社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－DBS－143 (15)
2007／7／2

文学作品成立論のためのデータベース構築に向けて

—坪田譲治作品による試み—

国島 丈生† 横田 一正† 山根 知子††

† 岡山県立大学情報工学部
〒719-1197 岡山県総社市窪木 111
†† ノートルダム清心女子大学日本語日本文学科
〒700-8516 岡山市伊福町 2-16-9
E-mail: †kunishi@acm.org, ††yokota@c.oka-pu.ac.jp, †††t-yama@post.ndsu.ac.jp

あらまし 文学作品研究において、作品成立の過程を明らかにすることは主要研究課題の一つである。この際、作者の草稿や出版物など様々な資料について多角的に関連や類似性を調べる必要があるが、今のところ、この作業は人手による部分が大きい。著者は現在、複数の文学研究者と共同で研究プロジェクトを行いつつあり、今回はそのうち、児童文学作家・坪田譲治の文学作品成立論の支援について、データ工学の視点から分析を試みる。

キーワード 近代文学研究, 作品成立論, エピソード検索, 注釈, XML

Toward the database systems for researches on formation of literature

—a test study by Joji Tsubota's literature—

Takeo KUNISHIMA†, Kazumasa YOKOTA†, and Tomoko YAMANE††

† Faculty of Comp. Sci. and Syst. Eng., Okayama Prefectural University
Kuboki 111, Soja, Okayama, 719–1197, Japan
†† Faculty of Literature, Notre Dame Seishin University
Ifuku-cho 2–16–9, Okayama, Okayama, 700–8516, Japan
E-mail: †kunishi@acm.org, ††yokota@c.oka-pu.ac.jp, †††t-yama@post.ndsu.ac.jp

Abstract One of the important topics on literature research is the process to form literature. In order to clarify the formation process, it is necessary to investigate relationships or similarities between various literature materials, including manuscripts or publications. The authors have been cooperating on computer support for clarifying literature formation process from the information science point of view. This report includes our test study by Joji Tsubota's literature, who is a Japanese writer of 1900's. We also study some possibilities for supporting literature researches by computer, especially by using data engineering technologies.

Key words modern literature research, literature formation process, episode search, annotation, XML

1. はじめに

文学研究において重要な研究課題の一つに「作品成立論」(もしくは「作品構成論」)と呼ばれるものがある。これは、特定の文学作品について、作家がその作品をどのように完成させていったかという過程を様々な資料から考察し明らかにしていくという研究であり、古典／近代、日本／海外を問わず、広く行われている。用いられる資料はその作品に関連するものはほぼすべてと考えられ、日本近代文学の場合、出版本、自筆の原稿、作家の日記や書簡などが対象となることが多い。これらの資料を互いに突き合わせて比較検討し、類似点や相違点を明らかにすることが作品成立論における基礎作業となる。比較検討の項目には、文章そのものの類似点や相違点、推敲の痕跡のほか、原稿用紙の種類、筆記具の種類、筆跡、書き損じの痕跡、出版本の組版なども含まれる。

作品成立論の一例として、宮沢賢治の「銀河鉄道の夜」の成立過程に関する研究が挙げられる [1]。この作品の自筆原稿はさまざまな種類の紙や筆記具を用いて執筆・推敲が行われており、

しかも多くの紙には番号が振られていない。そのため作品自体の構成を解読するのが極めて難しく、実際、昭和9年に最初に活字化されてから長い間、物語の構成が大きく誤った形で流布していた。昭和40年代に自筆原稿に関する大規模な研究調査が行われ、その結果、ほぼ決定版と考えられる成立過程と物語構成が明らかになった。その後の出版は多くの場合この研究成果を基に行われている。このときの研究調査の項目は、文章の類似点／相違点、自筆原稿の紙の種類、執筆や推敲に用いられた筆記具の種類や色、筆跡、原稿用紙に残されている折り目、書き損じ、裏面などに書かれた別の作品の文章、インクの染み写りなど、極めて多岐に及ぶ。これらの調査を自筆原稿84枚に対して行い、得られたデータから物語構成を決定するまでに5年程度を要している。

この例に見られるように、作品成立論では多くの資料について多面的に類似点や相違点を分析する必要があるが、少なくとも日本近代文学の分野では、現状でもその作業の多くを人手で行っており、作業に長い時間を要している。一方で、作品成立論には、作家論など他の文学研究のための基礎資料作りという側面もある。今回、作品成立論の作業のうち可能な部分についてはコンピュータによる作業の効率化を図りたいという要望を共著者が持っていることを知り、共同研究という形で可能性を探ることとなった。本稿では、日本近代文学の作品成立論をコンピュータで支援する可能性についてデータ工学の立場から（もしくはもう少し広く情報科学の立場から）分析を試みる。また、一部については支援の試みを行ったので、その結果や問題点についても報告する。

2. 近代文学作品成立論へのコンピュータの応用

作品成立論をコンピュータによって支援する方法には様々なものが考えられる。以下に我々のグループ内で議論を行ったときに話題となったものを列挙する。

（1） 自筆原稿や出版本からの文字認識・自動テキスト化
（2） 資料もしくは資料間に対する注釈付け
（3） 資料の検索
（4） 資料間の類似性の発見

このうち1は、出版本のように活字化されたものに対しては既存のOCR技術によって実現可能ではないかと思われるが、自筆原稿については、原稿用紙の升目に字が埋められているとは限らず、また手入れが複雑に行われていることが多いため、技術的にかなり難しいのではないかと考えている。以降の章では、残りの2, 3, 4について考察を行う。

2.1 作品成立論で扱われる資料

自筆原稿をはじめとして、文学研究における資料は貴重なものが多い。また文学作品は、出版されるたびに作家もしくは編集者によって手入れが行われることがあり、出版本間の差異も作品成立論の研究対象となる。そのため出版本も、極論すればすべての版が研究の資料となる。このような理由により、研究者がすべての資料の現物を所有していることは稀であり、復刻本や写真などの形で資料の複製を手元に置き、研究を行っていることが多い。現在では高精細画像としてコンピュータ上に複製を所有している場合もある。

自筆原稿は資料として貴重である一方で、活字化されていないため万人に読みやすい資料とは言いがたい。そこで、原稿の内容をそのまま活字化した資料を作成することが多い。これを「翻刻」と呼ぶ。出版本とは異なり、翻刻の目的は原稿の内容をそのまま写し取ることであり、原稿用紙の切れ目、文章に対する手入れなども何らかの記法で活字化される。現在は、コンピュータ上のテキスト文書として翻刻を作成する場合が多い。また、翻刻をコンピュータ上の文書とすることで検索や加工が容易となることから、出版本もテキスト文書として翻刻することが多い。この場合も、出版本の組版の様子が分かるようにテキスト化されるという点が、青空文庫[2]に代表される文学作品の電子化とは異なる。ただし、自筆原稿をコンピュータで文字認識させるのは技術的に難しく、現在、翻刻の作成は人手で行われている。

以下に示すのは、坪田譲治（1890-1982）の作品「笑顔のお地蔵さま」の翻刻の一部である[3]。坪田譲治は共著者の山根が積極的に研究に取り組んでいる作家の一人であり、岡山県出身であることから、自筆原稿や生家、小説のモデルとなった地域など、実際に残されている資料を参照しやすい。また、作品の手入れの過程が比較的複雑で、作品成立論支援を考える題材として適当であると考えられる。以上のことから、本稿では主に坪田譲治の作品を例として取り上げる。

〈第4葉〉
一つ彼の目をひくものがありました。それが、【そ→こ】の作品の題目になってゐる笑顔のお地蔵さんです。それは山裾をまわって、その村へ出るちょっと前、三本の杉の大木の下に立ってゐました。それらの木に囲まれ【(無)→てゐ】るやう【に→な】形で、【丁度蓮の花ビラのやうな→その一尺くらいしかない】お地蔵さまは立ってゐたのです。

この例では、【　→　】という記法によって原稿用紙上に見られる手入れを示している。矢印の左が手入れ前、右が手入れ後をそれぞれ表す。このように、手入れによって発生した文章の変化を「校異」と呼ぶ。校異のパターンに関する詳細な分析は2.2節で述べる。

自筆原稿を資料として扱う場合、もう一つ重要視されるのが手入れ前の文章（作品）と手入れ後の文章（作品）である。これは、それぞれの文章がある時点における作品の姿であると考えられるからである。共著者の山根はこれらに草稿の「形態」という言葉を与えており、手入れ前を第一形態、手入れ後を第二形態、というように名付け、それぞれテキスト化している。一般に手入れは自筆原稿中に複数箇所存在するが、筆記具の種類や筆跡その他の情報を基に、同時期に行われたと考えられる手入れについてはまとめて扱う。したがって、手入れ箇所の数だけ形態が存在するわけではない。また、いったん手入れを行った箇所にさらに手入れを行うことがあるため、同一の自筆原稿について形態が三つ以上存在することもある。

一つの文学作品が完成するまでに下書きや清書が繰り返され

るため、一般には一つの文学作品に対し複数の自筆原稿が存在する。これを「稿」と言い、年代順に一次稿、二次稿、三次稿などと呼ぶ。これらはすべて完成しているわけではなく、途中で中断されたものも「稿」と呼ばれることがある。

まとめると、作品成立論をコンピュータで支援する場合、次のような電子的資料を扱う必要がある。

（1） 自筆原稿…原稿用紙単位の画像。複数の稿が存在し得る。また、原稿用紙の順番が固定していない場合も考えられる。

（2） 出版本…ページ単位の画像。一つの作品が複数回出版されることがあり、そのすべてが研究対象となり得る。

（3） 自筆原稿の翻刻…原稿用紙単位のテキスト文書。原稿用紙の順番が明らかな場合には稿単位となることもある。校異、原稿用紙や筆記具の種類、筆跡、推敲を行った人物（本人か他人か、など）といった情報も何らかの形で表現される必要がある。

（4） 出版本の翻刻…ページ単位のテキスト文書。

（5） 自筆原稿の形態をテキスト文書化したもの。一般には原稿用紙単位だが、原稿用紙の順番が明らかな場合には稿単位になることもある。

2.2 「でんでん虫」を例とした校異パターンの解析

次に、これら電子的資料の間の関連について考察を行う。著者らのグループ内での議論で挙がった要求には次のようなものがあった。

（1） 画像とそれに対応するテキストを大体並べて閲覧したい

（2） 2つのテキスト間の差分を人手によらず求めたい

例えば出版本の場合を考えると、前節で挙げた電子的資料の4がページ単位のテキスト文書であるので、該当する2の画像と並べて表示するのは易しい。自筆原稿の場合も、3もしくは5が原稿用紙単位になっている、もしくは稿単位であってもその中に原稿用紙の区切りが書かれていれば、難しいことではない。従って、問題の多くはユーザインタフェース設計の工夫により解決できると考えられる。この要求の発展形として、複数の資料を同時に画面上で閲覧したいという要求も考えられるが、これもユーザインタフェース設計の問題と考えられる。ただし、資料の検索を行うときに原稿用紙もしくはページにまたがった文章を対象とすることがあるので、テキスト文書をどのように記述するかは考慮する必要がある。

2は、校異を求めたり翻刻を作成するとき2つのテキストの差分を人手で求めており、これに多くの時間を費やしていることから出てきた要求である。しかし、この要求は作品成立論の支援において非常に本質的な要求であると考えている。なぜなら、作品の成立過程を明らかにする場合、さまざまな時点での作品（文章）をさまざまな粒度で比較し、類似点や相違点を求めることが重要であろうと考えられるからである。

実例として、坪田譲治の「でんでん虫」(注1)の翻刻[4]を対象とし、この中でどのような校異が発生しているかを調査することとした。この作品を選んだのは、近年岡山市立中央図書館にて自筆原稿が発見され、この第一形態が「カタツムリ」、第二形態が「でんでん虫」という別の作品としてそれぞれ出版されていることが判明したことによる。このため、手入れが多数かつ複雑であり、校異のパターンを解析するのに適切であると考えたためである。

調査の結果、次のような校異パターンがあることが分かった。

テキストの置換　例：美代【チャン→ちゃん】は【病気で→熱があつて】ねていました。

テキストの挿入　例：「どうせう【(無) →。】」

テキストの削除　例：「善太【君→ (削)】、とっ来い。」

改行の変更　例：【(改行・一字空け) → (改行トル)】…改行と段落の一字下げの削除、すなわち二つの段落を一つにまとめたことを意味する。

テキストの入れ替え　例：【一人で⇔駈けて【行→い】き、門【から→の】中へ頭を突っ込【みました。そして→んで】】…この場合、3ヶ所のテキスト置換と1ヶ所のテキスト入れ替えが手入れされていることを表している。

この分析結果から分かることとして、改行の変化も情報として残すべきである、修正操作が多重に行われることがある、といったことが挙げられる。

また日本語作品に固有の問題として、傍点やルビの扱いがある。つまり文学作品の文章はおおまかには単なる文字列であるが、厳密に見れば傍点やルビといった注釈が付与された（構造を持った）文字列であると考えられる。テキストの差分を求めるアルゴリズムや編集操作に関する研究は、単なるテキストに対するもの（例えば[5][6][7]）、XMLに対するもの（例えば[8][9][10]）、いずれにもあり、歴史が非常に長い。どのような先行研究を参考とするかは今後の課題である。前節で挙げた電子的資料のうち3と5は、適切なデータ表現を用いれば、一方からもう一方を自動生成することも可能であると考えられる。第一形態のテキストに対して文字列の修正操作を繰り返すことで第二形態のテキストが得られる。このとき、第一形態と第二形態という二つのテキストを差分情報を含めて併合したものが翻刻と考えられるからである。

2.3 テキストや画像に対する注釈付け

原稿用紙や筆記具の種類、筆跡など、翻刻中に直接埋め込みにくい情報については、現在は「解題」と呼ばれる作品解説中に記述することが多い。しかし、本来はこれらの情報は画像やテキストに付随する属性であり、近年技術進歩の著しいメタデータや注釈付けを適切に用いれば、電子的資料そのものに埋め込むことができると考えられる。このうち原稿用紙の種類は原稿用紙全体に付けるべき情報である。一方筆記具の種類や筆跡は、原稿用紙全体ではなく一部、翻刻全体ではなく一部の文字（または文字列）に対して付けるべき情報である。

後者のようにテキスト中の一部に注釈を加える場合、注釈を含めたテキストをXMLとして扱えると、実装を考える上では都合が良い。しかしながら、一般に注釈を加える場所が互いに重なり合うことはまったく不自然ではない。したがって、要素の重なり合いを許さないXMLは、注釈付きテキストのデータ

(注1)：正確には「でん（繰り返し記号）虫」であるが、本稿では読みやすさを考慮して「でんでん虫」と記述する。

モデルとしては制限が強すぎる。マークアップ言語において要素の重なり合いを扱う方法は著者らも過去に提案を行った[11]し、先行研究も非常に多い[12]が、いずれにも一長一短がある。

別の問題として、同一の注釈を複数の箇所に付けることを考慮しなければならない。文学研究において資料に付けた注釈は不変ではなく、新たな資料が発見されたり、新たな事実が判明すると変更されていくものである。この点を考慮して、注釈をどう扱うか決定せねばならない。

図1に示すのは、注釈を付ける領域をTEIのmilestone形式[13]で指定し、注釈そのものは別のXML文書で記述する、という方式による記述例である。TEIのmilestone形式は空要素タグによって領域の開始地点と終了地点を示すもので、形式的にはXML文書である、要素のテキスト値（この例ではmanuscript要素のテキスト値）が正しく保存される、などの利点を持つ。これが最善であるかどうかは検討の余地が多分にあるが、一つの選択肢であるとは考えている。

2.4 エピソード検索

資料の検索に関して、山根から「エピソード検索」という手法の必要性が挙げられている。エピソード検索については[14]でも言及されているが、作品中のある出来事を指定して、それに類似する出来事が記述されている作品（の部分）を求めるという検索である。この検索は、坪田譲治が一つのエピソードを複数の作品で使い回すことが多いために、坪田の作品成立論を行う上で切実な要求となっている。ただし、エピソード検索の需要は、他の作家の文学研究でも存在するようである。

[14]では、エピソードの単位を段落とみなし、係り受け関係に類似性のある段落を類似エピソードとして求める、という手法を提案しており、実験の結果、ある程度正しいエピソード検索に成功している。しかし、類似した単語を用いたエピソードの検出や複数段落にまたがったエピソードの検出などに問題を残している。

一方、形態素解析や係り受け解析を行わず、文学作品を単なる文字列とみなして解析を行う手法も提案されている。具体的には、n-gramを用いた和歌の解析[15]、データマイニング技術を応用した和歌の解析[16]などの研究がある。近代日本文学と和歌や漢字文献とは文字列の長さなどの性質が異なるので、これらの手法がそのまま応用できるとは考えにくいが、[14]とは異なる可能性として検討に値する方向性だと考えている。

3. おわりに

本共同研究はまだ始まったばかりであり、現在は互いの語彙や背景知識を擦り合わせつつある段階である。したがって、本稿で述べた考察も非常に不完全かつ断片的なものにとどまっている。最終的には画像とテキストを統合して扱える支援環境にできれば理想であるが、まず、どちらかというと扱いやすいと考えられるテキストに対象を絞って研究を進めていくつもりである。

謝辞 本研究で使用した坪田譲治作品の一部は、坪田譲治研究データベース[17]の資料を参考にしている。また、2.2節、2.3節の内容については主に、平成19年3月に岡山県立大学大学院情報系工学研究科修士課程を修了した難波佐代氏（現在(株)システムエンタプライズ）の研究成果[18]に因るところが大きい。ここに謝意を表す。

文　献

[1] 入沢（編）：“「銀河鉄道の夜」の原稿のすべて”, 宮沢賢治記念館 (1996).
[2] “青空文庫”, http://www.aozora.gr.jp/.
[3] 山根：“坪田譲治　草稿「笑顔のお地蔵さま」（ノートルダム清心女子大学所蔵）– 解題と翻刻 –”, 清心語文, 8, pp. 81–91 (2006).
[4] 山根：“坪田譲治　草稿「でんでん虫」「妹とカタツムリ」– 解題と翻刻 –”, ノートルダム清心女子大学紀要, **29**, 1, pp. 30–43 (2005).
[5] E. W. Myers: “An o(nd) difference algorithm and its variations”, Algorithmica, **1**, 2, pp. 251–266 (1986).
[6] J. André and H. Richy: “Paper-less editing and proofreading of electronic documents”, EuroTEX'99 Proceedings (1999).
[7] H. Ogata, C. Feng, Y. Hada and Y. Yano: “Computer supported proofreading exercise in a networked writing classroom”, ICCE'99 (1999).
[8] Y. Wang and D. J. DeWitt: “X-diff: An effective change detection algorithm for XML documents”, ICDE2003 (2003).
[9] 鈴木：“XML文書と正規木文法との間の最適編集操作列の発見”, Technical Report PRMU2005-46, 電子情報通信学会技術研究報告 (2005).
[10] 久保山, 宮原：“木の編集距離を用いたwebページからの情報抽出”, Technical Report DC2004-32, 電子情報通信学会技術研究報告 (2004).
[11] K. Yokota, T. Kunishima and B. Liu: “Semantic extensions of XML for advanced applications”, Australian Computer Science Communications, Vol. 23, pp. 49–57 (2001).
[12] S. DeRose: “Markup overlap: A review and a horse”, Extreme Markup Languages 2004 (2004).
[13] M. Sperberg-McQueen and L. Burnard Eds.: “Technical Topics: Multiple Hierarchies”, chapter 31, TEI Guidelines for Electronic Text Encoding and Interchange, Text Encoding Initiative (2004).
[14] 川上, 劉, 北川：“坪田譲治の児童文学作品におけるエピソードの検索”, DEWS2005, No. 4A-i5 (2005).
[15] 近藤：“n グラム統計処理を用いた文字列分析による日本古典文学の研究 –『古今和歌集』の「ことば」の型と性差 –”, 千葉大学文学部『人文研究』, 29, pp. 187–238 (2000).
[16] 竹田, 福田：“古典和歌からの知識発見 – モビルスーツを着た国文学者 –”, 情報処理, **43**, 9, pp. 941–949 (2002).
[17] “坪田譲治研究データベース”, http://crane.mis.ous.ac.jp/.
[18] 難波, 国島, 横田, 山根, 劉：“近代文学での校異研究支援機能に関する考察”, 平成18年度電気・情報関連学会中国支部第57回連合大会, pp. 267–268 (2006).

```
<manuscript title="カタツムリ">
  <sheet kind="start" page="1" />
お母さんは<mark kind="start" id="1"/>お医者さんへ行ってゐました。<mark kind="end" id="1"/>
美代<mark kind="start" id="2"/>チャン<mark kind="end" id="2"/>は<mark kind="start" id="3"/>
病気で<mark kind="end" id="3"/>ねていました。
  <sheet kind="end" page="1" />
  <sheet kind="start" page="2" />
...
</manuscript>

<annotation>
  <region id="1"/>
  <region id="2"/>
  <region id="3"/>
  ...
  <description>ブラックインクによる手入れ</description>
</annotation>
```

図 1　TEL milestone 形式による坪田譲治「カタツムリ」のマークアップ例