論　文

コンピュータによる西夏語データベースの構築

小　高　裕　次

０．はじめに

０．０

近年のパーソナルコンピュータの発展には目覚ましいものがある。ＣＰＵ(Central Processing Unit 中央処理装置)および周辺機器にともない、パソコンの相対的な価格はますます低下している。筆者がパソコンを買った五年前と比べると、同じ金額で三ランクは上のシステムを揃えられるようになっている。また、ＯＳ(Operating System 基本ソフトウェア)をはじめ、ソフトウェアの進歩も著しく、パソコンをより使いやすく、身近なものにしている。

０．１

パソコンの利用で最も身近なものは、本誌でも行なわれているようなＤＴＰ(Desk Top Publishing)であろう。下書きと清書の区別がなくなり、レイアウトや校正をディスプレイ上で行なうことが可能になったことで、印刷までの行程が大幅に簡略化された。また、フロッピーディスクやハードディスクに保存された文書はコピーや加工が容易なため、引用や加工が楽にできるようになった。これだけでも論文作成支援ツールとしてのパソコンのメリットは計り知れない。ワープロ専用機では扱えなかった非ラテンアルファベット系文字でも、パソコンなら、現在使用されている言語で文字を持つものであれば大抵が扱えるようになりつつある。

コンピュータによるデータベースの構築も、ＤＴＰに匹敵するほど有用なパソコンの利用法である。コンピュータデータベースの一番のメリットは、検索が高速かつ容易なことである。例えば、カード型データベースで 1,000 件程度の検索なら、数秒で可能である。また、条件の組み合わせかた次第では、かなり複雑な内容の検索もできる。この機能を利用して、現在では様々な文献のコンコーダンス作成に威力を発揮している。また、文献を電子化することで、個人が入力したデータを他の人が共有したり、逆に、大勢で手分けして文献を入力し、それを持ち寄ることによって大きなデータベースを比較的短期間に作成することも容易である。研究者は、データの収集よりもその分析により労力を注ぐことができるようになるのである。

０．２

このように便利なコンピュータであるが、筆者が研究している西夏語の分野では、コンピュータを利用する環境はほとんど整えられていない。未だ西夏文字をコンピュータ上で表示できるソフトは実用化されていないし、西夏語文献のデータベースも作成されていないのである。そこで、筆者は西夏語研究の傍ら、西夏語をコンピュータで扱えるようにするために個人的に試行錯誤を繰り返してきた。現在、筆者のシステムはまだ未完成で不十分なものではあるが、それなりに機能しはじめている。本稿では、いかにしてコンピュータ上で西夏語を扱うかという問題について、筆者の経験をもとにして述べていきたい。第１節では、現行のシステムを変更せずに西夏語のデータベースを構築する方法を探る。第２節では、西夏文字そのものをコンピュータ上で表示する方法を探る。第３節では西夏語のコンピュータ処理が進むべき方向について筆者の意見を述べる。

１．西夏語データベースの構築

１．０

西夏語をデータベース化するに先立ってまず考えなければならないことは、コンピュータ上で西夏語をどのように扱うかということである。西夏語は約 6,000字からなる独自の表語文字体系を持っているからである。

１．１

まず考えられるのは、西夏語をアルファベット表記する方法であろう。例えば、シュメール語は約 600字からなる楔形文字で表記されていたが、学術論文では現在完全にアルファベットによって表記されており、そこで楔形文字を見ることはない[1]。西夏語も同様にアルファベット表記できればコンピュータで容易に扱えるようになるはずである。

けれども、西夏語のアルファベット表記には、いくつかの問題点がある。

まず、一番大きな問題は、研究者によって西夏語の推定音価が微妙に異なるという問題である。例えば、平声36韻·上声33韻の推定音価の代表的な例を挙げると、西田(1989)では -eɦ,-ʷeɦ,-eɦ²、Софронов(1968)では -ɪn、李范文(1986)では -e,-ẽ、黄振华(1983)では -iɛとなっている。標準となる推定音価とその表記方法が確立されるまでにはまだ時間がかかりそうである。

また、仮に表記方法が確立されたとしても、西夏語には13の普通母音（非緊喉母音）と12の緊喉母音があり、初頭子音だけでも30以上あるため、ラテンアルファベット26文字だけで表記することは困難であると思われる。

さらに、同音異義語の問題が挙げられる。四声を取り除けば場合によっては 100以上の同音異義語(文字)のある漢語ほどではないが、漢語と同様単音節語を基本とする西夏語にも同音異義語が多い。例えば、西田氏によって *pʷɪɣ (平声59韻·上声52韻)という音価を与えられている文字は、𗦳《威》·𗦴《量る、測る》·𗦵《尊ぶ》·𗦶《冠》·𗦷《宮殿》な

ど、8文字もある。西夏語の二つの声調である平声·上声の区別を行なっても、アルファベット表記だけではわかりづらいものとなる。これは、ローマナイズされた日本語や漢語を読むときのもどかしさを思い出せば、容易に理解していただけるであろう。

1．2

表語文字であり、しかも再構成音価が研究者間で統一されていない西夏文字をコンピュータ上で扱うには、文字を最小の単位としてとらえるほうがより実用的であると思われる。

そこで考えられるのは、西夏文字を数字によって表示する方法である。例えば、李范文氏の『同音研究』では、西夏人の書いた韻書『同音』に使用されている文字 5,814文字を部首別・画数順に並べ、通し番号をつけているので、この番号によって西夏文字を表わす、といったやり方である。例えば、𗏁 miɦ《否定の助詞》は、0624で表わされる。また、本来データベースのためのものとして考えられたものではないが、郑张尚芳氏は、西夏文字を構成要素に分解し、分解した部分を0から9までの数字で表示し、それを並べて一つの文字を表わすという方法を提案した[2]。それによれば、先に挙げた否定の助詞 𗏁 miɦ は、[11-3748·63]と表示される。

しかし、これらのやり方は、アルファベット方式以上にわかりづらいものであることになりそうである。前者は約 6,000文字分全ての番号を記憶することなど到底無理で、対応表を片手にコンピュータに向かわねばならないだろう。後者は西夏文字を分解して数字で表わすことは比較的容易だとしても、その逆の作業を頭の中だけで行なうためにはかなりの習熟が必要で、ディスプレイを前に紙と鉛筆で格闘するような状態がかなりの間続くのではないだろうか。どちらのやり方も大量の文献を処理するのには不向きであると思われる。

1．3

現段階でもっとも実用的なのは、表語文字である西夏語を同じく表語文字である漢字で代用する方法であると筆者は考えている。例えば、𗂧 mɯɦ《天》なら「天」、𗥃 ʈʷɨ《地》なら「地」、というように西夏文字を漢字で表示するのである。先に挙げたアルファベット化、数字コード化の二つの方法に比べてこの方法が優れているのは、目で見て直感的に理解することができ、漢字文化圏の人々には記憶しやすいという点である。また、欧米など非漢字文化圏の人々であっても、西夏研究者は漢語に堪能な人ばかりなので、彼らにとってもそれほど違和感はないはずである。この方法は中国でもすでに行なわれており、西夏語原文と漢語訳との間に逐字訳のかたちで西夏文字一字一字を翻字したものが付けられた論文が中国ではよく見られる。もちろん、西夏語と漢語は異なる言語であるから、西夏文字と漢字が必ずしも一対一で対応しているわけではない。そのことを忘れた逐次訳法の濫用は、西田氏(1957,1995b,1995c)の御指摘のように「西夏語は、漢語とまったく同種の構造をもった言語であって、その残存文献は、直ちに機械的に、解読し得るというごとき曲解をも招くことになる」(西田1957)恐れも確かにある。けれども、逐次訳法の限界を再

認識し、範囲を限定して使用すれば、この方法は大変有効である。逐字法には既存のコンピュータのシステムをなんら変更することなく利用できるというもう一つの大きなメリットがあるからである。

1．4

筆者は実際にこの逐次訳法を使って、市販のソフトを利用した個人的なデータベースを作成し、西夏語の研究に役立てている。けれども、西夏文字を漢字に置き換える作業をしていくうちに、様々な問題点が浮かび上がってきた。以下に、その問題点と筆者の解決法を紹介したい。

まず、逐字訳法による西夏語データベース作成の上で最も重要なことは、西夏文字と漢字を一対一で対応させることである。置き換えた漢字からもとの西夏文字を特定できなければ逐次訳法でデータベースの作成をする意味がないからである[3]。けれども、先述のように、西夏文字と漢字は必ずしも一対一の対応をするわけではないので、文字を置き換える際に工夫を必要とすることがある。

一つの西夏文字が状況によっていくつかの漢字に置き換えられる場合がある。例えば、𗍫 nɨ という文字は《至る》という意味を持つが、仏教教典では 𗍫𗅋 nɨ mbow² 《普く観ずる》のように《普く》と訳すべき場合も多い。だからといって、𗍫 nɨ を翻字する際に、その意味によってある時には「至」、またある時には「普」を用いていたのでは検索の際に混乱を生じる。したがって、このような場合は、文脈による意味の違いをあえて無視して、常に決まった漢字で置き換えるという規則を作っておかなければならない。筆者はこれまで仏典を中心に研究していたので、𗍫 nɨ を「普」で置き換えることにしている。

逆に、複数の西夏文字が同じ意味を持つ場合もあるが、この場合はそれぞれに異なる漢字を当てる必要がある。西夏語には系統の異なる二つの語彙層が存在し、同じ意味を持ちながら音形式も字形も異なるものがある。例えば、𗤓 kɑr² と 𗏁 mʷɛ はともに《目》の意味を持つ語であるが、筆者は、データベース上では前者には「眼」を当て、後者には「目」を当てることにしている。また、西夏語の動詞の中には、基本となるA形式と環境によって変化するB形式の二つの形を持ち、それぞれに異なる字形を有するものがある。データベース作成の際には、このＡＢ両形式も区別する必要がある。例えば《与える》という語はA形式が 𗖵 khɪ̌ɔn²、B形式が 𗗚 khɪ̌en である。この場合、二つの形式を「与ₐ」「与ᵦ」と区別することも可能であるが、筆者は、A形式の 𗖵 khɪ̌ɔn² に「与」をあて、B形式の 𗗚 khɪ̌en に「給」を当てることにしている。

その他、漢字への置き換えが困難な西夏文字もある。動詞接頭辞がその良い例である。例えば完了を表わす西夏語の動詞接頭辞は六形式があり、𗹦 rɨr 《完了＋外側へ》、𗍊 ʔa 《完了＋上方へ》などのようにそれぞれが方向の概念を含んでいる。これらの西夏文字に対応する漢字はないので、筆者はこれらを外字で処理している。また、氏族名や地名などの固有名詞を表わす文字や、陀羅尼などの音訳のために作られた文字も漢字への置き

換えが困難である。筆者は暫定的に 𗍫 tsɑñ を「薩」に当てるように漢字で置き換えるやり方と 𗱠 šɪa のように外字を作成する方法を併用している。

さて、本来ならば、このようにして作成した筆者のデータベースを、広く公開して大勢の人に利用して頂きたいのであるが、現時点ではまだそれができる段階には至っていない。というのは、実際の文献を読み、そこに現われる西夏文字について、一対一で対応させられる漢字を逐一検討しながら逐字訳を行なっているので、対応を確定させた文字が西夏文字全体の一割にも達していないからである。また、𗅋 nɨ を「普」に当てたような便宜的な逐字訳や、《与える》という語のA形式 𗗚 khɪɔn² に「与」を当て、B形式 𗍊 khɪen に「給」を当てるような、多少恣意的な逐字訳を行なっている場合もあるため、筆者以外の人には使いにくい点が多いと思われるからである。

2．コンピュータ上での西夏語表示

2．0

1．3節および1．4節で述べたように、データベース構築という目的に関して言えば、逐字訳法は充分実用的である。けれども、この方法には問題もあり、あくまで既存のシステムに手を加えず利用するための便宜的な方法の域を出られない。逐字訳法の問題点を解消し、さらにＤＴＰ化をも視野に入れるなら、どうしても西夏語を直接コンピュータで扱う必要がある。

2．1

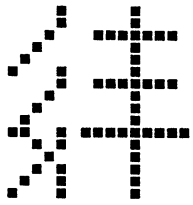
西夏語をコンピュータで扱おうとする試みは、これまでにも行なわれている。その一つが、以前京都大学において西田龍雄氏を中心に行なわれたプロジェクトである。このプロジェクトの経緯については西田(1991)に述べられているが、使用されたハードや開発が行なわれたプログラム言語などの詳細については不明である。また、未確認の情報であるが、北京大学ではMS-DOS上で動作する西夏語ソフトを開発したという話も聞く。筆者も、個人的にではあるが、西夏文字をコンピュータ上で表示し、印刷を行なっている。本稿で例に挙げた西夏文字がそうである。以下に、筆者の西夏語表示･印刷の方法を紹介する。

2．2

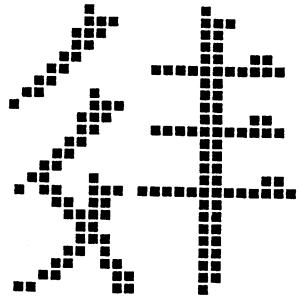
コンピュータ上で西夏文字を表示させるために現在筆者が採っているのは、市販のアプリケーションの一部を変更する方法である。筆者が使用しているアプリケーションはKOA-TechnoMate３中国語(以下　ＫＴＭ３中国語)である[4]。ＫＴＭ３中国語は、多言語ワープロであるKOA-TechnoMate３に中国語ツールを組み込んだものである。ＫＴＭ３中国語は、外字フォントを作成できるだけでなく、もとからある漢字フォントを変更することもできる。そこで、その漢字フォントを西夏文字に変更し、表示･印刷を行なっている。フォント変更のもととなる漢字と、変更後の西夏文字との間には、１．４節で説明した一対一の関係を持たせてある。つまり、「普」を「𗅋」に、「眼」を「𗏁」に、「目」を「𗏒」に

といった具合にフォントを変更している。ＫＴＭ３中国語の漢字フォントには中華人民共和国で用いられている簡体字のフォントと、台湾や香港で用いられている繁体字のフォントの二組があり、それぞれに画面表示用の16×16ドット、ドットプリンタ及び熱転写プリンタ用の24×24ドット、レーザプリンタ用の48×48ドットの三種類のフォントがある。筆者はそれらのフォントの内、繁体字の16×16ドットフォントと24×24ドットフォントを西夏文字に変更している。

図： 𗦟 thaŋ 《仏》の例



16×16



24×24

西夏文字を入力するには、繁体字入力モードにしてもとの漢字の拼音をキーボードから入力する方法、同じく繁体字入力モードでもとの漢字のＧＢ区位を入力する方法、日本語入力モードでもとの漢字を入力し、補助処理メニューで西夏文字に変える方法の３通りの方法がある。図に示した 𗦟 thaŋ を表示させるには、漢語「仏」の発音である「fu」を入力し、変換させるか、「仏」のＧＢ区位、2380を入力するか、日本語で「仏」を入力し、ＫＴＭ３中国語の補助処理メニューから日本字→繁体字変換を選び、「𗦟」に変換させればよい。第三の方法は一見手間がかかるようであるが、データベースソフトに入力したものをコピーして一括変換させることもできるので、筆者にとってはかえって効率的な方法である。

２．３

筆者は以上の様なやり方で西夏語の入出力を行なっているが、問題点も少なからずある。まず、ＫＴＭ３中国語というワープロソフトの中でしか西夏文字を表示することができないという点が挙げられる。データベースソフトでは西夏文字を使えないため、漢字に変換しなければならないが、その際には、逐字訳したものを一旦テキストファイルに保存し、ワープロソフトを終了させてからデータベースソフトを起動するという煩雑な手続きが必要である。MS-DOSには複数ソフトの同時起動ができないという制約があるからである。また、ＫＴＭ３中国語はNECのPC-98とその互換機でしか使用できない。そのため、NECのシェアが低い海外では使用できないのである。

３．コンピュータによる西夏語処理の将来

３．０

第１節·第２節の問題点に示したように、筆者が現在使用している西夏語データベースおよび西夏語表示のシステムは、一言で言えば筆者一人しか使えない「閉じたシステム」である。そこで、第３節では、誰もが使える「開かれたシステム」について考えていきたい。

３．１

西夏語ソフトを動かすＯＳは、シェアの面から考えると、Windows がベストであろう。もちろん「西夏語専用ワープロ」など無意味で、他言語との混在が絶対条件である。また、ワープロソフトだけでなくデータベースソフトの使用も考えると、日本語ＩＭＥ(Windowsにおける日本語入力システム)のような形態、「西夏語ＩＭＥ」とも言うべき形態にすることが望ましい。

「西夏語ＩＭＥ」に必要なものは、西夏語のフォントと、入力システムである。

多言語混在という面からみれば、西夏語フォントには独自のコードを持たせて日本語や中国語と切り替える方法が理想的であるが、技術的に可能であるのかどうか、筆者には判断できない。フォントを日本語Windows または中国語Windows のフォントと同じ規格にしておけば、Windows への組み込みも容易で、２節で紹介した筆者のシステムのような方法で西夏文字の表示ができるはずであるが、フォントの切り替えができないアプリケーションでは文字が化ける恐れがあるだろう。

入力システムについては、アルファベットによる変換を基本とし、複数の方式を併用できるようにすることが望ましい。アルファベットによる入力は、あくまで西夏文字入力のための便法と考え、論文などでは音価を別に表記することを前提として、ラテンアルファベット26文字のみによる入力としたい。母音については、西田氏の簡略表記(西田1991) [5] を基本とし、緊喉母音·非緊喉母音の区別は行なわないことにする。緊喉母音·非緊喉母音の区別をなくすと、変換候補が多くなり、煩雑になることが予想されるが、熟語辞書を充実させることによってある程度は解消できるはずである。子音については、現在荒川慎太郎氏(京都大学大学院)と筆者の二人で入力のための簡略方式を検討中である。併用する他の変換方式としては、番号による変換や部首による変換を考えている。前者は、先述の李范文氏による韻書『同音』の通し番号が最も多くの西夏文字を収録しており、これが適当であると思われる。また、後者は、西田氏による「西夏語小字典」(西田1966　に収録)を基本に部首の追加を行ない、部首の番号と画数で入力できるようにしたい。例えば、　[illegible] thaɦ 《仏》なら204-04、[illegible] khĭɔn^2 《与える》 なら261-04という具合にである。また、これとは別に、中国で行なわれている漢字入力方法のように、西夏語を筆画に分解して入力する方式 －例えば郑张氏の方式のようなもの－ も、中国の研究者には受け入れられやすいかもしれない。

4．おわりに

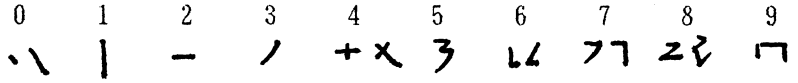
以上、筆者が数年来試行錯誤を繰り返しながら使用してきた方法を中心に、西夏語のコンピュータ処理について述べてきた。筆者の方法が常に既存のアプリケーションを利用したものであるのは、筆者がプログラミングに関して素人だからである。そのため、第３節では Windows への移行について述べてはいるものの、現状では筆者の手には負えない問題である。西夏語のコンピュータ処理、西夏語文献のデータベース化には、今後おそらく国際的な研究者間の協力が必要とされるであろう。特に、西夏語の逐字訳については、国際な統一ルール作りが急務であろう。

現在、西夏語研究者の数は少なく、コンピュータ化の恩恵を直接受ける人は限られている。けれども、西夏語の研究は蔵緬語研究のなかにおいても重要な位置を占めており、西夏語のコンピュータ化によって西夏語研究がさらなる発展を遂げれば、蔵緬語研究全体の発展にも必ず寄与するはずである。

註

1) シュメール語がローマナイズされた形で扱われているのは、一つの文字が複数の読みを持つという理由に依るところも大きい。

2) 郑张氏の方式では、西夏文字の字形を



の十種類に分類し、偏や傍ごとに数字で表示する。また、偏と傍の間は"·"で区切り、冠と足の間は"/"で区切る。

3) 中国で行なわれている逐字訳はデータベース構築を意識したものではないため、この点についての考慮はなされておらず、複数の西夏文字を同じ漢字で訳出している場合もあるので、残念ながらそのままデータベース化することはできない。

4) 本稿に記載したアプリケーションは、各企業が権利を有する登録商標·商標であるが、本稿ではそれらについて一々明記しない。

5) 西田氏の簡略表記は、次のようなものである。

ɪ → y　ʊ → v　ə→ö

ɨ → ii　ʉ → uu　ɛ→ee　ɔ→oo

ɑ → aa

参考文献

克恰諾夫(Кычанов, Е. И.) 李范文 罗矛昆(1995)：『聖立義海研究』宁夏人民出版社，银川

陈炳应(1995)：『貞觀玉鏡將研究』宁夏人民出版社，银川

李范文(1986)：『同音研究』宁夏人民出版社，银川

西田龍雄(1957)：「西夏小字刻文」村田次郎 編『居庸関』京都大学工学部，京都

—— (1964, 1966)：『西夏語の研究－西夏語の再構成と西夏語の解読』Ⅰ，Ⅱ，座右宝刊行会，東京

—— (1975, 1976, 1977)：『西夏文華厳経』Ⅰ，Ⅱ，Ⅲ，京都大学文学部，京都

—— (1986)：「西夏語『月々楽詩の研究』」『京都大学文学部研究紀要』25，京都大学文学部，京都

—— (1986)：「西夏語動詞句構造の考察」『京都大学文学部研究紀要』25，京都大学文学部，京都

—— (1989)：「西夏語」『言語学大辞典』第２巻，三省堂，東京

—— (1991)：『西夏文字のコンピュータ処理の研究　－西夏文字字典の編纂を目指して－』(昭和63年度·平成元年度·平成2年度科学研究費補助金(一般研究B)研究成果報告書)京都大学文学部，京都

—— (1995a)：『西夏文字的特性和西夏语的声调变化－西夏文字新考』首届西夏学国际学术讨论会提要，於银川

—— (1995b)：「林英津著『夏譯《孫子兵法》研究』」『東洋學報』第77巻 第1·2号，東洋文庫

—— (1995c)：「史金波・黃振華・聶鴻音著『類林研究』」『東洋學報』第77巻 第1·2号，東洋文庫

史金波 白滨 黄振华(1983)：『文海研究』中国社会科学出版社，北京

史金波 黄振华 聂鸿音(1993)：『類林研究』，宁夏人民出版社，银川

Софронов, М. В. (1968)：*Грамматика Тангутского Языка 1, 2*, Наука, Москва

田嶋一夫(1990)：「中国語文処理と文化　－漢字分解方式の背景と入力方式の課題－」『しにか』第１巻 第２号，大修館書店

郑张尚芳・冯蒸(1995)：『西夏语拟音和宋代汉语西北方音声母的若干问题』首届西夏学国际学术讨论会提要，於银川