

## 序

本報告書は、平成 17 年度科学研究費補助金(基盤(C)(2)企畫調査・課題番号 17630004)「中國近世白話文學の電子化の現況及び學術利用に有効なコーパスの設計に関する調査」の成果の一部をまとめたものである。

本企畫調査の組織は以下の通りである。

研究代表者： 笠井直美（名古屋大學大学院國際開發研究科助教授）

研究分擔者： 廣瀬玲子（専修大學文學部教授）

上田望（金沢大學文學部助教授）

研究協力者： 中塚亮（名古屋大學大学院文學研究科博士課程）

近年、中國古典文學の分野でも電子化の進展はめざましく、底本・校訂等を問わなければ主要な資料はほぼ電子化されているという状況になりつつある。

これらは大きくは二つのタイプに分類できよう。

一つは、現代人の娯楽・教養に供するもので、インターネット上で無料または比較的安価で公開されているか、或いは安価な CD-ROM の形で供給されているもの。このタイプのデータは、恐らくは流用・盗用が頻繁に行われているため、重複が多い。また、これらのデータは、學術的な利用を前提としていないため、電子化に使用した底本が明示されていない、校訂が不十分である、電子化の底本として使用されたと思われる排印本自體が學術利用に不向きである、知的財産權に關して配慮されていない、といった問題もまま見られる。とは言え、後述のように、學術的に價值の高いデータベースはしばしば高額であり、かつ、その高額なデータベースを購入しても比較的單純な検索しか使えない、という状況下では、「現代人の娯楽・教養に供する」タイプのデータの利用も、選択肢の一つとして重要性を失っていないと考える。

もう一つは、學術研究向けの目的で構築されたもの。多數の専門家が關與した大規模な電子化資料としては、以前は臺灣の「中央研究院漢籍電子文獻」<sup>1</sup>が唯一といった感じであったが、ほかにもウェブサイトで検索システム付きのコーパスを公開する研究機關が現れ、また、『四庫全書』『四部叢刊』『中國基本古籍庫』といった大規模な電子化資料が續々と發賣されており、今後も増えていくと予想される。ある程度信賴できるデータを具えた大規模な電子資料を検索できるようになったことは福音である。

しかし、近年學術研究向けに發賣されたり、大學や研究機關のウェブサイトで公開されるようになったデータベースにも、以下のような問題がある。

まず、こうしたデータベースのうち、販売されているものは高價であり、個人はもちろん、大學などの研究機關でも購入が難しいものもあること。

また、ウェブサイトで無料で公開されているものでは、用例の検索はできても、その出典が表示されない、検索結果が KWIC 表示されるが、その用例を含む段落全体を見ることができない、また、コーパスに入っている作品の朝代別の分布等、概況を示す文書はウェブサイト上で公開されているが、詳細な作品名や版本などの情報は公開されていない、といったものがある。（これは北京大學漢語語言學研究中心の「CCL 語料庫」<sup>2</sup>のことであるが、出典を表示しないのは、現代中國語のバランスド・コーパスである臺灣の中央研究院の「現代漢語平衡語料庫」<sup>3</sup>にも共通する。これは、英語をはじめとする他のメジャーな言語の研究者向けコーパスには見られない特徴である。学習者向けにはこれで十分ということなのかもしれないが、こうした第三者の検証を許さない「由らしむべし知らしむべからず」的な設計のために、研究のための利用価値が大きく減じているのは残念である。）この点では、中央研究院の「近代漢語標記語料庫」<sup>4</sup>が比較的良く配慮されていてありがたいが、これでさえも使用された版本についての情報はサイト上には明示されていない。また、いずれのサイトについても、版本の分岐状況がよくわかっている小説について、幾つかの検索結果をつきあわせれば、使用された版本がある程度推定できるが、採用されていると思われる版本が、こちらの目的と合わないものであることも多い（學術利用上問題があると思われる版本が採用されている場合もある）。

そして、最大の問題点としては、販賣されているものも、ウェブサイトで無料で公開されているものも、使いやすさを重視したインターフェイスと一緒に供給され、可塑性のあるテキストファイルの形ではないため、高度な操作・加工を施すことができず（検索結果のコピー&ペーストすらできないものもある）、電子化の恩恵を十分に受けられないことが挙げられる<sup>5</sup>。

もちろん、比較的単純な検索（and/or 検索程度）しかできないものから、ワイルドカードやパスの指定等によって、ある程度複雑な検索が可能であるもの（「CCL 語料庫」）、データに形態素によるマークアップが施されており、それを利用した検索が可能なもの（「近代漢語標記語料庫」）までであるが、検索対象となるデータも、検索システムも、「與えられたものをそのまま」利用する以外なく、利用者がカスタマイズする餘地がないことからくる制限は、研究を目的とする場合にはやはり大きい。

現段階では、英語や日本語を扱う場合（無料で公開されているテキストデータが大量にあり、Brill's Tagger や茶筌のような無料で配布されている形態素解析プログラムがあり、辞書をカスタマイズしたり、プログラム自体を改変することもできる）に比べると、かなり条件に差があると言えよう。

こうした条件下では、正規表現による検索や計量的な分析、研究目的に合わせた種々のマークアップを考えている研究者は、やはり目的に應じて自分で（或いは共同で）コーパスを構築する必要があると思われる。

本企画調査では、こうした状況を踏まえ、中國近世白話文學の學術研究向けコーパスの設計・構築に向けての基本的な情報の把握・整理を目的とし、（1）中國、臺灣をは

はじめとする各国における中国近世白話文学の電子化状況の調査、(2) 学術利用に供するために適切な電子コーパスの設計・形式に関する調査、を行った。

たとえてみれば、現在、資金力のあまりない研究者でも利用可能なコーパスは、いわば「TV ディナー」的なものであって、手輕な代わりに自由が利かず、様々に展開・應用していくには向かない、という状況下で、手作りの料理をする爲に必要な情報——食材に関する情報(1)と、料理法に関する情報(2)——の収集を試みた、ということになるのか。

こうした調査を行うこと、また継続的に情報を更新していくことは、個人の力では限界がある。例えば、東洋学全般にわたるリンク集「Kanhoo! 東洋学サーチ」<sup>6</sup>は、たいへんな労作であり、本企画調査もその恩恵を大いに受けているが、更新はしばらく止まっている。ある程度多数の研究者が共同で情報を共有し、なるべく少ない負擔で更新していく仕組み作りを考えていくべきだろうが、その場合にも「たたき臺」となるデータが必要である。この観点からは、2006年3月時点の暫定的な報告に過ぎないものでも、今後修正・増補されるべき「たたき臺」としてまとめておくことも一定の意義があると考え、本企画調査のうち(1)の部分の成果の一部をまとめ、報告書として刊行することとした<sup>7</sup>。

本報告書は、「雑劇篇(増訂版)」「戯文・傳奇篇」「白話小説篇」の三部から成り、それぞれについて、凡例・採録した電子化資料についての説明・正文(電子テキストのリスト)をつけた。既に雑劇に関しては、笠井直美「中国近世白話文学の電子化の現況(雑劇篇)」(『名古屋大学中国語学文学論集』第18輯、2006)において報告を行っている<sup>8</sup>が、本報告書「雑劇篇(増訂版)」では、その後発見した誤字を訂正し、紙数の関係で盛り込めなかった情報を増補したほか、「南戯・傳奇篇」「白話小説篇」に合わせた形式上の変更を行った。

本企画調査では、上述のように、(あまり資金力のない個人や組織が)研究上の目的に合わせてコーパスを構築する場合を念頭に置き、現段階でどの作品について、どのような形態の電子版にアクセスが可能であるかの大要を把握することを目標とし(網羅的な調査は不可能だし、また、そうする意義も大きくはないと考える)、萬人に開かれているもの、比較的安價でアクセスできるもの、テキストデータが入手できるものに重点を置いた(例えば、『中国基本古籍庫』は、多数の作品を収録していると思われるが、價格、テキストデータの取得のしにくさ等の點から今回の調査対象からは外した<sup>9</sup>)。畫像データは、また別の價值があるが、今回の目的からは外れるので、調査の対象とはしなかった。

また、電子化された資料は重複が非常に多いので、本報告書のリストでは調査結果の全てを掲出することはしなかった。ウェブサイトについては、上記の基準に加え、比較的多くの作品をまとめて公開している、テキストデータが取得しやすい、アクセスが容易である、比較的長期にわたって安定的に運営されている、といった観点から、サンプ

的に幾つかのサイトでの状況を示したにとどまる<sup>10</sup>（ただし、研究者が独自の校訂を行ったものや、生データに近いもの——木版本や鈔本などを OCR してテキストファイル化したデータなど——を公開しているサイトについては、なるべく掲出するよう心がけた）。また、CD-ROM については、収録されている作品名を明らかにできたもの（ほとんどは筆者が偶々入手できたもの）に限った。現段階でどの作品が電子化されているか、どのような形態の電子テキストにアクセス可能かの概況を示すのみとお考えいただければ幸いである。

調査に際しては、研究分擔者の上田望氏（金沢大學文學部助教授）、廣瀬玲子氏（専修大學文學部教授）、研究協力者の中塚亮氏（名古屋大學大学院文學研究科博士課程）をはじめ、多くの方々からご教示・ご助力をいただいた。厚くお禮申し上げたい。

本企画調査では、地方劇脚本・説唱には調査が及ばなかったなど、今後の課題も多く、また、調査を行ったジャンルに関しても、本報告書は、上述のように今後修正・増補されるべき「たたき臺」であって、誤りも多いことと思う。ご批正・ご教示をお俟ちする次第である。

2006 年 3 月

笠井 直美

---

<sup>1</sup> <http://www.sinica.edu.tw/~tdbproj/handy1/>

<sup>2</sup> [http://ccl.pku.edu.cn:8080/ccl\\_corpus/jsearch/](http://ccl.pku.edu.cn:8080/ccl_corpus/jsearch/)

<sup>3</sup> <http://www.sinica.edu.tw/ftms-bin/kiwi.sh>

<sup>4</sup> <http://www.sinica.edu.tw/ftms-bin/kiwi1/pkiwi.sh>

<sup>5</sup> これは、無法な海賊版が横行している現状ではやむを得ない面もあるが、英語における Project Gutenberg に相当するものが存在しないことは、中國古典文學を対象とする電子コーパスを利用した研究の發展を阻害する大きな要因となっていると考えられる。

<sup>6</sup> <http://www.jaet.gr.jp/kanhoo/>

<sup>7</sup> 電子版は以下のサイトで公開している。<http://www.gsid.nagoya-u.ac.jp/kasai/>

<sup>8</sup> また、本報告書のこの「序」は、この拙稿の「はじめに」と、内容が一部重複している。

<sup>9</sup> 『中國基本古籍庫』を開発した北京愛如生數字化技術研究中心のサイトには、比較的安價な年間利用料を支拂うことで、電子版の閲覧ができる「八十萬卷樓」もある

（<http://www.cn-classics.com/bashi/>）。「八十萬卷樓館藏一覽」によると、例えば戲曲作品に限っても、脈望館鈔本や李玉の作品など、他のサイトに見られないデータを収めている。この「八十萬卷樓」の電子版は、テキストデータではあるのだが、「読む」以外の目的には使いにくい縦書きデータのため、やはり今回の調査の対象からは外した。

<sup>10</sup> また、『國學宝典』（北京國學時代文化傳播有限公司）をリストから外したのも、現在入手不可能である・著作権上の問題があるほか、検索はできるがテキストデータの取得がしにくいことも理由の一つである。