人文科学とコンピュータ　26－9
（１９９５．　５．　２５）

パソコンを利用した日本文学作品の研究
－芥川龍之介作品のテキストデータベース化を通して－

上村 和美
（社会保険 神戸看護専門学校 非常勤講師）

私は博士論文『テキストデータベースによる色彩表現の研究－芥川龍之介作品への適用－』[1]の研究にあたり、パソコン[2]を利用し、テキストデータを作成、用例を抽出するという作業を行った。ここで「テキストデータベース」と読んでいるものは、分かち書きされていたり、品詞情報が付加されているものではなく、いわゆるプレーンテキストのことである。

パソコンを利用する以前には、用例の抽出作業は手作業で行っていた。しかし、手作業には限界があり、大量のデータの中から正確に用例を抽出するというのは困難であった。したがって、本来は研究の手段であった用例の抽出が研究の目的そのものにならざるを得ない状況であった。そこで、本研究では大量のデータから正確に用例を抽出することを実践するために、パソコンを利用した。

テキストデータの作成にあたっては、さまざまな問題が生じた。まず第一に入力の問題、さらには英語のように分かち書きがなされていない日本語を処理する上での問題等である。

本発表では、日本文学作品の中でも特に、近代文学作品研究におけるパソコンの利用および問題点を、実践した経験から述べてみたい。

Studies on Modern Japanese Literary Works by Personal computers
– A Databese of AKUTAGAWA Ryunosuke Novels –

Shakaihoken Kobe nursing vocational school

Kazumi Uemura

Lexicographical study of literary works has traditionally been performed manually using paper cards. This is time-consuming and collecting data itself has often been turned into the objective, not the means, of the research.

While studying color terms in novels by AKUTAGAWA Ryunosuke, I have constructed a database of his novels on a personal computer in order to facilitate the research. In doing so, I have encountered with a variety of problems concerning input itself and the processing of Japanese data.

In this paper, I would like to discuss issues concerning use of personal computers for studying Japanese literary works.

[1]大阪大学大学院言語文化研究科博士論文, 1994

[2]本発表で利用するコンピュータは、あくまでもパーソナルコンピュータであり、大型汎用機等の利用は考えていない。

1 テキストデータの作成

1.1 作成方法

自分の必要としている作品が、既にテキストデータ化されていれば、それを利用すればよい。しかし、現段階はまだ発展途上にあるので、必要とする作品は必ずしもテキストデータ化されていない。また、既にテキストデータ化されていたとしても、公開されていないものも多い。

よって、現状では**必要なテキストは自分でテキストデータ化の作業を行う**という可能性が高い。

通常、テキストデータ化を行う方法は、次の2つに大別することができる。

1. マニュアル入力による方法
2. OCR を利用する方法

私自身、修士論文作成の段階においては、OCR を使用せず、総てはいわゆる手入力＝マニュアル入力で行っていた。ごく初期のデータ作成段階ではワープロソフトを使用していた。[3]芥川は、短編の多い作家であるので、当初はさほど問題も起こらなかったのであるが、比較的長い作品の入力にとりかかった時に、1つの文書ファイルにおさまらないという問題が生じてきた。ワープロソフトでは、あらかじめ文書ファイルの容量が指定されているので、それ以上の容量のものは扱えなかったのである。また、処理スピードの面でも不満が出てきた。

そこで、テキストデータベースの作成時には、**エディタ**を利用するようになった。[4]

1.2 作成時の問題点

1.2.1 マニュアル入力の問題点

芥川作品の入力に関する最大の問題点は、作品が**歴史的仮名遣い**で表記されているという点であった。たとえば「云ふ」と入力したい場合、一旦「云う」と入力した後、「云ふ」に直すという作業を行うのである。この作業は、古典作品の場合には、さらに問題が複雑となるだろう。ワープロやエディタソフト上で使用する日本語フロントエンドプロセッサは、当然のごとく現代語の文法に従って構成されている。したがって、歴史的仮名遣いの作品では、次のような不都合が生じるのである。

- ワ行→ハ行
- そうだ→さうだ
- ようだ→やうだ
- だろう→だらう
- い・え→ゐ・ゑ
- っ→つ

1.2.2 OCR 入力の問題点

それでは、以上のような仮名遣いの問題は、OCR で解決できるのではないかということになる。

修士論文作成時点において、未入力であった作品に関しては、OCR[5]を利用して入力を行った。

実際に使用してみると、次のような問題点が出てきた。

1. 読み取り作業が総て終了するまで、OCR につきっきりになる。
2. オートフィーダー機能を利用する為には準備が必要である。
3. 日本語 OCR の認識率は英語のようには高くない。

1. の方法では、原典とする書籍をページ単位で読み込んでいくので、1ページが終われば次のページをという作業を続けることになる。2. ならば、用紙をセットしておくだけで、OCR は自動的に読み取り作業を行う。しかし、その為には、あらかじめ、テキストをゼロックスコピーしておき、各ページの図表やページ数など、ノイズに当たるものは、この段階で処理しておかなければならない。また、3. の問題点の具体例としては次のようなものがある。

- 時→峙
- 手→乎

[3] 当時は『一太郎 Ver.3』（株式会社ジャストシステム）を使用していた。

[4] 私は、Vz エディタに、ATOK8 を組み込んで使用している。

[5] 大阪大学言語文化部・言語工学部門の富士電機 XP-70S を使用した。

- 見→兄・兌

また、次の例では相互間でミスがあった。

- 向←→何

そして、色彩表現の抽出にあたって直接的なダメージを受けたのは、次のような例であった。

- 赤→亦
- 黒→無
- 白→自
- 緑→縁

また、機械ならではのミスもある。

- 。」→ピ
  (句点とかぎ括弧を1字とみなしている)
- 早→♀

以上のミスは、いずれも**字形の相似**が原因であると考えられる。ここに挙げたような、OCRの間違いの**法則性**を見い出すことができるようになってくれば、それらをエディタの置換機能で処理することにより、作業の能率化を図ることもできるだろう。

しかし、私の経験から言うと、芥川の歴史小説のように、漢字の多い作品の場合は、極端な場合、マニュアル入力を行ったほうが早いということもあった。ただし、入力に使用した書籍の状態やゼロックスコピーの状態に左右されることも確かである。全集のように活字のポイントが小さいものよりも、文庫本のように活字のポイントが大きいものを使用することにしたり、拡大コピーしたものを使用することで、ある程度の精度を上げることはできるだろう。しかし、国文学では、いわゆる「底本」に何を使用するかということが重要であり、通常、文庫本を底本にすることはない。また、拡大コピーを行った結果、文字がかすれる等の弊害も同時に起こっていた。

1.2.3 漢字の問題

作品中には、第二水準までに含まれていない漢字も使用されている。原典を忠実に再現するのならば、外字の作成も考えられるが、そうすれば、汎用性はなくなってしまう。特定の機種やソフト上でしか使用できないデータベースということになってしまうのである。よって、漢字については、第二水準の範囲内で代用することにする。

2 テキストデータの実例

テキストデータ化にあたっては、『筑摩全集類聚』の芥川龍之介全集を使用した(昭和46年初版発行)。全8巻・別巻1であるが、テキストデータ化の対象としたのは、1巻から4巻までの「小説」148編と、『侏儒の言葉』を加えた合計149編である。

『筑摩全集類聚』を原典とした理由は、詳細な脚注が付いていることと、第8巻に総索引が付いているという2点である。現時点でのテキストデータは、作品の本文のみを機械可読化状態とすることを目標としているが、後述するように将来的にはさらに発展させ、ルビや注を付けることなども構想にあるからである。

たとえば、『鼻』をテキストデータ化すると、次のような形になる。

```
ha01u06　禅智内供の鼻と云へば、池の尾で知らない者はな
ha01u07 い。長さは五六寸あつて上唇の上から顋の下まで下
ha01u08 つてゐる。形は元も先も同じやうに太い。云はば細
ha01u09 長い腸詰めのやうな物が、ぶらりと顔のまん中から
ha01u10 ぶら下がつてゐるのである。
```

各行頭の7カラムの英数字(1バイト文字)は、「**情報行**」を表している。それぞれは、

- ha—作品名の記号。
- 01—作品のページ数。各作品ごとに1ページから始まる。
- u—全集は2段組になっているので、上段はu、下段はdで表す。
- 06—行数。上段・下段ごとに1行目から始まる。

という内容を表している。この情報行があることにより、用例検索などを行った場合など、どの作品中のどの部分から抽出されたのかということがわかるのである。

白	面白い　明白　関白　表白　潔白 余白　白状　目白押し　告白　白日 敬白　淡白　白々しい　白痴　白ける 独白　飛白　腕白　白連　白銅
赤	赤子　赤さん　赤児　赤裸　赤裸々 赤坂 赤帽　赤穂　赤の他人 真赤な嘘　赤松　「赤と黒」
青	青年　刺青　青酸加里　青山 青侍　青女房
黄	黄昏　黄泉　硫黄
黒	黒板　黒檀　黒門　大黒　黒子　黒焼き

表 1: 用例検索中に抽出されたノイズの例

3　用例の抽出

3.1　抽出方法

次に作成したデータからの色彩表現の用例抽出を行った。

色彩表現の抽出に当たっては、該当行だけでは内容が判断できないので、ある程度の文脈を付加するという目的で、色彩表現を含む前後 2 行の文を取り出せるプログラムを Pascal[6]で作成し、用例の抽出を行った。

たとえば「白」を含む用例は次のような形で抽出される。

```
白
hy01d03
のつきる所から、土手の上をずつと向う迄、煤けた、
うす白いものが、重さうにつゞいてゐるのは、丁度、
今が盛りの桜である。言問の桟橋には、和船やボー
```

3.2　抽出にあたっての問題点

3.1 の方法では「白」という文字列を含むもの全てを抽出してしまう。その結果、たとえば、表 1のような「色彩」ではない語までが抽出されてしまう。

このようなノイズを排除するためには、3 つの抽出基準を設け、色彩表現のみを抽出した。[7]

[6]TURBO PASCAL(ver5.0) (株式会社ボーランドジャパン)

[7]詳細はここでは述べないが、「色彩を感じることができるか」「他の機能での表現が可能か」「文脈から判断する」の 3 点である。

4　テキストデータの充実

4.1　「注」の問題

機械可読化の状態にしたテキストは、当然ながら印刷された書物とは違う。

たとえば、『筑摩全集類聚』では、本文の他に、最下段には「注」がある。

今回、私がテキストデータ化の対象としたのは、本文の部分のみであるが、研究に利用する際には、「注」は必要なものである。

現時点では、書籍と対照させながら、使用しているが、次段階では「注」をテキストデータの中に組み入れることが必要となる。

4.2　「ルビ」の問題

また、同時に「ルビ」の問題がある。「ルビ」とは、言うまでもなく「ふりがな」のことである。しかし「ルビ」にも、2 種類あり、

1. 「読み方」としてのルビ
2. 意図的に読ませたい場合に付けるルビ

1. は、小学生ならば「薔薇」という漢字を学習していないから、読むことができないだろうという仮定のもとに、「薔薇(ばら)」のようにルビを振る場合である。

一方 2. の場合は、「希望」を「のぞみ」と読ませたり、「予定」を「スケジュール」と読ませたりするものである。2. の場合は、特に作家の創作意識を知ることができる、貴重な情報である。

前述の「注」や「ルビ」も、テキストデータとは独立させて、別個のデータベース、たとえば「芥川龍之介の用字法データベース」を作成し、本文データベースと相互参照することができれば、さらに充実を図ることができる。

4.3　SGML 方式の利用

現在は主に「2 テキストデータの実例」で挙げたような形で使用しているが、「注」や「ルビ」の問題を解消する為にも、今後は SGML（Standard Generalized Markup Language）方式による利用を検討していきたい。

5 まとめ

最後に、パソコン使用の長所と短所について次のようにまとめてみる。

- 短　　所
  1. テキストデータの作成に時間がかかる
  2. 文の意味までは判断できない
- 長　　所
  1. データの加工が容易である
  2. 膨大な量のデータも迅速かつ正確に扱える
  3. 一旦決定した規則・法則が忠実に適用される

特に、「文の意味が判断できない」という点については、ノイズまで抽出してしまうという実例がその短所性を物語っている。

しかし、これも表裏一体であって、長所にもなりうるのである。

つまり、長所の3. を忠実に適用した結果、人間が手作業で抽出した場合ならば、予め削除するような用例まで抽出してしまったのである。

人間が行った場合、その時々で判断が違ったり、見落としてしまう危険性もある。

それならば、ノイズをも含めた抽出結果から、必要な用例のみを選択したほうが、データの信頼度は高い。

また、長所1. の「データの加工」であるが、今回のように、テキストデータから「色彩表現」だけを抽出して、新たなデータベースを作成するのも加工であるし、私は、特に『鼻』『羅生門』『芋粥』の3作品については、fixseg[8]を使用して、品詞分類を行った。つまり、この3作品については、それぞれ

1. 加工を加えていないプレーンテキスト
2. 「分かち書き単位」のデータ
3. Fixseg により形態素解析した「形態素単位」のデータ

が存在するのである。1. のプレーンテキストは全ての作業の基本となるものであり、2.3. はそれに加工を加えたテキストデータである。2. の「分かち書き単位」のデータは、単に区切り記号を与えただけであるが、3. の「形態素単位」のデータには品詞番号も付与されている。

現時点での私の興味は芥川の「色彩表現」であるが、一旦テキストデータを作成しておけば、たとえば、「森」にはどのような色彩表現が使われているのかということを調べたくなれば、容易に検索することができる。また、興味の対象が色彩表現から動詞の用法や、擬音語・擬態語などに移ったとしても、利用することが可能である。むしろ、その時にこそテキストデータの利点が活かされるのである。仮に手作業で「情報カード」などに「色彩表現」の用例を書き出していたとすれば、また一から同じ作業を繰り返さなければならないのである。

また、芥川の研究者にとってはもちろんのこと、国語学や日本語学の分野においても、用例抽出の際に利用することも可能である。

以上のように、**他人との共有性・将来性**ということを考えれば、日本文学作品を研究する上において、テキストデータを作成し、利用する方法は、ぜひとも必要なものである。

[8] 筑波大学の荻野綱男氏が開発した形態素解析プログラム。ワープロカナ漢字変換用辞書 FIXSER を利用している。

参考文献

参照した文献の中で主だったもののみ以下に列記する。

[1] 青木　シゲル，『パソコンデータ活用法－ MS-DOS のテキストファイル互換法』，産能 大学出版部，平成4年

[2] 荒股　宏，『データベース夜明け前』，株式会社ジャストシステム，1992

[3] 石田　晴久他，『ワープロと日本語処理』，共立出版，1985

[4] 伊藤　鉄也，新・文学資料整理術『パソコン奮闘記』，桜楓社，昭和61年

[5] 荻野　綱男，『ワープロによる知的生産の方法』，岩波書店，1989

[6] DB-West（西日本国語国文学データベース研究会），『パソコン国語国文学』，啓文社，1995

[7] 三橋　一夫，『新・日本語入力術』，JICC 出版局，1988

雑誌特集号

[8] 『月刊しにか』2 月号「特集・古典とコンピュータ」，大修館書店，1992

[9] 『國語學』第 178 集「テーマ別研究・電子化テキストの国際的共有」，國語學會，平成 6 年

[10] 『人文学と情報処理』NO.1「特集　コンピュータ利用の現在」，勉誠社，1993

[11] 『日本語学』8 月号「特集　新しいデータ・新しい研究」，明治書院，1991

論 文

[12] 上村　和美，「芥川龍之介作品のテキストデータベース化とその利用方法について」，大阪大学大学院言語文化研究科修士論文，1991

[13] 上村　和美，「テキストデータベースによる色彩表現の研究－芥川龍之介作品への適用－」，大阪大学大学院言語文化研究科博士論文, 1994

[14] 荻野　綱男，「コンピュータと言語」，『言語』4 月号，大修館書店，1979

[15] 郡司　隆男，「日本語入力と編集」『日本語の特性と機械翻訳』，出版科学研究所，昭和 62 年