電子テクストとデータベース(1)
製品版テクスト集成の検証と評価

篠田勝英

はじめに

1996年に中世フランス文学のテクスト・データベース構築という構想が生まれ，そして翌1997年から1998年の二年間にわたって個人的なレヴェルでのプロジェクトを進め，その時点までの成果を「テクストの電子化とデータベース構築」[1] としてまとめてから早三年が経過した。2002年度にはその延長線上における研究が「中世フランス文学のテクスト・データベース作成の試み」として，さいわいにも文部科学省の科学研究補助金を受けることができた。具体的なテーマは，商品化された電子テクストの解析・評価と，すでに作成されている電子テクストとの比較である。データベース構築の試みという大きなプロジェクトにおいては，きわめて限られたテーマではあるが，現状確認という位置づけで中間報告を試みたい。

日本語環境におけるOSと文字コードの問題

対象とするテクストがアルファベット26文字と算用数字，それに若干の記号類（ハイフン，アステリスク等）だけで構成されていれば，独自のオペレーティング・システム（以下OSと略記）にもとづくワープロ専用機でも，それなりの文字列処理はできるし，いわゆる文字化けの問題も生じない。しかしアクサン記号・分綴記号の付いた文字（é, è, ê, ë等）や，合字（æ, œ等），さらにはフランス語固有のセディーユ（ç）などが用いられている場合には文字化けが生じるばかりではなく，検索・ソート等の操作にも支障をきたすのがふつうである。もっともこの問題は，フランス語に関するかぎり，英語版（フランス語版）のOS上ではまったく生じない。マッキントッシュであれ，ウィンドウズであれ，ユニックスであれ，あるいはメインフレーム上で動く特殊なプログラムの場合も，いわゆる欧文特殊文字も含めてすべての文字がカヴァーされている。しかし，日本語版OS上でのデータ操作，あるいは異なるOS間でのデータ移動（あるいは共有）になると，事情はまったく異なってくる。

もっとも異なるOS間の場合は，英語ないしフランス語版を用いる限り，テキストファイル形式のデータであれば，まず問題はない。まれに生じうる文字コード上の問題も，さまざまなコード変換ツール[2]が用意されているので，手間を惜しまなければファイル自体を移植するのはきわめて容易である。

一方日本語版OSの場合は，これまでさまざまな問題を生じさせ，利用者に多大な労苦を強いてきた。要するにほとんどの場合，JISコードとASCIIコードの重複が少なからぬトラブルを

ひきおこしてきたのである。ということは，逆にいえばコード問題が解決すれば，障害が一掃されることになる。事実，文字化けという問題は同一機種内のフォントの間で生じるものであれ，異機種間のデータ・コンヴァートで出現するものであれ，アルファベットを用いるヨーロッパ系言語に関しては，ユニコードの導入により，日本語ベースのOSにおいてもすべて解消されるはずである。したがって将来的な問題は，これまでに蓄積されたASCIIコードによる電子テクストのコード変換と，現存のアプリケーション類のユニコード対応の二点に絞られてくる。

現在，人文系の研究者が日常的に触れているコンピューター環境は，マッキントッシュかウィンドウズであろう。両者のどちらかに依拠していれば，ハードウエア性能の急速な向上により，よほど特殊な用途でない限り，個人レヴェルでワークステーションやメインフレームのお世話になることはないと思われる。したがって今後の研究環境を論じる場合，ハードウエアはいわゆるパーソナル・コンピューターで十分であり，OSに関しては普及の現状を見ると，(Linux等のUnix系のOSは視野に置きつつ，また性能と普及率が極度にアンバランスな「超漢字」=BTRONの存在にも目を配りつつ）マッキントッシュとウィンドウズに対象を限定せざるをえない。ところでこのふたつのOSの年季の入ったユーザーには，それぞれの優位性について，これまで信仰に近い思い入れが見られた。そして外国語の処理というわれわれの関心領域は，少数派のマック・ユーザーが声を大にしてそのOSの優位を主張する分野だった。事実OSレヴェルでの多言語対応は数年前までマックの独擅場であり，漢字・アルファベット以外の文字を用いる言語の研究者にとっては，マッキントッシュ以外の選択はありえなかったほどである。けれども数年前からその差は次第に縮まり，現在ではほとんど同レヴェル，場合によってはそれぞれ一長一短というところまで来ている。しかしながら両OSともに最新のヴァージョンではシステム内部でユニコードを取り入れているにもかかわらず，アプリケーションが十分には対応していない。そのためマッキントッシュの場合は日英ないし日仏のふたつのOSを組み込むことが容易という利点を生かしてのダブルブート方式，ウィンドウズの場合はひとつのOSで「入力ロケール」を切り替えるという，ある意味でより簡便な方法で「欧文特殊文字」を含むテクストを処理したり，あるいは日本語版OSでは使えないアプリケーションやCD-Romをインストールしたりする方式で，多言語対応（少なくともローマン・アルファベットやギリシア文字やキリル文字で表記する言語と日本語の共存）を実現しているのが現状であろう。ただしどちらの場合も切換には再起動が必要である。マシン性能の向上により起動にかかる時間はかなり短縮されてきたとはいえ，この操作はデータの保存を必要とするし，環境によってはパスワードの再打ち込みが求められるなど，思考の流れを中断することはなはだしい，やっかいな手続きであることにかわりはない。したがって日本語環境における欧文特殊文字を含むテクストの処理に関しては，個別のファイルはどのようにもコンヴァートできるにしても，これまでに商品化されたテクスト・コレクションないしデータベースの場合には，ユニコードの一般化で利用が容易になるという期待は持てないことを覚悟しておかなければならない。

こうした条件の下ではあるが，近代の作家についてはテクスト・データベース的な性格を持つCD-Romが，いわば個人全集のような形で次々に刊行されている。一方中世文学の場合は，需要が少ないせいか，商品化されたものはほとんど存在しなかったのだが，2001年に画期的なCD-Romが刊行された。Championによる *Corpus de la littérature médiévale* である。

Corpus de la littérature médiévale について

カタログによれば，このCD-Rom（以下 *Corpus* と略記）には「500以上の作品の全文」が収められているとのことだが，収録作品はすべて冊子体で刊行されたものばかりで，一覧をタイトルで数えると約190篇，その中には集成がかなりあって，たとえば抒情詩の一篇一篇を数え上げると，とても500ではきかないし，印刷本の巻数とも思えないので，「500以上」という数え方がよく分からない。紙の本と異なり，CD-Romは全貌が見えにくいので，中に入って行くしかないのだが，作品別のインデックスの項目数は879に達していて，500以上なのは確かだが，なぜ500という数字が出てきたのかは分からない。たとえばマリー・ド・フランスの『レ』は一作品としか数えられていないが，ファブリオは集成の形で収められているものの，個々の作品が索引項目として立てられている。またインデックスの最初の画面には書誌情報がまったく出てこないので，どのような作品のどの刊本が収められているかを知るには試行錯誤が必要である。ただし，多くの作品がフランス中世文学の代表的なコレクション，すなわちSociété des Anciens Textes Français (SATF), Classiques Français du Moyen Âge (CFMA), Textes Littéraires Français (TLF) 等の定評のある校訂版から取ったものだから，書誌学的知識に基づく予想である程度の見当は付けられる。しかしながらCD-Rom固有の階層構造はなかなか頭に入りにくいし，常に「検索をしている」という意識がともなうのは，この集成が読書ではなく研究の対象であることを明確に示していると感じられる。実際，相手にしているのは一冊の書物ではなく，かなり大型の，それも複数の書架なのである。

これまでに電子テクストを収めた製品版のCD-Romは近代作家のものがかなり大量に刊行されている。しかしその大部分は依拠したテクストについて言及がないか，著作権が問題にならない古いテクストを使ったものである。たとえばプルーストに関しては *Corpus* と同じChampionとGallimardの二社からCD-Romが刊行されているが，前者は1920年代のNRF版（Gallimard社のいわゆるBlanche版）のテクストを使った一種の電子本であり，後者は写真・解説・音楽・画像・書誌等の関連資料の集成に新Pléiade版のテクストを一部引用した「プルーストの世界入門」的な「マルチメディア」を謳う製品にほかならない。Champion版は一応読書の対象にはなるが，このテクストでは研究論文で引用することはできないし，コーパスとして検索するのにも使えない。バルザックの『人間喜劇』集大成と称するCD-Rom（ACAMEDIA刊）は，印刷すると70頁にもおよぶ詳細な書誌を付けながらも，肝心のテクストの成立については言及がない。要するに研究用の道具としては構想されていないのである。それに対して *Corpus* は，すでに述べたように中世の文学作品の校訂版を電子化したものであるから，テクストの素性ははっきりしているし，校訂のレヴェルも概して高いものが多い。未見ではあるが，同社の *Corpus de la littérature narrative du Moyen Âge au XX^e siècle* と同じく，研究用のツールとして使うことを前提とした製品である。ただし，校訂者の著作権の生きている刊本を電子化したテクストが多いだけにガードが堅いが，この点に関しては後に触れる。

Corpus の利用条件

さて，すでに内容に入りかけてしまったが，ここでいったん立ち止まって，利用上の条件（いわば外的な規制）を確認しておこう。*Corpus* は CD-Rom PC と明記された商品である。すなわちウィンドウズ専用であり，そればかりか，マック用の代表的なエミュレーター，Virtual PC 上でも使えないので，マッキントッシュではまったく利用できない。環境によってはこれがまず障害のひとつとなる。次に値段が高い。2002年12月現在，コンピューター一台で利用するものが税別で3,353.89ユーロ，五台までのアクセスを許すものが4,268.56ユーロであるから，なかなか個人利用はできない。商品の供給自体は註文等を必要とせず，ごくふつうに行われているので，購入までの障害はこのふたつである。

利用にあたっては，付属ソフトのインストールは容易であり，Win OS は95から2000までカヴァーされていて（Xp での作動は確認，Me は不明だが，おそらく問題ないと思われる），かつ日本語 OS でもスムースに作動する（ただし確認したのは Xp のみ）ので，特に英語ないしフランス語環境を用意する必要はない。また，この種のデータとソフトは辞書ソフトの場合と同様，ハード・ディスクに格納して使いたいものであるが，そうした利用法には原則として対応していない。しかし最近の高速ドライヴであれば，ストレスを感じないスピードで作動する。

テクストと付属ソフトウエア

Corpus は書物をディスプレイ上に再現することを原則としている。したがってテクストは元になった冊子本のページを再現し，ページ単位で表示させることしかできず，スクロールさせて全体を見ることはできない。その点では一般に流通している電子テクスト・ブラウザー（Voyger 社の BookBrowser や T-Time）に似ているが，非専門家にはファイル構造がよく分からず，作品とファイルの関係付けを確認するのがきわめて難しい。ただし通常の使用，特に作品をディスプレイで読む際には，ページを繰るのも，章を変えるのも，別の作品に移るのも一瞬であり，電子テクストならではの navigation が可能になる。しかし同一作品の内部で任意のページにジャンプするのは，栞 signet 機能を利用してあらかじめ目印を付けている場合以外はできないようだ。

もちろんこのテクスト集成は単なるコレクションではなく，コーパスであることを謳っているのだから，ソフトウエアが付属している。BABEL という名の付属ソフトは，コンコーダンス生成が主要な機能である。一種のハイパーテクスト機能を持っていて，テクスト中の任意の単語を右クリックして《Rechercher le contexte du mot》を実行すると，瞬時にテクスト中の全用例を文脈付きで（たとえば八音綴の作品であれば，当該行とその前後の二三の単語までを）左側の「検索結果欄」にずらりと表示する。この「検索」Recherche は狭い意味での作品としてのテクストだけではなく，タイトル・校訂者名等の付属情報にも有効であるし，利用者が個人的に付けることのできる註記 notes personnelles も対象にできる。さらに便利なのは，利用者が独自に複数のテクストを組み合わせてあらたなコーパスを作ることができ（fusion du corpus），そのコーパスに対して BABEL を適用するのも可能なことである。こうして作成したコーパスの組

み合わせに関する情報は当然保存できる。またこの場合も，検索はほとんど一瞬に行われ，スピードの点ではまったく問題がない。

検索結果の表示は，作者名・作品名・成立年代・刊行年・巻数・ページなどの書誌情報の後に，テクスト中の該当個所が，検索対象の語（群）をマークアップして示される。書誌情報をいちいち付けるのは検索結果を他のアプリケーションで利用する場合の便宜を考えてのことであろうが，*Corpus* の利用法としてもっとも頻度が高いと思われるコンコーダンス生成の場合には，各項目に同一の，しかも既知の情報が3-4行付くことになり，かえって煩わしい。しかも韻文作品の場合，肝心な該当個所の行数がそこに示されていないのは非常に不便である。ただし検索結果欄でマークアップされた該当語（群）をクリックすると，右側のテクスト表示がただちに対応箇所に飛んで，行数も確認できる。該当個所の数（いわゆるヒット数）は検索結果欄の上部に示されている。

こうした基本的な利用法のほかに，BABEL はいくつかの機能を備えている。十分に試しきれてはいないが，条件検索，その場合のワイルド・カードの使用，連続していない複数語の検索 recherche de proximité，イタリック，ボールド等の文字属性を含む検索 recherche typographique，ギリシア文字の検索も可能である。ただし動詞の活用形を網羅するような検索，あるいは複数の表記のある語を一度に拾い上げるあいまい検索などはできない。けれども全体として見ると，文字列検索ツールとしての BABEL はきわめて高速であり，使い勝手も悪くない。テクスト・データベースに付属するソフトとしては，一定の評価ができると思われる。しかしながら問題は別の形で存在する。

「誤植」の存在

検索が正しく行なわれるためには，ソフトが正確に機能するのはもちろんのこと，それ以前にコーパス自体に誤りがあってはならない。けれども電子テクストには印刷本と同じように「誤植」の生じることがある。それも当然であり，一般に電子テクストは手入力かスキャニングによって得られるものだから，誤入力の可能性もあればスキャニング時に一定の確率で生じる誤認識をチェックの段階で見落とすこともおおいにありうる。*Corpus* にもそれがある。システマチックに探したのではなく，たまたま目に付いたものだが，マリ・ド・フランスの『レ』のひとつ，『ギジュマール』に以下のような誤りがあった（電子テクストの元本は Jean Rychner による CFMA 版）。

v. 85 portëuns → portë uns
v. 147 laundë → laundë;
v. 275 aseëre → aseëre.
v. 438 chevalier, → chevalier.
v. 462 s'aër → s'aë

147，275，438行の誤植は句読点に関わるものだから，検索にはさほど影響しないし，特にコンコーダンス作成の場合は無視しうる。しかし85行と462行の場合は検索結果に直接影響する深刻

な誤りといえよう（スペースの脱落は不要なスペースの挿入より害が少ないとはいえるが)。また発生頻度が概して高い。上記の5例は『ギジュマール』のすべての誤植ではなく，全体の半分程度，冒頭から480行ぐらいまでのところで目に付いたものにすぎない。*Corpus*の大きさは512 MBほどである。このCD-Romはファイル構成が分かりにくく，テクスト・データの占める部分がどのくらいの大きさなのか，計算は難しいが，データの入っていそうなフォルダー容量を調べてみると，おそらく全体の三分の一以下であろうと思われる。したがって，もし『ギジュマール』の例と同じ確率で誤植が生じているとしたら，非常に荒っぽい，しかしかなり控えめな計算をしても，約35,000の誤りが存在することになる。これは無視しえない数字である。

改訂の可能性？

誤りは訂正すればいい。けれどもCD-Romに収められているソフト，データの場合，そうはいかない。発売元による更新，改訂に俟つしかないのである。一般にコンピューター・ソフトはユーザー登録をすることが多く，これは顧客管理とセールスという意味でソフト・ハウスにとって，サポートの保証という意味でユーザーにとって意義のあるものだが，*Corpus*の発売元Champion Électroniqueはユーザー登録の制度を設けていない。したがって特にユーザーとの回路は設けられていないのである。Corpusは発売後すでに1年になる。何らかの形でテクスト・データの「誤植」については報告がなされていることと思うが，改訂版の刊行が強く求められるのは当然であろう。

データのエクスポートとコピー

*Corpus*は，検索ソフトBabelをインストールするが，データ部分はいわゆるハードディスク格納型ではないので，CD-Romをドライヴに入れて利用する。けれどもテクスト・データのコピーと他のアプリケーション上のファイルへのペーストができないわけではない。ただし一種のプロテクトが働いていて，テクストの選択箇所のコピーが10回に限られている。また選択のための範囲指定はページの範囲内でしかできない。すでに述べたように冊子本のページを再現した作りになっているのだから，したがってある作品全体をエクスポートしようと思えば，そのページ数だけコピー・ペーストを行なわなければならないし，しかも10回ごとにソフト自体を再起動しなければならない。こうしたプロテクトは不正コピーの抑止を狙ったものであろうから，無碍に批判はできないが，前述のように一定の率で「誤植」がある場合には，テクスト・データの編集のために簡単にデータのコピーができる形にしておいてもらいたいという要望は当然生まれてくるだろう。

*Corpus*を構成する作品の元版の中には入手困難なものが少なくないと思われるが，利用者の多くは，電子テクストを使いたい作品の元版を所有していて，検索のためにのみ*Corpus*を購入したのではないかと思われる。その意味では，仮に*Corpus*のテクストデータが簡単にコピー・ペーストできたとしても，元版の需要にはさほど影響は出ないのではないだろうか。もちろん個人利用を越えた海賊版の作成が企図される可能性はあるし，*Corpus*自体の需要にも影響がある

だろうから，一切のプロテクトを排除することはできないだろう。しかし「誤植」の発生をゼロにおさえるよりは，発生した「誤植」を訂正できるようにしておく方が容易なはずである。そして希望者が共有し，情報交換によって誤りをただし，制度を高めていくことが可能になれば，印刷物というメディアの限界をこえることができるだろう。インターネットという環境では，それが実現するはずである。けれども商品として市場に出す以上，そうした形は取りにくい。製品版のテクスト・データベースの限界はそのあたりにありそうである。拙稿で紹介した「中世フランス電子テクストの会」のような，個人レヴェルでの物々交換による，汎用性のある電子テクストの共有が意義を失わない所以であろう。*Corpus* はそのような意味において，電子テクストならではの機能を備えながらも，まだ限りなく印刷本に近い性格を保っている。そういえばCD-Romの製造工程は印刷に他ならないのであった。

註

1）「テクストの電子化とデータベース構築」（*Lilia Candida* 白百合女子大学フランス語フランス文学論集，N°29, 1999年3月）

2）たとえばTower of Babel（サポート終了），PcConverter（旧ASCII Converter），MacWinText，Text-Converter，「欧字文字化け 修正キット」など（これらはソフト名をGoogle等の検索エンジンにかければ，簡単に所在が確かめられる）。

付記

本論考は，平成14年度科学研究費（萌芽研究）の支給を得て行なわれた研究に基く。