

国立国語研究所学術情報リポジトリ

Digitization of Typeset Books in Constructing the Corpus of Historical Japanese : The Case of the Shogakukan (SNKBZ) Edition of the Konjaku Monogatarishu

メタデータ	言語: jpn 出版者: 公開日: 2015-10-30 キーワード (Ja): キーワード (En): 作成者: 須永, 哲矢, 堤, 智昭, SUNAGA, Tetsuya, TSUTSUMI, Tomoaki メールアドレス: 所属:
URL	https://doi.org/10.15084/00000516

『日本語歴史コーパス』のための書籍活字の電子化

——小学館新全集『今昔物語集』を事例として——

須永 哲矢^a 堤 智昭^b

^a国立国語研究所 コーパス開発センター 非常勤研究員 [-2013.03]

^b東京農工大学 博士課程

要旨

国立国語研究所で計画されている『日本語歴史コーパス』の構築にあたっては活字書籍化された古典資料のコーパス化を基本とし、その際には国内規格 JIS X0213 文字集合を用いて活字を電子化することが予定されている。本稿では JIS X0213 を古典資料の活字書籍に適用した場合の効果を検証するため、小学館新全集『今昔物語集』での漢字活字を調査し、のべ字数にして 99.86% の活字が JIS X0213 でカバーできることを明らかにし、JIS X0213 の有効性を確認した。また、JIS X0213 では表現できない活字に関しては、コーパスとしての利便性を鑑み、「■」表示せず JIS X0213 の範囲内の別字で代用しつつ、原資料での字形の情報を保持する方針を考案した。別字代用によりほぼ 9 割の外字は解消されるが、「■」表示を完全になくすためには、文字レベルではなく、語の表記というレベルでの代用を考えなければならない。末尾には小学館新全集『今昔物語集』で代用処理の対象となる特殊活字の一覧を付した*。

キーワード：コーパス構築, JIS X0213, 外字処理, 今昔物語集

1. はじめに

国立国語研究所では『日本語歴史コーパス』の構築が構想されており¹、これが実現した場合、さまざまな古典資料が言語研究目的で電子化されていくことになる。電子化にあたっての原資料は、『小学館新編古典文学全集』など、活字化された紙媒体の資料を想定している。古典資料の電子化と言っても、既に活字化され、整形されたテキストをもとにするわけだが、それでも紙媒体としての利用を想定した書籍としての活字は、電子的利用を想定したコーパスでの符号化文字集合にそのまま移し替えることはできない場合がある。特に古典資料関連の書籍を電子化する際にはその底本の文字の特殊性や校訂方針等により、現代語書籍を電子化する場合以上に困難が伴うことも予想される。本稿は、古典作品の電子化にあたっての予備調査として、『小学館新編古典文学全集』（以下、新全集）版『今昔物語集』における活字調査を行い、それらを電子化する際における、符号化文字集合 JIS X0213 の有効性と限界を検証し、そのままでは電子的に表現が困難な文字についてはどのような処理がありうるか、その可能性を検討するものである。

* 本稿は NINJAL 「通時コーパス」プロジェクト・Oxford VSARPJ プロジェクト合同シンポジウム「通時コーパスと日本語史研究」（2012 年 7 月 31 日）での口頭発表「小学館新全集『今昔物語集』での漢字活字コーパス化のための調査と処理方針の検討」（須永・堤）の内容をもとにしている。

¹ 2013 年 6 月時点で、平安仮名文学 10 作品（古今和歌集・土佐日記・竹取物語・伊勢物語・落窪物語・大和物語・枕草子・源氏物語・紫式部日記・和泉式部日記、全て小学館新全集）が『日本語歴史コーパス 平安時代編』として先行公開されている。

2. 問題の所在

2.1 電子化に際しての文字処理の問題

『日本語歴史コーパス』の原資料となるものは、基本的に活字化された紙媒体の資料を想定している。中古や中世の古典資料と言っても、『新全集』など、現代において活字化されたものをもとに電子テキスト化を進めようというのである。そのため、原資料を電子化するという作業自体は、既に完成を見た『現代日本語書き言葉均衡コーパス』(以下、BCCWJ)等、現代語コーパスの場合と大きく変わるものではない。

言語資料を電子化する際に、主に問題となるのは大きく以下の2点である。

- (1) 原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(文字包摂の問題、粒度の問題)
- (2) 文字集合にない文字をどう扱うか(規格外字の問題、文字セットの規模の問題)

この2つの問題は、現代語資料を電子化する時点でも当然問題になるが、歴史的資料を電子化するにはさらに切実になってくる。また、対象とする原資料によって、問題の重点も変わってくる。

例えば、現代のものと異なる活字が使用されている場合、(1)の問題が浮き彫りになる。明治前期の雑誌、『明六雑誌』では図1に示すように当時の活字字形と現代の通用字形には差異が見られ(○で示す)、電子テキスト化にあたっては、これらを現代の通用字形で表現してよいか、あるいは外字とすべきかが問題となる(須永ほか2011)。

序 序 万 万 除 除

図1 『明六雑誌』に出現する「序」「万」「除」の字形(右側)

これに対して今回の調査対象となる新全集『今昔物語集』は、現代の活字を用いて表現されているため、(1)の問題は近代活字ほど切実にはならない。新全集『今昔物語集』での文字処理の問題は(2)規格外字の問題が中心となる。

2.2 紙媒体と電子媒体の違い

『日本語歴史コーパス』は活字化された書籍を電子化するという工程を想定している。前節(1)(2)の課題は、活字でない底本を活字化し、新全集という形で書籍化された段階で一度処理されていると言える。底本に存在したであろうさまざまな異体字や省画・増画の字を、どのような活字で表現するか(通用字として表現するか、通用字とは別の活字を用いるか、など)は、新全集の内部でも作品によって方針にばらつきがあるが、『今昔物語集』に関しては、字形差をかなり細かく表し分けようとしたことがうかがえる²。図2は新全集『今昔物語集』での「あける」と読

² 『日本語歴史コーパス 平安時代編』に収録されている10作品は、仮名文学であるために漢字の字形差な

む活字例だが、一般的な活字では表現できないようなこれらの字形差を逐一区別して表現していることがわかる。

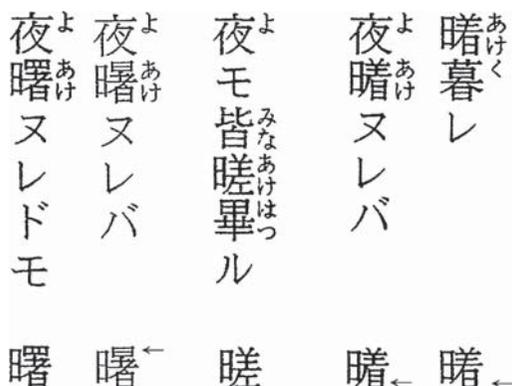


図2 新全集『今昔物語集』での活字例

図2のとおり、新全集『今昔物語集』では特殊な活字を用いて字形差を表現しようという姿勢がうかがえるのだが、このような処理が可能なのも、ユーザーの手に渡る形態が「印刷された紙媒体」であればこそである。電子データの状態での利用を考えるなら、一般的な各端末で使用・表示の可能な符号化文字集合の範囲外の文字は表現できない。つまり、図2のような活字字形は、あらかじめ規格化された電子的な文字集合になれば、電子的に写し取ることは不可能なのである。

2.3 符号化文字集合

2.3.1 符号化文字集合 JIS X0213

『日本語歴史コーパス』のための電子化にあたっては、国内規格「JIS X0213」に依拠して文字処理を行うことが想定されている。JIS X0213は、これに先立つJIS X0208（1978年第1次規格、1997年第4次規格）を拡張する形で2000年に制定（2004年、2012年改訂）された規格である。JIS X0208時点では、使用頻度に合わせて第一水準・第二水準の漢字が設定されていたが、地名・人名なども含め、現実に日本国内で使われている文字をコンピュータで表現するにあたっては不足であることも指摘されており、一般的な文書においてもしばしば外字処理の問題が生じていた。そこでその不足を補うべく、第三水準（1249字）、第四水準（2346字）を追加した規格がJIS X0213である。実際このJIS X0213は、BCCWJでの文字処理にも採用され、その際のべ99.96%の文字がJIS X0213で表現できることが確認された（高田ほか2009）。このような実績から、現

どがさほど問題にならないこともあってか、JIS規格外字は少ない。10作品全てを通し規格外字の予備調査を行ったが、外字は10作品全体でのべ2字、異なり5字に収まり、全てUnicodeでは表現可能である。これと比べると本稿での調査対象である新全集『今昔物語集』1作品での外字の多さ、活字字形の多様さがうかがえよう。

代日本語の一般的な文書の電子化に際しては、JIS X0213 を用いることで、外字問題はほぼ解消できると見てよからう。しかし、JIS X0213 が目指したのはあくまで現代における一般的な日本語文書への対応であり、古典資料を書籍化した新全集などの場合は、その性質ゆえに状況が変わってくることも予想される。そこで『日本語歴史コーパス』の構築にあたっては、新全集のような活字書籍に対しても JIS X0213 で電子化することが妥当なのか、その有効性と限界を検証したうえで、JIS X0213 外字となる文字についてはどのように処理すべきかの方針を画定しなければならない。

2.3.2 JIS X0213 と Unicode

符号化文字集合としては、国内規格 JIS X0213 の他に、Unicode もよく知られている。JIS X0213 に収録されている文字は Unicode にも含まれており、Unicode は JIS 漢字を含みつつ、より大きい文字集合であると言える。文字集合の規模としては JIS X0213 が 1 万 1233 字に対し、Unicode4.0 では 9 万 6477 字³と、Unicode の方が圧倒的に大きい。実際、図 2 に示した各活字のうち、JIS X0213 で表現可能なものは 1 字のみであるのに対し、Unicode4.0 では 2 字が表現可能である (表 1)。

表 1 図 2 のうち、JIS X0213, Unicode で表現可能な活字字形

JIS X0213	曙	
Unicode4.0	曙	嗟

文字集合が大きければ、当然表現できる文字も増えるため、その点に関しては Unicode の方が勝っているように見える。しかし Unicode を採用する際には別の問題が生じる。まず、Unicode は文字集合の規模が大きすぎるため、作業コストがかかる。簡単に言うなら、1 万字の中から 1 文字を探すコストと、その 10 倍の 10 万字の中から 1 文字を探すコストの差である。また、現時点では、JIS 外字だが Unicode では表現可能な文字があっても、現在の一般的な動作環境では正しく表示されない場合も多い。せっかく Unicode で表現したところで、環境によっては結局表示されないのであれば意味がない。このような事情を鑑み、また、BCCWJ において、Unicode を使わずとも JIS X0213 の範囲内で大抵の文字は表現できたという実績も合わせ、『日本語歴史コーパス』でも JIS X0213 の適用が妥当であろうと想定されている。ただし、『日本語歴史コーパス』の電子化に際しては Unicode の難点を補って余りあるほどに、JIS X0213 と Unicode の間で表示できる文字に差がある場合は、JIS X0213 の適用という方針自体も再考せねばなるまい。そこで、新全集『今昔物語集』での活字調査では、JIS X0213 ではどの程度の活字が表現され、どの程度が表現しえないかを調査するとともに、Unicode を用いた場合どうなるかも検証することとした。

³ 2012 年 9 月時点での最新版としては Unicode6.2 (収録文字数 110,182) が公開されているが、本稿では Unicode4.0 を参照している。



図5 「=箱」を用いた外字チェック

この外字チェック機能を利用して作成した<外字>タグ付きデータから<外字>タグの付与された文字 (JIS 外字, Unicode 内字) と、初期状態から活字番号等のタグ付きで「=」となっている特殊活字 (JIS 外字, Unicode 外字) を洗い出すことで、新全集『今昔物語集』内の JIS X0213 外字を数え上げることが可能となる。

3.3 調査結果

3.3.1 JIS X0213 カバー率

調査結果は表2のとおり、のべ字数にして総計 749,922 字中 748,903 字、99.86% の文字が JIS X0213 で表現できることが明らかとなった。BCCWJ での JIS X0213 のカバー率はのべ字数で 99.96% (高田ほか 2009) であり、新全集『今昔物語集』はそれを 0.1 ポイントほど下回る結果となった。

表2 新全集『今昔物語集』活字調査結果

文字区分	のべ字数	異なり字数
JIS X0213	748,903	2,426
第1水準	742,536	1,610
第2水準	5,482	646
第3水準	603	87
第4水準	282	83
外字	1,019	193
Unicode 内字	583	104
Unicode 外字	436	89
計	749,922	2,619
X0213 カバー率	99.86%	92.63%

3.3.2 Unicode との比較

『日本語歴史コーパス』の構築には JIS X0213 文字集合の使用を予定しているが、新全集『今昔物語集』での JIS X0213 外字は異なりで 193 字。この中には Unicode では表現可能な文字も多数存在する。JIS 外字・異なり 193 字中、Unicode を使用すれば表現できる外字は 104 字。実例として頻度順に 10 位までを表 3 に示す。

表 3 Unicode で表現可能な JIS 外字 (頻度上位 10)

順位	文字	読みなど	頻度	順位	文字	読みなど	頻度
1	嗶	ののしる かまびすし	145	5	鼻	くさし	22
2	嗟	(夜が) あく	64	7	閑	まら	17
3	掬	りょう	28	7	隙	ひま	17
4	勾	こつがい (乞丐) の「がい」	27	9	蝥	こ	16
5	奄	いおり	22	10	噀	いゆ	13

Unicode の方が大きい文字集合であるため、JIS X0213 よりも高いカバー率を得ることができるのは当然予想される。『日本語歴史コーパス』での符号化文字集合には、BCCWJ 同様に JIS X0213 を用いる予定ではあるが、この機会に今一度、可能性として Unicode を使用した場合との比較を通して JIS X0213 採用の妥当性・有効性を検証してみたい。Unicode 採用の場合のカバー率は以下のとおり。JIS X0213 より、のべ字数にして 0.08 ポイント、異なり字数にして 3.97 ポイントほど上昇する。

表 4 JIS X0213 と Unicode カバー率の比較

	のべ字数		異なり字数	
	JIS X0213	Unicode	JIS X0213	Unicode
内字	748,903	749,486	2,426	2,530
外字	1,019	436	193	89
カバー率	99.86%	99.94%	92.63%	96.60%

また、各文字集合が持つ文字総数に対し、実際使用した異なり字数がどの程度か、という稼働率を算出したのが表 5 である。JIS X0213 では文字集合の 5 分の 1 が実際に使われたことになるが、Unicode を用意した場合、実際に使用するのは 2～3% である。

表5 JIS X0213 と Unicode 稼働率の比較

	JIS X0213	Unicode
文字集合の文字総数	11,233	96,477
使用した異なり字数	2,426	2,530
稼働率	21.60%	2.62%

稼働率の面からは、やはり Unicode は規模が大きすぎ、効率の面で望ましいものとは言えない。それを補って余りあるほどにカバー率の差が出るということもなく、実際のところ JIS X0213 でも大抵の文字はカバーできるという事実があるうえに、Unicode を使用したとしても外字問題は完全には解消されず、結局実数としてのべ 436 字、異なり 89 字の外字が残ってしまう。さらには既述したとおり、現時点での動作環境の問題などを含め考え併せると、『日本語歴史コーパス』においても JIS X0213 文字集合が、必要十分な、妥当な文字集合と結論付けられよう。

4. JIS 外字のコーパス上の扱い

4.1 「■」表示と別字代用

JIS X0213 採用の有効性がある程度確認できたとはいえ、上の調査結果のとおり、規格外字がゼロというわけではない。新全集『今昔物語集』での規格外字は比率にして全体の 0.14% に過ぎないが、のべ字数にして 1,019、異なり字数にして 193 であり、実に 200 種近くの JIS 外字が、現代活字で印刷された新全集『今昔物語集』内に存在することになる。

これら規格外字の電子的表現はどのようにすべきか。一つの一般的な方法は「■」等の表示を用い、規格外字であることを示す、という方法である（図 6）。JIS X0213 に準拠、という点では厳格な処理と言えるが、研究資料としての利用を考えた際には、「■」が使われている時点で、文字列上は語として取り出せなくなってしまう（「にえどの」という語としてはヒットしない）。

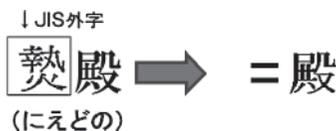


図6 「■」表示による処理例

そこで、「コーパス」という使用目的を鑑みるならば、JIS 規格内字に置き換える、という別の方法もありうる（図 7）。こちらの方針であれば、電子テキストとしても「読める」ようになり、文字列上も語として取り出すことが可能となる。ただし、別字代用を行うということは、原資料の字を改変するということになってしまう。

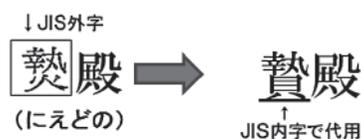


図7 別字代用による処理例

しかし、そもその話として、『日本語歴史コーパス』は、その性質上、テキストそのもの（ここでは『今昔物語集』そのもの）の文字研究に向くものではない。新全集として書籍化する時点で、文字の面だけでも相当手が加わっていることは事実であり、『日本語歴史コーパス』が原資料の活字を厳密に再現したところで、それは『今昔物語集』そのものではなく、あくまで「新全集版」の『今昔物語集』が再現されるだけである。そうであるならば、原資料の字形を厳密に再現しようとするものの意義はさほどあるとは言えず、のべ1,000字ほどの規格外字を「■」表示するよりは、別字で代用表示した方が、語彙・文法研究を主とするコーパスの用途からしても有用であろう。そこで、規格外字に対してはなるべく「■」表示にすることは避け、JIS規格内の別字で代用表示するという処理方針が妥当であると考えられる。

特に今回のように古典語を電子化しようとする場合、規格外字が大量に現れることが多く、規格外字となる文字の扱いをどうするか、また、規格内字で代用する場合、どの字で代用するかなどが常に問題となる。今回の調査対象である『今昔物語集』と時代も近く、規模も大きいものとしては、山田（1999）の『平安遺文』全文データベース化にあたっての外字調査が挙げられる。山田（1999）では、『平安遺文』を常用漢字を基本としてデータ化した場合の外字738字（異なり）が挙げられ、検索の便を考えるならば、やはり作字するか他の字に置き換えるかして「■」表示を解消していく処理が考えられると述べられている。また、古典資料のデータ化にあたってJIS外字をJIS内字で代用する際の対応関係を考える際の体系だった資料としては田嶋編（1980）の漢字シソーラスが挙げられる。田嶋編（1980）では、当時のJIS漢字を中心とした見出し字6567字の漢字に対し、関連字として『新字源』『大漢和辞典』に収録されている漢字が「本字、古字、別体字、譌字、同字、旧字、簡略字」として関係づけられて整理されており、データ量も豊富であることから、文字代用の参考として有用である。これらを参考にしつつ、規格内字での代用を通して、「■」表示を減らしていくことが、コーパスとしての有用性を高めることにつながると考える。

4.2 代用字の管理方針

規格外字はJIS内字で代用することを基本方針とするが、別字代用という処理を行うことで原資料の文字を改変してしまうということになるので、これに関しては何らかのカバーをしておく必要がある。そこで代用字に関しては、それが代用字であることを示すタグを付与することでコーパス作成時に手を加えたことを示し、さらにもとの字形を辿れるような情報を付与することとする。

さて、JIS X0213を新全集『今昔物語集』に適用した場合、規格外字は、大きく2種類に分け

ることができる。一つは JIS X0213 より大きな文字集合、Unicode では表現できるもの、すなわち「JIS 外字・Unicode 内字」、もう一つは Unicode でも表現できない「JIS 外字・Unicode 外字」である。

まずは Unicode では表現可能な Unicode 内字の場合。Unicode で表現可能であるならば、代用字であることを示すタグ内に Unicode での文字コードを記入しておくことで、本来の字形を知ることができるようにする (図 8)⁵。

蝨 (Unicode 内字 U+8745) を 𧈧 で代用した場合

→ <代用字 Unicode="8745">蝨</代用字>

図 8 Unicode 内字の代用字情報付与の例

これに対し、Unicode でも表現不可能な印刷所固有の特殊活字を用いて表現された Unicode 外字の場合、特殊活字の画像一覧をコーパス本体とは別に作成し、本来の字形を参照できるようにしておく。新全集『今昔物語集』の Unicode 外字については、本稿でも末尾にその一覧を掲げる。

𪗇 (Unicode 外字, 凸版印刷特殊活字)
を 漆 で代用した場合
→ <代用字 特殊活字連番="3">漆</代用字>

連番	特殊活字	代用字	活字画像一覧
1	猯	狗	
2	媚	媚	
3	𪗇	漆	

図 9 Unicode 外字の代用字情報付与の例

4.3 別字代用の種々

ここで別字代用の方針の概略を述べる。そもそも JIS X0213 文字集合にない文字を、多少の無理を承知で JIS 内字に置き換えていくという作業であるため、実際には置き換えが難しいものも多々見られる。置き換えの方針としては、まずは文字レベルでの置き換え、それが困難な場合は、語レベル、すなわち語の表記の変更というレベルでの置き換えるの可能性を考えている。

4.3.1 「文字」としての代用

別字代用という処理として典型的に想定しているのは、表示不可能な「文字」を、ほぼ等価とみなしてよい文字に置き換える、という処理である。しかし実際のところ、別字で代用すると言っ

⁵ ここでのタグの書式は、処理の理念を伝えやすくするための便宜上の書式であり、『日本語歴史コーパス』で電子化するための実際の書式そのものとは異なるものである。実際のタグの書式は『日本語歴史コーパス』全体のタグセットの問題に関わるので、ここでは触れない。

でも、どの字で代用するか、そしてその字で代用してよい根拠をどこに求めるか、といった問題が逐一生じる。古典資料を電子的文字集合で漢字処理をする際の体系立った指針となりうるものとしては、田嶋編（1980）が挙げられ、実際、新全集『今昔物語集』のJIS外字のうち、異なり14字に対しては田嶋編（1980）において代用すべきJIS内字を求めることができた。ただし、田嶋編（1980）に収録されている関連字はあくまで『新字源』『大漢和辞典』に採られるレベルの漢字であるため、日本独自と思われる異体字が多数出現する『今昔物語集』には対応しきれない面も多く、外字のうち、異なり179字は田嶋編（1980）でも見出すことができない。

そこで大部分の外字に対しては、個別に代用字の根拠を求めなければならない。まずは古辞書での異体字表記をあたるなどの手続きが考えられるが、文字処理は『日本語歴史コーパス』構築の作業工程全体の中では前処理の段階であり、現実的にはさほど時間がかけられない。そこで実際のところは新全集『今昔物語集』の注釈、『日本国語大辞典』での古辞書表記を確認する程度が限度である（《代用A》）。それでも代用字が定まらないものに対しては、外形上の類似性等から暫定的に代用字を決めておき、別案もありうる場合はコーパス完成までの間に再度検討することとする（《代用B》）。

《文字代用A》

新全集『今昔物語集』注釈等を参考に、JIS内字のある字と同字、またはJIS内字のある字の通字・俗字・異体字・誤字・省画・増画などとみなせるもの、また田嶋編（1980）の関連字に見出せるものについては、そのJIS内字で代用する。

新全集『今昔物語集』では、漢字の読みを定めるために、主に『和名類聚抄』（二十卷本）、『類聚名義抄』（観智院本）、『色葉字類抄』（上下：前田本、中：黒川本）、各種節用集が参照されており、読みを推定する際、その根拠としてこれら古辞書での異体字関係などが取り上げられている。新全集『今昔物語集』においてそのような注記が見られる場合は、それに従ってJIS内字を代用字とする。

褱 「褱」の異体字（新全集『今昔物語集』注、名義抄を根拠とする） ⇨ 褱 (1-74-71)

図10 文字代用Aの処理例1（新全集『今昔物語集』注、典拠：名義抄）

狛 「狛」の異体字（新全集『今昔物語集』注、字類抄「犬」の項を根拠とする） ⇨ 狛 (1-22-73)

図11 文字代用Aの処理例2（新全集『今昔物語集』注、典拠：字類抄）

また、田嶋編（1980）において、見出しJIS漢字との関係が整理されている字に関しても、それに従ってJIS内字を代用字とする。

蝻 田嶋編 (1980) 「蝻」の項に「俗字」として掲出 ⇨ 蚕 (1-27-29)

図 12 文字代用 A の処理例 3 (田嶋編 (1980) に掲出)

《文字代用 B》

注釈等を参照しても現代の通用字との関係が明らかな文字に関しても、本文に与えられた読みと同訓で、字形の近い JIS 内字（偏・旁いずれかの差異や有無など）があれば、その JIS 内字で代用する。根拠という面では弱いことは否めないが、データとしての利便性から「■」表示を減らすことを優先する。

捻 本文「攝津ノ=持寺」とあり、訳文では「総持寺」 ⇨ 総 (1-33-77)

図 13 文字代用 B の処理例

表 6 のとおり、《文字代用 A》《文字代用 B》を適用した段階で外字のべ 1,019 字中 910 字が処理可能となり、カバー率の面でも BCCWJ の 99.96% を上回るようになる。注釈などの根拠をもって異体字と認定できる《文字代用 A》自体はさほど多くはないが、《文字代用 B》のように代用の範囲をやや広めにとるだけで、JIS 外字の 9 割を解消することが可能となる。

表 6 文字代用を適用した場合の処理可能文字数の変化

	JIS X0213	+ 文字代用 A	+ 文字代用 B
処理可能文字総数	748,903	749,064	749,813
新たに処理できる文字総数 (のべ)	—	161	749
外字総数 (のべ)	1,019	858	109
カバー率	0.99864	0.99885	0.99986

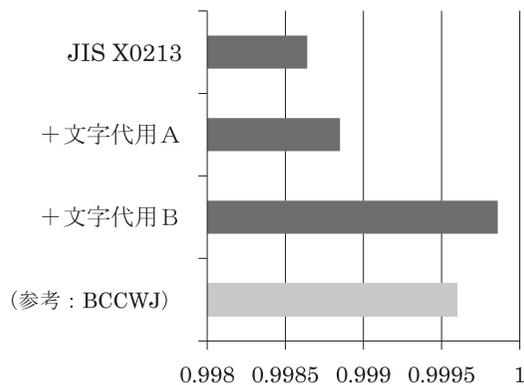


図 14 新全集『今昔物語集』カバー率の比較

4.3.2 「語」の表記としての代用

一般的な文字レベルでの代用として許されるのは上記《文字代用 A》《文字代用 B》までであろうが、新全集『今昔物語集』の外字の中にはこの処理では収まらないもののべ 109 字ほど残る。《文字代用 A》《文字代用 B》の処理を施しても外字として残るものの類型としては、表 7 のようなものが挙げられる。

表 7 別字での代用表示が困難な字形例

a	懨	おもしろし	1 字で表現できる類字なし。「おもしろし」と読むことが可能な字としては「賞」「怜」などあり。
b	幪	おびとり	1 字で表現できる類字なし。「おびとり」と読むことが可能な 1 字自体なし。
c	悌裕	もこう	2 字に対する読み。「もこう」の通常の表記は「抹額」で、大きく異なる。
d	濬	濬■ (しゅうちゅう) の「ちゅう」	代用字候補なし。字音。
e	壽	ひひめく	代用字候補なし。和語。

一般的な処理としては、これらに関しては外字として「■」表示すべきであろうが、言語研究の資料としての利便性を鑑み、可能な限り文字を充て、一切「■」表示はしないというデータ作成を目指す場合は、これらに関する代用処理は、文字のレベルを越え、「語の表記」というレベルでの代用を行うことになる。以下では、表 7 に示されたようなタイプの外字をも何らかに代用表示するなら、どのような処理になるのか、その可能性を検討する。

表 7 の a は「おもしろし」と読む字だが、JIS 内字には代用できるような類字が見当たらない。データ上「おもしろし」と読めさえすればよいというところまで譲歩するならば「面白」などを充ててしまうという方法がありうるが、この処理はもはや「文字レベルの代用」ではない。このような代用をするのであれば、これは「語に対する表記を換える」という、語の表記レベルでの代用ということになる。

また、このように思い切った代用をするにしても、出来ることならデータ上の文字数を変えないで済ませたい。そこで、読みをそろえるだけなら「面白」でも構わないが、極力 1 字で表現できる場合は 1 字での表現を優先すべきであろう。実際はデータ化の作業工程上、さまざまな古辞書にあたるということは困難であろうが、「おもしろし」に関しては『日本国語大辞典』第 2 版での古辞書表記を確認するだけでも、「賞」「怜」など、1 字での代用候補を得ることができる。適切な代用字を探すためにさまざまな古辞書をあたりたい場合もあるが、文字処理はコーパス作成という流れの中では前処理の工程にあたり、コーパスの完成を目指すならここに必要以上の時間はかけられない。そこで作業の実際上は『日本国語大辞典』第 2 版での古辞書表記の範囲内で代用字を探す、というのが現実的であろう。これに対し表 7 の b「おびとり」は、『日本国語大辞典』

第2版での古辞書表記をあたっても、1字での代用字を充てることはできない。このような場合、「帯獲」など、2文字での代用表記しか方法がなく、これらまで代用表記するならば、元のデータと文字数を変えざるを得なくなる。

表7のcは2文字で「もこう」と読むタイプである。このような場合、「もこう」の通常の表記「抹額」に、2文字合わせて置き換えることになる。

表7のd, eはそれぞれ、漢字で代用すること自体が困難な事例である。dは「ちゅう」と読む字であるが、これに代用できるような漢字はJIS内字には存在しない。また、e「ひひめく」も、この語自体をJIS内字で漢字表記することはできない。これらまで「■」にはせず、「読める」状態で電子化するならば、「ちゅう」「ひひめく」というように仮名で表示するしかないだろう。

以上、文字代用を適用した上で残る外字についても、「■」表示しないとするならどのような表示方法がありうるかを検討した。処理方針は以下のようにまとめられる。

【「語の表記」レベルでの代用指針】

仮に「■」表示を一切しないテキストを完成させねばならないのであれば、文字代用の適用外となる文字に関しては、「語の表記」というレベルでの代用を行う。

- (1) 表記の代用に関しては、漢字表記で代用することを優先する。かつ、漢字1字に対しては漢字1字で代用できるものを優先する。
- (2) 漢字で代用表記が不可能な文字に対しては、仮名表記に開く。

文字数の変更や、漢字から仮名への変更など、通常の文字代用よりもかなり思い切った変更が加えられることになるが、仮にこれらの処理までを行うとすると、新全集『今昔物語集』の「■」表示はようやくゼロとなる。

5. おわりに

以上、新全集『今昔物語集』での活字調査の結果、JIS X0213のカバー率はのべ字数にして99.86%を実現しており、BCCWJを0.1ポイント下回る程度であるという事実からは、『日本語歴史コーパス』の符号化文字集合としても、JIS X0213が妥当な文字集合であると結論付けられよう。また、JIS X0213で電子化した場合に外字となる1,019字に対しては、別字で代用という手法で「■」表示を減らし、コーパスとしての実用性を高める方法を検討した。一般的な文字代用のレベルで、JIS外字の約9割は解消できる。残る1割に対しては、語レベルでの代用表記を認め、文字数の変更も認めるのであれば「■」表示をゼロにすることが可能となることを示した。

ただし実際には、代用字に選ばれる文字は必ずしも最初から一つに決まるとは限らない。別字代用といっても、《文字代用B》の場合は代用字の候補が複数ありうる場合も存在する。また、特に「語レベルでの代用」に関しては、そのような処理自体の妥当性に関して、理念の面でも、実用性の面でもさらに検討していく必要があるだろう。現在の作業段階は、代用という方針で限界まで外字処理を減らす可能性を探る、という方針のもと、新全集『今昔物語集』のコーパス用デー

タを試験的に整備している段階であり、代用字に設定した文字にも暫定的なものも含まれる。今後、代用を適用する範囲を確定し、そのうえで最適な代用字も確定しコーパスの完成を目指す、というのが課題となる。

また、別字代用という作業は、今回の新全集『今昔物語集』だけの問題にとどまらず、『日本語歴史コーパス』全体で統一的に検討すべき課題である。例えば現在は予備調査の段階だが、新全集『日本霊異記』においても相当数、JIS 外字となる異体字が見られ、その中には今回の新全集『今昔物語集』調査で見られたものと同形のものも複数見られる。将来的にはそれら全体を異体字辞書として管理していく必要がある。また今回の調査だけでも田嶋編 (1980)、山田 (1999) にも見られる字形が複数現れた (表 8)。今後さらに事例を蓄積して、古典資料を電子化する際の文字代用のありかたを、体系的に整備していきたい。

表 8 新全集『今昔物語集』外字のうち、田嶋編 (1980)、山田 (1999) に見られるもの

	田嶋編 (1980) 掲載	山田 (1999) 掲載	(新全集『今昔物語集』JIS 外字)
Unicode 内字	12	13	(104)
Unicode 外字	2	5	(89)
計	14	18	(193)

末尾に資料として、JIS 外字のうち Unicode でも表現不可な字形と、その代用案を掲げる。

参考文献

- 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』381-388.
- 田嶋一夫 (編) (1980) 『データ処理システムの為の漢字ソース [試作版]』東京: 「計算機による日本語文字システムの実用的処理」班 (文部省科学研究費による特定研究「言語」研究代表者 山中光一).
- 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). 東京: 国立国語研究所.
- 堤智昭・須永哲矢・高田智和 (2012) 「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」—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論文集 2012』171-178.
- 山田邦明 (1999) 「『平安遺文』全文データベースと外字」『人文学と情報処理』25: 63-74.

関連 Web サイト

- 現代日本語書き言葉均衡コーパス (国立国語研究所) http://www.ninjal.ac.jp/corpus_center/bccwj/
- 日本語歴史コーパス (国立国語研究所) http://www.ninjal.ac.jp/corpus_center/chj/

Digitization of Typeset Books in Constructing the Corpus of Historical Japanese: The Case of the Shōgakukan (SNKBZ) Edition of the *Konjaku Monogatari*shū

SUNAGA Tetsuya^a

TSUTSUMI Tomoaki^b

^aAdjunct Researcher, Center for Corpus Development, NINJAL [–2013.03]

^bDoctoral Student, Tokyo University of Agriculture and Technology

Abstract

Digitizing characters not included in the standard set is an urgent problem for electronic corpora of historical documents. Such non-standard characters have hitherto been replaced with the symbol “■” in digital corpora, which is quite inconvenient for users. In constructing the Corpus of Historical Japanese, the current Japanese standard for character codes, JIS X0213, will be adopted for the digitization of printed documents. This paper first examines the efficacy of JIS X0213 for typeset versions of old texts. A thorough investigation of the Shōgakukan (SNKBZ) edition of the *Konjaku Monogatari*shū found that JIS X0213 covers 99.86% of the total character tokens. The paper then proposes a substitution system for the remaining 0.14% of the characters not covered by JIS X0213. The idea is to replace these non-standard characters with similar characters that are included in JIS X0213 while retaining information about the original characters for reference. All the non-standard characters in the Shōgakukan (SNKBZ) edition of the *Konjaku Monogatari*shū are listed at the end of the paper along with their replacements.

Key words: construction of electronic corpora, JIS X0213, non-standard character processing, *Konjaku Monogatari*shū

資料 新全集『今昔物語集』Unicode 外字となる特殊活字一覧

No	活字字形	読み他, メモ	代用字	No	活字字形	読み他, メモ	代用字
1	嗟	あける	晞	17	墜	おそう	壓
2	曙	あける	曙	18	壓	おそう	壓
3	晡	あける	晞	19	癩	おそう	壓
4	杵	あてる	充	20	壓	おそう	壓
5	踪	あと	跡	21	癩	おそう	壓
6	蚣	虻	虻	22	肺	おそれる	怖
7	掃	あわける	褫	23	幘	おびとり	帯獲
8	矜	あわれむ	矜	24	慈	おもしろい	賞
9	稜	いなづか	積	25	廻	おもねる	洵
10	狗	いぬ	狗	26	匄	丐	丐
11	媚	うつくしい	媚	27	匄	がい	丐
12	洩	漆	漆	28	搏	かこい	搏
13	蝨	えびら	蠶	29	補	蒲	蒲
14	咲	えむ	咲	30	衤	かみ	衤
15	啞	おうし	啞	31	勁	かみがき	髮
16	癩	おそう	壓	32	獺	かり	獺

No	活字字形	読み他、メモ	代用字	No	活字字形	読み他、メモ	代用字
33	ケ	くさ	艸	49	咲	しょう	咲
34	𪗇	くさい	臭	50	搯	じょう	掾
35	𪗇	くさか	日下	51	裝	装束の「しょう」	装
36	憤	くちばしる	𪗇	52	鐵	せん	錢
37	燻	煙	煙	53	率	卒	卒
38	劔	劍	劔	54	慥	たしか	慥
39	𪗇	こくわ	こくは	55	慥	たしか	慥
40	𪗇	こぼれる	泛	56	憑	たのみ	憑
41	薦	こも	薦	57	壇	壇	壇
42	賽	賽	賽	58	揅	ちやく	揅
43	崎	崎	崎	59	𪗇	つか	𪗇
44	醒	さめる	醒	60	𪗇	つか	𪗇
45	𪗇	しぐれ	𪗇	61	裹	つつむ	裹
46	𪗇	しぐれ	𪗇	62	勒	つとめ	勒
47	𪗇	しごつ	𪗇	63	𪗇	つび	開
48	𪗇	しゅうと	姑	64	揅	つむ	採

No	活字字形	読み他, メモ	代用字	No	活字字形	読み他, メモ	代用字
65	蕨	なつめ	棗	78	壽	ひひめく	ひひめく
66	恐	なまじい	愁	79	慕	ぼ	暴
67	愁	なまじい	愁	80	瓮	ぼん	盆
68	熨	にえ	贄	81	殺	まつわる	纏
69	慥	にくむ	悪	82	媼	姪	姪
70	慥	にくむ	悪	83	裕	もこうの「こう」	額
71	肺	はぎ	脛	84	怙	もこうの「も」	抹
72	啟	はげます	勵	85	嬪	やもめ	孀
73	枅	はじかみ	枅	86	寡	やもめ	寡
74	稜	稜	稜	87	寡	やもめ	寡
75	魘	ぱつ	魘	88	昧	らい	耒
76	扱	はらえ	扱	89	篋	わく	柶
77	頽	ひたい	額				