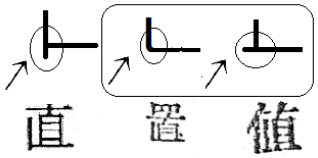


図 14　『国民之友』における「直」「置」「値」

さらに，同様の差異は図 15 のような活字にも見られるため，包摂規準としては「十」の部分を除いて一般化した方が適用範囲上，望ましい．そこで図 13 に示した近代活字用包摂規準「近代 16」は図 16「近代 16 改」に修正した．

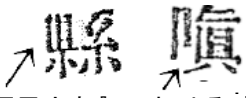


図 15　『国民之友』における「縣」「嗔」

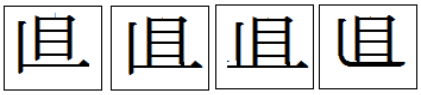


図 16　近代活字用追加包摂規準「近代 16 改」

同様の事例として，図 17 に示す「近代 15」の修正も行った．『国民之友』には活字「号」は現れず，この規準自体は稼働しなかったが，図 17 矢印部分のような差異に関しては，図 18 に示すように，『国民之友』における活字にも広く見られた．そこで図 19 に示すように修正した．

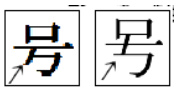


図 17　近代活字用追加包摂規準「近代 15」

図 18　『国民之友』における「汚」「朽」（左）

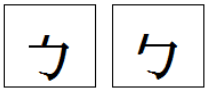


図 19　近代活字用追加包摂規準「近代 15 改」

以上のような，近代活字用包摂規準の修正・一般化という作業に加え，規準自体の新設も行った．

まず，『明六雑誌』での活字調査をもとに近代用包摂規準を設定した時と同様，既存の JIS 包摂規準に，新たに包摂字体を追加する場合．図 20 は JIS 包摂規準連番 116, 117 であるが，『国民之友』には図 21 のような異体活字が存在する．そこで図 22 のように包摂字体を追加し，これを近代活字用包摂規準「近代 40」「近代 41」とした．

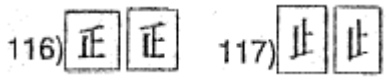


図 20　JIS 包摂規準連番 116, 117

図 21　『国民之友』における「延」「政」

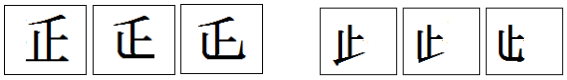


図 22　近代活字用追加包摂規準「近代 40, 41

また，図 23 は JIS 包摂規準連番 183 であるが，『国民之友』には「掫」に限らず，図 24 のような異体活字が存在する．そこで図 25 のように「てへん」を除き一般化し，これを近代活字用包摂規準「近代 54」とした．

図 23　JIS 包摂規準連番 183

図 24　『国民之友』における「取」

図 25　近代活字用追加包摂規準「近代 54」

JIS 包摂規準連番のいずれかを増補する場合にとどまらず，包摂規準自体を新設する場合もある．『国民之友』では図 26 に示すような活字が多様に見られる．既存の JIS 包摂規準連番 189 でも似たような差異を扱っていることからも，このような差異も包摂規準として設定するのが妥当と判断し，図 27 のような近代活字用包摂規準を設定した．

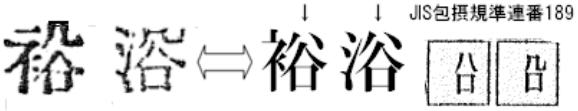


図 26　「裕」「浴」の活字字形と JIS 包摂規準

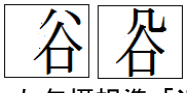


図 27　新設した包摂規準「近代 31」

上記のような過程を経て，近代用追加包摂規準のうち 3 つを修正，32 を新設した．前掲図 5 に対し，『国民之友』での検証を経て修正・新設した近代用追加包摂規準を図 28 に示す．

図では近代活字用包摂規準連番，包摂字体，参考にした JIS 包摂規準連番，適用対象活字例，の順に示した（『明六雑誌』で設定した近代 1～28 のうち，修正されなかったものは図 5 参照）．

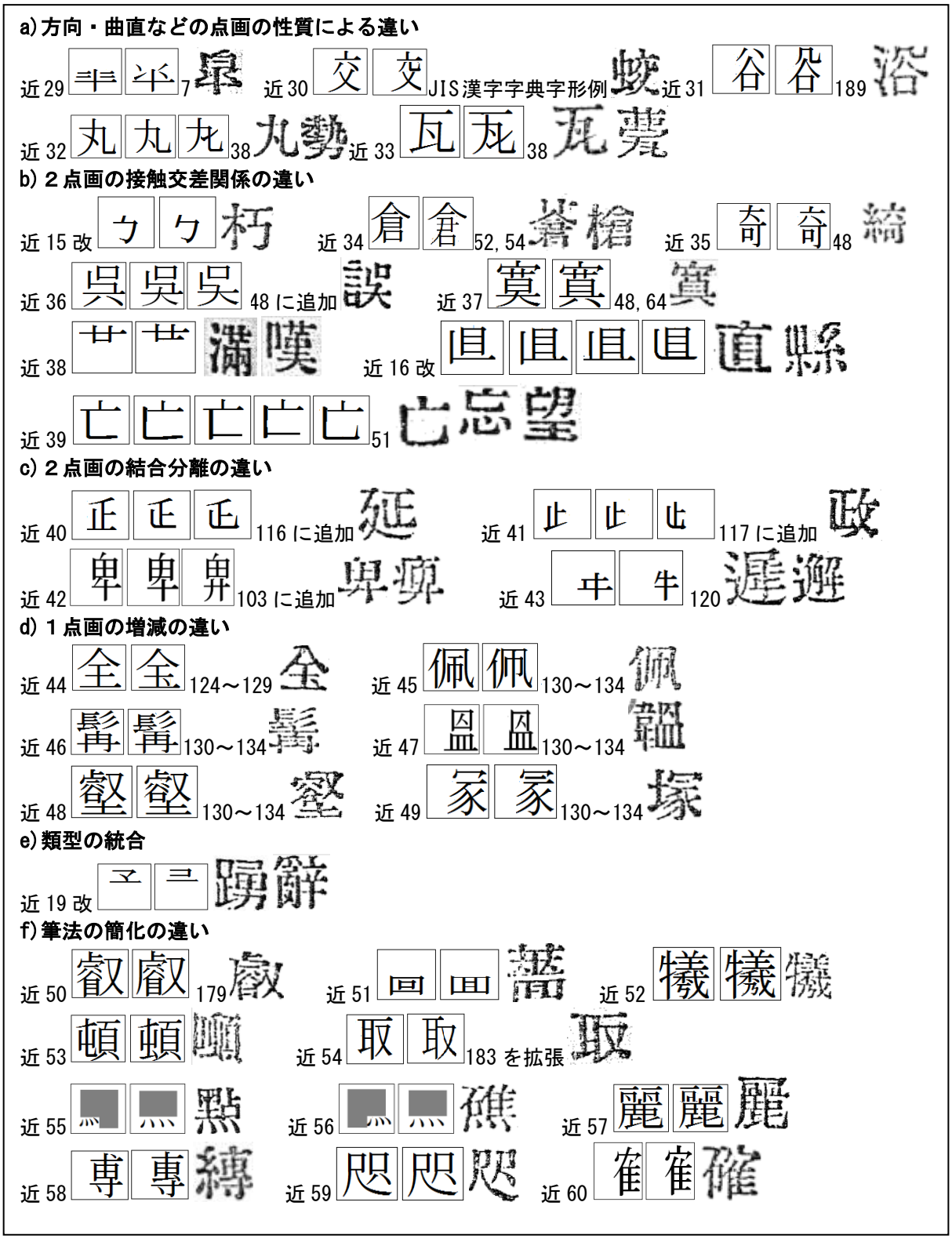


図 28　修正・新設した近代活字用追加包摂規準

７．３　修正版包摂規準の適用結果

『明六雑誌』をもとに設定した近代活字用追加包摂規準（全 28）に，『国民之友』での活字調査をもとに図 28 に示したような追加・修正を施した修正版包摂規準（全 60）を『国民之友』に適用した結果，表 4 に示す通り，カバー率は 99.98%に上昇，『現代日本語書き言葉均衡コーパス』をやや上回る結果を得ることができた．

表 4 修正版包摂規準の適用結果

	JIS X0213	『明六雑誌』追加包摂 近代 1～28	『国民之友』追加包摂 近代 1～60
処理可能字（のべ）	1,627,078 字	1,639,324 字	1,642,382 字
外字（のべ）	15,713 字	3,417 字	409 字
カバー率	99.04%	99.79%	99.98%
『明六雑誌』カバー率	98.48%	99.76%	
『現代日本語書き言葉均衡コーパス』カバー率	99.96%		

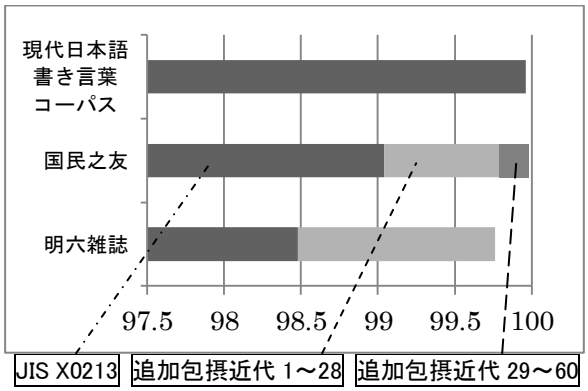


図 29　カバー率の比較

８．おわりに

以上，本稿では『国民之友』での活字調査をもとに，近代活字用追加包摂規準の検証・増補を行った．『明六雑誌』での活字をもとに設定した近代活字用の追加包摂規準は，『国民之友』においても有効であり，これを適用することによって近代の活字でも，現代活字と遜色ない程度に JIS 漢字の範囲内で表現できることが確認された．

今回『国民之友』での活字をもとに包摂規準の増補・修正を試みたが，調査する文字が増えれば増えるほど，目に留まる差異も増えるというのはある意味当然であり，追加包摂規準は『明六雑誌』時点では 28 であったのに対し，本研究の結果 60 に倍増した．今後，調査の度にやみくもに規準が増えるというのも実用性の面では望ましいことではなく，今後は事例の蓄積のみならず，整理も必要になってこよう．本研究は「コーパス作成の作業のための指針作り」という実地面での要請からスタートしていることもあり，入力作業者が迷うもの，気になる差異は「包摂規準」として明示する，という方針を取ってきた．しかし，近代活字用の追加包摂規準の中には，実際は「包摂される字体差」というより近代活字の「デザイン差」とみなした方がふさわしい差異も含まれよう．現代活字における「デザイン差」の適用範囲が明確でなかったこともあり，本研究までは作業上の効率を考えて，気になる差異に関しては「包摂規準」として明確化していくという方法で，ここまで蓄積を行ってきた．ある程度事例が蓄積されてきた現在，今後の整理のあり方としてはここまでの追加包摂規準，および「作業者に気にとめられなかった，さらにわずかな差異」の事例蓄積まで含め，「近代活字特有のデザイン差」を設定し，現在「追加包摂」一本である処理方法を再整理を目指すべきではないかと考えている．『明六雑誌』『国民之友』と調査を重ね，事例も度蓄積されてきたので，今後の課題としたい．

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]芝野耕司　編著(2002)『JIS 漢字字典』日本規格協会

[3]須永哲矢，堤智昭，高田智和：明治前期雑誌の異体漢字と文字コードー『明六雑誌』を事例としてー，人文社会とコンピュータシンポジウム「じんもんこん 2011」論文集，pp. 381-388 (2011)．

[4]須永哲矢，堤智昭，高田智和：明治前期の漢字活字とＪＩＳ漢字包摂規準ー『明六雑誌』活字字形への，包摂規準適用実験ー，第 95 回人分科学とコンピュータ研究発表会 (2011)．

[5]須永哲矢:近代語文献を電子化するための異体字処理,『近代語コーパス設計のための文献言語研究成果報告書（国立国語研究所共同研究報告 12 - 03)』.

[6]高田智和，小林正行，間淵洋子，大島一，西部みちる，山口昌也：JIS X0213:2004 運用の検証，国立国語研究所内部報告書 LR-CCG-09-01 (2009)．

[7]田中牧郎：漢字の実態と処理の方法,『『太陽コーパス』研究論文集ー（国立国語研究所報告 122)』，博文館新社，pp. 271-292 (2005)．

[8]堤智昭，須永哲矢，高田智和：コーパス用テキストを対象とした文字処理支援ツール「＝箱(げたばこ)」-文字校正・処理情報付与作業の効率化-，人文科学とコンピュータシンポジウム「じんもんこん 2012」論文集，pp. 171-178 (2012)．