『源氏物語大成』のフルテキストデータベース

上田裕一　○上田英代
樺島忠夫　村上征勝

The Full-Text Database of *Genji Monogatari Taisei*

Y. Ueda　○H. Ueda
T. Kabashima
M. Murakami

The quantitative analysis of sentences, the study of patterns formed in the process of linguistic encoding of information, has been applied to manyimportant documents in foreign countries. However, it was first applied to Japanese documents only in the middle decades of the 20th century. The mainreason for this delay is the following characteristic of the Japanese language.

Japanese words are not separated by spaces as in English. Thus it is difficult for the computer to recognize word boundaries.

The purpose of this study is to build a useful full-text database of Genji Monogatari for use in quantitative analysis. Using the *Genji Monogatari Taisei* published by Chuokoron-sha as a textbook, we divided all the sentences of Genji Monogatari into words to which were attached codes for parts of speech.

In this paper we report how to build such a database and what difficulties we encountered in this process.

1．はじめに

統計的手法を用いた文献の計量分析学は、諸外国では古くから行なわれているが、日本語文献の場合には、まず分かち書きをしてからデータとしなければならないために歴史は浅い。筆者等は、計量国語学的視点から「源氏物語」の特徴を明らかにするために、『源氏物語大成』（中央公論社）を単語分割し、品詞情報付きフルテキストデータベースとして作成を試みた。１９９０年６月には『源氏物語大成』の本文入力を終え、１９９１年に自動単語分割のプログラムを完成し、１９９２年夏に５４帖すべてを単語分割して、１２月に自動品詞付けを終了した。現在はそのデータの細かい修正を行いつつ、統計的手法を用いて多方面からデータ解析を行っている。ここでは、テキストの選定から本文入力、自動単語分割、自動品詞付けの作業過程について報告し、いくつかの問題点とその解決方法について述べる。

2．テキストの選定と本文入力

本文の入力方法についてはいくつか考えられる。影印本を画像認識させてそのまま画面に出し、タグを着けていく方法、活字文献をＯＣＲ（Ｏｐｔｉｃａｌ　Ｃｈａｒａｃｔｅ

ｒ　Ｒｅａｄｅｒ）等で読み込み機械可読化してゆく方法などが一般的である。どの方法を用いるかは、データの使用目的によって決まる。

本研究では、①人間が行うと単位認定に揺れが生じやすい単語分割をできるだけ機械化し、正確な分割が敏速に行える方法を開発する、②「源氏物語」を、文の長さ、単語の量、意味の面からのみ計量分析するのではなく、各種の文法的側面、即ち品詞の頻度や接続関係の面からも分析するため、すべての単語に自動的に品詞コードを付ける方法を開発する、③「源氏物語」だけでなく他作品との比較を行うため、機械化できる作業はできるだけ取り入れ、計量分析できるデータにしていく、などの目的があったため、活字文献をＯＣＲで読み込むこととした。

「源氏物語」の活字文献は多種類あるが、本研究ではテキストとして『源氏物語大成』を選んだ。その理由は、校異が精密であり、ひらがな表記が多く、ルビ行がないこと、語彙索引が完備している事による。本文の入力は、まずＯＣＲ（富士電気ＸＰ－５０Ｓ）で読みとり、手作業で修正を行なった。読みとり作業は延べ２３時間ほどであった。修正作業の主なものは、用いたＯＣＲが漢字第二水準に対応していないためその漢字を修正入力する事、踊り字を入れる事、繰り返し記号にも字数分の記号を入れることなどである。

３．自動単語分割

まず「源氏物語」の１～８巻までを手作業で単語分割し、異なり単語を集めて単語集をつくり、分割用辞書とした。次に『日本古典文学大系』を参考にして、テキストに句点を入れ、分割用辞書を用いて自動単語分割し、手作業による分割との同一性を確認した。分割結果に多少のズレはあるものの、プログラム上は問題ない事がわかった。

１９９１年９月、『フロッピー版古典対照語い表』が入手できたので、その中の「源氏物語」使用語彙を分割用辞書に加えることにした。『古典対照語い表』の「源氏物語」の使用語彙は、『源氏物語大成総索引』より採られているので好都合であった。この『語い表』は、自立語のみ所収で、活用する語は終止形だけ載っている。見出し語は、濁音、半濁音を含んですべてひらがな表記となっている。『大成』の本文は、濁音、半濁音がないのでそれらをすべて清音に直した。『古典対照語い表』の中の「源氏物語」使用語彙は、１１４２１語であった。

次に、活用する語のすべてに活用形をつけ、手作業分割による１～８巻までの単語集と、『古典対照語い表』を合成した辞書を作成し、一巻毎に自動分割していった。合成辞書による分割結果が、＜図１＞である。

<=いつれ-><=の-><=御時-><=に-><=か->。<=女御-><=更衣-><=あまた-><=さふらひ-><=給-><=ける-><=なかに-><=いとやむことなき-><=き-><=はに-><=は-><=あら-><=ぬか-><=すくれ-><=て-><=時めき-><=給-><=ありけ-><=り->。<=はしめ-><=より-><=我-><=は-><=と-><=思あかり-><=給へる-><=御方〻〻めさましき-><=もの-><=に-><=おとしめ-><=そねみ-><=給->。<=おなし-><=ほと-><=それ-><=より-><=下らう-><=の-><=更衣-><=たちは-><=まして-><=やすからす->。<=あさゆふ-><=の-><=宮つかへ-><=に-><=つけて-><=も-><=人-><=の-><=心を-><=のみ-><=うこかし-><=うらみ-><=を-><=おふ-><=つもり-><=に-><=やありけ-><=む-><=いと-><=あつしく-><=なり-><=ゆき-><=もの心ほそけに-><=さとかち-><=なる-><=を-><=いよ〻〻-><=あかす-><=あはれなる-><=物-><=に-><=おもほし-><=て-><=人-><=の-><=そしり-><=を-><=も-><=え-><=は-><=ゞ<=から-><=せ-><=給はす-><=世-><=の-><=ためし-><=に-><=も-><=なり-><=ぬ-><=へき-><=御もてなし-><=也->。<=かん

＜図１＞

＜＝　－＞で区切られた部分が辞書の単語とマッチし分割された単語である。このプログラムは、長い単語から先に区切ってゆき、一度＜＝　－＞で区切られた後は、その中は区切らないという仕組みになっている。桐壺の巻を自動分割するのにＵＮＩＸマシンのＮＥＷＳで１時間２７分かかった。この自動分割プログラムでは、一文字或いは二文字の単語分割は不正確なので、自動分割の後、手で修正した。手修正で正確に分割された巻で異なり単語集をつくり、元の分割用辞書にない単語を追加していった。固有名詞や初出単語などが増加してゆくわけである。一巻ごとの辞書用単語の元辞書への追加は、自動分割の正確さを増してゆくことになった。最終的に合成辞書による自動分割で、「夢の浮橋」の巻は、８０％の正確さであった。今後、単語の前後関係から判断して分割箇所を認定するプログラムの開発がのぞまれる。

４．自動品詞付け

正確に分割されたテキストに、自動的に品詞付けをしてゆくために、まず品詞付き辞書を作成する。ここでも『古典対照語い表』を利用した。「源氏物語」使用単語を品詞付きで取り出し、活用語のすべての活用形を加えて元辞書とする。同音異義語で同一品詞のものは一語だけ採り、異なる品詞のものは一つの語に可能性のある品詞をすべてつけ、複数の品詞を付けた多品詞語とした。この辞書を使って、巻１から自動品詞付けを行なった結果が＜図２＞である。元辞書中に該当単語が収録されていないときは、その単語には、品詞が付かない。この出力結果に、複数品詞が付いてる単語は適切な品詞を選択し、品詞無しのものには品詞を付けた。＜図３＞　このようにして一巻が完成すると自動単語分割の時と同様に、その巻の品詞付き異なり語集をつくり元辞書に新異なり単語を順次加えていった。一巻終わるごとに異なり語が増えてゆくことになる。次の巻を追加済み辞書で、自動品詞付けを行なう。修正の後、再びその巻の異なり語集を作り新異なり語を元辞書に加え、次の巻の自動品詞付けを行なう。この作業を５４帖続ける。

/いつれ〔　　〕/の〔助詞〕〔名詞〕/御時〔　　〕/に〔助詞〕〔助動〕〔動詞〕〔名詞〕/
か〔助詞〕〔代名〕〔名詞〕/。/女御〔　　〕/更衣〔　　〕/あまた〔副詞〕/
さふらひ〔助動〕〔動詞〕〔名詞〕/給〔助動〕/ける〔助動〕/なか〔動詞〕〔名詞〕/
に〔助詞〕〔助動〕〔動詞〕〔名詞〕/いと〔副詞〕〔名詞〕/やむことなき〔形容〕/
きは〔名詞〕/に〔助詞〕〔助動〕〔動詞〕〔名詞〕/は〔助詞〕〔助動〕〔名詞〕/
あら〔動詞〕/ぬ〔助動〕〔動詞〕/か〔助詞〕〔代名〕〔名詞〕/すくれ〔動詞〕/
て〔助詞〕〔助動〕〔名詞〕/時めき〔　　〕/給〔　　〕/あり〔動詞〕〔名詞〕/けり〔助動〕/。

＜図２＞

いつれ〔代名〕/の〔助詞〕/御時〔名詞〕/に〔助詞〕/か〔助詞〕/。/女御〔名詞〕/更衣〔名詞〕/
あまた〔名詞〕/さふらひ〔動詞〕/給〔助動〕/ける〔助動〕/なか〔名詞〕/に〔助詞〕/
いと〔副詞〕/やむことなき〔形容〕/きは〔名詞〕/に〔助詞〕/は〔助詞〕/あら〔動詞〕/
ぬ〔助動〕/か〔助詞〕/すくれ〔動詞〕/て〔助詞〕/時めき〔動詞〕/給〔助動〕/あり〔動詞〕/
けり〔助動〕/。/はしめ〔名詞〕/より〔助詞〕/我〔代名〕/は〔助詞〕/と〔助詞〕/
思あかり〔動詞〕/給へ〔助動〕/る〔助動〕/御方¥〔名詞〕/めさましき〔形容〕/
もの〔名詞〕/に〔助詞〕/おとしめ〔動詞〕/そねみ〔動詞〕/給〔助動〕/。

＜図３＞

5．『源氏物語大成』と同じページと行を付ける

検索作業を容易にするために、品詞情報付きデータに自動的に『源氏物語大成』と同じページと行番号をつけた。単語分割する前で、『大成』と行対応しているテキストの行末5文字を品詞付きテキストでサーチして、そこに改行マークをいれ、ページと行番号を付ける。＜図4＞

```
0005-01
いつれ〔代名〕/の〔助詞〕/御時〔名詞〕/に〔助詞〕/か〔助詞〕/。/女御〔名詞〕/更衣〔名詞〕/
あまた〔副詞〕/さふらひ〔動詞〕/給〔動敬〕/ける〔助動〕/なか〔名詞〕/に〔助詞〕/いと〔副詞〕
/やむことなき〔形容〕/きは
0005-02
〔名詞〕/に〔助詞〕/は〔助詞〕/あら〔動詞〕/ぬ〔助動〕/か〔助詞〕/すくれ〔動詞〕/て〔助詞〕/
時めき〔動詞〕/給〔動敬〕/あり〔動詞〕/けり〔助動〕/。/はしめ〔名詞〕/より〔助詞〕/我〔代
名〕/は〔助詞〕/と〔助詞〕/思あかり〔動詞〕/給へ〔動敬〕/る〔助動〕/御方
0005-03
￥〔名詞〕/めさましき〔形容〕/もの〔名詞〕/に〔助詞〕/おとしめ〔動詞〕/そねみ〔動詞〕/給
〔動敬〕/。/おなし〔形容〕/ほと〔名詞〕/それ〔代名〕/より〔助詞〕/下らう〔名詞〕/の〔助詞〕
/更衣たち〔名詞〕
```

＜図4＞

6．今後のデータベースの利用

今回、『源氏物語大成』の品詞情報付きフルテキストデータベースを完成したことによって得られる成果は、計り知れない。成立過程に関する諸説や複数作家説や物語音読論等々の詳細な検討が、文法的側面からも、使用単語の面からも行えるし、「源氏物語」の文体を構成する諸々の要素についても一つ一つ検証してゆくことができる。現在このデータベースから得られた結果の主なものは、次の項目についてである。

イ、単語の長さのヒストグラム
ロ、各巻の文の長さと文の数
ハ、単語の前後関係
ニ、ある品詞が文頭、文末にある割合
ホ、単語別の度数とヒストグラム
ヘ、各品詞別、度数付き単語集
ト、各品詞の度数と出現率
チ、品詞の相対出現率
リ、単語の出現回数
ヌ、各品詞ごとの接続関係。

同様に、「紫式部日記」や各種の文献のデータベースを作成し、データ解析と利用方法の開発を同時に行なってゆく予定である。