# 科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　２６　日現在

機関番号：　１３９０１
研究種目：　挑戦的萌芽研究
研究期間：　2011～2013
課題番号：　２３６５２０７４

研究課題名（和文）中国近世白話文学コーパスへの文学理論研究に資する情報付与の基礎的研究

研究課題名（英文）Corpus Annotation of Pre-modern Vernacular Chinese Literature for Literary Theory Studies

研究代表者
笠井　直美（Kasai, Naomi）

名古屋大学・国際開発研究科・准教授

研究者番号：９０２５１３８９

交付決定額（研究期間全体）：（直接経費）　　2,700,000円、（間接経費）　　810,000円

研究成果の概要（和文）：本研究では、中国近世白話文学（戯曲・白話小説等）の電子コーパス構築における情報付与について検討を行った。
１）コーパス構築の基礎として必要な電子テキスト整備の一環として、明中期の雑劇作品24種を排印本に基づき電子テキスト化し、うち9種は明中期版本の調査に基づいた電子テキストも作成した。又、批評つき白話小説1種を電子テキスト化した。
２）文学及び隣接分野の研究の為に中国近世白話文学コーパスにどういう情報を付与するのがよいか検討を行い、１）の一部にサンプル的にメタデータ・批評に関する情報を付与したコーパス、及び現代語向け形態素解析ソフトによるPOSタグ付きコーパスを試作し、その有効性について検討した。

研究成果の概要（英文）：This study aims to explore useful annotations in the corpora of pre-modern vernacular Chinese literatures for literature studies and relevant disciplines.
1. As preliminary arrangements for building the corpora, I edited digital texts of 24 zaju dramas (which have woodblock print versions in the middle of the Ming period) from typographical print versions. In addition, 9 of them were also edited based on the research of woodblock print versions in those days. Besides drama texts, I edited a digital text of a vernacular fiction with interpretative comments of a critic.
2. I examined what kinds of annotations in digital texts of pre-modern vernacular Chinese literatures could be useful for literature studies and relevant disciplines. For trial I made two small corpora: one with annotated metadata as well as interpretative comments of a critic and theother with POS tag by using a morphological analyzer for Modern Chinese. Through these samples, I investigated their effectiveness.

研究分野：　人文学

科研費の分科・細目：　各国文学・文学論

キーワード：　白話小説　戯曲　コーパス

１．研究開始当初の背景

中国古典文学の電子テキストは、「高価で良質だが、テキストデータもブラックボックス化されていて、加工や計量的研究に利用できないデータベース」と「無料・安価だが学術研究には不向きな一般向けテキスト」が多く、研究者に開放された信頼できるプレインテキストという、最も基本的なものが不足していた。こうした状況に鑑み、申請者らは、科研費基盤研究「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」において、学術利用向き電子テキストの作成及・共有システムの構築を進めた。また、中国都市芸能研究会、中国・首都師範大学の周文業教授のプロジェクト等でも白話文学電子テキストの作成・公開を進めており、状況は改善されつつある。

しかし、英語等と異なり、中国古典文献の情報付与済みコーパスは少なく（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究」、第七届中國古代小說文獻暨數字化國際研討會論文、2008)、「電子コーパスは、（語学研究にはともかく）文学研究には用例検索や版本比較に使える程度」という認識が支配的であった。

POS タグ（part-of-speech tag: 品詞タグ）等の言語学的情報に比し、ディスコース情報等、文学研究に役立ちそうな情報の付与は自動化が困難なためか、こうした情報を付与したコーパス・そのコーパスを利用した研究は多くないが、中国古典文学研究でも視野に入れるべき手法と考えられ、本研究の着想に至った。

２．研究の目的

中国近世の白話文学（戯曲・白話小説・説唱等）の電子コーパス構築における情報付与（annotation）について、特に、作品に附けられた批評（総批・夾批・眉批）に関する情報やディスコース情報（談話情報）等、文学理論を中心とした文学研究及び隣接分野に役立つ情報の付与方法（内容・形式・方法）を検討し、こうした情報を付与したコーパスを試作・試用して、ディスコース分析・ナラティヴ理論をはじめ、関連する諸分野における新たな展開の可能性を探ることを目的とする。

３．研究の方法

(1)情報付与の対象とする電子テキスト（プレインテキスト）の整備

本研究に先行する科研費基盤研究「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」や、「研究の背景」の項で言及したプロジェクトにより既にプレインテキストが準備できている作品に加え、研究上重要で、現在まで良質な電子テキストがまだ公開されていないものについて、適切な版本・鈔本を選定し（影印本未刊行の版本については、必要に応じ、版本の実地調査・マイクロフィルム版入手を行う）、版本のスキャニング・OCR（専門業者への外注）を行い、電子テキスト化・校訂を進める。

(2)文学研究及び隣接分野の研究に有効な情報付与項目・方式の検討

中国近世白話文学を対象とした文学理論的研究を中心に、中国近世白話文学を資料として利用している隣接分野（中国語史、中国史、中国思想史等）の研究もできるだけ広く参照して、文学研究及び隣接分野の研究に有効と思われる（が、従来、中国古典文献の電子テキストには付与されてこなかった）情報付与の項目（メタデータ、作品自体に付加されている批評に関する情報、ディスコース情報、品詞情報等）及び付与方式の細部について検討する。

(3)タグつきコーパスの試作・試用

(2)で検討した情報付与の項目の一部につき、サンプル的にタグ付けしてコーパスを試作し、試用して、有用性や問題点を検討する。コーパス試作に際しては、簡単な Perl スクリプトを用いた半自動化と手作業を組み合わせたタグ付けを試みるほか、品詞情報の付与については現代中国語向けの形態素解析ソフトを試用し、近世の白話文献に利用した場合の精度等についても検討する。

４．研究成果

(1)情報付与の対象とする電子テキスト（プレインテキスト）の整備

・朱有燉雜劇

明の皇族、朱有燉（周憲王：1379-1439）作の雑劇脚本は、周藩刊本（その多くは作者自身による宣徳年間（1426-1435）の「引」が附されている）が現存しており、これは現存するテキストとしては『元刊雜劇三十種』に次いで古いものである。また、『元刊雜劇三十種』が曲辞を主として載せ、賓白（せりふ）をかなり省略しているのに対し、「全賓」をうたい、まとまった量の賓白を収めているため、口語資料として言語学的にも注目される資料と言える。周藩刊本は、中国国家図書館（旧北京図書館）、中央研究院傅斯年図書館、京都大学等に収蔵されている。しかし、京都大学蔵の三種（『周憲王樂府三種』（京都大學漢籍善本叢書第十五卷、同朋舍出版、1981年））を除き影印本が刊行されておらず、また、マイクロフィルム等の複製も部分的にしか許可されていない。

そこでまず、戯曲理論家でもあり実作者でもあった呉梅（1884-1939）が、蒐集した周藩本（現在は中国国家図書館蔵）に基づき校訂排印した『奢摩他室曲叢第二集』を外注 OCR によって電子テキスト化し、初歩的な校訂をおこなった後、周藩本の収蔵館で（マイクロフィルムまたは画像ファイルと）目睹対校して修正を進め、9 種の作品につき周藩本にできるだけ忠実な電子テキスト「周藩本原貌

版」も作成した。
　作成の過程で、『奢摩他室曲叢』第二集排印本は、文言については周藩本にかなり忠実な排印本であるが、明らかな誤字の修正や異体字を正字とする校訂のほか、曲牌・襯字の認定・断句などについて（おそらく呉梅自身の戯曲理論に基づいて）周藩本から少なからぬ変更を行っていることが確認された。

・金聖歎評『水滸伝』
『水滸伝』には著名な文人の名を冠した批評を付した版本が多く刊行されているが、文学批評史・文学理論などの観点からも注目されてきた、明末の文人金聖歎の批評が付された貫華堂本『水滸伝』、中華書局影印本に基づき、批評を含めて OCR するよう外注の際依頼して電子テキスト化を行った。この一部については、東北大学高等教育開発推進センター非常勤講師の井上浩一氏が校訂をお引き受け下さり、その成果を提供してくださった。

(2)(3)文学研究及び隣接分野の研究に有効な情報付与項目・方式の検討及びコーパスの試作・試用
　メタデータ（テキストそのものに関するデータ）、作品に附された批評（総批・夾批・眉批）、ディスコース情報（地の文・会話・心的思惟等の別、話し手の性別・地位・職業、戯曲の場合、“脚色（役柄）”、等）、品詞情報等の付与項目の詳細および付与方法について検討を行った。以前よりのプロジェクトで作成したプレインテキスト及び(1)で作成したプレインテキストの一部をサンプル的に用い、以下のタグ付けを行ったごく小さなコーパスを試作・試用し、その有用性や問題点について検討した。

・メタデータ及び批評に関する情報を付与したコーパス：
　メタデータについては、中国近世白話文学の特徴として、作者・作品の成立時期等が必ずしも明確でないものが多い点を考慮し、電子テキストの底本に関する書誌学的なデータをやや詳細に採り、生に近い形のデータと、データベース内で操作するための（生データから推定を行って算出した）データを分けて記載することとした（例えば、当時の書籍は、作品が成立したと推定できる時期、序に見える年月、刊記に見える年月が大きく異なっていることがままあり、またいずれかの情報しか無い、いずれの情報も無い、或いははっきりしないといったことがよくある。通常のように「出版年」の項目を一つ立てるだけだと、書籍のどの部分の情報に基づいて推定を行ったかという情報が消えてしまい、誤解が生じやすいので、必要に応じて生のデータ（序に記載されていた文言、刊記に記載されていた文言等）を参照できるよう、項目を別に立て、分けて記載する）が、なお検討の余地があろう。
　また、批評については、總評・眉批・夾批を区別してマークアップした。
　テキストは XML 形式とし、批評部分のみ／正文のみを取り出す XSLT スタイルシートを作成し、分析の補助とした。

・品詞情報を付与したコーパス：
現代語向け形態素解析ソフト ICTCLAS / NLPIR（http://ictclas.nlpir.org/）を使用し、POS タグ付けを試行した。形態素解析の結果については、資料によりばらつきがあるもののおよそ 70%前後の精度が見込めることを確認した。試用した ICTCLAS/NLPIR（2011 年版）では、筆者の力不足のためかインストール・セッティングで不具合が多く起き、辞書のカスタマイズを行うことができず、また、POS タグ付けを行った上でのディスコース情報の付与までは作業を進めることができなかったが、将来的には（一定のカスタマイズを行った上で）近世白話の文献にも利用可能ではないかと考えられる。

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　0　件）

〔学会発表〕（計　0　件）

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔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
(　　)

研究者番号：

(3)連携研究者
(　　)

研究者番号：