テキストの電算処理

——電子テキストと解析ツール——

小栗栖　等

はじめに

電子テキストをデータベース・データとして扱うことで、いかなる可能性が開けることになるのかについては、別の機会に述べた[1]。その際には、理論的な側面から様々な可能性を論じたが、本論では、実際的側面、すなわち、電子テキストを活用するためのアプリケーションを論じる。

筆者は、HpConc、Numérotation、LEXICA、LEXICA - Maker、FileManager、VersionManager といったアプリケーションを、ここ数年間で公開した[2]。最初の二つは電子テキストを扱い、次の二つは電子データを扱い、最後の二つはファイルを管理する。これらのうち、どう利用したら良いのかわからない、という意見がよせられるのは、電子テキストを扱うアプリケーション、とりわけ、HpConc である。また、アプリケーションそのものではないが、東京都立大学の岡田真知夫氏と筆者が作成している、TCAF（『古仏語動詞活用表』）の利用法についても質問を受けることがある。これは、LEXICA で検索できる電子データである。本論では、特にこの二つのアプリケーションの機能と利用法を紹介しつつ、電子テキストの電算処理一般に関わる諸問題を論じる。

電子コンコーダンスと電子索引

テキストを電算処理することで、どのようなメリットがあるかを垣間見たのは、Duggan の *A Concordance of the "Chanson de Roland"*[3] と Wathelet-Willem の *Recherche sur la Chanson de Guillaume*[4] の全語彙索引を見た時である。いずれの書物も、まだコンピュータが身近になかった時代に刊行された。私がこの書物を見たのも、1980年代後半のことで、ワープロ専用機が百万円以上もする時代のことであった。十数年後に小型コンピュータが家庭に普及しはじめると考えた人は、当時ほとんどなかったに違いない。実のところ、Wathelet-Willem の全語彙索引は、コンピュータを使わずに作られたものであることが、本文に目を通して判明した。が、おそらく、組版にはコンピュータが使われているし、語彙集の精密さは、後の時代にコンピュータを利用して作られたものを思わせる。

ワープロ専用機が普及しはじめた時代に、筆者は最初の電子テキストを作成した。Moignet 版の *Chanson de Roland*[5] を全て手入力したのである。Duggan のコンコーダンスがあるのに、それほど手間のかかることをしたのには、理由があった。当時、筆者は武勲詩の冒頭詩行(vers d'intonation)に関心をもっていた。詩節の第一行目(冒頭詩行)が他の行とは異なった形式的特徴を持っている、というのは多くの研究者が信じるところである。その真偽を確かめようと考えたのである。行の特徴を記号で分類し、冒頭詩行と他の行に有意な偏差が見られるかどうかを確かめるためには、全行が電子化されている方が便利だった。とはいえ、当時のワープロ専用機では、たった4000行ほどのテキストを八つの文書に分割せねばならなかった。また、複数の文書を同時に開くことも不可能だったため、全行を分類した後、ソートするためだけに数週間を要した。現在なら、数秒で終わる作業である。しかし、辛い作業が終わり、論文を書き上げた後[6]、電子

テキストは様々な役に立った。特に重宝したのが、コンコーダンスでは不可能な、後方一致検索が可能になったことだった。通常の検索では、指示された文字列が含まれる単語を、すべて抜き出す通常一致検索しかできない。筆者は、全単語の前後を-で挟むことにより、er- を検索すると、er で終わるすべての単語を抜き出せるようにしたのである。もちろん、当時は、八つの文書にわたって、検索を繰り返す必要があったし、また、複数の文書を開くことができないので、検索結果を記録するには、手書きで行番号を書き写さねばならなかった。

上記の経緯を考えれば、コンピュータと電子コンコーダンス作成ソフトを手に入れた時、筆者がどれほど喜んだかは、容易に想像されるだろう。フリーソフトの Conc は、コンコーダンスと全語彙索引をわずか数秒で作ってくれる。前者は Duggan のコンコーダンスに、後者は Wathelet -Willem の全語彙索引にそっくりの書式である。

当時、筆者は、ラテン語版『聖ウスタッシュ伝』の言語的特徴に興味をもっていた。早速、作品を電子化し、コンコーダンスと全語彙索引を作って、活用しようとした。しかし、期待は完全に裏切られた。索引もコンコーダンスも使い物にならなかったのである。まず、出来上がったコンコーダンスや索引は、画面の上で見るには大き過ぎる書類であった。スクロールするのがひどく煩わしい。さらに困ったのが、目当ての単語を見つけるのが非常に手間だということだった。周知のとおり、ラテン語は語尾が大きな文法的機能をになっている。ところが、たとえば、andum で終わる単語は、コンコーダンスでも索引でも、巨大な文書のそこここに散在している。しかも、それらの単語を見つけるには、巨大な書類をスクロールして目で探さねばならない。止むなく、索引をプリントアウトすると、膨大なページ数になり、しかも、コンピュータ上ではもはや作業できなくなる。筆者がとった解決手段はワープロ時代に帰ることだった。すなわち、すべての単語を-で挟み、後方一致検索を可能にした電子テキストを検索したのである(7)。

HpConc

長々と昔話を続けたのは、思い出にひたるためではない。重要なのは、上記の経験に基づいて開発されたのが、HpConc だということである。今にして思えば、Conc で作ったコンコーダンスや索引がさほど便利でなかった理由ははっきりしている。Conc の書式が印刷本のコンコーダンスや全語彙索引と瓜二つだったということが、正しくその理由なのである。

書籍とコンピュータは静的か動的かという点で決定に違う。

書籍は原則として、最初からすべてを提供しなければならない。すなわち、コンコーダンスでも索引でも、独立した一個の書物とするには、すべての行、すべての単語を網羅していなければならない。一方、コンピュータは、行や単語を、必要に応じて、取り出す能力に優れている。

また、書籍は容量に限りがあるから、データの整理法も極めて限定される。実際、コンコーダンスや索引が可能にするのは、前方一致検索と厳密一致検索のみである。特定の文字列で始まる単語や特定の単語のみを探すには便利だが、一定の語尾をもつ語を網羅的に取り出す（後方一致）には恐ろしく手間がかかる。一方、コンピュータはデータを与えさえすれば、多様な方法で目当ての単語を探し出してくれる。前方一致、厳密一致、後方一致はもちろんのこと、特定の文字列を含む単語を見つけたり（通常一致)、複数の検索法を組み合わせること（複合検索）もできる。

結局のところ、コンピュータは、1)必要なデータのみを取り出す、2）データの整理法を必要に応じ変更する、という点で動的であり、書籍とは決定的に異なる。したがって、コンコーダンス、索引といった、書籍時代の書式にとらわれるのは、無意味どころか、有害なのである。

ワープロ時代から筆者がとった方法、すなわち、全単語を-ではさんだ電子テキストを検索するという方法では、-pro を検索すれば、pro で始まる

単語を、um- なら um で終わる語を、-and- なら and を含むすべての語を見付けることができる。現在のテキストエディタやワープロは、一括検索の機能を備えている。それゆえ、同じ方法をとるとしても、かつてのワープロ専用機を使うより、ずっと手早く目当ての単語を取り出すことができるだろう。だが、問題がないわけではない。たとえば、40-50種類の電子テキストから、pro で始まり、tion で終わる語を、一括して見つけたい場合などが、それである。多くのアプリケーションでは、一括検索では、単語そのものだけでなく、その単語を含む行を取り出す。だから、pro の検索結果を ionem で再検索しても、pro で始まり、ionem で終わる語を見付けることはできない。行内の他の単語が ionem でヒットするからである。

筆者が開発した HpConc は、上記のような問題を避けつつ、複数のテキストを一括して検索する。原理はワープロ専用機時代と変わらない。ただし、電子テキストを直接検索するのではなく、電子テキストの索引を検索し、索引から必要な行を取り出すという仕組みになっている。また、索引自体も検索直後には表示されず、ヒットした見出し語の一覧がリスト表示される。リストは再検索が可能なので、pro で始まり ionem で終わる語も簡単に見つけることができる。索引はリスト上で見出し語をクリックすることで得られる。したがって、必要な単語だけの索引を動的に生成できる。索引は見出し語と参照行番号からなるが、行番号を選択すると該当する行が表示される。リスト、索引表示フィールド、本文表示フィールドは独立しているので、索引表示でリストの内容が、行表示で索引の内容が失われることはない。また、本文表示フィールドには、単語を含む行だけではなく、その前後の行も好きな数だけ同時表示できる。

HpConc はハイパーコンコーダンスの省略であるが、その実体はコンコーダンスでも索引でもない。それらとは異なり、必要なデータを適宜取り出して動的にデータを整理する。HpConc データ自体は、電子テキスト本体と電子テキストの全単語索引からなるが、通常、それらのデータの全体を、利用者が目にすることはない（必要とあれば、テキストは全体表示

が可能)。

ところで、リスト-索引-本文というような段階を踏むと言うと、索引はただ一つの見出し語のみ、本文はただ一行のみ（と前後の行）しか取りだせないように見えるかも知れない。だが、リスト内の全単語の索引を一括生成したり、索引の全参照番号を自動検索することもできるし、索引フィールドや本文フィールドの表示モードを切り替えて、検索結果を累積することもできる。

HpConc では、リストやフィールドの検索結果は、すべてテキストファイルとして書き出すこともできる。設定さえ怠らなければ、本文フィールドの内容は、論文引用に適した書式に自動変換される。また、いちいち単語を入力して検索するのが面倒な場合には、単語をリストに読み込ませて一括検索することも可能である。

筆者は様々な場面で HpConc を利用している。次に、そうした利用法を紹介しておこう。

すでに述べたとおり、筆者は共同作業で『古仏語動詞活用表』を作成している。活用形態そのものは、主な文法書から採集しているが、無批判に形態を収録するというわけではない。怪し気な形態に出くわした場合、可能な限り用例を探し出す。文法書には参照なしで収録された動詞形態が少なからずあるが、50種類以上の電子テキストを HpConc で検索し、実在を確かめた形態も少なくない（逆に、実在しないことを証明できた場合もある)。

テキスト校定でも HpConc は頻繁に利用する。筆者は現在『ロランの歌』の諸写本を校定中だが（オックスフォード本は暫定版を配布中)、とりわけ、ある形態を修正するかどうかといった問題に直面した時や、正体不明の語に出くわした時に、HpConc が役に立つ。

語法や慣用表現を調べる際にも、HpConc は便利なツールとなる。筆者が参加する中世仏文学のメーリングリスト、doc-et-doil で、"donner telle"

という表現の telle の後に省略されている語は何かということが問題になったことがある。その時にも、“colee”ではないかという、筆者の直感を裏付ける用例を、幾つも採集することができた。

HpConc は、本文フィールドの検索結果を再検索できる。慣用表現を調べる際には、これが威力を発揮する。たとえば、donner telle に対して、donner tel もあったはずだとなると、当然、確かめたくなる。この場合、tel を検索して表示させた本文を don や dor の前方一致で検索すれば、donner tel の用例を得ることができる。電子テキストを直接検索するとすれば、連語検索は絶望的に困難な作業となるだろう。語順の変化や、別の単語が二語の間に挟まっている可能性を考慮しなければならないからである。

さらに、HpConc が行番号でも検索できることを指摘しておきたい。研究書を読む際、引用箇所を書籍のなかで探し当てるのは厄介な作業だが、電子化されたテキストなら、HpConc で該当箇所を即座に表示させることができる。また、辞書を引いた場合など、用例の前後の文を確認したい場合は少なくないが、そういう時にも、行番号検索は役に立つ。

HpConc は Macintosh 上でも Windows 上でも使用できる。HpConc データそのものは、HpConcMaker で電子テキストを加工することによって作成できる。HpConcMaker も、Macintosh、Windows のいずれの OS 上でも使用できるが、Windows 版では、別途、Perl がインストールされている必要がある[8]。

電子辞書

電子ブックの登場とともに様変わりしたのは、辞書だろう。ディスプレイ上での読書に否定的な人でも、電子辞書の便利さは多くの場合認めている。電子辞書は、書籍とは異なった書式を採用することで、コンピュータ

の特性を引き出した好例の一つである。書籍の辞書では、コンコーダンスや索引と同様、前方一致と厳密一致でしか検索できないが、電子辞書は後方一致、通常一致、複合検索、さらにキーワード検索が行える。また、複数の辞書を一度に検索することもできる。

辞書検索ソフトには定番があり、Windows 上では DD-Win が、Macintosh 上では Jamming が（現在は Windows 版も活躍している）、ほぼ独占状態を保っている。どちらも安価なシェアウェアであるだけでなく、機能の面で市販ソフトを上回っているからである。

Jamming は市販の辞書だけではなく、ユーザーが作成した辞書も電子辞書として検索できる。ユーザー辞書の歴史は古く、少なくとも十数年以上前から、NiftyServe 上では、「英辞郎」という英和辞書がダウンロードできた。Jamming は、市販の電子辞書なみのスピードで、ユーザー辞書を検索する。そればかりでなく、見出し語そのものではなく、検索語を検索するので、複数の形態から同一の見出し語を探し出すことができる。たとえば、fait や faisons などの活用形を検索すれば、不定法形態の faire がヒットするようにできる。つまるところ、Jamming ユーザー辞書は、速度・仕様の面で、市販の電子辞書と同等の性能を有している。同種のアプリケーションの存在を許さないほどの圧倒的優位を、Jamming が長らく保ったゆえんである。

しかし、研究に利用しようとした場合、幾つかの問題がある。もちろん、それは Jamming の欠点ではない。辞書が単語の語義を調べるためのものだという、当然の前提を、このアプリケーションが有するに過ぎない。しかし、電子辞書という形式が、それを越えた利用価値を持っているのも事実である。

たとえば、カルガリー大学の D. C. Walker 氏は、Altfranzösisches Wörterbuch（A. Tobler & E. Lommatzsch）の全見出し語を電子データ化し、検索できるサイトを公開している(9)。ここでは、「電子辞書=語義を調べるためのデータ」という図式は通用しない。語義は収録されていないか

らである。だが、そのことがこのサイトの利用価値を損なうことはない。とりあえず、ある語が収録されているかどうかを調べることができれば、大部な辞書を手に取るまえに、その辞書を利用するかどうかを決めることができる。Altfranzösisches Wörterbuch に準拠して語彙集の見出し語を整える場合にも、非常に役に立つ。さらに、ある特定の語尾を持つ名詞や形容詞を網羅的に探したりといったことも簡単にできる。

筆者は Walker 氏からデータを譲り受けて電子辞書化した。見出し語とキーワード（品詞）を検索できるようにしたのである。当初、Jamming 辞書に仕立てたのだが、満足できるものとはならなかった。Jamming は、他の辞書検索ソフトと同様、検索結果をまずリスト表示し、リスト上で単語を選択すれば、語義が表示されるという仕組みになっている。そして、語義の方は保存できるが、リストの内容は保存できない。たとえば、ion で終わる名詞を検索する（後方一致とキーワード検索を組み合わせる）ことはできても、その一覧表は保存できないのである。

LEXICA と TCAF

Jamming が辞書検索ソフトとして非常に優れているのは疑いようもない。しかし、語義を調べるという目的から外れたとたん、使いづらくなるのも事実である。筆者がユーザー辞書検索に特化したソフト、LEXICA を作成したのは、そうした理由による。LEXICA は、語義を調べるための辞書だけではなく、Walker のデータのような、見出し語だけの辞書も快適に利用できるようになっている。

LEXICA はユーザー辞書、つまり、個人が自分用に作った辞書の検索を可能にする。とはいえ、その辞書を別の人が使うことも、もちろんできる。実際、筆者が共同で作成している TCAF(『古仏語動詞活用表』)もユーザー辞書であり、Jamming 版と LEXICA 版がある。

TCAF データの本体は、動詞の活用表である。表示は動詞単位ではなく法・時称（tiroir）単位である。たとえば、avoir を検索すると、avoir と法・時称名からなる見出し語が並ぶ。そのうち、avoir ind. prés. をリスト上で選択すれば、avoir の直説法現在の活用表が表示される仕組みである。もちろんことながら、活用形態からの検索も行える。たとえば、et という活用形を検索すると、Jamming の場合、リストには AVOIR subj. prés. という見出し語が表示され、LEXICA の場合、subj. prés. AVOIR 3P という見出し語が表示される。いずれのソフト上でも、リスト上で選択を行えば、AVOIR 一接続法現在の活用表が表示される。LEXICA の場合には、活用表のなかで、et という形態が色文字で表示される。

現代フランス語で、Je suis Marie. を文脈を考慮せずに、「私はマリだ」と訳す学生は多い。suis が être だけではなく、suivre の直説法現在一人称単数形でもあるということを知らないためである。古仏語ではこの手の勘違いが続出する。書記法も活用形態も安定しなかった時代にあっては、一つの形態は、しばしば複数の動詞の活用形態に相当する。たとえば、先ほどの et という形態は、haïr の直説法現在三人称単数形でもある。しかも、現在なら通常一つしか存在しない活用形態に、複数の綴りや形態が対応する。たとえば、avoir の接続法現在三人単数形では、aiet, ait, eit, et, aist などといった具合である。そのため、古仏語の初学者にとって、動詞活用は、常に「躓きの石」であり続けた。TCAF は、第一に、そうした学習上の障害を軽減するために作られたものである。とはいえ、TCAF は、専門家にとっても、便利なツールとなりえる。

まず、テキスト校定を行う際、ある動詞形態を語彙集に収録するかどうか、その形態に註でコメントを加えるかどうかを判断する目安となる。TCAF は文法書に収録された動詞形態を収集したものなので、TCAF に収録された形態である限り、読者は語彙集や註以外のもので、その形態を確かめることができる。TCAF には、時に、かなり特異な形態も収録されている。それゆえ、目安に過ぎないことは強調しておかねばらないが、た

とえば、「よほど特殊な形態でないかぎり、TCAF に収録された形態は語彙集から除外する」といった方針をたてることは可能だろう。

また、TCAF のトレマは、文法書の引き写しではなく、独自の方針で可能な限り一貫性を保って付されている。Foulet と Speer の規則[10]に若干の変更を加えたのが我々の方針である。したがって、動詞活用形態に関しては、TCAF はトレマの使用法に関する一定した指針を提供している。活用形態の音節数を確認するためにも TCAF は利用できるが、形態によっては音節数そのものが変動する場合もあることは断っておかねばならない。

さて、本項では、電子辞書を、語義データの集積ではなく、見出し語の集積にリンクした諸情報の集積と定義し直すことで、生じて来る可能性を説明した。辞書と言うのは、そもそも、見出し語と語義のペアの集積ではないかと思う人があるかも知れない。だが、辞書の見出し語は、語義を見つけるためのラベルの役割を果たしているに過ぎない。実際、単語そのものを見つけるために辞書を使おうとしたとたんに、状況は一挙に困難になる。仏和辞典で日本語の単語に対応する仏語単語を探そうとすることを想像すれば良い。そして、電子辞書でさえも、見出し語がラベルに過ぎなかったのは、Jamming 以外の検索ソフトでも、ヒットした見出し語そのものを一覧表として書き出す機能を持たないことから、明らかである。

最後に、ソフトの仕様に関して、再度まとめておく。LEXICA には Windows 版と Macintosh 版がある。ただし、日本語環境の Windows ではアクサン付きの文字が表示できない（したがって、TCAF の Windows 版ではアクサンは使用されていない。英語版 Windows 用の TCAF も別途用意されている）。LEXICA 辞書は、LEXICA-Maker により作成することができる。Macintosh でも Windows でも利用することができるが、Windows 版は Perl がインストールされている必要がある[11]。

TCAF は Macintosh 版（Jamming 用と LEXICA 用）、日本語 Windows

版（Jamming 用と LEXICA 用）、英語 Windows 版（LEXICA 用）がある。TCAF を利用するためには、メーリングリスト、doc-et-doil に参加する必要がある（参加無料）。doc-et-doil の詳細については、管理者の岡田真知夫氏のサイト Isle d'Avalon[12] で御確認いただきたい。

電子辞書データの概念の拡張——マルチメディア辞書——

LEXICA が、語義に付されたラベルの地位から、見出し語を解放するということはすでに述べた。見出し語と語義が同等の価値をもつということは、電子辞書がよりデータベース・データに近付くということを意味する。実際、見出し語と語義からなるデータを、市販のデーターベース・ソフトで開いてみれば、見出し語と語義の扱いにまったく違いが生じないことがわかる。すなわち、見出し語でデータを並べ直すことができるのと同様に、語義でデータを並べ換えることもできる。語義だけを書き出すこともできれば、見出し語だけを書き出すこともできる。LEXICA は検索機能に特化したデータベースだと言える。実際、LEXICA も、最近のデータベースソフトと同様、マルチメディア素材を扱える。すなわち、文字情報に関連づけられた映像、音声、画像を表示できる。これにより、たとえば、単語の発音を聞かせたり、見出し語と関連する映像を見せたりする、マルチメディア辞書を検索することができるのである。

筆者は目下、二つのマルチメディア辞書を準備している。

一つは、Hilaire Van Daele の *Petit dictionnaire d'ancien français*[13] の画像を、文字情報化した見出し語とリンクしたものである。この辞書は評価が高いものの、現在では入手困難となっている。そこで、誰でも簡単に手に入れることができるよう、電子データ化してしまおうというわけである。とはいえ、辞書全体を電子化するとなれば、五百頁ほどとはいえ、困難が大き過ぎる。そこで、すでに触れたメーリングリスト、doc-et-doil 上で協力者をつのり、見出し語だけを入力することにした。見出し語から

該当ページの画像を呼び出せるようにすれば、辞書として十分に利用できるからである。現在、百頁分のデータが一応の完成をみている。また、四百頁分の見出し語入力が完了している。見出し語だけではなく、品詞まで入力する予定なので、まだかなりの時間が必要だが、ここ一、二年のうちに利用可能にするつもりである（Van Daele の著作権期限が不明なので、完成の時点で一般配付するかどうかは未定)。

もう一つは、ホームページ上で随時拡充している、古仏語音声学の辞書である。音声学では様々な発音記号を使用するが、テキストファイルで発音記号を扱うのはかなり難しい。利用者のコンピュータに発音記号フォントが準備されていなければならないだけではなく、通常のテキストと混在させた場合、様々な不具合が生じる。実際、発音記号フォントと通常のフォントが入り交じったテキストファイルを、Macintosh から Windows へ移植するのは不可能だろう。一方、画像を呼び出すための文字データを準備し、発音の記述そのものは画像として定着すれば、その困難は消滅する。

マルチメディア辞書と言うと、どうしても市販の百科事典のようなものを想像しがちであるが、上記のように、文字情報と文字以外の情報をリンクしたデータは全てマルチメディア辞書である。

画像は、書籍の電子化や和仏混在文書、発音記号フォントと通常フォントの混在文書などが引き起こす障害を大幅に軽減してくれる。索引と本文とを頻繁に行き来しなければならない、音声学や形態学の研究書を、筆者が計画する Van Daele 辞書のような形式で電子化できれば便利だろう（もちろん、著作権には十分配慮すべきだが)。また、現代仏語の動詞活用表をデータベース化し、収録された形態のすべてに発音音声をリンクすれば、仏語学習者に歓迎されるに違いない。校定テキストと写本画像をリンクさせることもできる。LEXICA 辞書は教育と研究、二つの分野にまたがって、多様な使い道を考えることができるツールなのである。

今後の展望

前節までは、いわば、テキストの電算処理という分野における、筆者の活動報告である。この最終節では、これまでに開発したプログラムを、より大きな文脈の中に位置づけ、今後のプログラム開発および、執筆活動の展望を示すことにする。

現在でも、筆者が開発したプログラムの意味に理解を示してくれる研究者は、数少ない。コンピュータに習熟するには、かなりの手間がかかることがその原因だろう。様々な電算処理を実際に行い、困難に直面した経験がないと、あるプログラムが、どのように役に立つのかを理解するのは極めて難しい。プログラムが便利だと感じられるのは、それまで苦労していた作業が、簡単にできるようになった時が多いからである。

しかしながら、コンピュータに詳しくない人に理解してもらえないのは、仕方がないことだとあきらめてしまうのは、怠慢にすぎないだろう。ワープロソフトやデータベースソフトの便利さに気づいて、コンピュータを始めた人は、少なくない。それまでに、コンピュータとは離れた文脈で、文書の処理や、カードの整理に苦労していた人々は、修正が容易で、書籍のように印字ができるワープロソフトや、カードの保管場所に頭を痛めずに済むばかりか、データの並べ替えが簡単にできるデータベースソフトの存在を知った時、それらのソフトの便利さが直感的に理解できたはずである。同様に、LEXICA や HpConc の有用性も、コンピュータ上での作業という文脈だけではなく、文学研究という文脈の中でも説明することができる。

ここで問題とするのは、文学研究の中でも、古仏語校訂本の作成である。LEXICA や HpConc が、校訂本の語彙集の作成と深く関わるのは、論を待たないだろう。HpConc は、語彙集の網羅性への強迫観念を軽減する。テクスト内で使用された全ての単語が語彙集に収録されていると、非常に

便利だということは間違いない。すでに言及した Wathelet-Willem の語彙集や、Bédier の *Chanson de Roland* に収録された、Lucien Foulet の手になる語彙集が、どれほど多くの人に利用されたかは、想像もつかないほどである(14)。しかし、語義や文法的機能により、作品内での単語の用例を全て整理し、その出現箇所を指示するのだから、こうした語彙集の作成には、大変な手間がかかる。HpConc の利用を前提にすれば、全ての単語の出現位置を指示する必要はなくなる。検索しようと思いつきもしないような、特殊な綴りなどだけを、指示しておけばよい。むろん、HpConc により、網羅的語彙集が無用の長物と化すと言いたいわけではない。しかし、あらゆるテクストに網羅的語彙集を付けるわけにはいかないのも事実である。単語の用例が動的に取り出せるのであれば、静的な書籍本の体裁で語彙索引を整える必要性は大幅に軽減される。

LEXICA により、語彙集を電子辞書化できるというのは、誰もが考えつくことだろう。とはいえ、そうした読者の負担軽減は、LEXICA データを利用することの、最終的な結果に過ぎない。たしかに、LEXICA データは、電子辞書を作るためのデータである。だが、データ作成者の負担を軽減するためのデータ形式が LEXICA データの本質だというのも事実なのである。電子辞書を作るためのデータそのものは、テキストエディタでも準備できる。しかし、LEXICA データ作成のための専用ソフト、LEXICA-Maker を利用することにより、データ作成にかかる手間が格段に減少する。

LEXICA-Maker は LEXICA データを作成するための、専用データベースソフトであり、辞書作成の便宜をはかるための工夫が凝らされている。たとえば、データ作成の中途段階でも、擬似的に電子辞書として検索することができる。テキスト校訂を進めつつ語彙集を準備するといった平行作業を行っている場合、それまでに準備した語彙集データをテキスト校訂の作業の中で利用することが可能になる。逆からいえば、テキスト校訂

を進めている途中で、気になる単語があれば、ごく簡単な作業で、そうした単語を次々に電子辞書の見出し語としてエントリーしていくことができるのである。また、品詞名の略語などの一貫性の保持は、かなり気を遣う作業だが、LEXICA-Maker では、一貫性の保持ということを意識する必要もない。品詞名の略語を登録してしまえば、後は、ポップアップメニューの選択で、略語を入力することができる。最後に、LEXICA データは簡単に印刷本の書式に書き換えられる。一つのプログラムを実行するだけで、LaTeX 書式に変換し、ほぼ印刷本の体裁で印字することができるのである。実は、筆者が現在開発中の電子校訂本作成システムは、標準書式の一つとして、この LaTeX を採用している。

筆者が開発中のシステムは、校訂本のデータを、コンピュータ上で合理的に扱いつつ、市販本の体裁でのデータ出力を可能にする。コンピュータ上で扱い易いデータが、印刷した場合、必ずしも読みやすいものとならないのは、多くの人が知るところである。余分な文字修飾や書式設定が全くない、プレインテキストが、電算処理には、もっとも適している。しかし、そのデータをそのまま印字すると、かつてのワープロ専用機以上に、お粗末な体裁となる。筆写が提供しようとするのは、そうしたプレインテキストの電子テキストと印刷本の垣根をなくすプログラムである。むろん、どんな電子テキストでも印刷本のように印刷できるわけではない。簡単なルールに従ってデータを整えねばならない。しかし、そうしたルールに従う限り、ボタン一つ押すだけで、プレインテキストを LaTeX 原稿に変換できるのである。

実を言えば、上のようなプログラムそのものを作るのは、それほど難しいことではない。だが、多くの人に受け入れられるルールを策定することは、それなりの困難を伴う。校訂作業の特殊性を踏まえつつ、複雑になりすぎず、また、常識からも大きくかけ離れることのない書式を、非常に限られた数の文字、すなわち、LowAscii 文字で実現しなければならないから

である。これについては、別の機会に論じることとして、いくつかのルールを紹介しておこう。たとえば、[ ]は写本にない文字を校訂者が補ったことを、( )は逆に写本内の文字を校訂者が削除したことを意味する。最後に¥と｜で挟まれた文字は、略語を解釈した結果だということを示す。

一見話の本筋からはずれるような、電子校訂本作成システムに言及したのは、上に述べたルールに、HpConc が完全に対応しているからである（LEXICA も内部的には、対応している）。たとえば、HpConc で estre を検索した場合、est[re]、estre(l)、es¥tre｜といった形態もヒットするようになっている。むろん、それらの形態は、別々の見出し語となる。ある estre が、そのまま、写本の中に見つかるのか、略語を解釈した結果得られるものなのか、それとも、校訂者の修正の結果なのか、を電子辞書のユーザーが、取り違えることは絶対にない。一方で、est [re] , estre(l) が、LaTeX 書式で印字した場合、そのまま出力され、es¥tre｜の¥｜で挟まれた部分がイタリックで出力されることも、申し添えておこう。こうすることにより、プレインテキストの電算処理とデータの印刷出力結果は、常に一定の対応関係を保ち続けることになる。

HpConc と LEXICA は電算処理のためのツールであるが、単に電算処理という文脈中でのみ存在価値を持つのではなく、それを越えた、文献学的意味でのデータ処理という大きな文脈の中に位置づけられ得るのである(15)。

注

1. 「文学と電子テキスト――電子テキストの可能性、技術的問題について――」（『LUTECE』第30号、大阪市立大学フランス文学会、2002年）
2. http://hp.vector.co.jp/authors/VA026513/
3. Joseph Duggan, *A Concordance of the "Chanson de Roland"*, Ohio State University Press, Columbus, Ohio, 1969. なお、Joseph Duggan 氏は現在、http://ies.berkeley.edu/frenchsp/duggan.htm にホームページを開設している。
4. Jeanne Wathelet-Willem, *Recherche sur la Chanson de Guillaume*, Les Belles Lettres, 1975.

5．Gérard Moignet, *Chanson de Roland*, Bordas, 1989.

6．この作業によって得た結果は次の論文にまとめた。
《Les vers d'intonation de la *Chanson de Roland*》(『TLLMF』第4号　大阪市立大学大学院文学研究科森本研究室、1993、pp.21-30)
「冒頭詩行の存在性格――形式的研究の限界に関する一試論――」(『LUTECE』第23号、大阪市立大学フランス文学会、1993、pp.1-21)

7．この作業によって得た結果は次の論文にまとめた。
「ラテン語版『ユスタッシュ』の言語――その俗ラテン語的な特徴についての覚え書き――」(『TLLMF』第5号、大阪市立大学大学院文学研究科森本研究室、1994、pp.40-57)
「韻文版『聖ユスタッシュ伝』における固有名詞の格変化――「韻文版」とラテン語原本の関係についての一試論――」(『TLLMF』第7号、大阪市立大学大学院文学研究科 TLLMF 研究会、1996、pp.25-32)

8．最新版は現在試用版の段階で、公開は2004年3月末日を予定している。現在のところ、Windows XP 上では不具合があるが、正式公開までには対応する予定。MacOS X には正式公開の際にも、未対応のままになる。Windwows 98、MacOS 9.2 上では現在でも大きな問題なしに動作する。

9．http://www.ucalgary.ca/%7Edcwalker/

10. Alfred Foulet and Mary Blakely Speer, *On Editing Old French Texts*, The regents press of Kansas, Lawrence, 1979.

11. LEXICA の OS 対応状況は HpConc と同じなので、注8を参照されたい。

12. http://www.comp.metro-u.ac.jp/%7Eokadamac/

13. Hilaire Van Daele, *Petit dictionnaire de l'ancien français*, Garnier, 1939.

14. Joseph Bédier, *La Chanson de Roland* (*Commentaires*), Piazza, 1968.

15. 現在、「コンピュータを使ったテキスト校訂」という仮題で、原稿を執筆しており、この最終節で触れた、いくつかのプログラム、およびデータ作成のルールは、すべて、その原稿の中で詳述する。「コンピュータを使ったテキスト校訂」では、スキャナを使って、マイクロフィルムから写本画像を電子化する方法、LaTeX 書式を利用したディプロマティック版の作成法、注釈、語彙集、固有名詞索引、書誌などを含めたクリティック版の作成法を論じる。これらの各部分で扱われるのは、校訂法そのものではなく、校訂を行う際に、効果的にコンピュータを使う方法である。また、紹介されるプログラムは、すべて、筆者が作成した Perl スクリプトであり、原稿を発表する段階では、Macintosh 版と Windows 版の両方を提供する（現在、Macintosh 版の一部のみが完成している)。なお、原稿は、html 形式で作成しているが、最終的な公開形式は未定である。