様式 C-19

科学研究費補助金研究成果報告書

平成２１年　４月３０日現　在

研究種目：基盤研究(B)
研究期間：2006～2008
課題番号：18300082
研究課題名（和文）　中国近世白話文学の電子化状況情報及びコーパスの共有基盤の構築に関わる基礎的研究
研究課題名（英文）　How to share and change information on developing classical Chinese corpora?: a fundamental study

研究代表者
笠井　直美　(KASAI Naomi)
名古屋大学・大学院国際開発研究科・准教授
研究者番号：90251389

研究成果の概要：
本研究は、学術的利用に適した中国近世白話文学コーパスを構築する為の基礎的研究として、文学及び隣接分野の研究上、必要と思われる電子テキストの形式、校訂、異体字表、複数の版本やテキストに付された批評の扱い、情報付与の方法等について基礎的な検討を行い、それに沿ってサンプル的に、明代の木版本・鈔本に基づく電子テキストの作成・校訂を行うと共に、電子テキストを利用した新たな文学研究のアプローチを試みた。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	3,000,000	900,000	3,900,000
２００７年度	2,200,000	660,000	2,860,000
２００８年度	1,300,000	390,000	1,690,000
総　計	6,500,000	1,950,000	8,450,000

研究分野：中国古典文學
科研費の分科・細目：情報学・図書館情報学・人文社会情報学
キーワード：コーパス、戯曲、白話小説

１．研究開始当初の背景

本研究は、先行する 2005 年度基盤研究(C)(2)(企画調査)「中国近世白話文学の電子化の現況及び学術利用に有効なコーパスの設計に関する調査」(研究代表者・笠井直美)を承けて開始された。

上記企画調査の結果、中国古典文学においても電子化の進展はめざましく、主要な資料は、底本や校訂を問わなければ概ね電子テキスト化されているが、無料または安価なものは学術的な使用に耐えうる水準でないことが多く、専門家の関与した水準の高いデータベースは、可塑性のあるテキストファイルの形ではなく、高度な操作が施せない形態で供給されている等、問題も多いこと、この分野で最も不足しているのは、実はごく基礎的なもの、「広く研究者に開放されている、信頼できるプレインテキスト」であることが明らかになった。

信頼のおけるプレインテキストが十分でない状況下では、初歩的な検索に基づいた分析さえも正確は期しがたく、形態素解析を初めとする標識付けのためのプログラムや、計量分析のためのプログラムも、効果を発揮することができない。

学術的利用に適した電子テキストを、研究者コミュニティ全体で共有しあえる仕組みや、テキストの電子化の状況について情報を

継続的に共有する仕組みが必要と考えられる状況が、本研究の背景として存在した。

２．研究の目的

(1)中国古典文学・言語学・文献学・歴史学等の研究に有用で、研究者コミュニティ全体で共有可能な、学術的利用に適した中国近世白話文学コーパスを構築する為の基礎的研究を行う。

(2)中国近世白話文学（白話小説、戯曲、説唱等）の電子化状況に関し、情報を広く研究者間で共有でき、容易に更新を行って継続しうるような方法を追求する。

３．研究の方法

(1)学術利用に適した中国近世白話文学コーパスの設計に関する基礎的研究

①電子テキストの校訂、設計等に関わる問題の検討

信頼できる電子テキストをある程度大量に集積するために、研究者が相互に電子テキストを提供し合う、バザール方式の採用が考えられる。この場合、電子テキストの校訂ルールや形式が、多くの研究者に共有されうるものである必要がある。本研究では、こうした共有ルールとなりうるものを念頭に、文学・版本学・文献学・中国語史・歴史学等の角度から必要な情報・形態は何か、校訂、異体字・誤字の扱い、異体字表、複数の版本や批評の扱い、情報付与の方法等について基本的な問題を洗い出し、検討を行った。

先行する研究を参照・検討すると同時に、関連の諸分野において、従来手作業で行われてきているが、電子テキスト・コーパスを利用することでより効率的な分析が可能になると考えられるアプローチや、電子テキスト・コーパスを前提とすることで可能となるアプローチを想定し、その場合に必要となる情報・形態を洗い出す方法を採った。

②サンプルとしての電子テキストの構築

サンプル的に少数の電子テキストを構築する（校訂は専門知識のある研究者の手作業なので、計画立案時には１～数篇を予定していた）ことを通じ、①に関する検討・修正、新たな問題点の洗い出しを試みた。

具体的には、

〈1〉学術的に重要な木版本・鈔本を OCR 外注して基礎となる「たたき台」テキストファイルを作成する。

〈2〉本研究のメンバー及び協力を申し出てくれた専門家によりこれを校訂する

〈3〉校訂ルールや電子テキストの形式に関する問題点、提案、要望等に関し、随時情報交換を行う。（ファイル共有及び情報交換には WebDAV システムと UTF-8 化した Wiki システムを利用）

〈4〉完成した校訂済みテキスト、及び OCR による「たたき台」テキストファイル作成までで止まっており、校訂する予定がないテキストをウェブサイトで公開する。

といった手順を採った。

(2)中国近世白話文学（白話小説、戯曲、説唱等）の電子化状況に関する情報の集積・共有

①2005 年度基盤研究(C)(2)(企画調査)「中国近世白話文学の電子化の現況及び学術利用に有効なコーパスの設計に関する調査」の成果を承け、引き続き中国近世白話文学の電子化状況に関する情報の修正・増補を行った。

②能率的に更新が行われるシステムの追求

研究開始年度においては、①の基礎データを SQL サーバーに組み込んでデータベースを構築し、一方で、アクセス制限付きの Wiki サイトで研究者間の情報交換を行って、修正を行うという方向で作業を進めたが、中国でも少しずつ「広く研究者に開放されている、信頼できるプレインテキスト」を作り、公開しようというプロジェクトが出てきて、①のデータが主たる対象としている、次善の策としての「無料・安価だが信頼性の低い電子テキスト」の必要性が相対的に低下してゆく状況が見込まれることとなった。このため、このデータに関しては小規模の増補・修正にとどめ、代わりに、(1)②の電子テキストの校訂・コーパスの構築に関し、研究者間で情報・成果の共有・交換・更新を容易に行うシステムを追求することとした。

４．研究成果

(1)学術利用のための中国近世白話文学電子テキストの校訂、設計等に関する諸問題

①電子テキスト構築の基本方針：目的に応じた複数バージョンの作成

中国近世白話文学に関しては、（英語等のコーパスでは既に常識と成りつつある種々の情報付与を行う以前に）信頼のおけるプレインテキストが必要であることは、「１．研究開始当初の背景」で述べたとおりであるが、このジャンルの作品群の大きな特徴として、「1 人の『作者』による『オリジナル』なテキスト」という概念があまり有効ではないものが多いこと（作品は多くの人々のかなり大胆な改変を経、多種多様な形で享受されていたのであり、現存する多種のバージョンや改作は享受の状況を考察する上ではそれぞれに価値がある）に留意し、底本とする木版本や鈔本に（異体字等も含め）できるだけ忠実なバージョン（「原貌版」）を作成するのが適切という結論に達した。

理屈としては、一つの作品について、現存するできるだけ多くのバージョンの「原貌版」電子テキストが作成されるのが理想的と

言えるが、これは予算の関係で無理であり、また、作業量を考えれば有意義とも言えないので、版本研究や受容に関わる研究の中で特に重視されている版本、メジャーな版本と分岐が大きい版本を中心にOCR底本を選択した。汎用性、継続性、様々なプログラムで利用する場合の便を考慮し、UTF-8 プレインテキストファイルを作成することとし、文字ブロックとしてはCJK 統合漢字、CJK 統合漢字拡張A､CJK統合漢字拡張Bまでの範囲を利用した。

この「原貌版」電子テキストは、フリーのテキストファイル比較プログラムを利用することにより版本研究を助けることが期待でき、また、通用字の採用状況を手がかりに、版本・作品の成立時期を推定するアプローチや、一つの作品の中での用字・語彙・語法的特徴の偏りから作品の成立過程を探るアプローチ等への応用にも役立つと期待できる。

「原貌版」は原則として句読点がついていない（原本に傍点や圏点が付されている場合にはそのままOCRされるが、現代的な意味での「句読点」以外に、声調を示すための圏点なども含まれていることがある）ため、検索の際には、検索文字列が複数の文にまたがって含まれている場合にもヒットしてしまう（＝本来検索したい語彙以外のものも大量に検出してしまうことがある）といった缺点もあり、句読点を付したバージョンも必要である。出版されている校訂本や、それに基づいた電子版は、古い白話では珍しくないがその後淘汰された文字遣いや語彙をも現代語に近い形に「標準化」しているものが多く、学術利用上は、こうした文字遣いや語彙は本来の状況を保存しつつ、句読点を付したバージョン（「原貌分句版」）が有用な場合もある。

一方、中国古典文学・哲学等の分野では、重要な一次資料への校注・訳注が研究上の基礎作業の一つであり、良質なそれは重要な研究業績と見なされてきた。やや分野の離れている研究者が、版本には特にこだわりを持たずにある作品の「標準的」テキストを利用したい場合を想定すると、上記の「分句版」よりさらに進んで、校訂者の「読み」に従い一定の標準化が行われた、「(繁体字)校訂版」も有用と考えられる。この「校訂版」は、(異体字表を組み込んだプログラムではなく）テキストエディタやワープロソフト附属の検索機能等を利用して単純な検索を行った場合にも、意図した検索結果を得やすいという利点もある。

こうした諸種の事情を勘案し、どのような標準化を施したどのような版を作るかについては、校訂者が研究の目的に応じて選択することとし、その代わり、校訂の方針をやや詳しく説明した文書を付すこととした。ただ、「原貌版」は原則としてなるべく作っていただくことにした。

以上のように、本研究では、種々の学術利用の目的になるべく合うよう、複数の底本はそれぞれにつき電子テキストを作成し、また、各々について、「(OCR したままの）たたき台版」、「原貌版」、「原貌分句版」、「校訂版」等の複数のバージョンを作成・公開することとした。

②校訂ルール

「原貌分句版」や「校訂版」の校訂ルールに関しては、(2)で述べる、実際の電子テキスト構築作業からのフィードバックを基に、大枠はできる限りシンプルにし、細部については、研究上の必要に応ずるよう校訂者各自に任せ、その代わりに校訂の細則に関する説明を同時に公開することとした。できあがった校訂版テキストという「結論」だけでなく、作成者の狙いや作成経過を明らかにし、透明化することにより、他の研究者が異なる目的で利用する場合に注意すべき点が推測しやすくなると考えられる。

また、改行を底本のまま保存する／しない、ページの区切りやページ数を入れる／入れない、誤字の訂正や衍字の示し方等、機械的な置換によって他の形式に変換可能なものに関しては、無理に統一せず、簡単な変換スクリプトを利用できるようにすることで対応することとした。

③異体字表

本研究の初年度にOCRを行った電子テキストの一部は、電子版『四部叢刊』『四庫全書』で使用されているフォント HT_CJK を使用しており、これには Private Use Area も利用している。この部分に関しては、このフォントが無い環境においては文字化けが起きるため、異体字表（フォント開発元からは公開されていない）の作成とそれに基づく置換作業が必要となる。この異体字表の作成を行った。この表は、電子版『四部叢刊』『四庫全書』の検索結果を利用する場合にも有用となる。

また、「原貌版」に対して効率的な検索を行うには､CJK統合漢字､CJK統合漢字拡張A、CJK 統合漢字拡張 B の範囲内での異体字・通用字表を用意し、それを組み込んだ検索システムを構築するのが良いと考えられ、このための異体字表作成の作業も進めたが、これの完成・異体字表を組み込んだ検索プログラムの開発は今後の課題となった。

④情報付与

関連する様々な分野（中国古典文学、語学、歴史学、思想史等）において付与するのが有

用と思われる情報としては、所謂メタデータが挙げられる。ただし、実際の付与に際しては、「『作者』による『オリジナル』なテキスト」という概念があまり有効ではないというこのジャンル特有の事情から、特に「作者」「時代」等の扱いには注意を要する。

中国近世白話小説や戯曲には、批評や注釈がテキスト本文に挿入されている（挟み込む場所・形式により、総評、眉批、夾批等がある）ものが相当数あり、作品の受容や文学批評史、思想史との関連等から重視されている。本研究でも、こうした批評を含めて電子テキスト化した版本があり、適切な形で情報付与を行えば、有用な資料となると考えられる。

また、語学の分野において特に有用と思われるPOSタグの付与は、適切な形態素解析ソフトが無かった（自由に利用できる形で公開されていなかった）ため難しいと考えられてきたが、本研究最終年度に現代中国語の比較的高性能な形態素解析ソフトがオープンソース化され、自由にダウンロードできる体制が整えられたため、今後近世白話向けにカスタマイズを行えば実現する可能性がでてきた。これは、上述の情報付与の細部にわたる検討、xml 形式でタグ付けを行う場合の DTD の決定等とともに、今後の課題となった。

(2)中国近世白話文学電子テキストの構築

(1)で検討した点に留意しつつ、サンプルとして、これまで電子化されていなかった・または信頼できる電子テキストが広く供給されてこなかった明代鈔本・木版本の戯曲・白話小説17種21バージョンを選び、OCRして「たたき台」テキストファイルを作成、2種2バージョンを除き公開した。

上記のうち、8種10バージョンにつき、専門家（うち2種3バージョンについては本研究のメンバー以外に、大東文化大学・中川諭氏、東北大学・井上浩一氏の協力を得た）による校訂作業を開始し、5種6バージョンにつき第一次校訂済みテキストを完成・公開した。残りについても、校訂が済み次第、本研究で構築したサーバに順次公開の予定であり、今後も可能な限り、新規の電子テキストを継続して追加してゆく予定である。

(3)中国の類似プロジェクトとの連携

本研究と同時期に、それまで『三國演義』などの簡体字排印本の電子テキスト化を進めてきた、中国・首都師範大学中国伝統文化デジタル化研究センター長・周文業教授が、白話小説の各種木版本の電子テキスト化に範囲を広げ、精力的にプロジェクトを進行させつつあった。周教授のプロジェクトでは、まず四大奇書の代表的版本を対象とし、OCR済みのテキストをCD-Rの形で研究者に頒布し、プロジェクト側では校訂にタッチしない点、中文Windowsベースの版本比較プログラムを開発・頒布する点が本研究とは異なり、相互に補完・協力が可能と考えられた。

2007年8月、周教授の日本訪問に合わせ、本科研プロジェクトで学術講演会を開き、「中国的古籍数字化及其応用研究情況（中国における古典籍のデジタル化とその応用研究）」と題する講演を行っていただくと共に、情報交換を行い、その後も引き続き情報交換・協力しつつ電子テキスト化を進めることが可能となった。

(4)文学研究への電子テキストの活用

中国文学研究では、電子テキストは主として「検索に便利」という点が評価されており、せいぜい版本研究の補助に役立つ程度という認識が多かった。

これに対し、研究代表者の笠井は、英語・日本語・フランス語などのコーパス構築状況と研究への応用を参考にしつつ、電子コーパスを活用して行いうる文学・語学研究上の各種のアプローチ法、コーパス構築や分析の際に有用なツール、中国古典籍のコーパス・データベースを構築・活用するに際し、障害となる構造的な問題点等について、整理・紹介を行った。また、研究分担者の上田望氏は電子テキストを活用した計量的アプローチで文体論に新たな視点を開くことが可能であることを示した。

(5)複数の研究者間で容易に情報共有・交換を行うしくみの追求

具体的には、ファイルの共有には WebDAV、情報交換にはUTF-8化したFSWikiを用いた。いずれもアクセス制限を行い、少人数の研究者間の情報共有・交換手段としてはそれなりに有用だったと評価できる。

しかし、電子コーパスはある程度大量なデータがあってこそ効果を発揮するので、将来的には、（研究成果を静的に公開するだけでなく）更新や情報交換を行う部分も広く公開してゆくことが望ましいが、スパム対策やセキュリティの関係から、技術的なハードルがかなり高くなる点は大きな問題である。

また、汎用性、継続性、セキュリティ、各種プログラム言語の利用の便、予算の節約等を考え合わせると、Unix系のOSにメジャーなオープンソースソフトウェアやツールを利用し、汎用性の高いデータ形式を用いることが適切で能率的と考えられる一方、中国近世白話文学テキストの校訂に携わりうる研究者の多くはWindows環境に慣れており、見慣れぬデータ形式やツールに抵抗を感じる人も少なくない。両者の異文化とも言うべき溝をどのように埋めていくかも、今後検討すべき課題の一つである。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１　件）
①▪廣瀬玲子「変奏されるドラマーー元雑劇『魔合羅』『勘頭巾』試論」『専修人文論集』第 83 号、2008 年 10 月、1-36 頁。査読なし

〔学会発表〕（計 2 件）
①笠井直美「純文本(plain text)和開源軟件(open source software)於古代文獻研究上的運用：借鑑於英、法、日語語料庫研究」第七屆中國古代小説文獻與數字化研討會（第 7 回中国古典小説文献とデジタル化検討シンポジウム）（マカオ大学、2008 年 8 月 27-28 日）

②上田望「淺析《三國志演義》的寫法：從對白分析明清小説的語體特點」第七屆中國古代小説文獻與數字化研討會（第 7 回中国古典小説文献とデジタル化検討シンポジウム）（マカオ大学、2008 年 8 月 27-28 日）

〔図書〕（計 1 件）
笠井直美編『中国近世白話文学の電子化状況情報及びコーパスの共有基盤の構築に関わる基礎的研究』(2006-2008 年度科学研究費補助金報告書；近刊)

〔その他〕
ホームページ等
http://dicom3.gsid.nagoya-u.ac.jp/bhwiki/

６．研究組織
(1)研究代表者
笠井　直美　（KASAI Naomi）
名古屋大学・大学院国際開発研究科・准教授
研究者番号：90251389

(2)研究分担者
上田　望　（UEDA Nozomu）
金沢大学・文学部・准教授
研究者番号：90293331

福田ムフタル（FUKUDA Muhtar）
名古屋産業大学・環境情報ビジネス学部・准教授
研究者番号：20283517

廣瀬　玲子（HIROSE Reiko）
専修大学・文学部・教授
研究者番号：90238410