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−

堤 智昭* 須永 哲矢† 高田 智和†
東京農工大学 電子情報工学専攻* 国立国語研究所†

コーパス作成時の文字校正処理を支援するため，校正作業を効率化する支援ツールを設計，実装した．本ツールは，校正テキストに作業用の XML タグを付与し，そのタグを用いて校正対象文字をリスト化し，見やすく使いやすい簡単な GUI を用いた校正作業環境を実現した．また，作業用 XML タグを最終的にコーパスの入力仕様に変換することで，校正と構造化を同時に支援した．本ツールを用いて『明六雑誌』の校正作業を行ったところ，ツールを使用しなかった場合に比べて，作業時間を最大約 10 分の 1 にまで効率化することができた．さらに，校正作業のミスを削減する効果も確認された．

Support Tool for Proofreading and Structuration for Corpus Text

Tomoaki TSUTSUMI Tetsuya SUNAGA Tomokazu TAKADA
Faculty of Information Science
Tokyo University of Agriculture and Technology
National Institute for
Japanese Language and Linguistics

We design and implementation of support tool for proofreading and structuring characters in corpus text. The tool has 3 main functions; adds XML tags of character to corpus text. , product the list of characters to check, and text proofreads and structures corpus text by the corpus specifications. The tool also has an easy-to-use GUI interface. In case of used the tool for constructing the "*Meiroku zassi*" corpus, we confirm the efficiency of the tool. The tool made proofreading time decrease to 1/10, and reduced human errors.

１．はじめに

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，これが実現した場合，近代の活字資料が言語研究目的で電子化されていくことになる．近代の活字資料を電子化し研究資料とする場合，電子化の仕様を正確に定め，その仕様に従って統一的な処理を実現する必要がある．文字に関する処理に限定しても，どの符号化文字集合を用いて電子化するか（例えば JIS か Unicode か，また，そのどのバージョン，どの領域までを使うかなど）を定め，使用する符号化文字はその範囲内に収めねばならない．さらに，特に近代の活字資料では原資料に出現した文字と符号化文字との間に字形差がみられるものも多いため，どの程度の差異までを包摂し，どの活字を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う処理方針も正確に規定したうえで電子化処理の統一を図らなければならない．言語研究用として利用できる電子テキストの室を保証するためには，一度入力した電子テキストに対して入力仕様に定められた符号化集合に収まっているか，また，文字包摂の処理が入力仕様に沿う形で統一されているか等を確認する校正作業は必須である．こういった作業は，原文と照らし合わせて入力文字を逐一確認していかざるを得ないという，きわめて煩雑な作業である．また，入力仕様との兼ね合いで，本来の活字とは別の符号化文字で置き換えて表現するなど，電子化時に原文資料に対して改変を行う場合もあるが，その際はどのような改変を行ったかといったメモ情報などを本文データに影響を与えないように付与しておきたいという要請もある．このような文字校正作業と，処理情報の付与は非常に煩雑であり，人の手のみでこれらの作業を行った場合，時間がかかるうえに，精度の面でも校正漏れ等，作業ミスが発生する可能性が高い．現在公開されているさまざまな言語研究用テキストに関しては，その構築過程においてどのような校正処理作業を行ったのか明らかでないものが多いが，国立国語研究所のコーパスを例にとると，『太陽コーパス』（2005）ではすべて手作業で校正が行われ，『現代書き言葉均衡コーパス』（2011）では，各処理工程ごとに用途を限定された専用ツールが使用されていた[3]．近年の大規模電子テキスト構築では，それぞれの現場で何らかに作業支援ツールが作成・利用されていると考えられるが，公開されているものは少なく，またそれらは，用途・操作方法の面でも，

きわめて限定的なツールと言わざるを得ないものである．そこで，本研究ではコーパス用テキストの文字校正作業の作業ミスを減らし，高効率化するための作業支援ツールをコンピュータ操作にさほど長けていないユーザにも使用でき，なおかつ現時点でのツール適用対象となっているテキスト以外にも転用しやすい形で設計，実装した．本稿では，明治時代の学術誌『明六雑誌』でのツール適用例を紹介する．

1.1 言語研究用電子テキストでの文字処理

紙媒体の活字資料を電子テキストに写し取ってコーパスを構築する場合，誤入力等を見つけ出す，通常の意味での「校正」ははもちろん必要であるが，電子テキスト化に当たってはさらに前述のとおり，「文字の表現の仕方そのものの仕様統一を図る」という別種の校正が必要になってくる．近代の活字資料，『明六雑誌』（1874～1875）を例にとると，近代の活字では図１のような字形差が見られる．

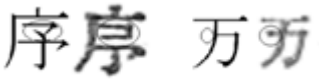

図１　『明六雑誌』に出現する「序」「万」の字形（右側）

これらの字形差に対して，入力作業者によっては「序」「万」を入力するか，外字として「=」を入力するかの揺れが生じうる．また，第一次入力段階では差異の存在そのものが見落とされる可能性もある．入力対象とする原資料のどの文字に，現在の符号化文字集合で表現される通用字形との差異があるかをあらかじめ知ることはできないため，一次入力時点で問題となる文字を経験的に洗い出したうえで処理方針を確定し，次の校正段階で確認，統一化を図るということになる．

1.2 言語研究用という目的に応じた文字処理方針と，そのための作業

「近代語コーパス」における文字処理方針の概要を紹介する．

1.2.1 拡張包摂

図１のような差異に関して，「序」「万」を入力すべきか，「=」とすべきかは，その電子テキストの使用目的による．言語研究用のテキストとしては，読めること，語が語として取り出せることが望ましいため，「=」はなるべく少なく，「序」「万」のように文字として読める状態のテキストの方が，有用性が高い．字体差異の処理として，例えば JIS 漢字では「漢字の字体の包摂規準」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ができる（図２）．

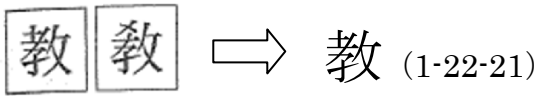

図２　JIS 包摂規準の例

図１のような差異を包摂してよいことを示す包摂規準は設定されてはいないが，既存の包摂規準と照らして，包摂規準の拡大解釈ないし拡張で，同一字とみなしてよいものと判断し，「序」「万」を入力する．

1.2.2 別字代用

図３は『明六雑誌』に出現する「すう」と読む字である．これは Unicode では表現可能だが，「近代語コーパス」での文字集合は JIS X0213 が採用されており，JIS 内字としては電子的に表現できない．

図３　『明六雑誌』出現漢字（「すう」）

「すう」に当たる字としては「吸」が一般的で，研究資料として「読めること」を優先するならば「吸」で表現したいところであるが，「吸」は図３とは字形差が大きすぎ，同一字とみなせるような差異ではない．よって，理念上「包摂」という処理は当たらない．このような場合も，研究資料としての有用性を考え，電子テキスト上は「=」にはせず，同訓の通用字（ここでは「吸」）で表現することとし，こちらを「別字代用」とする．結果的には読める文字が入力されるという点では「拡張包摂」と変わらないが，理念として同一の文字とみなしたか（拡張包摂），別字であるとしたうえで，テキスト上の表現として文字を置き換えたか（別字代用）という，認定面での区別である．

1.2.3 文字処理情報の付与

上記「拡張包摂」「別字代用」といった処理は，コーパス構築作業上，使用目的から要請された臨時的な処理であり，文字処理一般に通用している処理方針とは言えない．そのため，そのような処理をした文字に関しては，ただ文字を入力するだけでなく，タグの形で処理内容の情報を残しておくことが望ましい．

【例】

図１→＜包摂＞序＜/包摂＞

図３→＜外字　代用＝”1” Unicode =”564F”＞吸＜/代用＞

＝　→＜外字　代用＝”0”＞＝＜/外字＞

1.3　校正支援ツールの必要性

コーパス化にあたってどの程度の差異までを同一字とみなす（包摂する）か，また，どの字をどの字で代用するかといった指針は，原資料全体を見渡してからでないと確定させることができない．そのため，一次入力作業中に，処理上問題となる文字を洗い出し，次の工程である校正時に統一を図るという順序になる．

ここでの文字処理作業で，特に難しいのは以下の２点である．

（１）該当文字だけでなく，多岐にわたる情報をタグの形で付与したい．

「拡張包摂」「別字代用」の処理を経て入力された文字に関しては，逐一その旨をタグとして記入しなければならない．また，研究資料としての有用性を考え，原資料はどのような字であったかの情報も残すためには，Unicode で表現可能な文字に関しては Unicode 番号を記入しておくなど，タグ内に様々な形で注記をせねばならない．

（２）一括変換はできず，原資料との目視確認が必要である．

原資料に使用されている活字が全て均質であるとは限らない．そのため注意すべき字を確定し，その字に対する処理が決まったとしても，一括変換はできない．

図４　『明六雑誌』における「敵」活字字形

『明六雑誌』には，図４（A）のように，通用字「敵」とは異なり，右旁が「欠」となっている活字が出現する（U+6B52 で表現可能．ただし「近代語コーパス」では使用しない文字コード）．しかし，全ての「敵」が（A）の活字で表現されているわけではなく，より通用字に近い（B）の活字も出現する．このため，一次入力されたテキストの「敵」のすべてを，例えば「＜代用　Unicode ＝”6B52”＞敵＜/代用＞」などと一括で変換するわけにはいかず，言語資料としての質を鑑みるならば，一文字ずつ原資料と照合を行いながら確認していかなければならない．

（１）（２）の作業は手作業では多大な時間を要する上に，ミスも生じやすい．そこで，要確認文字の原資料照合・目視確認を援助し，「拡張包摂」「別字代用」などの処理方針に従い電子テキスト上の文字置き換え，およびタグ情報付与を行いやすくする支援ツールを設計・実装することで，作業の効率化を図った．

近代の活字資料の電子化作業工程と，本ツールの位置づけを図５に示す．

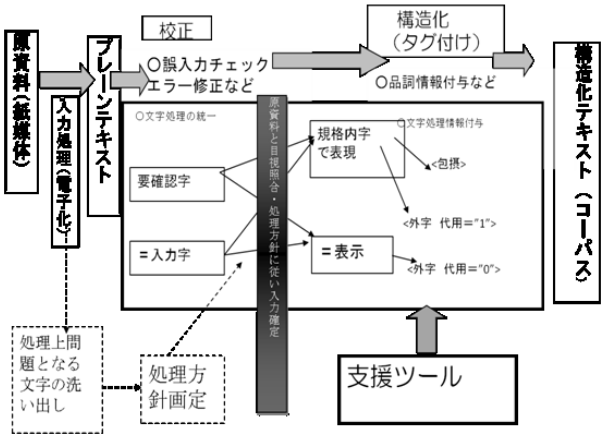

図５　作業工程とツールの位置づけ

２．システム概要

本ツールは，UTF-8 で記述された XML ファイルを読み込み，対象文字を抽出し効率的な校正作業を支援するツールである．本ツールの目的は，文字校正作業の効率化である．ツールによる自動化・均質化により校正作業における操作量の削減を実現するとともに，警告機能等により，誤入力の削減を行い，作業時間の短縮と作業ミスの減少を実現する．また，ツールを介した入力を行い，入力候補の提示等を行うことで，複数の作業者が校正作業を行った場合でも，作業内容が統一化されやすいように支援する．

2.1　機能概要

本ツールの持つ主な校正機能は以下の 3 つである．

（１）対象文字の抽出・作業画面にリスト化（目視照合支援）

コーパス作成用に電子化された XML 形式のテキストデータから，確認対象となる文字を抽出，リスト化する．また，対象文字の原資料活字との照合支援として原資料を撮影した PDF データにおける該当箇所を自動で表示する．

（２）対象文字の置き換え（校正入力支援）

抽出した対象文字に対して，必要に応じて置き換えを行う．

（３）タグ付け（構造化支援）

「拡張包摂」「別字代用」など，行った文字処理の内容に応じてタグ付けを行う．また，それ以外にも研究資料として利用するために残しておきたい情報を，XML タグの属性情報として記述する．

2.2　処理方式概要

本ツールが，実際に上記作業内容を支援する流れは，以下のとおりである．その設計および実際の作業画面は次節を参照されたい．

（１）テキストに確認用タグを付与

校正対象テキストから要確認字を抽出し，＜確認＞タグを付与する．
（２）要確認字をリスト化して処理作業用画面に表示
テキストから＜確認＞タグが付与された文字を拾い出し，リストとして表示する．
（３）作業者がリスト上で校正作業
作業者は表示されたリスト画面上で，文字の置き換え，作業メモの記入等を行う．リスト上で行われた文字置き換えの処理は，実際のテキストに反映される．この際，作業を効率化するための支援機能をさまざまに用意した．
（４）作業内容の情報を＜確認＞タグ内に記録
Unicode 番号や，別字代用処理を行ったか等の作業メモは，テキストの本文ではなく，＜確認＞タグ内に記録される．
（５）＜確認＞タグを入力仕様に合わせた書式に整形
＜確認＞タグ内に記録された文字処理情報を，そのコーパスでの入力仕様に合わせた書式に変換する．例えば，拡張包摂した文字には＜包摂＞タグ，別字代用で処理した文字には＜外字＞タグを付与する，という入力仕様であれば，代用包摂・別字代用で処理された文字の＜確認＞タグを，それぞれ入力仕様に合わせて＜包摂＞タグ，＜外字＞タグに書き換える．なお，確認対象字として＜確認＞タグを付与され，リスト上に拾いだされたものの，確認の結果問題がなく，何の処理の必要もなかった文字は，最終段階の＜確認＞タグ書式整形時に，＜確認＞タグそのものが消去される，という処理がなされる．

以上のように，リスト上に確認対象字を拾い出す目印として，校正対象テキストにいったん作業用＜確認＞タグを付与し，それを足場に校正作業を進めるが，その作業用のタグを使い捨てにせず，最終的には入力仕様の書式に合うよう変換することで，文字校正と同時に構造化の支援までを行う，というのが本ツールの特長である．

３．設計・実装

3.1　対象文字の抽出方式

１章でも述べたように，活字資料を電子テキストとして入力する場合，一次入力時に字形の差異がある可能性があり，要確認となる文字が洗い出される．本ツールでは，校正作業の実行前の一次入力時に洗い出された要確認文字一覧データを用いて，校正作業時に確認対象となる文字の抽出を行う．抽出は，一次入力データに対して一文字ずつ文字コードを確認し，該当文字に＜確認＞タグを付与する．＜確認＞タグ付与の基準となる要確認文字一覧データは txt データとし，校正作業者が作業方針・対象テキストに応じて自由に書き換え可能とした．

3.2　GUI 設計

今回，文字校正作業者は Windows ユーザ，かつそれほどコンピュータの扱いには長けていないユーザを想定し，マウスクリック等簡単な操作で校正作業が行えるように GUI を設計した．本ツールの画面例を図６に示す．

本ツールの GUI は主に，図６に示すように

① 設定やファイル読み書きを行うメニューバー
② 校正対象の文字をリスト表示する画面
③ 読み込んだファイルをテキスト形式で表示する画面

の３つから構成される．

図６ツール画面例

3.3 目視確認機能

本ツールで支援する校正作業では，電子化されたテキストと，原文テキストの両方を照合しながら校正作業を行う必要がある．そこで以下の２つの目視確認機能を実装した．

（1）電子テキスト目視確認

図６中の②の，リストの「テキスト位置」「現在のタグ」セルのいずれかを左クリックし選択した場合，③のテキスト表示画面において選択したタグを赤文字で示し，タグの存在する行へ自動スクロールする機能を実装した．

（2）原文テキスト目視確認

原文テキストは PDF 形式のデータとして保存されているものとする．原文テキストの目視確認支援のため，「テキスト位置」「現在のタグ」セルをダブルクリックすることで，原文テキスト PDF の対象文字があるページを自動で表示する機能を実装した．対象の文字があるページの判断には，電子化の１次入力作業時にページ番号を入力しておき，そのデータを利用する．

3.4 リスト設計

本ツールは図６において②で示した，リストに対して操作を行い，校正作業を行う．リストの設計について主要なものを表１に示す．

表１：リスト設計

行名	内容
テキスト位置	校正対象テキストが，読み込んだ XML ファイルの何行目にあるかを表示する
現在のタグ	校正対象の XML タグを表示する
対象文字	校正対象の文字を表示する
Unicode	Unicode 番号をメモとして入力することができる．ここに入力された値は XML の属性値に記述される
代用	校正を行った文字が「別字代用」の方針に従って処理されたものであるか否かを表示する．値は「○」「×」の２種類である．ここに入力された値は XML の属性値に記述される．「包摂」行の値とは排他である．
文字種類	校正を行った文字の種類を表示する．文字種類は「記号」「合字」「漢字」「カナ」「絵文字」の５種類とした．ここに入力された値は XML の属性値に記述される
メモ	残しておきたいメモを表示する．ここに入力された値は XML の属性値に記述される
包摂	校正を行った文字が，「拡張包摂規準」をもとに包摂されたか否かを表示する．値は「○」「×」の２種類である．ここに入力された値は XML の属性値に記述される．「代用」行の値とは排他である．
更新日時	その行が最後に操作された時刻を表示する．形式は「西暦/月/日（曜日）時：分：秒」である．
自動	自動置き換え機能の対象であるか否かを表示する．自動置き換え機能については３．６参照．

3.5 校正機能

3.5.1 クリックを利用した入力手法

3.2 でも示したように，本ツールではリスト画面を操作して校正作業を行う．校正作業においては，直接入力する必要がある内容は，対象文字の置き換えと，Unicode の入力，メモの入力の３つのみである．代用，包摂の有無については○（行ったか）×（行なっていないか）以外は入力する必要がなく，また文字種類についても同様に，幾つかのそれほど多くない文字種類を入力するのみである．そのため，クリックすることで自動入力されるようにした．代用，包摂については○，×をクリック毎に入れ替え入力する．文字種類については，「記号」⇒「合字」⇒「漢字」⇒「カナ」⇒「絵文字」の５種類を１クリックごとに順番に入れ替わるようにした．これにより例えば，「<確認>文字</確認>」となっている部分を「<確認　代用＝”1” >文字</確認>」と書き換える作業を，１クリックで実現できる．また，代用，包摂両方に○がついている場合など，ありえない内容が入力されていた場合には保存時に警告を出すことで，入力ミスの削減を図った．

3.5.2 対象文字置き換え手法

校正対象文字の置き換えは，リストの「対象文字」セルの文字を置き換えることで行う．「対象文字」セルは直接入力により文字を書き換えることができる．この時，通常の IME による文字変換では入力しづらい文字を使用する場合も考えられる．それらに対しては， IME の文字パレットを使用して入力する方法が考えられるが，文字パレットから対象の文字を探し出し入力を行う作業は煩雑で，時間がかかる．そこで，Unicode 番号を入力してその Unicode 番号の文字を入力できる機能を実装した．「対象文字」セルをダブルクリックすると図７に示すような画面が表示され Unicode を指定して文字を入力することができる．たとえば，「U+3042」と入力すると対象文字列に「あ」が表示される．ここで入力した Unicode は「Unicode」セルにも自

動で入力され「現在のタグ」セルにも反映される．

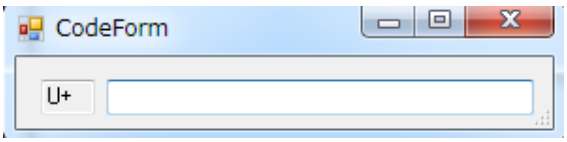

図７：Unicode 番号入力画面

「対象文字」セルに入力できる文字コードの範囲は，外部ファイルを用いて指定し制限することができる．外部ファイルでは，一行に一文字ずつ「人　\t　U+4EBA」のように「文字　\t Unicode」（\t はタブを表す）という形式で利用可能文字を指定する．このファイルに記述されていない文字が入力された場合は，図８のように入力不可能であることを表示し置き換えを行わない．これにより，校正時に指定文字セット外の文字が入力されるのを防ぐことができ，校正ミスを減らす効果が期待される．

以上のように，XML タグを直接書き換えるのではなく，整理されたリストのセルに対して操作を行い，XML 属性名や代用の有無等の毎回同様の内容を入力する作業を自動化することで校正作業の効率化とミスの削減を図った．

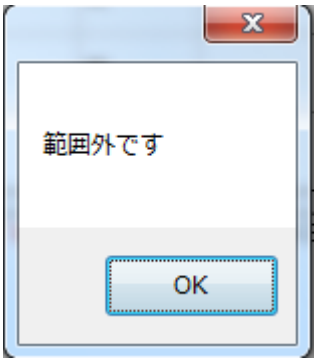

図８：校正対象文字の文字コード範囲警告

3.5.3　自動置き換え機能

一つのファイル内に，同一の校正対象となる文字が複数回出現する場合が考えられる．そのような場合，入力文字からメモ内容まで，一度入力した内容と全く同じ内容を，何度も入力することは煩雑である．そこで，さらなる作業効率化のため，対象文字の自動入力機能を実装した．自動入力を行う文字は外部ファイルに保存可能とし，「餐　\n　〈確認　代用="1"　memo="沃＋食"　〉」のように「対象文字　\n　置き換え後のタグ」（\n は改行を表す．改行コードが\r\n であった場合も同様に改行とみなす）という形式で記述する．自動入力は，外部ファイルのデータと対象文字とを比較し，同一の文字だった場合現在のタグセルと対象文字セルの値の置き換えを行う．ただし 1.3 でも示したように，同一の文字でも前後の文脈や目視確認の結果によって校正対応が変わる場合も考えられるため，一つの文字に複数の入力内容があることが考えられる．そこで，一括で全て置き換えを行うのではなく，図９に示すように置き換え候補のタグを確認する画面が，置き換え候補の数だけ繰り返し表示される．また，自動置き換えの対象となるか否かは図 10 に示すようにリスト表示の「自動」列に「無」「有」で表示される．

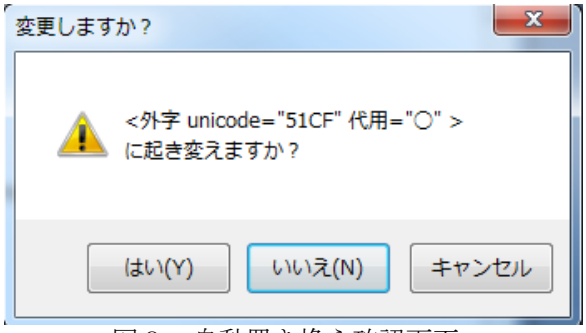

図９：自動置き換え確認画面

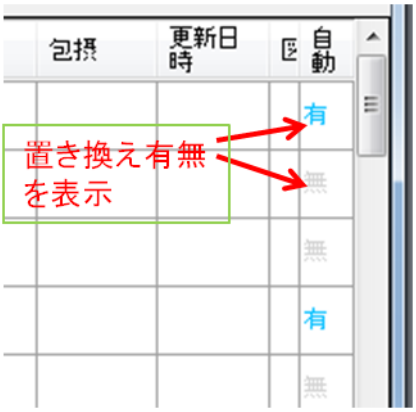

図 10：自動置き換え有無の表示

3.6　保存形式

本ツールでは，３種類の保存形式を実装した．

（１）　中途保存

作業を一時中断したい場合のため，編集中の XML ファイルを書き換えることなく作業内容のみを保存する．それにより，XML ファイルはそのままに，作業中断前とまったく同じ状態で作業を再開することができる．作業内容保存データは txt 形式で保存される．

（２）　確定保存

校正作業が完了し，保存する場合のための保存方式である．本ツールを用いて校正した内容を XML ファイルに適用し保存する．この時，コーパスの入力仕様に合うように，＜確認＞

タグを変換する．例えば，包摂欄にチェックが入っていれば＜包摂＞タグに，代用欄にチェックが入っていれば＜外字＞タグにするといった整形を行ったデータを保存する．

（3） タグ消去保存

本ツールで利用している校正用タグを全て消去したい場合のための保存方式である．編集している XML ファイルから，本ツールの校正用タグである＜確認＞タグ等を全て消去し，保存する．

3.7 ネットワーク通信機能

3.7.1 自動更新

本ツールは複数人で校正作業を行うことを想定している．校正内容を統一化するため，全作業者が同じ環境で作業することが望ましい．そこで，作業者全員にツールの更新や，校正仕様の変更を周知する機能が必要となる．そのため，本ツールではツール起動時に，ネットワーク経由でツールの自動更新を行う機能を実装した．データの送受信には FTP を用いた．ツール更新時には，図 10 に示すような画面で，更新内容，仕様変更等の連絡事項を表示することで，作業者全員での情報共有を可能とした．

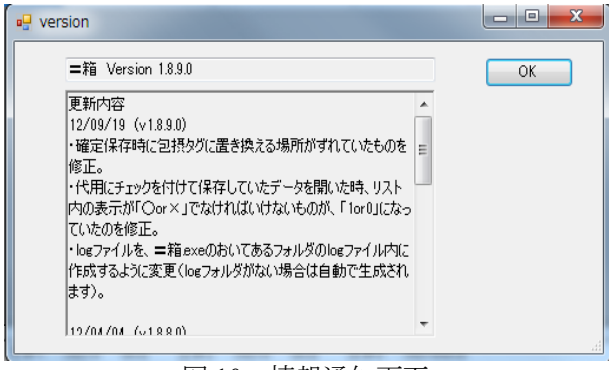

図 10：情報通知画面

3.7.2 作業情報の記録

本ツールは，複数人数で校正作業を行うことを想定している．そこで，情報を共有するため，誰がどういった校正を行ったかを記録しておく必要がある．記録した情報は，一箇所に収集し，効率的に管理を行いたい．そこで，本ツールでは記録した校正内容をネットワーク経由で自動的にサーバに送信する．リスト表示する画面へ行った作業内容を保存する．保存データは「リストの行，列，変更前のセルの値，変更後のセルの値，時刻」といった「，」区切りの CSV 形式で記述される．保存された CSV ファイルは，ツールごとに割り振られた ID，作業を行なっている XML ファイル名，及び時刻で管理を行う．データの送信には FTP を用いた．また，ネットワークに接続できない環境下での作業も想定し，送信データと同様のデータをローカルフォルダに保存する機能も実装した．また，記録した作業情報を利用して，一つ前の作業を取り消す機能を実装した．

4．評価実験

本ツールの有効性を検証するため，本ツールを用いた校正作業を行った．今回校正を行う資料には，『明六雑誌』を用いた．同じ資料に対してツールを用いた場合と，用いなかった場合の 2 通りの実験を行い，作業時間と発生した作業ミスについて比較を行った．作業条件は以下のとおりである．

① 『明六雑誌』6，12，18，24 巻の一次入力データを使用
② 要確認文字のリストと，各文字の処理方針一覧を配布し，それをもとに校正
③ XML タグ入力仕様を配布
④ 原資料のコピーを配布（紙媒体と電子媒体）

表 2 ：校正対象資料の内訳

巻数	確認対象文字	総文字数
6 巻	29 文字	7152 文字
12 巻	17 文字	4742 文字
18 巻	10 文字	3247 文字
24 巻	19 文字	4770 文字

（1） 作業効率化効果

校正作業にかかった時間を表 3 に示す．実験結果から，ツール使用時と，非使用時では，7～10 倍程度の時間差が生じることがわかった．校正対象ファイル内容により多少の差はでるものの，明らかにツール使用時には作業時間が短縮されているため，作業が効率化されたと言える．

作業時間が短縮された理由としては，校正作業で必要な入力をクリック一つで行えるなど，簡略化したことがあげられる．また，本ツールを使用しなかった場合には，電子テキストから校正対象文字を目視で探しださなければならない．しかし，本ツールを用いた場合，校正対象文字はリストとして表示され，該当箇所周辺の目視確認の支援もあるため，短時間で効率的に校正作業を行えたと考えられる．

表 3 ：校正作業時間

巻	ツール無し	ツール使用
6 巻	148 分	15 分
12 巻	103 分	11 分
18 巻	33 分	05 分
24 巻	50 分	06 分

（2）作業ミスの削減効果

本ツールを用いなかった場合に発生した作業ミスを表４に示す.

表４：作業ミス内訳

ミス内容	該当巻
漢字入力ミス	6巻
包摂入力の見落とし	6巻
「包摂」と「外字」の入力間違い	12巻
「外字」タグを不要な箇所に付与	18巻
「外字」タグを不要な箇所に付与	18巻
範囲外文字の見落とし	24巻

今回，校正を行う電子ファイルには，コーパスとして用いるため，本文以外にも様々な情報が記述されている．そのため通常の校正以上に人間にとっては読みづらいファイルになっており，校正を行う場合，雑誌本文を確認し，確認対象文字を探し出し校正を行う作業には，見落としミスが発生しやすい．ツール非使用時には２件の見落としミス，２件の不要な箇所への外字タグ付与が発生したが，ツール使用時には０件であり，作業ミスの削減効果があったと言える．これは，予め作業対象文字をリストとして表示し，該当箇所の表示等をツールにより支援したことにより，校正作業の必要な箇所がはっきりしたためであると考えられる.

12 巻で発生した包摂タグと外字タグの入力間違いについては，本ツールを使用しても起こりうる作業ミスであると考える．しかし，ツール内のリストに，包摂であるかどうかをチェック欄として表示する等，作業内容と操作が必要な箇所を整理して表示しているため，ツール非使用時に比べて作業ミスは起こりにくく，また作業後の見直しもしやすい．実際にツール使用時にこのようなミスがなかったことからも，こういった誤入力にの削減効果があると考える.

また，ツールを使用しなかった場合に範囲外文字の見落としミスが発生した．『明六雑誌』24 巻の電子ファイルには，Unicode の U+7358 に該当する「奭」という，文字コード使用範囲外の文字が含まれていた．ツール無しで校正作業を行う場合文字コードに関するチェックを行うことは非常に難しい．使用文字コード範囲外の文字を発見するには，表示されている文字の文字コードを調べ，その文字が範囲外であることを判断しなければならない．大量の文字の中から，使用文字コード範囲外の文字の候補をもれなく，文字を見ただけで判断することは多くの知識と注意力を必要とし，ほぼ不可能であると考えられる．このような場合でも本ツールでは対象文字を発見し，校正を行うことができるため，有用なツールであると考えられる.

５．まとめ

今回，コーパス作成時の文字校正処理を支援するため，校正作業を高効率化する支援ツールを設計，実装した．本ツールでは，文字校正と構造化を同時に行うことができる．コンピュータの扱いにあまり長けていない人でも効率的な校正作業を行えるように，簡単な操作で行える GUI を実装し，対象文字の抽出・確認，対象文字の置き換え，構造化といった作業を支援した．本ツールを用いた作業では，入力方法をクリック一つで行えるなど簡略化し，自動的に入力できるようにすることで作業時間の短縮と作業ミスの削減を実現した．さらに，複数の作業者が作業する場合の支援として，ネットワークを介したツールの更新，効率的な情報共有と収集を行った．実際に本ツールを用いた校正作業を行い，ツールを使用しなかった場合に比べて，作業時間を短縮し，作業ミスを削減できることを確認し，本ツールの有効性を示した.

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-『明六雑誌』を事例として-,じんもんこん 2011 論文集, pp.381-388, 2011.

[2] 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－『明六雑誌』活字字形への，包摂規準適用実験－, 第 95 回人文科学とコンピュータ研究発表会, 2012.

[3] 田島孝治，高田智和:JIS　X　0213 文字セット運用のための文字処理支援ツール 特定領域研究「日本語コーパス」, 平成 21 年度公開ワークショップ予稿集 ,pp.77-84, 2009.

【付記】

本研究は，平成 22 年度～平成 25 年度日本学術振興会科学研究費補助金基盤研究（Ｂ）「漢字字体変容の原理―敦煌文献から現代日本戸籍漢字まで―」（研究代表者：高田　智和，課題番号：22320087），および国立国語研究所共同研究プロジェクト「近代語コーパス設計のための文献言語研究」（プロジェクトリーダー：田中牧郎）による成果の一部です.