論文

日本語—英語対照「源氏物語」のテキスト・データベースの作成に関する基礎的研究†

長瀬 真理††

本論文は、この程2年がかりで完成した紫式部著の「源氏物語」の日本語版(漢字仮名まじり文)と英語版(E.G. サイデンステッカー訳 "Tale of Genji")の対照テキスト・データベース・プロジェクトの成果報告である。東京大学大型計算機センターでの公開と、オックスフォード大学電算機センターでの供託サービスを目前に控え、入力を始めとする、具体的な方法論や制作過程を説明するとともに、出来上がったテキスト・データベースの利用や研究動向を紹介する。同時にコピーライトやサービス体制等、テキスト・データベース作成のかかえる様々な問題点を検討する。

1. はじめに

電算機の文字処理能力の高度化と容量の増加と共に、人文系の学問研究におけるコンピュータの利用は近年増加の一途を辿っている。とりわけ、研究の基礎となるテキストや文献・資料の全文を入力するテキスト・データベース[注1]の開発が活発になっている。既にイギリス、アメリカ[注2]を始め、ドイツ、フランス、イタリアおよび北欧諸国は、積極的に大量の文献を機械可読な形にしている。完成したテキスト・データベースは、国際的なネットワークを利用してアクセスできるものから、磁気テープやフロッピィ・ディスクなど様々な形で提供されており、自国の研究者のみならず広く国外の研究者にも安価にサービスしている。そのほか商業ベースでのテキスト・データベースの販売も始まり、シェイクスピア全集のフロッピィ版も登場した。

わが国でも海外で作成されたテキスト・データベースを各自の専門研究に利用する研究者が現れている。又、国産のテキスト・データベースも作られつつある[注3]。しかしその数は上記の各国に比して、非常に少なく、近年輸入超過の非難も聞かれるようになった。

このような状況に鑑み、1988年秋、紫式部著の日本語版「源氏物語」と英語版の対照テキスト・データベースを作成するプロジェクトが企画された。国内の公開サービスは東京大学大型電算機センターに依頼した。しかし残念ながら日本には、国際的なサービスを行なう機関が無いため、オックスフォード大学電算機サービス(OUCS:Oxford University Computing Service)に供託サービスをお願いすることにした。OUCSは既に長年に亙って英国内ばかりでなく世界中の研究者にデータベースをサービスしている。

特にOUCS内部にはテキスト・データベースを専門に扱うものとしてOTA(Oxford Text Archive)があり、既に25ヶ国920冊の機械可読テキストを世界各国の研究者に提供している。しかしこれまでは日本語のテキストは供託されておらず「源氏物語」が最初のものとなる。

いずれにせよ世界中の研究者が「源氏物語」のテキスト・データベースの恩恵に預かることが出来るようになる訳で、遅蒔きながらこの分野での国際的な学術交流へ多少とも寄与出来るようになった。

本論文では、このパイロット・プロジェクトの作成経過並びに試用結果を中心に、テキストの選定、入力、作業経過、コピーライト、サービス、出来上ったテキスト・データベースの利用、問題点などを順次項目に添って解説し、今後のテキスト・データベースの展望と実用性を検討する。

† Project to develop machine-readable texts of English and Japanese versions of "The Tale of Genji" by Mari Nagase

†† 東京女子大学情報処理センター

2. テキストの選定と入力方法

2.1 何故「源氏物語」が選ばれたか？

「源氏物語」が選ばれた最大の理由は、この作品が我が国が誇る第一級の文学として、日本文化・日本語に大きな影響を与えたばかりでなく、海外の文化にも多大な貢献をしており、内外での研究者の数も非常に多く、データベースの需要が大きいと考えられたからである。

例えば、単語の数だけを見ても「源氏物語」は平安朝の物語の中で一番多く、1万2千語の語彙がある。これは万葉集のおよそ五倍の言語量に相当し、その後の日本語・日本文学に計り知れない影響を与えた。

海外でも、複数の英訳を始め、仏訳、独訳、中国語訳、スウェーデン語訳、オランダ語訳、イタリア語訳、フィンランド語訳など各国語に翻訳され、国際的な研究集会も開かれている。

そのため、今回のプロジェクトでは日本語のテキスト・データベースのみならず、英語版のテキストのテキスト・データベースの作成も同時に企画された。しかも単に二つのテキスト・データベースを作成するだけでなく、日本では初めての試みとして、両テキストにクロス・レファレンスを付け、相互に参照可能な形式のデータベースの構築が検討された。

2.2 日本語・英語テキストの選定

いうまでもなく、「源氏物語」は日本で最も著名な古典文学作品であり、青表紙本をはじめ、河内本、別本等、いくつかの権威ある写本があるばかりでなく、源氏・河海抄を始め注釈書も鎌倉時代からあり、その数は膨大である。又、校訂本として、出版されているものも多い。

今回は入力に印刷用文字の光学読み取り装置を使用するため、原本としては出版されたテキストが使われることになった。検討の結果、版も重ね、専門家より高い評価を得ている小学館の古典文学体系に収められている「源氏物語」が選ばれた。編集は源氏研究者として高名な阿部秋生、秋山虔、並びに今井源衛の三氏である。

一方、英文翻訳に関してはアーサー・ウェイリーのものが古くから知られているが、今回の入力にはE.G. サイデンステッカーのテキストを選んだ。

ウェイリーは20世紀初頭未だ研究のない時期に、ヴァージニア・ウルフ、E.M. フォースターや、ケインズのグループ、いわゆるブルームスベリー・グループが使った優雅な英語で翻訳を出版した。広範な読者を獲得し、この優れた訳によって、「源氏物語」が20世紀の批判にも耐える古典である、との高い評価が確立した。しかし残念なことに「すずむし」の巻が抜けていたり、意訳や省略が非常に多い。

サイデンステッカーの訳は、戦後の研究も踏まえた原作に忠実なもので、その意味ではウェイリー訳よりすぐれているとの評価を得ている。他に平安朝文学では「かげろう日記」を、近代文学では川端康成ほか多くの作品の翻訳を手がけている[1]。同書は1982年にペンギンブックス廉価版となり入手しやすくなっている。

2.3 入力方法

入力については、これまではテキスト・データベースに限らず、文字を主体にした情報のデータベース構築は手入力が主流であった。しかし近年、印刷漢字用の光学読み取り装置の開発もかなり進んでいることから、その実験も兼ねて機械入力の方式をとった。

日本語については、現代語の読み取りでは既に90%以上の読み取り率を誇る富士電気総設の富士OCR システム[2]を採用し、入力を委託した。文字認識のアルゴリズムは原稿をイメージ・データとして取り込み、文字線の境界に波を伝播させ、その波頭をから文字の高次の特徴を抽出するという手法を採用している。

英語版の入力は、既に各国で使われ定評のあるKurzweil 4000を使用した。機械入力の有効性等については、4. の校正及び7. の問題点の検討のところで詳しく検討する。

3. 検索ソフトと入力

ところで入力に入る前に、出来上がったテキスト・データベースの解析に利用する検索ソフトについて予め検討しておかなければならない。

3.1 検索ソフトについて

折角テキスト・データベースを作成しても検索ソフトがなければ、利用の効果は余り期待できない。今回のプロジェクトではソフト開発は含まれていないため、既存の検索ソフトの利用を想定しなければならない。

都合の良いことに、供託先のOUCSでは様々なテキスト・データベースを作成すると同時に、政府からの基金援助を得て、パッケージ・プログラム

の開発も盛んに行なっている。そのなかに、オックスフォード・コンコルダンス・プログラム (Oxford Concordance Program:略 OCP) と呼ばれる文章解析プログラムがあり、供託されたテキスト・データベースの整備やスペル・チェクにも利用されている。国内で公開の引き受け先である東京大学大型計算機センターにも、1987年より導入されている。

又、パソコン版 Micro-OCP[注4)] が作られ、日本語の文章解析も出来る為、今回のプロジェクトのように日本語・英語のテキスト・データベースの処理には最適なソフトであると考えられる事から、OCP が解析プログラムとして採用された。

一般に多くの機械翻訳や文章解析用ソフトは、テキスト・データの分かち書き、キーワードや読みの付与、形態素分析を前処理として必要とする。しかし Micro-OCP の場合はその必要がなく、ベタ書きのテキスト・データをそのまま処理できるという利点がある。

3.2 入力形式

入力形式については固定長形式を始め、開始符号形式など様々な方法があるが、英文、和文共に検索システムとして OUCS が開発した Micro-OCP を使用を予定しているため、COCOA 形式を採用した。この形式は、OCP が最も容易に取扱うテキスト・データ形式で、アトラス研究所時代から使われている。ちなみに COCOA の名前は word COunt COncordance generation on Atlas の大文字部分から採られている。現在では機械可読テキストの入力形式として広く採用されている。

3.3 入力仕様

3.3.1 COCOA 形式

COCOA 形式では、処理対象となる本文と別に、表題、ページ、版、章、著者、登場人物などの情報をあらかじめ参照部として宣言しておく。そうすることにより、版、ページ、登場人物によって検索したり、並べ変えることが可能になる。又、インデックスや索引作りにも利用できる。

参照部以外は、ほとんどオリジナルの本の形式と変わらないという利点がある。

以下表 1-a,b に入力された日本語及び英語の参照部の定義、実際の宣言、冒頭部を示し解説する。

表 1-a 日本語参照部の定義

```
<W 紫　式部> Writer's name
<T 源氏物語> Title's name
<V No.> Japanese volume(1-6)
<C きりつぼ> Japanese Title of Chapter
<E No.> English Chapter(1-54)
<S No.> Japanese subheading
<F No.> Page references of the English texts
hbox to 4em (英文箇所に対応)
<P No.> Page references of the Japanese texts
<A 帝> Names of Speaker(Actor or Actress)
```

表 1-b 入力された冒頭部

```
<W 紫式部>
<T 源氏物語>
<V 1>
<C きりつぼ>
<E 1>{The Paulownia Court}
<S 1> { 桐壺更衣に帝の御おぼえまばゆし }
<F 3>
<P 93>
いづれの御時にか、女御更衣あまたさぶら
ひたまひける中に、いとやむごとなき際に
はあらぬが、すぐれて時めきたまふありけ
り。はじめより我はと思ひあがりたまへる御方々、めざまし
きものにおとしめそねみたまふ。同じほど、それより下臈の
更衣たちは、ましてやすからず。朝夕の宮仕につけても、人
の心をのみ動かし、恨みを負ふつもりにやありけん、いとあ
つしくなりゆき、もの心細げに里がちなるを、いよいよあか
ずあはれなるものに思ほして、人のそしりをもえ憚らせたま
はず、世の例にもなりぬべき御もてなしなり。上達部上人な
ども、あいなく目を側めつつ、いとまばゆき人の御おぼえな
り。唐土にも、かかる事の起りにこそ、世も乱れあしかりけ
<P 94>
れと、やうやう、天の下にも、あぢきなう人のもてなやみぐ
さになりて、楊貴妃の例も引き出でつべくなりゆくに、いと
はしたなきこと多かれど、かたじけなき御心ばへのたぐひな
きを頼みにてまじらひたまふ。
　父の大納言は亡くなりて、母北の方なむ、いにしへの人の
よしあるにて、親うち具し、さしあたりて世のおぼえはなや
かなる御方々にもいたう劣らず、何ごとの儀式をももてなし
たまひけれど、取りたてて、はかばかしき後見しなければ、
事ある時は、なほ拠りどころなく心細げなり。
<S 2> { 更衣に皇子誕生、方々の憎しみつのる }
<F 3>
前の世にも、御契りや深かりけん、世にな
くきよらなる玉の男皇子さへ生まれたまひ
ぬ。いつしかと心もとながらせたまひて、
急ぎ参らせて御覧ずるに、めづらかなるちごの御容貌なり。
　一の皇子は、右大臣の女御の御腹にて、寄せ重く、疑ひな
きまうけの君と、世にもてかしづききこゆれど、この御にほ
(中略)
<P 98>
そと、心づかひして、皇子をば止めたてまつりて、忍びてぞ
出でたまふ。
　限りあれば、さのみもえ止めさせたまはず、御覧じだに送
らぬおぼつかなさを、言ふ方なく思ほさる。いとにほひやか
に、うつくしげなる人の、いたう面痩せて、いとあはれとも
のを思ひしみながら、言に出でても聞こえやらず、あるかな
きかに消え入りつつ、ものしたまふを、御覧ずるに、来し方
行く末思しめされず、よろづのことを、泣く泣く契りのたま
はすれど、御答へもえ聞こえたまはず。まみなどもいとたゆ
げにて、いとどなよなよと、われかの気色にて臥したれば、
いかさまにと思しめしまどはる。敲戸の宣旨などのたまはせ
ても、また入らせたまひて、さらにえゆるさせたまはず。
<A 帝>「限りあらむ道にも、後れ先立たじと、契らせたまひけるを
さりともうち棄てては、え行きやらじ」<A>とのたまはするを、
女もいといみじと見たてまつりて、
```

日本語の原本は縦書きで、上段に語彙の註、中段に本文、下段は現代語訳と三段に分かれている。又、所々に挿絵が挿入されている。

テキスト・データベースでは本文のみが横書きで入力され、註、現代語訳、挿絵は除かれた。又本文中に現れる註番号や振り仮名も取り除かれた。なお行の切れ目は原本と同じである。

参照部として今回採用された情報は、作者名(W)、表題(T)、巻数(V)、章(C)、英語の章(E)、小見出し(S)、小見出しに対応する英文ページ(F)、日本文ページ(P)、登場人物(A)である。これらの情報はカテゴリーと呼ばれ、それぞれの後ろの()内に示されたようなアルファベットの一文字変数として鍵括弧(<>)でくくって宣言される。

本の場合と違って英文との対応を示す幾つかの情報が付加されている。先ず章に関しては、日本語も英語も各々54章に分かれているため相互に対応する。しかし日本語の章見出しには数字が無い。その為英文から日本文への参照を容易にするため、あえて英文の章番号を挿入した。更に細かいレベルでの対応については、英文には無いが、日本文では章が又いくつかに分けられ、番号順に小見出しが付けられている。この点に着目し、日本文の小見出しに対応する部分についてのみ英文のページを挿入することにした。その為小見出しの冒頭の部分は相互に対応させることが出来るようになった。

更に本では、総てではないが会話部分について、話者の情報が挿入されている。それを生かすために、登場人物として話者マークのカテゴリー<A>を導入した。例えば上記のテキストの98ページの後ろから3行目あたりでは、本の場合同様、帝の会話部分が「」で示されているが、更にその回りが<A 帝>と<A>で囲われている。このように参照部で宣言した話者マークを各話者ごとに挿入することにより、特定の話者の会話部分のみを処理対象に指定することが可能になった。

原本で挿入のような取扱になっている小見出しの数字に続くタイトル状の解説はでくくり、コメント(注釈)とした。又英文の各章のタイトルも注釈扱いとした。OCPは注釈を本文とみなさないが、こうしておけば解析時に処理対象として宣言することも、あるいは逆に削除することも容易である。(表2-a,bを参照)

英文の場合も、入力されたのは本文のみで、脚注や各ページの先頭中央のタイトル及び挿絵は省かれた。

参照部は、一見して明らかな通り日本語より短くなっている。両者の大きな違いは、日本語の解説時にも触れられたが、小見出しが無いことにある。もう一つ、英文では会話部分はダブルコーテーション(")で囲まれているが、一体誰が話しているか、と云った情報が原本にはない。その為話者マークのカテゴリーもない。

日本語と英語の対照として日本語の巻の情報を入れたのは、日本語は全部で6巻に分かれており、各巻1からページ付けがおこなわれていることによる。他方英文は2巻に分かれているが両者には通しページが付いている。

なお既に解説したように、日本文の小見出しの冒頭部のページが参照部として入力されている。番号順にはなっているが、小見出しに続く節が長い場合には、後半部の対応に大きなずれが生じている。

3.3.2　日本語の構成

日本語の本文を構成する基本文字は平仮名と漢字であり、句切り文字は句読点、括弧(「、」)(『、』)である。只、JISの第2水準にもない漢字が若干あり、その場合は類似のものを代入し、暫定文字としてその文字の後ろにアスタリスク(*)を挿入した。

次行に続く漢字語句の場合は行末にハイフンを挿入した。漢字のみとしたのは、日本語用のMicro-OCPでは平仮名は処理上句切り文字として扱われるため、たとえハイフンを挿入しても、コマンド・レベルで指定しない限り、平仮名の検索は一切しないからである。又現実問題として、日本語の表記法の解釈は様々で語彙をどこで区切るかは必ずしも専門家の間でも意見の一致がみられていない。むしろ各人の自由な解釈にまかせ、それぞれ表記法のもとで語彙の形が設定されるほうが望ましいと判断した。

ハイフンで一番困ったのは“御”の処理である。“御簾”、“御覧ず”、や“御息所”のように他の漢字と切り離せない場合もあれば、“御心”、“御気色”、“御消息”のように切り離して使われる場合もある。今回はいずれの場合も区別せず、2行に亙る時は機械的にハイフンを付ける処置をした。

3.3.3　英語の構成

英文の基本構成文字ははアルファベットであり、その他句切り文字として句読点及びコロンが使われている。会話箇所の開始と終了にはダブルコーテー

表 2-a 英語の参照部

<W Murasaki Shikibu> Writer's name
<T The Tale of Genji> Title's name
<K No.> Japanese volume(1-6)
<C No.> Chapter(1-54)
<N No.> Page references of the Japanese subheading（日本文に対応）
<P No.> Page references of the English translation

表 2-b 入力された冒頭部（英語）

<W Murasaki Shikibu>{Translated by E.G.Seidensticker}
<T The Tale of Genji>
<K 1>{Japanese Volume}
<C 1>{The Paulownia Court}
<N 1>{Japanese Sub-chapter}
<P 3>
In a certain reign there was a lady not of the first rank whom the emperor loved more than any of the others. The grand ladies with high ambitions thought her a presumptuous upstart, and lesser ladies were still more resentful. Everything she did offended someone. Probably aware of what was happening, she fell seriously ill and came to spend more time at home than at court. The emperor's pity and affection quite passed bounds. No longer caring what his ladies and courtiers might say, he behaved as if intent upon stirring gossip.

His court looked with very great misgiving upon what seemed a reckless infatuation. In China just such an unreasoning passion had been the undoing of an emperor and had spread turmoil through the land. As the resentment grew, the example of Yang Kuei-fei was the one most frequently cited against the lady.

She survived despite her troubles, with the help of an unprecedented bounty of love. Her father, a grand councillor, was no longer living. Her mother, an old-fashioned lady of good lineage, was determined that matters be no different for her than for ladies who with paternal support were making careers at court. The mother was attentive to the smallest detail of etiquette and deportment. Yet there was a limit to what she could do. The sad fact was that the girl was without strong backing, and each time a new incident arose she was next to defenseless.
<N 2>

It may have been because of a bond in a former life that she bore the
<P 4>
emperor a beautiful son, a jewel beyond compare. The emperor was in a fever of impatience to see the child, still with the mother's family; and when, on the earliest day possible, he was brought to court, he did indeed prove to be a most marvelous babe. The emperor's eldest son was the grandson of the Minister of the Right. The world assumed that with this powerful support he would one day be named crown prince; but the new child was far more beautiful. On public occasions the emperor continued to favor his eldest son. The new child was a private treasure, so to speak, on which to lavish uninhibited affection.

The mother was not of such a low rank as to attend upon the emperor's personal needs. In the general view she belonged to the upper classes. He insisted on having her always beside him, however, and on

ション(")、行にまたがる単語の継続記号としてハイフン(-)を、更にアポストロフィ(')が使われ、これらは原文とほぼ同様である。

その他、以下の特殊記号が使用されている。

音標文字:
- ティルデ(~) → 長母音(例:Rokujo~)
- ドルマーク($) → 鋭アクセント(ex.Ide$)
- パーセント(%) → 文音記号(ex. nai%ve)

暫定文字:
- アスタリスク(*) → 注
- プラス(+) → 詠み人知らず
- シャープ(#) → 異本
- 二重ハイフン(--) → ダッシュ
- アンダーバー(_) → イタリックス部分の始めと終り

3.3.4 ファイル構成及び容量

英文、日本文共に原文通り54ファイルとし、日本語の本の巻に合わせて2巻づつを組に、それぞれ3枚の2HDフロッピィに収録した。OUCSはIBMコンピュータの2DDまたは磁気テープの作成を主要としていたが、近年2HDが読める機械を設置した。

4. 校正

上記のように仕様を決定して入力を依頼し、出来上がったものから順次校正にとりかかった。結局英文・日本文共に予想以上に時間がかかり、最終的に一回しか出来なかった。英文に関しては、本の文字が小さいため、拡大コピィを利用し、出来上がったテキスト・データベースをスペル・チェッカー用のプログラムにかけてもらった。その結果、単語のスペル・ミスは少なくなった。しかし、大量の入力量であった為、重複箇所やインデントのミスが多かった。又句読点の取りこぼしもかなりの量であった。文字の間違いとしては数字の"0"(ゼロ)とアルファベットの"O"に、"1"(数字)を"l"(エル)や"I"(アイ)にする場合が多々みられた。

一番大変だったのは、イタリックス部分の開じまりと終わりに挿入するよう設定したアンダーバーの入力であった。もともと原本にない記号だった為、後から原本と照らし合わせながら手で入力しなければならなかった。

日本語文字読み取りのミスは英文より多く、代表的なものを以下にまとめる。

(誤)		(正)
ぽ	→	ぼ
ほ	→	ぼ
住	→	ま
位	→	泣
佳	→	袿
膿	→	濃
借	→	惜
厨	→	廚(5724)
賽	→	覧
來	→	萊
腸	→	帳
東	→	梟
タ(カタカナ)	→	夕

上記から明らかな通り、濁点の間違いや偏の間違いが多くみられる。既に手書き漢字でも、部首情報から漢字を推定する研究がでており、「偏」が解ると72%の単語や語の決定が出来る、との報告がある[3]。OCRの今後の改良も「偏」の読み取り率を上げることにより、効果が上がるのではないかと思われる。

その他のミスとしては、やはり英文同様、大量の入力量のため重複箇所が何ケ所かみつかった。細かい点としては、段落の頭の1文字スペース(全角)と、その他の部分の半角スペースの間違いや、話者マークと話者名の間の半角のスペース等、間隔のとりかたにミスが多く発見された。

話者マークは英語のアンダーバー同様、後から手入力したためミスが多かった。地の文で話者が明記してある場合は、それをそのまま重ねて話者マークとして採用したので、誤りはほとんどなかった。しかし会話や和歌のなかには、煩雑さを除くためにもしばしば話者の挿入が省かれている。こういった場合は内容を読んで、校正時に入力しなければならず、場所によっては話者の決定が困難なこともあり、校正する側の解釈によっては相違が生じることも多々あった。

なお日本語版において、JISの第1水準および第2水準にない漢字が11個あり、先に示したように、別な漢字を代用し後ろにアスタリスクを挿入した[注5]。

5. コピーライト

供託する場合、そのテキストのコピーライトの状態を詳しくOUCSに報告しなければならない。

なぜならテキスト・データベースの作成と同時に制作者にコピーライトが生じるからである。そのためOUCSがテキスト・データベースを再供託する場合、供託を受けた側がそれを無断で第三者にコピィすることは許されない。OUCSでは、厳しい指定条件を作成しており、テキスト・データベースの利用者は全員この基準内容を示した書類(Condition of Use)に署名しなければならない。書類は、供託者の権利を守ることを目的としており、テキスト・データベースが学術研究にのみ使用されることを保証し複製の禁止が明記されている。

OUCS自身が作成したテキストの大部分は古典であり、著者の死後50年以上経過したものがほとんどである。その意味では既にパブリック・ドメインになっており、本をテキスト・データベースにする段階でコピーライトの問題は生じていない。

しかし、日本では一般に出版社が作家の著作権を買い取ることが少ないことから、出版社の中には編集著作権の他に複製権を主張する場合が少なくない。そのため、校訂本の全文をOCR等で入力をするのであれば、出版社に事前に了解を得る必要がある。

いまのところ、ソフトでは色々問題が起こっているが、テキスト・データベースに関しては、コピーライトが問題になった例はない。外国ではこういったトラブル専門の職業としてLiterary Executerがある。

現在では機械可読されたテキストがある方が本の売行きも伸びる、と判断する出版者も増えている。

日本語の「源氏物語」に関しては、小学館の理解を得ることが出来、研究者の利用に限ることや、営利目的にしない点、サービス体制の確立等の項目を盛り込んだ覚書が交わされた。

英文に関しても、E.G.サイデンスティッカー教授の了解を得ることが出来た。

6. アクセス

利用の方法について述べる。国内のサービスを引き受ける東京大学大型計算機センターでは、センターの登録者に対しいくつかのデータベースの公開を行なっている。「源氏物語」のテキスト・データベースも同様のサービスに供される。なお登録は大学等に所属する研究者に限られている。

OUCSの場合は、サービスされているテキスト・データベースのリストは小冊子の印刷物や電子メール等で知ることが出来る。それには著者名、タイトル、言語、容量のほか、U、X、Aの三種類の記号からなる利用条件についての情報が付いている[4]。

Uは一般利用(UNIVERSAL ACCESS)を意味し、このコードの付いたテキスト・データベースの利用者は先のコピーライトで述べられた利用条件の確認書に署名をすれば、希望のものの複製を受け取ることが出来る。

Xは非公開利用(EXCLUSIVE ACCESS)を意味し、このコードが付いたテキストはOUCSの登録者以外は使えない。これは東京大学大型計算機センターのサービスの条件と同じである。

Aは制限付き利用(RESTRICTED ACCESS)を意味し、このコードの付いたテキストは、利用者からの問い合わせは速やかに供託者に転送され、その許可のもとでOUCSは複製を作成する。又OUCSの光学読み取り装置によって作成されたテキストは、12ヶ月間このカテゴリーが付された後Uのカテゴリーに移される。

7. 研究例

ここではコンピュータを使った若干の検索結果及び研究例を紹介する。

7.1 Micro-OCPを使った検索

表3はMicro-OCPを使った検索例で、コマンド・セット、出力結果並びに語彙数等の統計量を示す。第一章の「桐壺」の巻にでてくる“かぎり”、“限り”、“聞*”、“きこえ*”、“きこゆ*”の五つの語のKWIC(用語索引)を作成している。この内、先の二つと、後の三つはそれぞれ同じ語の漢字表記と平仮名表記である。

“かぎり・限り”の場合は、編者の一人である秋山氏によれば、両者に区別はなく、漢字にするかどうかは、読み易さに依存する。この語は、“美しい”、“はずかし”、“おかし”などのように、現代と古代で意味の異なる語の一つである。我々は一般的に使用しているが、古代では重みのある言葉で、一種の絶対的ともいえる限界状況を意味する場合も多い。上記の例では一番最初の検索例が相当する。

一方、“聞*”、“きこえ*”、“きこゆ*”のグループでは、漢字表記と仮名表記では、語り手が上下関係のある二人の話者のどちらを尊敬するかで意味が違っている場合の例である。これは多出する“たまふ”と“給ふ”等にもあてはまる区別である。

表 3 Micro-OCP を使った検索の例

コマンド

```
*input      text hyphen "-" {and stop at record 100} .
            comments between "{" to "}".
            references cocoa "<" to ">" and on P set L="1".
         {  select where A = "帝".}
*words      punctuation with japanese.                    { 日本語 }
            alphabet "かぎり きこえ きこゆ".
*action     do concordance.
            references W=2,V=1,C=4,S=2,P=3,L=2.
            pick words "限* かぎり* 聞* きこえ* きこゆ*".
         {  sort in alphabetical order }
            maximum context span L.
*format     layout length 80 and depth 55.
         {  context size 1 and no printing character specified }
*go
```

出力結果

```
                                                かぎり      5
紫  1  きり  4   99   1                    かぎりとて別るる道の悲しきにいか
紫  1  きり  8   108  4                鈴虫の声のかぎりを尽くしても長き夜あかずふ
紫  1  きり  9   109  9    、忍びやかに、心にくきかぎりの女房四五人さぶらはせた
紫  1  きり  9   113  2   たれば、陪膳にさぶらふかぎりは、心苦しき御気色を見たて
紫  1  きり  9   113  3   。すべて、近うさぶらふかぎりは、男女、いとわりなきわ

                                                きこえ     17
紫  1  きり  2   95   15  らはしう、心苦しう思ひきこえ
紫  1  きり  2   96   1              御蔭をば頼みきこえながら、おとしめ、疵を求め
紫  1  きり  8   109  4   、いと うしろめたう思ひきこえたまひて、すがすがともえ参
紫  1  きり  10  114  7    まふ。年ごろ馴れむつびきこえたまひつるを、見たてまつ
紫  1  きり  11  115  5   びぐさに、誰も 誰も思ひきこえたまへり。　わざとの御学
紫  1  きり  13  118  1     母后世になくかしづききこえたまふを、上にさぶらふ典侍
紫  1  きり  13  118  14  子たちの同じつらに思ひきこえ
紫  1  きり  13  119  5    でたく、人もえおとしめきこえたまはねば、うけ ばりてあ
紫  1  きり  13  119  6    し。かれは、人のゆるしきこえざりしに、 御心ざしあやに
紫  1  きり  14  120  2    心地にいとあはれと思ひきこえたまひて、常に参 らまほし
紫  1  きり  14  120  9    まつる。こよなう心寄せきこえたまへれば、 弘徽殿女御、
紫  1  きり  15  122  15  、ともかく もあへしらひきこえ
紫  1  きり  15  124  6    しきまで、もてかしづききこえたまへり。いとき びはにて
紫  1  きり  15  124  7    ゆゆしううつくしと思ひきこえたま へり。女君は、すこし
紫  1  きり  17  125  8   まを、たぐひなし と思ひきこえて、さやうならむ人をこそ見
紫  1  きり  17  126  2   して、い となみかしづききこえたまふ。御方々の人々、世の
紫  1  きり  17  126  11  いふ名は、高麗人のめできこえて、つけたてまつ りけると

                                                きこゆ      5
紫  1  きり  2   94   15  君と、世にもてかしづききこゆれど、この御
紫  1  きり  6   101  4    しと、人々もてわづらひきこゆ。　内裏より御使あり。三
紫  1  きり  8   104  1                仰せ言伝へきこゆ
紫  1  きり  8   109  1                    出できこゆれば、とく参りたまはんこと
紫  1  きり  8   109  1    まはんことをそそのかしきこゆれど、かくいまいましき身
```

限 12

紫 1 きり 2 95 3 ほしかしづきた まふこと限りなし。 はじめよりおしなべ
紫 1 きり 4 98 3 びてぞ 出でたまふ。限りあれば、さのみもえ止めさせた
紫 1 きり 4 98 13 限りあらむ道にも、後れ先立たじと
紫 1 きり 4 99 12 なきに、なほい ぶせさを限
紫 1 きり 6 100 10 限りあれば、例の作法にをさめたて
紫 1 きり 9 111 6 じき絵師といへども、筆限りありければ、いとにほひすくな
紫 1 きり 10 114 1 、さばかり思したれど、限りこ そありけれ、と世人も聞こ
紫 1 きり 10 114 5 たこれを悲しび思すこと限りなし。皇子 六つになりたまふ
紫 1 きり 12 116 11 る句を作りたまへるを、限りなうめでた てまつりて、いみ
紫 1 きり 14 120 4 おぼえたまふ。 上も、限りなき御思
紫 1 きり 15 121 3 居起ち思しいと なみて、限りあることに、ことを添へさせ
紫 1 きり 15 124 3 も数まされり。なかなか限りも なくいかめしうなん。 そ

聞 23

紫 1 きり 4 98 6 みながら、言に出でても聞こえやらず、あるかな きかに消
紫 1 きり 4 98 9 ま はすれど、御答へもえ聞こえたまはず。まみなどもいとた
紫 1 きり 4 99 3 息も絶えつつ、聞こえまほ しげなることはありげ
紫 1 きり 4 99 7 聞こえ急がせば、わりなく思ほしな
紫 1 きり 4 100 1 聞こしめす御心まどひ、なにごとも
紫 1 きり 7 102 10 遣はしつつ、ありさまを聞
紫 1 きり 8 103 1 を掻き鳴らし、はかなく聞こ え出づる言の葉も、人よりは
紫 1 きり 8 106 5 をだに、はるく ばかりに聞こえまほしうはべるを、私にも、
紫 1 きり 8 108 9 る雲の 上人 かごとも聞
紫 1 きり 9 112 3 とすさまじう、ものし と聞こしめす。このごろの御気色を見
紫 1 きり 9 112 4 どは、かたはらいたしと聞きけり。いとおし立ちかどかどし
紫 1 きり 10 114 2 こ そありけれ、と世人も聞こえ、女御も御心落ちゐたまひぬ
紫 1 きり 12 115 11 しこ き相人ありけ るを聞こしめ して、宮の内に召さむこ
紫 1 きり 13 117 15 、御容貌すぐれたまへる聞こえ高
紫 1 きり 13 118 8 心とまりて、ねむごろに聞こえさせたま ひけり。 母后
紫 1 きり 13 118 15 聞こえさせたまふ。さぶらふ人々、
紫 1 きり 13 119 3 まつりたまへり。藤壺と聞こゆ。げに御容貌ありさま、 あ
紫 1 きり 14 120 1 典侍の聞こえける を、若き御心地にいと
紫 1 きり 14 120 5 まひそ。あや しくよそへ聞こえつべき心地なんする。なめし
紫 1 きり 14 120 7 聞こえ つけたまへれば、幼心地に
紫 1 きり 14 120 14 な るを、世の人光る君と聞こゆ。藤壺ならびたまひて、御お
紫 1 きり 14 120 15 れば、かかやく日の宮と聞
紫 1 きり 17 125 13 のをりをり、琴笛の音に聞こえ通ひ、ほのかなる 御声を慰

統計量

TOTAL WORDS READ	=	1552
TOTAL WORDS SELECTED	=	1552
TOTAL WORDS PICKED	=	62
TOTAL WORDS SAMPLED	=	62
TOTAL WORDS KEPT	=	62
TOTAL VOCABULARY	=	5

話者同士の上下関係や男女の区別をするために敬語の果たす役割は大きい。しかし量が多く総合的に様々な用例を研究するのは困難であった。しかしコンピュータを使えば敬語のみならず、様々な語彙やフレーズを容易に文脈ごと取り出すことが出来るわけで、敬語のみならず大量の用例の抽出を必要とする研究に効果を発揮しそうである。

なお、アスタリスクは複数の文字のワイルド・カー

ドを意味し、語幹の変化形を、このように総てコマンド・レベルで宣言しておけば、平仮名、漢字を問わず自動的にピック・アップしてくれる。

7.2 統計的手法を使った研究例

戦後まもなく武田宗俊は54帖配列の謎をめぐっての明快な成立論を展開し一躍脚光を浴びた[5]。彼は源氏物語の第一部(源氏を三部構成とし具体的には33巻までの光源氏の将来についての予見が完全に実現するまで)を17帖と16帖の2系列に分け、前者を紫上系、後者を玉鬘系と命名した。「紫上系の物語は独立し、完全な統一を持つものとして後者に無関係である。即ち源氏物語第一部から玉鬘系16帖を除いても何等欠けることのない物語となっている。玉鬘系の物語は紫上の物語を背景とし、その系の物語を取り入れているが、それはただ影を落とすのみで、その物語を玉鬘の巻に於いて発展させてこれを紫上系にかえすことはない。玉鬘系の物語は松にからみついた藤のように、外見は一体をなしているが、その本を異にし、付加的結合で、有機的な融合とはなっていない。」と主張した。更に紫上系17帖が先ず構想記述され、これに付加して玉鬘系が後記挿入されたのだと推定したのである。

この武田の源氏物語成立説は定説としては受け入れられていないが、学界に大きな影響を与え、現在でも賛否両論の活発な論議があり重要な研究領域になっている。

この観点に着目して安本美典[6]文献は紫上系と玉鬘系の「比喩多用型か比喩節用型」どうかであるかどうかという観点から因子分析を行い、武田氏の説が妥当なものであるとの結論を出している。

7.3 「宇治10帖」について

「源氏物語」の最後の10帖と他の44帖との関係、いわゆる“宇治10帖問題”については、長年議論がなされている。

安本氏は「歌物語」「作り物語」といった因子や、名詞の頻度、和歌の用いられている度合いなどを調査して、この問題にも挑戦している。そして10帖は他の44帖に比べ「作り物語」的で且つ又「比喩節用型」であり、両者の文体には差があるとの興味深い結論をだしている。

他方コンピュータを使わない研究も勿論盛んで、現代語訳で有名な円地文子氏などは別人の説をとっている。他方石垣謙二氏は助詞の「は」「の」の優れた研究家であるが、助詞の使い方だけでは別人の筆であることを立証するのは困難であると主張する。

又サイデンスティッカー[7]は、シェイクスピア＝ベイコン説にならって、この問題を“Genji-Baconian Theory”あるいは“Genji Baconians”と命名し興味深い議論を展開している。氏は、42帖の「匂宮」の冒頭で源氏が亡くなるところで物語は終り、42とそれに続く43の「紅梅」、44の「竹河」の3帖を変わり目、あるいは経過の部分とし、45以降から新しい物語がはじまると主張する。

この問題はまだまだ論議が続くと思われるが、テキスト・データベースが公開されれば、コンピュータを利用した研究はもっと盛んになると思われる。

8. 問題点

次に作成されたのテキスト・データベースの問題点の検討に移る。先ず、日本文と英語文の対応についてである。今回は54帖の章の対応と、各章の中の小見出しの頭の部分を相互に参照できるような仕様を作成した。しかし、文、段落あるいはページの対応は考えられなかったのであろうか？

例えば「聖書」の場合、旧約はヘブライ語、新約はギリシャ語で書かれている。そして個々の文には全て番号が付いている。又多数の言語による翻訳が出版されているが、それぞれ原文と同じ文番号がついており、各言語間で一対一の対応がついている。

ところが「源氏物語」の場合は、日本語と英語の各文にはそれぞれ番号もないし、もともと対応するように翻訳されてはいない。日本語ではしばしば主語が省かれる。又関係代名詞が無いため、英文とでは文の長さに大きな差がでてくる。

又サイデンスティッカーが翻訳の際、原典としたのは岩波古典体系所蔵の「源氏物語」であり、加えて青表紙本を典拠としたいくつかの現代語訳などを参照している。そのため、段落のとりかたや主語等、編者の解釈に応じて小学館の版とは随分異なっている。その結果、文のレベル、段落のレベルで両テキストを対応させることはほとんど不可能であった。さらにページ対応ということも考慮されたが、テキストがあまりに長く、作業量が多くなると判断され見送られた。検討の末、結局日本語文の小見出しを対照の最小単位とする案に落ち着いた。しかし小見出しが細かく挿入されている場合は良いが、中には四から五ページに亙る場合もあり、テキスト・データベースの対応が荒くなる部分もしばしば生じた。

今後は少なくともページごとの対応ができるような工夫が必要であると思われる。

もう一つは話者マークの挿入である。入力作業をとりわけ困難にしたのは、話者のはっきりしない場合や複数の話者が想定される場合である[注6)]。

入力作業の煩雑さを覚悟してまでも、話者マークの挿入にこだわったのは、これらに関しては、話者別に会話文の中身だけを解析対象にする場合を想定したからであった。また、従者や無名の女官などのマークも挿入しておけば、宮廷人と世間一般の人々の言葉の違い等の研究にも利用できると考えられた。もともと解析ソフトがこのような分析に利用されていることが分かっており、既にシェイクスピア等の戯曲の分析で成果をあげていることがヒントになっている。しかし話者マークの挿入は全て手入力を必要とするため、やはりミスの多発を防ぐことができなかった。むしろ中途半端に話者を入れるよりは、まったく別のテキストとして話者マーク付きと、そうでないものをつくる方が楽だったかもしれない。又、あえて話者を挿入したために実際の研究を束縛する恐れもあるわけで、出版された形を可能な限り保持した方が良かったのではないかとおもわれる[注7)]。将来のテキスト・データベースの作成ともからめて今後の検討課題として残りそうである。

9. 将来の展望

欧米では新たに TEI(Text encoding Initiative) や Centre Computing in the Humanities(Toronto University) の研究者グループが中心になって、SGML(Standard Generalized Markup Language: 汎用マークアップ言語) 方式によるテキスト・データベースの研究が盛んなりつつある。SGML は文書の論理構造、例えば著者や表題等のデータ項目を、後から項目別に判別できるようタグを挿入しながら作成する方式で、ハイパー・テキストの作成にも効力を発揮する。既に ISO や EC で、そのタグの標準化が進められている。COCOA 形式に似ていることから、OUCS が開発した OCP のバージョン・アップも SGML 対応が第一課題になっている。日本では現在、学術情報センター[8)]が中心になって研究を開始した。

今後「源氏物語」のテキスト・データベースも SGML 方式による作成が可能であり、それによって益々国際的にも利用価値の高いものになると思われる。

例えば、古典を研究するためには、索引や辞書が必要であるばかりでなく、内在資料の検討や様々な背景への理解を助ける外在的資料にあたることが不可欠である。「源氏物語」では登場人物の呼び名や官職がしばしば変わる。主人公の源氏は、光、君、若、おとど、六条院、大将、大殿、院、男君、など 30 以上の呼び名を持つ。又中国の文献からの引用も多い。そのたびごとに、後ろの付録の家系図などをいちいちひっくり返して調べながら読み進まなければならない。こういった項目ををタグとして挿入しておけば、コンピュータの画面でそれらの情報の解説へ飛ぶことも容易になる。又呼び名の変遷を通してテキストの順番や書かれた年代の推定等の研究もできる。あるいは和歌や注釈のタグを付加することも可能であろう。そうすれば読者は和歌だけをとりだして研究することもできる。又現在「源氏物語の絵巻物」が出版されている。将来は、絵巻と今回削除した平安時代の服装や小物等の挿絵を一緒に画像データベースとして組み込むことも可能になるであろう。更に、源氏は優れた音楽演奏者で、宴の場面や楽器の演奏場面もよくでてくる。古代の楽器の奏でる音楽もテキストに組み込むこともミュージック・ソフトの技術を導入すれば可能であろう。将来は文字と音と絵と組み合わせたハイパー・テキストの作成も夢ではない。

海外では既に SGML は辞書作成で大きな成果をあげており、既に OED(Oxford English Dictionary) の CD-ROM 版にも採用されている。この電子化辞書版の OED を使うと、語源がアラビアあるいはヘブライ語の言語を 18 世紀のテキスト中から捜す、あるいはデイケンズのテキストから捜しだす、といった昔なら一人の研究者が何年もかかるような作業があっという間に出来る。辞書を一番使うのはいうまでもなく研究者である。このような辞書をみると末端のユーザーである文学研究者の意見がいかに良く反映されているかが解る。欧米では辞書のステイタスは学者・研究者の使用に足るものかどうかで決まる。質の高い辞書学 (lexicography) の長い伝統の蓄積と最先端のコンピュータによるデータベース作成技術である SGML の結合の見事な成功例といえるだろう。実際、優れた古語辞典もない日本の現状を考えると驚嘆の念を禁じえない。言葉は数字と違って曖昧である。重要な語句、キーワード程、意味も

広く、多様で有り、沢山の用例研究を必要とする。日本文化の重要なキーワードである、“わび”、“さび”なども、時代の流れ、解釈により意味は多様化している。これらの語の用例が日本の古典の中から全て検索することが可能になれば、意味の変遷の研究に大いに役立つであろう。「源氏物語」の場合でも基調をなす“もののあわれ”は、「伊勢物語」や「古今集」等でも使われている。これらのテキストの用例を網羅した辞書が出来れば、研究者のみならず多くの人々が恩恵を受けるであろう。今回のテキスト・データベースのプロジェクトを通して切実に思ったのは、テキスト・データベースの作成だけでなく、日本でもOEDのように文学や歴史資料の用例を総て入力した優れた「古語辞典」が必要だということである。

10. 結語

以上今回のプロジェクトの経過をたどりながら、その間得られた様々な知見、問題点、将来の展望などを検討してきた。今後多くの人々に利用して頂き、良きアドバイスに従ってバージョン・アップをはかると同時に、新たなテキスト・データベース作成にも挑戦していくつもりである。特にSGML方式を使えば、制作者側はテキスト・データベースに様々な付加価値をつけることが可能になり、他方使用する側にとっても、テキスト・データベースを自分の関心にあわせて多面的に利用できる、という利点がある。今後この種のテキスト・データベースの作成が益々盛んになり、世界の文化活動への貢献を期待したい。

出来ればこのような研究が刺激になって、テキスト・データベースのみならず、優れた古語辞典の電子化辞書のプロジェクトが企画されることを願ってやまない。

最後に、今回一番残念だったのは、日本に世界に機械可読テキストをサービスする機関がない為、やむなくオックスフォード大学電算機センターに海外サービスを依託しなければならなかったことである。現在世界各地で日本語・日本文化への関心が高まっており、テキスト・データベースの需要は今後増すと予想される。もはや日本に行かなければ日本研究ができないという時代ではないであろう。将来、日本国内から各国の日本研究機関や研究者に対して、我が国の古典や文学のテキスト・データベースのサービスが可能になる日が来ることを願ってやまない。

謝辞　このプロジェクトは千葉大学の加藤尚武、坂井昭宏両教授の御尽力により、社団法人「東京倶楽部」の文化活動補助を受けることが出来た。又、完成までに多くの人々の御協力を賜った。とりわけ、東京女子大学教授の秋山虔先生にはテキストの内容その他について多くの事を御教示戴いた。東京大学大型計算機センター教授の石田晴久先生、Oxford University Conputing Service の L.Burnard 氏と S.Hockey 氏にも供託等の面で大変御世話になった。ここに記して皆様に深く御礼申し上げる。

注

1) “データベース”という言葉は一般には“検索可能な状態にある種々の情報”と定義されている。近年、文学あるいは哲学の関係者の間では、書誌データや本を丸ごと入力したテキストについて、先の一般的なデータベースと区別して、“テキスト・データベース”という呼び名がしばしば使われている。その他、単にデータベースと呼ばれたり、フル・テキスト、電子化テキスト、あるいはテキスト・データと省略して使われることもある。このように、全文を丸ごと機械に入力したテキストについて日本では統一した呼び名がない。英語圏では、こういったテキストは“機械可読テキスト(Machine-readable Texts)”と呼ばれている。本論文は“テキスト・データベース”で統一している。

2) 機械可読テキストを作成している世界の主だった機関のリストを以下に紹介する。

   (a) African Text: School of Oriental and African Studies, University of London

   (b) Dutch: Postbus 132, 2300 AC, Leiden

   (c) French: Institute la Langue Francais, 44 ave de la Liberation C.D.33 10, F-54 014 Nancy-Cedex

   (d) German: Institute fur Kommunikationforschung und Phonetik, Bonn, Institute fur Deutche Sprach, Manheim

   (e) Greek: University of California, Irvine

(f) Hebrew: Academy of Hebrew Language, Jerusalem

(g) Icelandic: Armagnaen Institute, Copenhagen

(h) Italian: Lessico Intellettuale Europeo, Rome

(i) Latin: Batiment 16A, Louvain la Neuve

(j) Norwegian: Bergen University

(k) Swedish: Logotheque, Goteborg

(l) Welsh: University College of Wales

(m) English & others: Oxford University Computing Service

ほとんどがヨーロッパに集中し、自国の言語で書かれたテキストは、やはりその国が中心になって作っているのがわかる。例外的に大きいものとして、5番目のカリフォルニア大学アーヴァイン校の付属機関であるシソーラス・リンガエ・グレカエ (Thesaurus Lingae Graecae) がある。一般には TLG の略称で呼ばれ、既に紀元前から紀元後 600 年迄のギリシャ語のテキストが入力されており、その数は 4,500 冊以上にものぼっている。13 番目にオックスフォード電算機センター (Oxford University Computing Service) が挙がっている。これも一般には OUCS の略称で知られている。規模は TLG と比べると、920 冊と少ないが、サービス体制がしっかりしており、他の機関で作成されたテキスト・データベースの紹介や供託サービスも行なっている。

3) 既に公開されているものの一部と製作者を以下に挙げる。

(a) 独語:
Thormas Mann, Gesammelte Werke in 13 Baden(s. Fischer 版)
Goethes Werke, Hamburger Ausgabe in 14 Baden
Goethes Werke, Weimarer Ausgabe in 143 Baden(九州大学:樋口 忠治)

(b) 日本語 (漢字):
日本書紀、続日本紀 (京都大学:星野 聡)

(c) 英語:
中世イギリス研究資料 (東京大学:久保内 端郎)

(d) ラテン語:
デカルト「省察」「反論答弁」(山口大学:村上 勝三)

(e) サンスクリット語:
仏典 (東北大学:塚本 啓祥)

4) OCP は 2 年間に亘る文科系研究者のアンケート調査の後、最初アトラス研究所が開発に成功した。後に OUCS が引き継ぎ、更に 2 年がかりで汎用化を行なった。現在欧米はもとより、日本各地の大学の大型電算センターで使われている。主な機能を以下にまとめる。

(a) OCP の基本出力形態は、ワードリスト、インデックス、KWIC に代表される用語索引の 3 つ。これに組み合わせて使用単語総数、語彙数やそれらの比率などテキストに関する統計量の出力も出来る。

(b) 単語の検索は、指定がなければテキスト内に使用されている単語を全て一挙に出力。

(c) 特定の綴りの単語を指定しての検索だけでなく、単語の長さ、出現頻度、辞書範囲を指定した検索も可能。

(d) また単独に現れる単語の他にフレーズの検索や、分離動詞、共出語等の抽出も出来、これらの綴りの指定にはワイルド・カード (任意の文字、文字列の指定) も使用可能なので語尾変化等を気にする必要はない。

(e) KWIC に代表される用語索引は、文脈一行の出力だけでなく複数行に亘る出力も可能。また文脈の出力範囲を句読点の出現位置までに限定することも。

(f) 文脈中のキーワードの出力も中央揃えの他、左揃え、右揃えが選択ができる。

(g) 出力の見出しについては、接頭語、接尾語の付いた単語を付かない単語と同列に扱ったり、動詞の変化形 (例えば am,is,was,are,were 等) をまとめて一つの見出しの下に統合することもできる。

更にアクセントの有無や、大文字小文字の別により分離、統合することも可能である。

(h) 結果の出力順序については、アルファベット順や、逆アルファベット順、また単語の語尾からのアルファベット順序等も指定可能。また単語の出現頻度順や、単語の長さによる順序、参照部を利用してテキストの内容毎に順序を決定することも、これらを組み合わせて使用することも可能である。

(i) 複合文字を含む文字や記号は、出力の段階で全く別の文字や文字列に置き換えることが出来る。字体を変えてプリンターに出力したり、文書を他人に解読できない秘密文書に暗号化することもできる。

5) 日本語版「源氏物語」において、JISの第1水準および第2水準にない漢字のリスト及び入力に際して代用している漢字を以下に示す。

第1巻	p.115	臚	→	炉*
第2巻	p.182	縑	→	謙*
	p.199	竄	→	究*
	p.205	棊	→	某*
第3巻	p.160	黶	→	瘴*(616F)
	p.236	瘖	→	音*
	p.284	籙	→	録*
	p.365	笄	→	笄*(6422)
	p.413	綷	→	淡*
第4巻	p.49	泔	→	柑*(343B)
	p.87	黶	→	瘴*(616F)
	p.92	綷	→	淡*
	p.256	蓆	→	席*
	p.269	黶	→	瘴*(616F)
第5巻	p.148	縑	→	謙*
第6巻	p.53	泔	→	柑*(343B)
	p.62	泔	→	柑*(343B)

6) 地の文はもとより会話を直接話法で始まったものが、終わりの部分では間接話法になっていたり、一体どこから会話文が始まっているかわからない部分があり、これらは色々な写本により解釈が分かれている。

7) "The Machine-Readable Texts of Wittgenstein" by Prof.A.McKinnon and Prof.H.Kaal of McGill University はクリーンな版と文法情報等を付加した編集したテキスト・データベースの両方を作成している。

文 献

1) 井上英秋: 源氏物語の英訳をめぐって, 言葉の諸相, 笠間書房 (昭和57年).

2) 富士 *OCR* システム, ユーザーズマニュアル, 富士電気株式会社

3) 梅田三千雄: 漢字部首情報からの日本語単語の推定, 情報処理学会人文科学とコンピュータ研究会報告, Vol. 89, No. 102 (1989).

4) *Text archive–Notes for descriptions of Machine-Readable Texts*, Oxford University Computing Service (1987).

5) 武田宗俊: 源氏物語の最初の形態, 文学, 昭和25年6・7月号 (1950).

6) 安本美典: 文学作品を因子分析する, 計量国語学, No. 47, pp. 21-32 (1968).

7) E.G.Seidensticker: *How Many People Wrote The Tale of Genji*, An Invitation to Japan's Literature, 財団法人日本文化研究所 (1974).

8) 根岸正光: フルテキスト・データベースの実用化における諸問題, 情報処理学会情報学基礎研究会報告, No. 14 (1989.7).

(1990年 1 月 29 日受付)
(1990年 2 月 28 日採録)