コーパス用テキストの文字校正支援ツールの設計と実装

堤　智昭（東京農工大学）†
須永　哲矢（国立国語研究所）††

Design and Implementation of the Support Tool for the Proofreading of Corpus Text

Tomoaki Tsutsumi(Tokyo University of Agriculture and Technology)
Tetsuya Sunaga (National Institute for Japanese Language and Linguistics)

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コンピュータにテキストとして入力する場合、近代の活字には入力仕様で規定された符号化文字集合に存在しない文字や符号化文字集合での文字と字形差がある文字が多く存在する。そのため言語研究用として保証できる質の電子テキストを得るためには、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等を確認する校正作業は必須である。具体的には、原文と照らし合わせて入力文字を確認し「=」として入力するか、存在する文字に包摂して入力するという対応が考えられる。また、電子化する時に、原文資料に対して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する必要がある。このようなテキスト入力とその文字校正作業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、時間がかかり校正漏れ等作業ミスが発生する可能性が高い。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設計、作成した。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、「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。

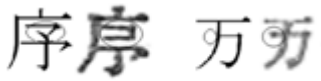

図１　『明六雑誌』に出現する「序」「万」の字形（右側）

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を見落としている可能性もある。入力対象とする原資料のどの文字に、通用字形との差異が

† t_tsu77@ninjal.ac.jp
††tsunaga@ninjal.ac.jp

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として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、読める文字として「序」「万」として表示されるテキストの方が、有用性が高い。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）。

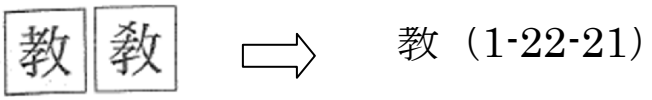

図２　JIS 包摂規準の例

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。これはコーパスで使用する文字集合になく、電子的に表現できない。「すう」に当たる字としては「吸」があるが、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ため、「包摂」という処理は当たらない。このような場合にも、研究資料としての有用性を考え、「**=**」にはせず、同訓の通用字（ここでは「吸」）で代用することとする。

図３　『明六雑誌』出現漢字（「すう」）

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。

校正において、「包摂」「代用」の処理を統一的にしつつ、見落としの可能性まで含め確認作業を行うというのは非常に煩雑であり、作業上のミスも生じやすい。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。

「包摂」「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、注意すべき字が確定し、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、右側が「欠」となっている活字が出現する（U+6B52 で表現可能。ただし「通時コーパス」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）の作業は極めて煩雑であり、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、「包摂」「代用」などの処理方針に従っての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。

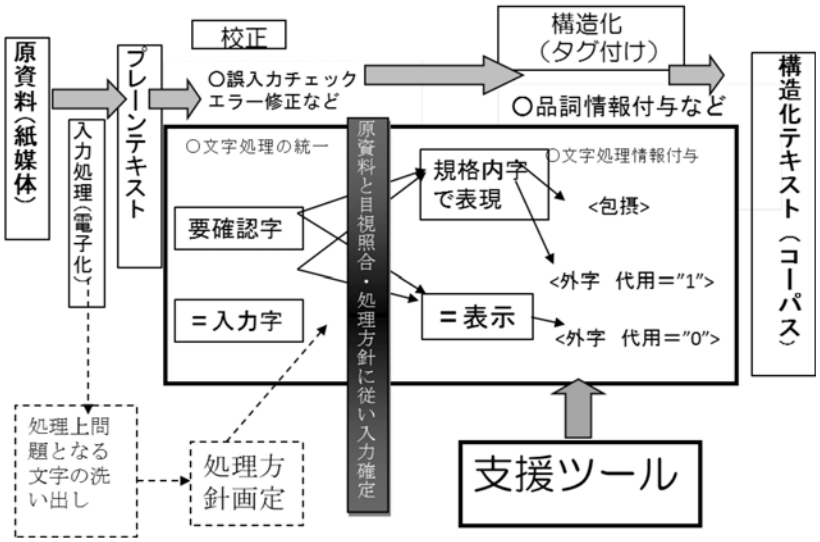

図５　作業工程とツールの位置づけ

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。今回は、コーパス用テキストの文字校正作業を行う作業者を対象ユーザとし、操作しやすい GUI をもつツールを設計した。本ツールのもつ主な校正用機能は以下の３つである。

（1）対象文字の抽出・確認

xml 形式で記述された文章データから、確認対象となる文字を検索しリスト形式で表示する。また、テキストデータの該当箇所、及び原文 PDF の該当箇所を自動で表示し、目視照合支援も行う。

（2）対象文字の置き換え

抽出した対象文字に対して置き換えを行う。書き換える文字の入力方法は、IME を用いた日本語入力に加え、unicode 番号から入力することも可能とした。同じ文字が複数個抽出され、同一の書き換えを何度も行う場合は、記憶した内容で自動的に置き換えることも可能とした。

（3）メモ機能

「記号」「合字」といった文字種類や unicode 番号等のメモを、xml タグの属性として記述する。

これら 3 つの校正用機能に加え、本ツールの評価実験、管理を効率化するためにネットワークを利用した情報収集、保守機能を設計、実装した。

3. 設計・実装

3.1 GUI 設計

本ツールのメイン画面例を図 6 に示す。メイン画面は次の 3 つから校正される。

① 設定やファイルの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（以下、リスト画面）
③ 読み込んだファイルをテキスト形式で表示する画面（以下、テキスト画面）

校正作業は、主に②のリストを操作して行う。リストは操作が行われると、行の色が白から茶色に変更される。これにより、編集済みか未編集かが判断しやすくなる。また、最終更新時刻も記録する。③のテキスト画面では、対象の XML タグを赤文字で強調し、それ以外を黒文字で表示している。②のリストと③のテキスト画面は連携しており、リストからある行を選択すると、③のテキスト画面でも選択した行のタグがある箇所に自動でスクロールする。

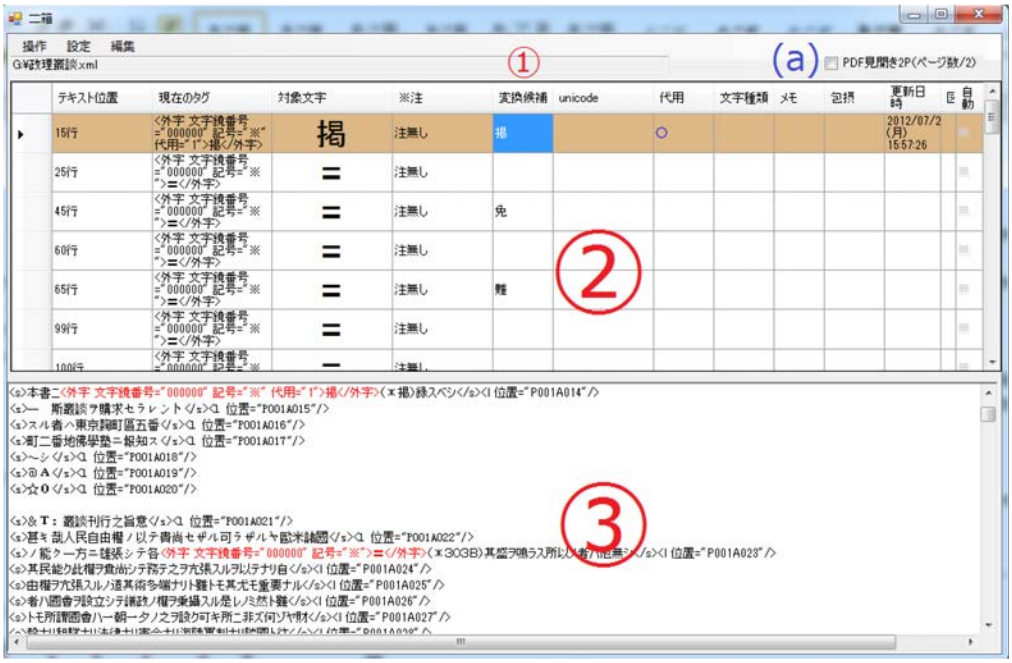

図 6 ：メイン画面

3.2 目視確認機能

本ツールでは、3.1 で示したように校正対象の XML データをテキスト画面に表示している。それに加えて、原文 PDF を表示することで目視確認を補助する。原文 PDF は、リスト画面の『現在のタグ』セルをダブルクリックして表示する。XML データには予め、行ごとに原文の何ページに存在するかという情報が記述してあるものとし、この機能はその情報を読み取り実行する。PDF を表示する時には、対象の文字があるページの番号を XML データから取得し、表示する。原文 PDF の作成方式によっては、PDF のページ１ページにつき原文の見開き２ページで作成されている場合も考えられる。その場合、原文のページと原文 PDF のページにずれが生じる問題が考えられる。そこで、図６の(a)で示したチェックボックスを作成し、チェック時には見開き２ページとしてページ計算を行うことで対応した。自動で原文 PDF を開くためには、PDF データのあるフォルダの場所を、設定→詳細設定から設定する必要がある。

3.3 対象文字のリスト化

本ツールでは、＜外字＞タグ、及び＜確認＞タグのついた文字を検索し昇順リスト形式で表示する。例としては「<外字 記号=″※″>=</外字>」のようなものがあげられる。この場合=が対象文字となる。リストの列は表１のように設計した。

表１：リスト設計

行名	内容
テキスト位置	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	対象の XML タグを表示する
対象文字	対象の文字を表示する
※注	タグの後に注がついている場合、その内容を表示する。
変換候補	タグの後に（＊文字）という記述があった場合、＊以下の文字を変換候補文字として表示する
Unicode	Unicode を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代用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、漢字包摂基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 XL 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。
図	対象文字のフォントを表示する。表示には、フォントデータを個別に用意する必要がある。
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後述する。

3.4 校正機能
3.4.1 対象文字の置き換え
（1）直接入力
リストの中から、対象文字セルを指定してキーボードから直接文字を入力することができる。
（2）unicode 入力
『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」セルに反映される。入力できるコードの範囲は、外部ファイルで指定可能である。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コード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、『対象文字』セル、『unicode』セル、『現在のタグ』セルへの入力は行われない。

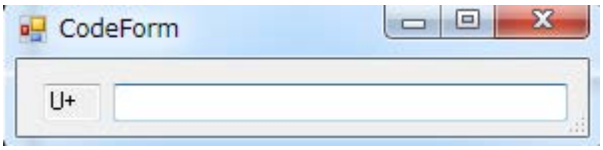

図 7 ：unicode 入力画面

図 8 ：unicode 入力範囲警告

（3）変換候補入力
　変換候補文字がある場合、『変換候補』セルをダブルクリックすると変換候補文字が『対象文字』セルに自動入力される。

3.4.2　付属情報入力
（1）unicode
『unicode』セルに値を入力すると、現在のタグの属性値に「unicode=”unicode セルの値”」の形式で入力される。
（2）代用
代用セルを左クリックすると、「○」「×」が自動入力される。1 クリックごとに○×は入れ替わる。代用セルの値は、現在のタグの属性値に「代用=”0or1”」の形式で入力される。○ならば 1、×ならば 0 が入力される。
（3）文字種類
『文字種類』セルも『代用』セルと同様に左クリックすると値が自動入力される。1 クリックごとに値は入れ替わる。値は「記号」「合字」「漢字」「カナ」「絵文字」の 5 種類

である。文字種類セルの値は、現在のタグの属性値に「moji=”値”」の形式で入力される。
（４）メモ
『メモ』セルを指定して文字をキーボードから直接入力することができる。このセルには任意の値を入力することができる。『メモ』セルの値は、現在のタグの属性値に「memo=”値”」の形式で入力される。
（５）包摂
『包摂』セルも、『代用』セルと同様に左クリックすると「○」「×」が自動入力され、１クリックごとに○×が入れ替わる。

3.4.3 校正補助機能
（１）戻る、進む機能
校正作業中に変更した内容を元に戻したい時のために、校正によるデータの変更を全て記録することで操作を一つ戻る、または戻した操作を一つ進める機能を実装した。データの変更は、変更されたセル、変更前の値、変更後の値、変更した時刻の４つのデータを１セットとし、リスト形式で保存する。
（２）自動入力機能
一つのファイル内に、同一の校正対象となる文字が複数存在する場合が考えられる。そのような場合、一度入力した内容を何度も入力することになる。そこで、作業効率化のため対象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。¥r¥n にも対応する）という形式で記述する。リストの校正対象全てに対して自動入力を行う方法と、選択した一つの行に対してのみ自動入力を行う方法の２種類を実装した。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、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、置き換え実行前に置き換えるかどうかの確認をする画面を表示する。また、自動置き換えの対象となるか否かはリスト表示の「自動」列に「無」「有」で表示される。

3.4.4 データ保存
本ツールでは、３種類の保存方式を実装した。
（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、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。
（３）タグ消去保存
全ての作業が終わり、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＜外字＞タグ、及び＜確認＞タグを全て消去し、保存する。

3.5 ネットワーク機能
本ツールは情報収集及びメンテナンスを行うために、ネットワークを介してサーバと通信を行う。サーバには FTP サーバを利用した。
（１）作業情報収集機能
本ツールの有効性を示すため、行った校正作業内容を記録する必要がある。校正作業は、複数人で行う為記録した校正内容の収集を効率的に行いたい。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を保存し、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。

（２）自動更新機能

（１）でも示した通り、本ツールを用いた校正作業は複数人で作業を行う。ツールの更新時にはタグの仕様変更など、全作業者で共有するべき更新が行われる場合も考えられるため、全員で同じツールを使用することが必要となる。しかし、作業者全員が常にツールの更新を確認することは非常に手間がかかる。そこで、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、更新内容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。

4. まとめ

今回、コーパス用テキストの校正作業を効率化するための作業支援ツールを設計、製作した。本ツールでは、対象文字の抽出・確認、対象文字の置き換え、メモといった作業を効率化するための機能を、簡単な操作で行える GUI と共に実装した。さらに、校正データの保存方式を３種類用意し、作業の状態に応じて柔軟に対応できるようにした。また、ネットワークを介したツールの更新や、有効性を示すための作業情報収集の効率化を図った。今後は、現在行なっている校正作業の作業情報を収集、解析し本ツールの有効性の確認を行う予定である。

文　献

田島孝治、高田智和（2010）「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」、pp.77-84 平成 21 年度公開ワークショップ予稿集

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−」、pp.381-388、じんもんこん 2011 論文集

須永哲矢、堤智昭、高田智和(2012)「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