活字資料のコーパス化における外字チェックと処理

須永 哲矢
昭和女子大学 人間文化学部

堤 智昭
東京農工大学 工学府

歴史的作品の活字資料からコーパスを作るときの文字処理の方法を定めた．活字資料の電子化にあたっては，外字処理・字体包摂の２つが中心課題となるが，両者をまとめて処理できるツールを利用し，作業方式を確立することにより，もとになるテキストの特性によらず，統一的な処理を可能にした．本作業のために開発したツールと本稿で提案した一連の作業手順は，コーパス構築という作業のみならず，活字研究にも適用可能である．研究利用の例として，小学館新編日本古典文学全集『日本霊異記』等の漢字活字を調査し，JIS X0213　や　UniCode でどの程度再現できるかを明らかにした．

Extracting and processing external characters upon constructing corpora of printed documents

Tetsuya Sunaga
Showa Wemen's University

Tomoaki Tsutsumi
Tokyo University of Agriculture and Technology

The paper proposes a new processing procedure of external characters included in printed historical texts, which is essential to constructing an electronic corpus of historical documents. Digitization of printed historical documents so far has two major problems to be dealt with: representation of external characters and establishment of unification standard. We present a solution to the problems, introducing a new software tool which handles the two problems altogether. By applying the tool, the characters can be processed uniformly, regardless of the document variation. Furthermore, the processing tool and a series of procedures or our proposal can also be applied to character research. In the paper, we present a small sample investigation on the external characters of SNKBZ, Shogakukan, revealing what percentage of the total printing types JIS X0213 and Unicode respectively can represent.

１．はじめに

言語研究資源として，紙媒体の活字資料を電子化する際には，元のテキストの活字をどのように処理・表現するかということを常に考えねばならない．発表者らは時代・ジャンルとも多岐にわたる資料のコーパス化を行っているが，その作品の時代や電子化素材となる資料の形式（現代の活字で出版されたものか，100 年ほど前の活字本か，など）によって，文字処理に注がれる力点は変わってくる．しかし，各時代を横断して研究できる資源としてのコーパスを構築するのであれば，資料ごとの時代や原資料の形式の差異を想定したうえで，すべての資料が統一された処理方針で電子化され，共通規格としての文字処理がなされることが望ましい．本発表では，発表者らがコーパス構築の際にこれまで試みてきた文字処理作業の統一工程をまとめ，その作業工程が活字研究にもつながることを報告するものである．

２．原資料の活字を電子化する際の問題

発表者らが構築している『日本語歴史コーパス』『近代語コーパス』等では，活字化された資料を原資料として電子化を行う．よって，本研究における電子化では，手書きの文字を対象とすることはなく，紙に印刷された活字を電子的な符号化文字に置き換える，という作業となる．平安期を中心とする古典資料は，現在書籍形態で出版されている活字本を原資料とする．近代の活字資料に関しては，当時の活字本そのものを原資料とする．

一般に，紙媒体の文書を電子化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，主に２つの問題と向き合わねばならない．１つめは，資料に出現した文字を文字集合のどの区点位置に対応させるべきかという問題（字体包摂の問題，粒度の問題）であり，もう１つは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）である．

これら２つの問題のうち，どちらがどの程度重い問題になるかは，電子化対象のテキストの性質による．

近代の活字本を原資料とする場合，現代の活字とは字体・字形に差異がみられる場合がある（図1）ため，それらを現代の文字集合のどの符号位置に対応させるか，あるいはさせないか，といった問題に逐一指針を与えねばならない．

図 1　近代の活字（左）と現代の活字

また，現代の文字集合にはそもそも含まれていないような規格外字が多様に出現するため，それらの扱いにも検討を要する．

一方で，古典作品と言っても現代出版されている活字書籍を原資料とし，活字書籍の状態が再現できれば良しとする場合，写し取るもとの文字が現代の活字字形であるため，近代活字資料のように字体包摂の問題は特に深刻化しない．外字問題に関しても，現代の日本語資料を現代活字で表現した，一般的な文書であれば，国内規格 JIS X 0213 で必要十分であることがある程度立証されており，高田ほか（2009）[5]によれば，およそ 5,800 万字の現代日本語コーパスで，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．ただし，これはあくまで現代語を現代活字で表現した場合であって，古典資料を現代活字で表現した書籍の電子化となると，事情が異なってくる．例えば古典資料の書籍版では，原資料の文字を表現するための出版社固有の活字体が用いられることがあり，現行の符号化文字集合では対応できない文字も多い．須永・堤（2013）[4]では，小学館新編日本古典全集版の『今昔物語集』での活字調査を行い，JIS では表現不可能な外字が異なりにして 193 字，うち 89 字は UniCode でも表現不可能なことを明らかにした．

図 2　JIS 外字となる現代の印刷活字（小学館書籍）

このように，現代活字で書籍化されている古典資料を原資料にする場合には，近代活字資料を対象にする場合ほど字体包摂の問題は深刻化しないにせよ，一般的な現代語テキストを電子化するよりも多くの外字問題に直面することになる．

以上のように，『日本語歴史コーパス』『近代語コーパス』では，字体包摂の問題，規格外字の問題と向き合いながら電子化を行わねばならず，原資料が現代活字の場合は主に後者の問題，原資料が近代の古い活字であれば前者・後者両方の問題が顕在化する．

3．コーパス化のための文字処理方針

『日本語歴史コーパス』『近代語コーパス』の文字処理方針は，基本的には以下のとおりである．

(1)使用する文字集合は JIS　X0213 とする．符号化文字集合 JIS　X0213 は，『現代日本語書き言葉均衡コーパス』でも実用性を実証されたこともあり，『日本語歴史コーパス』『近代語コーパス』でも JIS　X0213 に依拠して文字処理を行う．

(2)JIS 包摂規準に従い，字体包摂を行う．

基本的な処理方針は以上のとおりだが，コーパス，すなわち研究資源としてのテキストとしては，これだけでは不十分である．これまでに表 1 に示すような各テキストの文字処理を行ってきたが，上記(1)(2)の方針のみで文字処理を行うと，「=」表示で読めなくなる外字が大量に出ることが明らかになっている．例えば『明六雑誌』では(1)(2)の適用のみで文字処理を行った場合，のべ 13 万 7897 字中、2100 字が外字「=」表示となってしまうことが明らかになった（須永・堤・高田 2011[1]）．

表 1　文字処理を試みたテキスト

現代活字化されているた書籍資料	小学館新編日本古典文学全集(『今昔物語集』『日本霊異記』『徒然草』『方丈記』『沙石集』『十訓抄』ほか)
近代活字資料（発行当時の活字のまま）	『明六雑誌』『国民之友』

言語研究用の資料としては，原資料の厳密な再現以上に，語が語として取り出せること，すなわち「読める」ことの方が望ましい．

まず，(2)字体包摂の問題に関しては，特に近代活字資料では，JIS 包摂規準では想定されていない差異があり，それらにもある程度の類型を認めることが可能である．そこで，図 3 に示すように，近代活字も視野に入れた包摂規準の拡張案を制定し，拡張した包摂規準に則って字体包摂を行うことで，外字を減少させることとした．

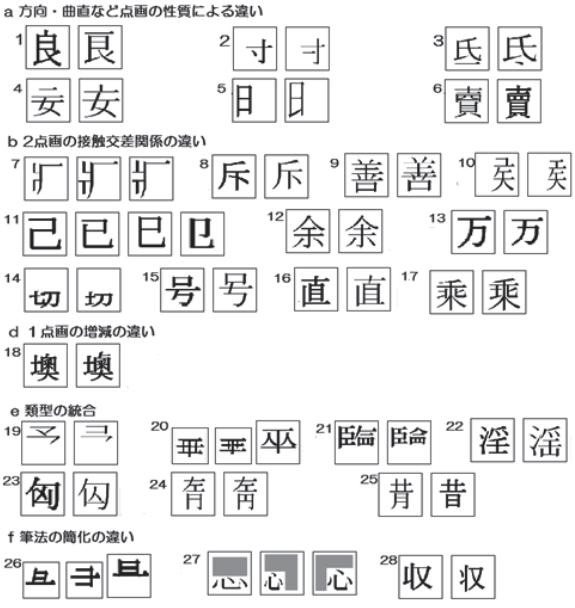

図 3　追加設定した包摂規準の例（須永ほか 2011 [3]

(1)の規格外字の問題は，より大きな文字集合，例えば JIS ではなく Unicode を適用すれば解消される可能性があるように見えるかもしれないが，実際には既存の文字集合で歴史コーパス構築に適した文字集合は存在しない．そもそも現代とは文字体系の異なる歴史的資料を電子化する場合，現代の文字集合では対応しきれないことも多く，書籍版の『今昔物語集』を例にとると，JIS X0213 では異なりにして 193 種の外字が出るが，このうち Unicode を使えば解消できるのは 104 種までで，結局 89 種もの文字が表現できないまま残ってしまう（須永・堤 2013[4]）．そのため，作業コストや現時点でのユーザー環境を考慮しても，依拠する文字集合は JIS X0213 内に収め，その代わりそのままでは外字となってしまう文字は，可能な限り JIS 内字で代用する，という方針を定めた．

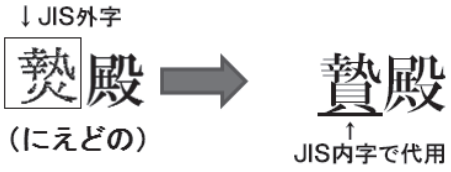

図 4　別字代用の例

これらの処理により，「=」表示を大幅に減少させ，言語研究としての利用に耐えうるテキストが構築できると考えるが，追加設定した包摂規準や，別字代用というのはあくまでコーパス内での臨時的措置である．そのため，このような臨時的な措置を施した箇所には，タグの形で，処理を行ったという情報を付与しておく．

4．電子化に際しての文字処理工程

以上紹介した処理内容を含んで，コーパスの構築を行う．

コーパス構築にあたっての文字処理の手順を以下に示す（図 5）．

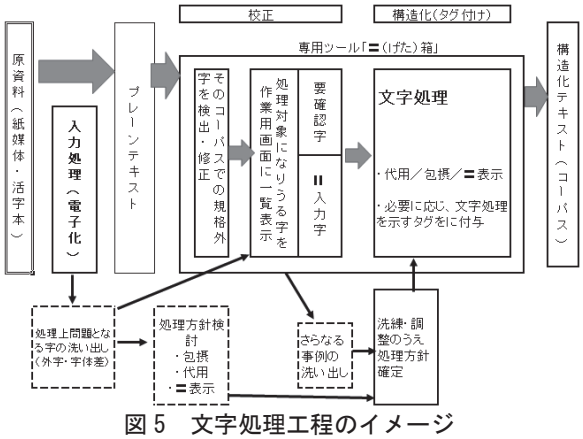

図 5　文字処理工程のイメージ

以下，文字処理工程で行う作業を，内容的に大きく 5 つに分けて紹介する．

(a) 1 次入力（プレーンテキスト）

まずは紙媒体・活字本の原資料をに対し，電子化したプレーンテキストを得ることから始まるが，この段階では入力された文字の処理はさほど気にしない．例えば特に近代活字資料に多様に見られる異体漢字等に関しては，どの程度までを包摂し，どの程度から別字とするかを，1 次入力の前にあらかじめ決めておくというのは現実的ではない．異体字ははじめからそのバリエーションがわかっているものではなく，入力しながら作業者がさまざまな差異を見つけていく，という性質のものだからである．しかも，ある時点である字の差異に気付いたとしても，それ以前に入力された同じ文字にも同様の差異があったのかを，入力途中で遡って確認するのは非効率である．そこで，1 次入力段階では，入力しながら，差異のある可能性がある字を洗い出す，という作業を重視し，処理の統一化はこの段階では行なわない．よってこの段階では，原資料で全く同じ活字字形である文字が，作業者の判断や見落としによってある箇所では「=」表示，別の箇所では内字で表現，というような不統一を持つ場合があるが，それらの統一は(b)の検討を経て，(d)の工程で統一化を図る．

(b)処理方針の検討

1 次入力が終了した時点で，異体字・外字の事例が蓄積されるため，これらについてコーパスでの利用という目的に合致するような形で包摂・別字代用といった処理方針を定める．

(c)確認・処理対象字の洗い出し

(a)での検討課題の採集，(b)での検討を経て，1 次入力されたデータから，確認・処理対象となる文字を洗い出す．洗い出す文字は=」入力された文字，および 1 次入力段階で確認が必要とされた文字で，この 2 種がまず確認・校正の対象となる．異体字や規格外字は，事例として目立ちやすいため意識的に処理されやすいが，見落とされやすいのが，データ上に使用されている文字の全てが規定の文字集合（ここでは JIS　X0213）内に収まっているかである．入力段階で JIS 外字ではあるが，Unicode 内字ではある文字などは，JIS 外字であることが気付かれないまま入力されてしまうことも多いため，1 次データ完成時に，「=」入力字のみならず，文字として入力されている文字の中で，規格外字となるものはすべて洗い出し，内字代用の対象とする．逆に，1 次入力時に，実は JIS 内字で表現できるにもかかわらず，見馴れない文字であるために外字だと判断され，「=」入力されてしまっている文字がある可能性もあ

る．入力作業には一定数ミスも伴うので，このようなミスを発見し，修正するのも，校正・統一処理段階での重要な作業である．校正・統一段階で洗い出される字は表 2 に示した通り，「=」入力字のほか，内字入力された文字のうち異体字の可能性のあるもの，規定の文字集合外の文字の 2 種となる．

表 2　校正・処理対象となる文字のタイプ

1次入力時点で内字	作業ミス①異体字の見落とし（→コーパス上は拡張した包摂規準適用／別字代用で内字表現拡張した包摂規準適用で内字表現）
	作業ミス②使用する文字集合外の文字を入力（JIS 外字で Unicode 内字，など）（→コーパス上はJIS内の別字で代用）
1次入力時点で外字（=表示）	作業ミス③入力者が内字であることに気付かず=入力（→内字に入力訂正）
	作業ミスなし（本来，外字と判断されるのが正しい字）（→コーパス上は拡張した包摂規準適用／別字代用で内字表現）

「=」入力された文字，および 1 次入力段階で確認が必要とされた文字，入力されてしまった規格外字に対し，包摂・代用処理などを行いを，処理の統一を図るとともに，必要に応じて処理内容をタグに記録する．

(e)文字集合内に収まっていることの最終確認

文字の校正，処理の統一化を行った最終段階で，テキスト内のすべての文字に対し，規定の文字集合（ここでは JIS　X0213）内に収まっているかを確認する．

以上の工程を経ることで，統一規格で文字処理がなされたコーパスが完成する．なおこのうち(c)以降の作業には，専用に開発した校正ツール「=（げた）箱」を利用することで，作業の効率化・正確化を図っている．

5．校正ツール「=（げた）箱」

上記(c)～(e)，すなわち確認・校正対象文字の抽出から実際の書き換え・情報付与，さらに文字集合内に収まっていることの最終確認までの一連の作業には，このために開発された専用ツール「=箱(げたばこ)」（図 6）を利用する．「=箱」は，指定された文字をテキスト内から抽出してリスト化，リスト画面での操作によって元テキストの書き換え，タグ付与までが行えるツールであり，原資料を pdf 化しておけば，原資料の該当箇所を参照することも可能である．このツール利用により，煩雑でミスも多い文字処理作業が，作業時間にして手作業比で 1/10 前後に短縮でき，作業ミスも減らせることが確認されている（堤・須永・高田 2012[7]）．

図 6　校正用ツール「=（げた）箱」操作画面

「=箱」では，抽出対象として指定する文字を自由に指定することができる．「=」および，「入力中に異体字が発見され，確認が必要な文字（1 次入力作業中に一覧を作成しておく）」を指定することで，処理すべき文字を集中・一括して洗い出し，各字の確認・校正・タグ付けまでの作業を同時に行なうことが可能となる．また，指定した文字集合に「含まれない」文字を抽出することも可能であり，こちらの機能を用い，「JIS X0213 文字集合」外の文字を抽出するよう指定することで，Unicode で入力されているが JIS 外字となる文字を抽出し，修正することで規定の文字集合内に収まったテキストを作成することが可能となる．文字集合の指定は，文字ごとに追加・削除が自由に可能であるため，「JIS X0213 のうち，康熙別掲字のみ使用しない」「常用漢字に抑える」など，必要に応じて柔軟に指定することができる．

また，「=箱」は，処理対象文字の抽出と校正を独立に行えるため，処理対象文字の抽出を行った結果，処理すべき文字が極めて少なく，手作業で済む場合には校正を手作業に切り替えることも可能である．校正作業自体は，作業量に応じて「=箱」利用の場合と手作業の場合とが選択されるが，いずれの場合でも，最終的には「=箱」の外字抽出機能を利用し，テキストに使用された文字が JIS X0213 に収まっているかを確認する．

6．小学館新編全集の漢字活字

「=箱」はコーパス用文字処理ツールではあるが，その機能を用いて作業をする中で，現代の活字・符号化文字集合の実効性を知ることもでき，研究利用も可能なものといえる．須永ほか(2013)[3]までは，主に近代活字の実例に関して報告してきたので，今回は最後に補足的な例示とし

て，現代活字書籍の古典作品の場合の実例を紹介したい．歴史コーパス構築の際の原資料書籍とされる，小学館新編日本古典文学全集をコーパス化する過程で明らかになった漢字活字のありようを報告する．

小学館新編日本古典文学全集版の古典資料に関しては，以前に『今昔物語集』に関して調査を行っており（須永・堤 2013[5]，表 3，表 4），現代の活字印刷資料であっても，原資料が古典作品の場合，一般的な現代語資料よりも外字が多くみられることが明らかになった．

表 3　小学館新全集『今昔物語集』と JIS, Unicode

	JIS 内字	JIS外字		計
		Unicode 内字	Unicode 外字	
異なり	2,426	104	89	2,619
のべ	748,903	583	436	749,922

表 4　小学館新全集『今昔物語集』カバー率

	JIS X0213 カバー率	Unicode カバー率
異なり	92.63	96.60
のべ	99.86	99.94

今回は新たに『日本霊異記』『徒然草』『方丈記』『沙石集』『十訓抄』について，「＝箱」の外字抽出機能を用いて文字処理を行ったが，この中では，特に外字が多かったのが『日本霊異記』であった．そのため，ここでは主に『日本霊異記』の活字を中心に報告する．

小学館新全集版『日本霊異記』は，現代の活字で書籍化さたテキストであるが，平安初期の作品であるため現代では見慣れない漢字が多く含まれる．出版にあたっては印刷所固有の活字を用いればよいが，これを電子化となると符号化文字集合は限られる．本作品のコーパス化・文字処理を「＝箱」を用いて行ったが，JIS 外字と＝表示の抽出を行ったところ，表 5，表 6 のとおり，『今昔物語集』ほどではないにせよ，一般的な現代語資料よりも多くの外字が存在することが明らかになった．

表 5　小学館新全集『日本霊異記』と JIS, Unicode

	JIS 内字	JIS外字		計
		Unicode 内字	Unicode 外字	
異なり	2,367	88	13	2,468
のべ	150,965	363	48	151,376

表 6　小学館新全集『日本霊異記』カバー率

	JIS X0213 カバー率	Unicode カバー率
異なり	95.99	99.51
のべ	99.73	99.97

のべ字数としては，JIS X0213 でもカバー率 99.73%を実現しており，現代語での 99.96%よりはやや劣るものの，文字集合の規模，作業コストを考慮すると，JIS X0213 が必要十分な文字集合と言えよう．もちろん Unicode を利用すればカバー率はさらに上昇するものの，結局外字ゼロにはならず，異なりにして 13 字の外字が残る．『日本霊異記』の JIS 外字のうち，Unicode でも外字となる実字形の一覧を以下に掲げる．『日本霊異記』の Unicode 外字はのべ 13 字であるが，うち 5 字は『今昔物語集』にも出現する字であることが明らかになった（なお，小学館新編古典文学全集の両者の印刷所は異なり，『日本霊異記』は図書印刷，『今昔物語集』は凸版印刷である）．

表 7　小学館新全集『日本霊異記』Unicode 外字

実字形	読みほか	『今昔物語集』
	よる，たのむ	あり
	にくむ	あり
	＝離で「となかる」	
	おもしろい	あり
	こむら	
	ひらめく	
	しりぞく	
	ささめく	
	とじ	
	おもねる	あり
	くぼ	あり（『今昔』では「つび」）
	「ひき」の「き」	
	はいや	

『日本霊異記』の他に調査した新編全集のテキスト,『徒然草』『方丈記』『沙石集』『十訓抄』についても外字抽出・文字処理を行ったが,『今昔物語集』『日本霊異記』ほど外字は見られず，ほとんどが JIS 内字であり，わずかな JIS 外字も,『沙石集』を除いて Unicode で表現できるものばかりであった．なお，作業面においては，これら 4 作品に関しては，外字抽出後の書き換え・タグ付けは手作業の方が早いと判断し，手作業で行った．4 作品のうち唯一 Unicode 外字を含んでいた『沙石集』での外字を，参考までに表 9 に掲げる．

表 8　小学館新全集，ほか 4 作品の外字（異なり）

	徒然草	方丈記	沙石集	十訓抄
JIS 外字	0	1	10	3
Unicode 外字	0	0	5	0

表 9　小学館新全集『沙石集』Unicode 外字

実字形	読みほか	『今昔物語集』
棹	ちりとり	
愍	なまじい	あり
旃	せん	
匂	がい	
旱	おわる	

7．おわりに

以上，ここまでの作業を踏まえ，活字資料のコーパス化にあたっての文字処理の統一工程をまとめた．上記のような工程で作業することで明らかになってくることも多いが，それは同時に実際の作業面では課題に直面し続けるということでもある．近代活字では，異体活字の問題が大きく，ここまで事例を蓄積してきたが（須永・堤・高田 2011[1]，須永ほか 2013[3]など），包摂規準の整備，近代活字のデザイン差の具体化等，作業面での検討課題はまだまだ残されている．また，今回見たように，現代活字資料であっても，古典作品をもとにした資料では外字が多く存在し，複数作品に同時に見られる外字も存在する．将来的には字体参照の便も考慮したうえで，外字の一覧の整理・管理の在り方を検討していかねばならない．

今後とも事例の蓄積を重ねながら，コーパス化における文字処理の在り方を検討し続ける予定である．

付記

本研究は，日本学術振興会科学研究費基盤研究（B）「和漢の両系統を統合する平安・鎌倉時代語コーパス構築のための語彙論的研究（24320086, 研究代表者：田中牧郎）」による成果の一部である．

参考文献

[1]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）．

[2]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）』.

[3]**須永哲矢，堤智昭，近藤明日子，木川あづさ，服部紀子**：明治中期雑誌の異体漢字と JIS 漢字ー『国民之友』を事例としてー-，人文科学とコンピュータシンポジウム「じんもんこん 2013」論文集（2013）．

[4]須永哲矢，堤智昭：『日本語歴史コーパス』のための活字書籍の電子化―小学館新全集『今昔物語集』を事例として―，国立国語研究所論集

[5]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）．

[6]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）．

[7]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）