漢字文献情報処理研究

第6号

漢字文献情報処理研究会 編
日本中国語CAI研究会 編集協力

好文出版

漢字文献情報處理研究　第6号

目　次

在ベルリン吐魯番漢文文書とその電子化

――その現状と課題・展望――

小口　雅史（おぐち　まさし）

はじめに

今年と来年は「日本におけるドイツ年」にあたり、国内各地はドイツ関係の各種催し物でにぎわう。世界遺産・博物館島の「ベルリンの至宝展」も去る 4 月 5 日から上野の東京国立博物館で、またその後 7 月 9 日からは神戸の神戸市立博物館で開幕となった。これはドイツ統一 15 周年を記念するものでもあるが、ベルリン博物館島でも、東西ドイツ分裂の爪跡がまた一つ、ようやく消えようとしている。統一後 15 年たってもなお、分裂時代の負の遺産は完全には解消されていない。

筆者は 2002 年 9 月から翌年 9 月にかけて、フンボルト大学の客員研究員として迎えられ、一年有余のあいだドイツ連邦共和国の首都ベルリンにあって、彼の地に長期滞在しなければかなわない様々な研究活動に専念する機会を得た。その一つの柱としていたのが、在ベルリンの、20 世紀初めに発掘によって中央アジアからドイツに将来された、吐魯番コレクション（Berliner Turfan Sammlung、以下 BTS と略称する）中の漢文古文書類の研究である。おりしも筆者が渡独した 2002 年は、ドイツ探検隊の第一回吐魯番調査以来 100 年という節目の年にあたっており、ベルリンに到着してすぐに Turfan revisited という国際シンポジウムも開催された。吐魯番出土の古文書は、敦煌出土の古文書とならび、シルクロード時代の研究素材として国際的に著名な古文書群である。言語的にはおよそ 25 種の文字による 15 ～ 17 ほどの言語で書かれた典籍・文書を含むと言われている。筆者は日本古代史を研究する者であるが、上記諸言語のうちとくに漢文文献である、中央アジア出土の同時代の古文書類は、同じ古代中国（律令制）の影響を受けた西端地域を日本＝東端地域と比較史的に研究できるという観点からも、実に重要にして興味深い史料群である。

しかしながら日本にあっては、中央アジア出土の古文書といえば大英図書館のスタイン・コレクション、フランス国立図書館のペリオ・コレクションこそ著名であるものの、在ベルリンの吐魯番コ

ベルリン国立図書館（旧西ベルリン）正面入口

レクションはさほど知られてはいないようである。その理由は、英仏のコレクションに比して、古文書類があまりに断片的すぎるということにつきよう。私も比較研究のためにスタイン・コレクションやペリオ・コレクションに含まれる漢文文書を閲覧すべくロンドンやパリに赴いたが、見慣れたBTS中のそれに比して、事前に知識としてはもっていたこととはいえ、あらためて実物に接すると、完全な古文書の形を保つその見事さにはやはり圧倒された。どうしても日本の歴史学研究者のBTSへの関心は低くならざるをえない。数年前まで、コレクション中の漢文文献の全貌を詳細に教えてくれる整備された目録すら存在しなかったのである。

また長期にわたった東西ドイツ分断も、当然のことながら、その研究の大きな障害であった。統一後の混乱も忘れてはならない。筆者の手元には1991年4月15日付けの朝日新聞夕刊記事「流転――トルファン・コレクション」③の切抜があるが、そこにはこのコレクションの流転の経過と当時の旧東独出身の研究者のとまどいが赤裸々に記されている。もっともここに登場している三人の研究者とは、在独中に何度かお会いする機会があったが、今ではすっかり研究環境も落ち着き、しかもなお第一線で研究しておられることは嬉しいかぎりである。

さて近年になって故百済康義氏（仏教学）や西脇常記氏（中国思想）らの長年の努力によって、ようやくBTS中の漢文文書の本格的な目録が刊行され[1]、その全貌がしだいに明らかになりつつある。ただ実物に即して詳細に検討すると、歴史学を専門とする立場からみて、それらの目録にはなお訂正ないし追加すべき点もあり、今後さらに継続して目録の整理を続ける必要があるように思われる[2]。

またかつては出土品中の小さな文書断片について、それがいかなる典籍の一部であるかについては、担当者に職人技的な該博な知識が求められたが、今や、中国の主要古典籍あるいは仏典は、かなりのものが電子文献化されていて、数文字を頼りにその典籍名を割り出すことはかつてほど困難ではない。このコレクションが小さな断片の集合体であることこそ、電子化が急がれる理由にして、かつまたその恩恵を確実に享受できる理由なのである。今後、これら漢文文書についてそれを電子化した上で詳細なカタログを整備していけば、研究が立ち後れているBTS中の典籍・古文書類の内容分析の進展も大いに期待され、またその結果として様々な分野で新しい研究素材が提供される可能性をも秘めているのである[3]。

回 1.BTSの伝来

BTSは1902年から1914年にかけて4回にわたって実施されたドイツ（当時はプロイセン）隊によって将来されたものである。調査は考古学者・美術史家であったAlbert Grünwedelと、東洋学者であったAlbert von Le Coqに率いられて実施された[4]。ドイツに持ち出した文物としては、第二次世界大戦で焼失した大壁画が著名であるが、文書も多く含まれていた。だが、大戦中はいくつかに分け疎開し、戦後は多くがソ連や東ドイツのものとなった。そうした環境の下、ウィグル語や古トルコ語などヨーロッパ諸語に関わる言語の研究は進んだが、数千点とされる漢文文書は、地元に研究者がほとんどいないため、仏典をのぞけばほとんど手つかずの状態だった。

東西統一後、BTS中の典籍・文書類は基本的にベルリン国立図書館（ポツダム通り、旧西ベルリン）のオリエント部門の管轄下に統合された。ただしウィグル文の文献（片面に漢文文書を有する

ベルリン＝ブランデンブルク科学アカデミー（右から２番目）

ものを多数含む）は、ベルリン＝ブランデンブルク科学アカデミーのトルファン研究部門に預けられている。トルファン研究部門は当初はウンター・デン・リンデンに面したフンボルト大学・旧東ベルリン国立図書館に隣接したアカデミー分室にあったが、今では通りをはさんで反対側のアカデミーの本館（イェーガー通り）内に移転した。東西分裂時代の名残で、ベルリンには何でも複数の似たような施設があるが、国立図書館もそうで、旧西ベルリンのそれは、道路を挟んで向かい合うベルリン・フィルのホールとおそろいの外装で、瀟洒な建物であるが（すぐ近くには富士山をかたどったソニーセンタービルがある）、旧東ベルリンの国立図書館は、戦前からの伝統的な重厚な建物。目抜き通りであるウンター・デン・リンデンは、東西ドイツ統一に貢献し、先に逝去されたローマ法王ヨハネ・パウロⅡ世が、分裂と統合の象徴としたブランデンブルク門から西に一直線に伸びる道路である。

このように、筆者の研究対象である漢文文献の大半は、現在ではポツダム通りの国立図書館と、アカデミーのトルファン研究部門との二カ所で閲覧することになる。欧州では図書館利用が有料であることは珍しくないが、学者の地位が高いドイツ故であろうか（もちろん知人の図書館幹部職員の斡旋があったせいもあるが）、私には一年有効の利用証が無償で交付され、長期にわたって多くの古文書類を実見することが出来たのは望外の幸せであった。BTS 中の古文書類は、経典その他の大きな文書をのぞけば、原則として二枚のガラス板に挟まれて、その四辺を黒い粘着テープで密封されて保管されている。原文書にはドイツ隊の調査時期や採集場所が明記され、ガラス板には記述されている言語によって分類された番号を記したラベルが貼られている。こうしたガラス挟みの手法は日本の正倉院でも「玻璃装」と称して、明治以来、古裂・文書等、断片化しているものの整理法として一般的に通用している。その最大のメリットは、両面から見られることであり、また古文書の扱いになれていない閲覧者が扱う場合でも、比較的リスクが小さいこともあって流行したのであるが、しかしガラスからのアルカリ分の放出や、挟まれた中に汚れ・黴等が発生した場合のメンテナンスの難しさ、ガラス自体の危険性などが懸念され、無条件によいことばかりとはいえない。もっともベルリンは日本とは異なり湿気についてはほとんど心配がない。またガラス板に挟んで立てて収納すると、その集積度は高く、収蔵にあたってスペースの節約になることは確かである。ドイツの研究者の関心が漢文文献に向けられていないこともあって、漢文文献は BTS の中ではもっとも冷遇されてきたものであったといっても過言ではないが、逆にそれ故、漢文そのものを研究対象とする日本人研究者の貢献度は抜群である。故藤枝晃氏を中心とした京都大学・龍谷大学の諸研究者による漢文仏典の整理があり（詳細な仏教関係文書目録は二冊目まで東ドイツ時代に公刊されている）、また既述したように、漢文仏典の整理途上で先年急逝された百済康義氏による漢文文献全体にわたる簡易目録が刊行され、一方で西脇常記氏による世俗文献の目録もドイツで刊行された。近時刊行されたこれら先学による貴重な目録に導かれて、研究者は比較的容易に目的の文書に到達することができるのである。

回 2.日本古代史とトルファン文書

日本とトルファンとは、同じく唐の律令体制の影響を受けた東端と西端の国である。律令制が東西に同心円的に伝播していく過程で、唐から見た二つの「辺境」社会ではその受容にどのような違いがあるのか、逆にどのような類似点があるのか。比較史的にみてこれは大変興味深い問題である。BTS もそうした研究のための重要な素材となりうるものである。

近年、古文書類の画像ファイルによるインターネット上での公開が世界中でいろんな企画として進められており、BTS 古文書も順次画像がアップされつつある（詳しくは次章）。ただし表裏とも漢文の文献についてはまだその準備が整っておらず、現時点ではベルリンに赴くしか文字の表記、文書の状態を知る方法がない。しかしやはり実物

に接するだけの価値はある。例えば発掘品であるために文書が二枚重なり合っていることが指摘されてきたものがあるが、実物に接してじっくりと調べると、中には三枚以上重なり合っているものがある。また経典の中には、横の界線は普通の墨色だが、縦界を極薄の、あるかないかわからないような白い色の線で引いたものがあった。対して我が正倉院の聖語蔵の経典類においても、一般に文書と経巻とを比較すると、経巻は界線の色が淡い傾向にあり、やはり時にあるかないか、といってよいものもあるという。つまり日本とトルファンとで経巻の料紙などには一定の類似性がある。こうした所見は実際に手にとってみなければ分からないことである（おそらく写真でも不可能）。

さらに比較史的に興味深いのは正倉院戸籍・計帳との比較である。いまさら言うまでもなく、中国中原の王朝では秦漢以来、人間の登録主義が完徹している。その一方で、史籍を伝存するがゆえに、その編纂のための素材となった原史料は早くに湮滅するのが通例であって、これはちょうど日本とは逆の関係にある。したがって籍帳類についても中国中原では、実物によらない間接的知見にとどまっていたのが、20 世紀になって一気に始まった内陸アジアの考古学的調査によって籍帳類の実物が発見され、もって比較研究が可能になった。ただ欧州各地に分蔵されていることや、本稿でも触れてきたように必ずしも目録が完備していないことなどもあって、比較研究はまだ十分とは言えない。

それでも古くから両者の比較史的研究はなされてきている。ここで取り上げるのは、上記の 5 ～ 6 世紀のものとして数少ない残存戸籍である 3 点、すなわち①西涼建初 12 年（416）敦煌県籍（British Library S113）・②北涼承陽 2 年（426）高寧県籍（BTS Ch6001v）、③西魏大統 13 年（547）瓜州効穀郡？籍？（計帳？）（British Library S613 背）と、④唐代の戸籍（BTS Ch1034、Ch1212、Ch3810 等）である。これらを日本の正倉院古代戸籍と比較する研究は、早く曾我部静雄氏によってなされている [5]。それによれば①②の特徴として家族数統計記載の存在が、また③の特徴としては田土班給・公課記載の存在が挙げられるが、唐代になると、田土班給記載のみとなり、公課記載が消滅するという（②④は今回私がベルリンにて調査対象としたもので、曾我部氏の挙げた例ではないが特徴としては変わらない）。一方日本の正倉院戸籍の中には家族数統計記載を持つものがある。こうした書式・記載事項の変化から曾我部氏は、日本の古代戸籍が、唐代の戸籍ではなく、それ以前のものに似ていることを主張し、その淵源を唐以前にさかのぼらせ、またもって大化改新詔の造籍記事を事実と見るのである。大化改新詔の評価については別としても、近年、日本古代の律令制を構成する様々な要素において、唐よりも古い時代の制度の流入が指摘されている。この戸籍の様式論もこうした近年の研究を踏まえて、なお追究すべき価値があるように思われる。

ベルリン国立図書館（旧西ベルリン）書庫内トルファン文書の保管状況

3.BTS の電子カタログ作成に向けて

現在、敦煌・吐魯番文書関係の電子カタログとしては、国際敦煌プロジェクト [6] が著名であり、また BTS についても順次、貴重な古文書類の画像ファイルによるインターネット上での公開が進められつつある [7]。しかしそのテキストファイルまで全面的にアップしているところはまだない。BTS の古文書群のうち純粋漢文文書は画像すらまだアップされていないので、筆者は在ベルリンの間に精力的に写真収集を行った。それをもとに電

子的なテキストファイル作成を準備中である。電子的なテキストファイルさえ準備すれば、そのテキストがどういった文献の一部であるか、パソコンによる瞬時の解析が可能となる。というのは中国典籍は、国家的な支援のもと（例えば台湾の中央研究院の漢籍電子文献 [8] など）、驚くべき速さでテキストの電子化が進んでいるからである。また漢文文書のかなりの部分は仏典であるが、いまさら言うまでもなく、大正新修大蔵経をはじめとして、国際的な仏典の電子化もやはりすさまじいペースで進んでいる [9]。

断片文書が一体どういった典籍の一部であるかについては、従来は長年の研鑽を積んだ碩学による神業的な文書の比定に頼るところが多かった。しかしこうした電子化のめざましい進展は、数文字さえ判読できれば、それがどういった文献の一部であるのかパソコンによってつきとめることをだれにでも可能にしたのである。断片的古文書の集合体で、人間の目では正体を突き止めがたいものばかりの BTS の古文書群こそ、電子化にふさわしいのであり、またその恩恵は計り知れない。

いま試みにわかりやすい例を挙げよう。Ch582r という断簡がある。1 行目は右側が半分以上欠けているので判読不能であるが、2 行目は「而專謹戒律執志」と読み取れる。これを中央研究院漢籍電子文献（**http://www.sinica.edu.tw/ftms-bin/ftmsw3**）の、「人文資料庫師生版 1.1」に入って「檢索條件」にそのまま入力して（もちろん JIS コードのまま入力してよい）「執行」ボタンを押すと、一秒もかからずにそれが【大正新脩大藏經】／二〇五九 高僧傳（十四卷）／卷六／義解三／慧永三　の一部であることが判明する。

さらにこうして断片の正体が確認されれば、次のステップとして、断片相互の接続の検討も可能になる。完全に接続しなくとも同一典籍の一部どうしであることが判明する価値は計り知れない。膨大な断片群をなんらかの形でグループ化できる可能性も生じてくる。また私的な文書類だと電子テキストによる判定はもちろん不可能であるが、表裏使用のものが多いので、片面が典籍であると、その典籍面での判定によって、反対側の私的文書類の接続の判定も可能になるケースが考えられる。このようにまずは単純にテキストファイル化するだけでもその効用は大きいのである。

当面の目標はベルリン吐魯番漢文文書の画像とテキストファイルとをセットにした公開である。筆で書かれた古文書類の文字情報を完全に電子的に伝えることはまだできない。そうであるならばテキストファイルはできるだけシンプルにして、それに画像を添えることで利用者の便をはかるのが、少なくとも現時点では有効な方法である。こうすれば、いまなおデジタルテキストに不信感を抱く研究者にも受け入れられやすい。

しかし従来、これを難しくしてきたのは画像ではなく逆にまさにテキストファイル側であった。それはパソコンで表現できる文字の少なさに起因する。しかし近年のパソコンにおける多言語環境の急速な整備進展は、パソコン上での自由にして忠実な文字表記を実現しつつある。Unicode の急速な普及がその第一であり、それを超える文字セットも、まだデファクトスタンダードと言えるものはないにせよ、それなりに整備されてきている。もちろん誤写などに起因するものをはじめとしてなお電子化できない文字があることはあるが、それらは例外として処理できるものであり、もはやデジタルテキストを作成することを躊躇する理由は何もないといえる。

なお BTS 中の文書をはじめとして吐魯番文書を国際的にデジタルテキスト化するこころみがようやく本格化するという情報が最近伝えられている [10]。本稿執筆時点で詳細は未公表であるが、国際的なこうした動きは大いに歓迎したい。その際、相互に協力して、できるだけ効率的に作業が進むことを期待する。

回 おわりに

以上ふれてきたように、BTS 中の漢文文書は、いわばコレクション中の日陰者であった。しかしこれまでふれてきたような電子化を中心とした新しい研究法の導入によって、そこには無限の可能性が秘められているといっても過言ではない。

断片文書故の悩ましさは、その年代比定においてもある。このわずか数文字の断片はいったいいつのものなのか。断片故に紀年部分が残ることはまず期待できない。そこで威力を発揮するのが AMS^{14}C の利用による紙片の年代分析である。BTS 中のある種の古文書断片については、この分析がすでに在フランクフルトの玉井達士氏らの尽力によってキール大学にて始まっている。その成果も今後、注目される。これまで報告されたところでは rund AD 600 und rund AD 850-900 といった結果が伝えられている。

電子化をはじめとして、今後の各分野での研究の進展が期待される楽しみな分野である。

注

[1] 百済康義編『ベルリン所蔵東トルキスタン出土漢文文献総目録（試行本）』（西域研究会、2000）、*Nisiwaki Tsuneki, Chinesische und Manjurische Handschriften und Seltene Drucke Ⅲ Chinesische Texte Vermischten Inhalts aus der Berliner Turfansammlung,Stuttgart:Franz Steiner Verlag 2001*。ただ百済康義「マインツ資料目録――旧西ベルリン所蔵中央アジア出土漢文仏典資料――」（龍谷紀要 21-1、1999）に記されているような事情によって、両者の研究協力がなされないまま、百済氏がご逝去されたことは実に残念なことである。なお中国では栄新江氏による目録化がなされている。「徳国 “吐魯番収集品” 中的漢文典籍與文書」（『華学』3、1998）。

[2] とりあえず私が編集した電子的な暫定目録（CDR 版）として、*Oguchi Masashi*：*Katalog chinesischer weltlicher Textfragmente der Berliner Turfan-Sammlung, 2004* 参照。先行目録を訂正した箇所は、色付文字で表示されるようになっている。たとえば西脇氏の目録での recto と verso の別については今後なお検討の余地があるように思う。

[3] 本稿は拙稿「解読進むトゥルファン漢字文書」（朝日新聞夕刊 2005 年 5 月 2 日）のために準備した素案を、拙稿「古代アジア世界の東と西：在ベルリン吐魯番文書と正倉院文書の語るもの－その研究の歴史と一断面」（『国際日本学』2 号、法政大学国際日本学研究所、2005 年 3 月）を加味して再編成したものである。

[4] *M.Yaldiz, Archäologie und Kunstgeschichte Chinesisch-Zentralasiens, Leiden: E.J.Brill 1987* 他参照。

[5] 曾我部静雄「西涼及び両魏の戸籍と我が古代戸籍との関係－附、課役問題の現状」（『法制史研究』7、1957 年。後に同『律令を中心とした日中関係史の研究』再録）。

[6] **http://idp.bl.uk/**

[7] **http://www.bbaw.de/bbaw/Forschung/Forschungsprojekte/turfanforschung/de/DigitalesTurfanArchiv**

[8] **http://www.sinica.edu.tw/~tdbproj/handy1/**

[9] **http://www.l.u-tokyo.ac.jp/~sat/japan/** や、**http://www.cbeta.org/index.htm** など。

[10] 四年ほど前にも、死語となったトハラ語について、BTS 中の古文書類の電子化が試みられていた。**http://www.bsb-muenchen.de/mdz/dfgprojekte/berlin_turfan.htm** 参照。

「Katalog chinesischer weltlicher Textefragmente der Berliner Turfan-Sammlung」のレーベル面

中国の人文情報処理企業の最新動向

中易中標と創新力博

千田　大介（ちだ　だいすけ）

はじめに

筆者は本誌第２号で中国の古典文献デジタル化の現状について書同文社を中心に紹介し、中国への古典文献デジタル化委託という方法を紹介した。それから４年、中国が急速な経済発展を続ける中、古典文献デジタル化などの人文情報処理ビジネスに関する状況も大きく変化しつつある。

例えば、2003 年の SARS 騒動はオンライン学習に絶好の機会をもたらしたし、2004 年頃より全国各地の図書館による独自のデータベース構築も本格化している。おそらくはこのような状況の変化を受けて、近年、人文情報処理ビジネスを手がける企業が増えてきている。ユーザーの立場からすれば、文献デジタル化委託先の選択肢が増え、提供されるソリューションの種類が増えるのであるから、このような状況の変化は大変望ましいものである。

そこで本稿では、近年、人文情報処理企業としてめざましい動きを見せている中易中標社と、設立からまだ間もないこの分野のホープ・創新力博社とを取りあげ、それぞれの企業の概要、主要な製品、技術的特色などについて、特に中国学情報化に必要不可欠な文献デジタル化業務に注目しつつレポートする。以上を通じて、中国人文学情報処理の最新動向の一端を明らかにしたい。

中易中標

◈ 歴史と概要

我が国において、中易中標という企業名はほとんど知られていない。しかし、本誌の読者であれば大半の人が同社の製品を使っているはずだ。中易は、マイクロソフト中国と提携し、Windows に標準搭載フォント、すなわち宋体（SimSun）[1]・黒体（SimHei）等を提供しているフォントベンダである。また、Unicode 3.2・4.0 の規格書印刷用フォントを提供したのも同社である。中国語フォント、特に最新規格大規模漢字フォントの制作においては世界のトップ企業なのである。

筆者は本年三月下旬と八月上旬の二度、北京北四環路恵新北橋側の北奥大厦 19 階にある中易社を訪問し、同社市場部の藍飛氏より同社の沿革や概要・製品についてお話を伺った。

それによると、中易中標 [2] の設立は、1989 年。主要な業務はフォントおよび IME（すなわち鄭碼 IME）の開発である。同社はこれまでに、漢字十万字を使用可能な出版サポートプラットフォーム『全漢橋 2000』のような大規模漢字システムをリリースしており、近頃、後で紹介する Unicode の Ext.B までの約７万字の漢字の全てをサポートする Windows 向け漢字処理システム「中

易漢神 e」を発売した。また、昨年には西夏文字のフォントと鄭碼 IME とをパッケージングした西夏文字処理システムを発売している。

大規模漢字フォントに強いため、同社のフォントは大量の僻字を必要とする中国の戸籍管理システムや軍事 GPS システムにも採用されているという。

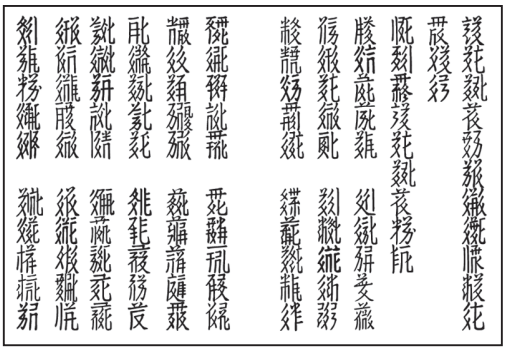
また同社は近年、人文情報処理分野に力を入れており、これまでに開発した製品に、『漢語大詞典』2.1 ネットワーク版がある。香港商務印書館の『漢語大詞典』2.0 版は Big5 コードのみの対応であり、必ずしても全ての漢字が収録されていないのに対して、この 2.1 版は CJK 統合漢字 Ext.B に対応しており、『漢語大詞典』の全文を完全にデジタル化したものであるという。[3]

❖「中易漢神 e」

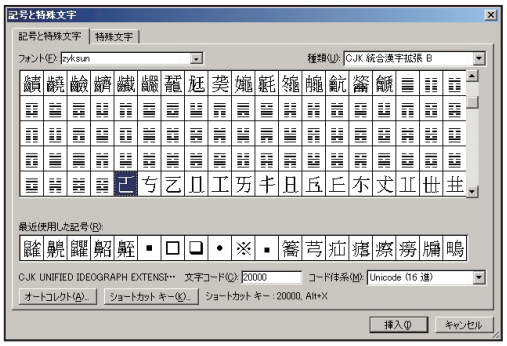
漢神 e は、中易中標社がこのほど発売した Windows 2000/XP 用の大規模漢字処理システムである。内容は、CJK 統合漢字 Ext.A・B 完全対応フォント（zyksun.ttf）と、それらの漢字の入力に対応した IME である。ここでは、その概要を紹介し、機能をレビューする。

漢神 e はセットアップの過程でユーザー認証が必要となる。ネットにつながる環境であれば、まず中易の Web サイト [4] に接続して「注冊」ページでユーザー登録し、製品番号等を記入し、キー ID を取得しなくてはならない [5]。また、漢神 e の IME を使用するには、USB の認証キーをセットしなくてはならない。認証キーを接続しない場合は、Ext.A・B 領域の漢字を入力できなくなる。

zyksun.ttf フォントは、ファイルサイズが約 18Mb。Office の「記号と特殊文字」などの文字コード表で開いてみると、同フォントでは Ext.A と Ext.B が六十四卦をはさんで断続しており、CJK 統合漢字が登録されていないことがわかる。グリフのデザインは SimSun と共通であるから、相互補完して使えばよいということなのであろう。また、Windows においては一つのフォントに収録できるグリフ数は約 64,000 字が上限であるとされるので、その対策でもあろう。

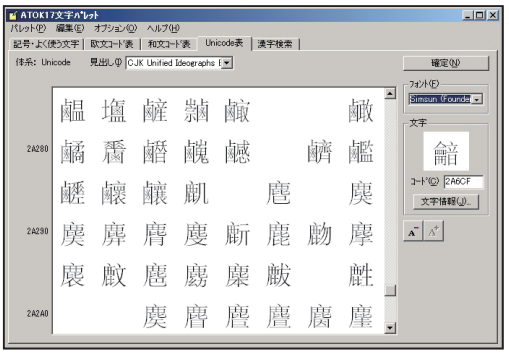


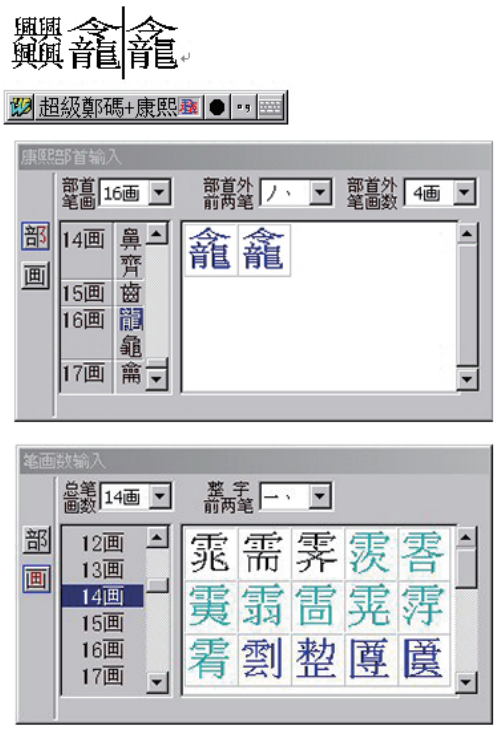
西夏文字フォントの使用サンプル（中易社サイトより引用）



zyksun を記号と特殊文字で開く

zyksun（上）と Simsun (Founder Extended)（下）で Ext.B を表示





超級鄭碼 IME の漢字パレット（上：部首検索、下：画数検索）

グリフデザインは、SimSun と同様、非常に洗練されている。ただし、中国の規格・法規に基づいてデザインされているため、例えば「骨」をパーツとして含む文字は、Unicode の中国提案漢字については全て「骨」、それ以外は「骨」になっているという、いつもながらの問題がある。

Ext.A・B 用の IME として用意されるのが、超級鄭碼輸入法（スーパー鄭コード IME）である。鄭碼とは字形コード化による入力方法の一種で、Windows にも標準搭載されている。開発者の鄭易里氏は、中易社の名誉社長である。

字形コード系 IME は、種類が多い上に入力規則を覚えるのが面倒でもあり、発音記号入力に慣れた日本人にはなかなかなじみにくいものであるが、超級鄭碼 IME には、CJK 統合漢字・Ext.A・B に対応した漢字検索パレットも備わっている。

パレットは、部首検索と画数検索に対応している。部首検索は、部首の画数を指定してリストから部首を選び、部首以外の始め二画の形（一丨丿乙丶）更に部首外画数を必要に応じて指定することで候補文字を絞り込む、というものである。例えば「龕」という文字であれば、16 画「龍」を選び、次いで「丿丶」を選択し、更に部首外画数 4 画を指定する（本来 5 画のはずであるが、データに誤りがあるようだ）。画数の場合は、総画数および初めの二画の形で文字を絞り込む。印刷画像では見にくいと思われるが、超級鄭碼の漢字パレットでは、CJK 統合漢字・Ext.A・Ext.B がそれぞれ色分け表示される。特殊な知識を必要とせずに目的の漢字に到達できる点は、非常に便利である。

問題点としては、漢神 e が Windows のシステムに影響する点がある。漢神 e をセットアップすると、zyksun.ttf 以外の Ext.B 対応中国語フォント、例えば方正超大字符集（Simsun (Founder Extended)）などが使えなくなってしまうのである。両者ともに宋体系フォントであるので、一方だけが使えれば実用上問題ないということのようだが、しかし、例えば Simsun (Founder Extended) を使用した Word ファイルのやり取りなどで問題が発生する危険性があるし、エディタなどの表示では、CJK 統合漢字と Ext.A・B が 1 つにまとまったフォントがどうしても必要になる。また今後、さまざまな書体の Ext.B フォントが登場した際にもなんらかのトラブルの種になりかねない。同社のエンジニア・朱人傑氏によると、これは Windows のバグに起因しており、マイクロソフトがパッチを出さない限り解決が困難であるという。いずれにせよ、早急に改善していただきたいものである。また、漢字検索パレットは便利ではあるが、ウインドウが小さく 15 文字しか表示できない。拡大できるといいのだが。

漢神 e の価格は 2,300 元、三万円強である。海賊版がはびこる中国国内では、正規版ソフトには政府機関や企業などのユーザーしか見込めないために値付けが高くなる傾向があり、この価格もその反映だと思われるが、日本の感覚では、いや中国の感覚でもフォントと IME だけでこの価格は少々高く感じられる。

しかし、zyksun は現在市販されている唯一のフル Ext.A・B フォントであり、しかも SimSun

と同等のデザインレベルを持っているのであるから、ヘビーな多漢字ユーザーであれば購入以外の選択肢は無かろう。

※『康熙字典』全文検索版

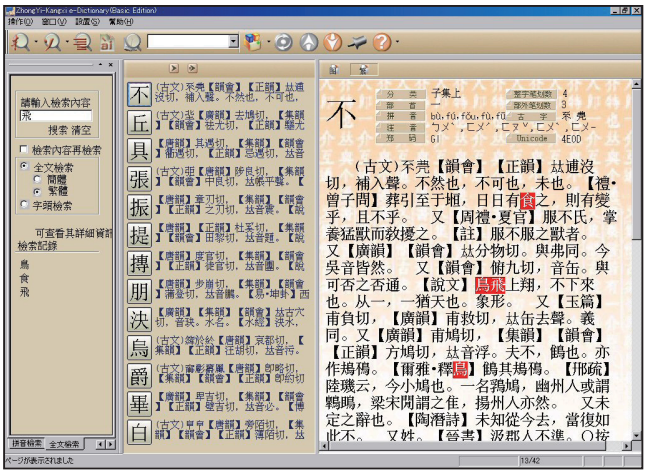
『四庫全書』所収のものを除いた『康熙字典』の単体デジタル版としては、書同文社が開発し、日本語版が三省堂から発売されているバージョンがあるが、本文は画像データであって親字しか検索できない。それに対して、中易の『康熙字典』全文検索版は、全文をUnicode 4.0 規格に従ってデジタル化したものである。

中易版『康熙字典』には、ネット版（企業版）・普及版（個人版）の二種類があり、後者では原書画像のコピーはできるが、本文テキストがコピーできない。電子テキスト化の底本には、北京師範大学の校点本が使用されており、文字コード番号などの情報が独自に付加されている。

中易版『康熙字典』では五種類の漢字検索方法が提供される。即ち、部首・総画数・康熙字典目次・ピンイン／注音・字句検索である。このうち、部首と総画数検索は、超級鄭碼の漢字パレットと同等である。字句検索は、自由にキーワードを設定して検索するものだが、検索結果を対象とした再検索を何度でも繰り返すことができる。引用書名による絞り込み検索など、さまざまな応用が考えられる。本文は簡体字・繁体字で表示可能であり、また検索箇所の原文画像も、ワンクリックで呼び出すことができる。

ただし、異体字同一視テーブルは搭載していないため、全文検索に際しては、例えば「説」と「說」を厳密に区別して入力しなくてはならないといった問題があり、十分に使いこなすには、文字コードの知識、あるいは中国語IME のスキルが必要となろう。

いずれにせよ、中易の全文検索版『康熙字典』の便利さは、従来の画像版に比べて圧倒的である。言語学などを専門としていて『康熙字典』を使いこなす必要があるのであれば、購入の価値は十分にあろう。



中易版『康熙字典』：「鳥・食・飛」と絞り込み検索

※ 文献デジタル化サービス

漢字規格の制定やフォントの制作は、文字学・版本学をはじめとする人文学研究としての一面を持っている。その意味で、中易の業務は本質的に人文学に近しいものであると言えよう。

1995 ～ 6 年には、中易は韓国サムスンの高麗大蔵経デジタル化事業に協力し、フォント・IME などを提供しているが、これが大規模古典文献

中易版『康熙字典』：原文画像と対照表示

データベースに関与した初の例であるようだ。

中易中標の人文情報事業進出の象徴的プロジェクトと言えるのが、中国国家図書館と共同で構築された地方志オンラインデータベースである。2003 年に構築がはじまった同データベースは、中国歴代地方志の総合データベースであり、第一期分として地方志 744 種、総文字数約 20 億字が完成している（外部未公開）。

中易ではこのような大規模プロジェクト以外に小口の文献デジタル化も引き受けおり、日本から発注することも可能である。

中易では文献デジタル化に、手入力と OCR とを併用している。現代の活字本のように版面が明晰であれば OCR、版本などの場合は手入力を使うことになる。筆者は現在、民国時期に出版された排印本を影印したテキストのデジタル化を委託しているが、横線が鮮明でないために手入力を利用するとのことであった。

入力スタッフは常時 300 人ほどそろっているとのことで、百万字単位の文献であれば一ヶ月もせずに完成することができる。Unicode Ext.B までに対応したデータを作成可能で、未収録文字については、フォントベンダの利点を生かして外字を無料で作成してくれる。データフォーマットは、委託側が作成した DTD に基づく XML 形式のほか、原本の版面を再現した PDF の作成も可能である。両方頼んだ場合でも、一千字あたり 2 元ほどの価格差しかない。

デジタル化コストは原本のコンディションによって上下するが、数百万字規模の刻本で 20 元／千字程度になる。規模が大きくなれば、更にディスカウントできるという。また、現代の活字本であれば、OCR が使えるのでもっと安くできるとのことである。日本語にも対応可能とのこと。

刻本のデジタル化であれば、コストは書同文と大差ない。しかし、書同文では文字コードに独自規格の CJK+ を採用しているためにデータ完成後に外字の処理が問題になること、また入力スタッフが 30 ～ 50 人程度と少なく、データ完成までに一年あまりの時間がかかることもあることを考えると、中易のアドバンテージは大きい。

また、中易は地方志データベースのようなオンライン全文データベース構築のソリューションを有している。中易の全文データベースは Windows サーバ上で動作するが、Tomcat ベースの JAVA アプリであるそうなので、Linux サーバなどへの移植も可能であると思われる。技術的には、Web データベースの中易、CD-ROM の書同文ということになろう。

同社が作成したデータの品質などについては、現在筆者が委託しているデータが完成した後、あらためて報告することとしたい。

現在の書同文ホームページ

創新力博

※ 創新力博の創立と書同文の変質

創新力博社 [6] は 2003 年に成立したばかりのベンチャー企業である。しかし、同社の設立には書同文社の動向が大きく関わっている。

現在、書同文のホームページを開くと、ポップな JAVA 広告が表示され、「書同文彩書」のアイコンが目を引くようになっており、従来の古典文献製品とミスマッチなことこの上ない。「彩書」は「彩信」、すなわち中国版写メール向けに、メッセージ添付用中国古典風画像を生成・有償提供するサービスである。書同文は近年、この種の携帯電話向けコンテ

ンツの開発に力を入れているのである。このため、書同文の技術開発は現在この方面ばかりに注がれており、もはや文献処理方面の技術スタッフはほとんど残っていないという。

もっとも、古典デジタル化製品の販売・開発は継続されるようで、昨年来、『十通』『中国歴代石刻史料彙編』といった文献 CD-ROM が発売されているし、現在は『四部備用』全文検索版の開発が進められているという。

このうち『十通』は、『通典』を初めとする歴代の儀礼制度をまとめた十種の書籍のデジタル版であり、歴史学研究においては貴重な資料となろう。ただし、『十通』は中国国内向けのイントラネット版しか設定されていない（日本語版 Windows XP のスタンドアロン環境でも問題なく使用できる）。このため、価格は少々高めである。

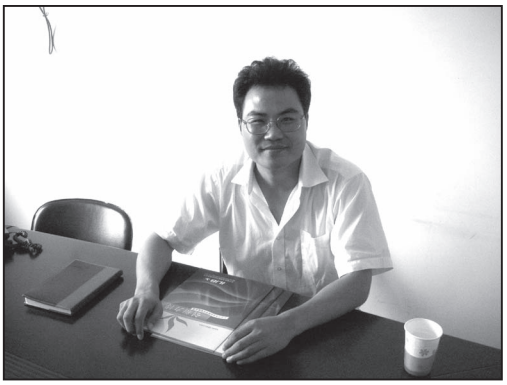
『中国歴代石刻史料彙編』は中国国家図書館が館蔵の拓本資料を歴代の金石史料をもとに編纂した叢書であり、歴史学のみならず文学研究などにおいてもさまざまに利用可能な資料である。こちらはスタンドアロン版も発売されている。なお、両ソフトとも、Windows 2000/XP のみの対応で、Windows 9x 系はサポートしていない点、注意されたい。

これらの文献デジタル化製品の開発は、既存の枯れた技術の応用に過ぎず、例えば Ext.B への対応といった技術改良は放棄されたようである。

創新力博社を設立したのは、元書同文の技術主任・王暁波氏である。王氏は弱冠 32 歳であるが、OCR 技術などを専門とするプログラマで、独力で二年ほどの時間をかけて書同文の文献デジタル化システムを書き直し、機能をも大幅にアップさせたという。それが後で紹介する青典システムであり、また、この技術力と書同文の変質とが創新力博の創立を後押ししたものとのだと言えよう。

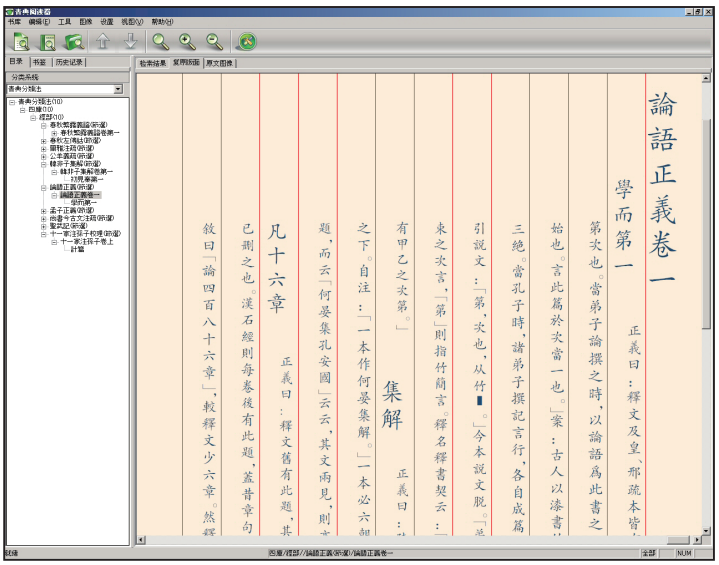
※ 創新力博の技術

設立間もないだけに、創新力博にはまだ目立った文献デジタル化製品がない。おそらく現在香港迪志文化出版社の委託により 2006 年中の完成を目指して作業が進められている『四庫全書』全文検索版アップグレード版が、同社にとって初の大規模製品となろう。これは、従来の『四庫全書』全文検索版のシステムを全面的にブラッシュアップするとともに、現行版未収録の図表類などを補完するものである。注目されるのは、文字コードが従来の CJK+（CJK 統合漢字＋ Ext.A ＋外字）から Unicode4.0 に変更され、Ext.B に対応することである。膨大な『四庫全書』のデータを国際標準規格で利用できるようになるわけであり、研究の際の資料引用や、『四庫全書』データを引用した文書の組み版などに大きな恩恵をもたらすものと期待される。書同文設立の契機ともなった記念碑的製品、『四庫全書』全文検索版のアップグレードを創新力博が行う、このことは同社が人文情報処理企業としての書同文の DNA を受け継いでいることを、如実に物語っている。


王暁波氏

同社の文献デジタル化システムは書同文のものと非常に近い。すなわち、OCR と強力な校正システムによる文献デジタル化、強力な漢字関連づけ検索機能による異体字問題への対応、という方法を採用している。ただし同社の OCR エンジンは、王氏が再設計したもので、書同文のものに比べて効率が 1.5 倍にアップしているという。また、Ext.B への対応も特筆される。

王氏の開発になる同社のデジタル化システムは、「青典数字化系統」と呼ばれ、OCR・校正システム・検索システム・サーバ・電子書籍生成システムなどを一体化したものである。書同文「数碼太師」との大きな違いは、Web サーバシステムを取り



青典ビューワ

込んだ点にあると言えよう。青典システムを導入すれば、独自に文献デジタル化・CD-ROM の作成、さらには Web 上に全文検索システムを設置することも可能になるのである。

◈ 青典閲読器

しかし、青典数字化系統は数十万元という非常に高額なソフトであり、個人ユースには適さない。このため､創新力博では個人向けに青典閲読器(青典ビューワ）を開発している。

青典ビューワでは、XML テキストと画像とを関連づけパッケージ化した、独自形式のファイルを利用する。それを青典ビューワで開くことで、全文検索・異体字関連づけ検索・版式復元 XML の表示･原文画像との対照表示･校定などなど､『四部叢刊』電子版などと同等、あるいはそれ以上の文献処理を行うことができる。また、ローカルに保存した電子ファイルをツリー構造に分類・管理することもできる。全体的なイメージとしては、ユーザーが無償のビューワをダウンロードして、自ら作成したり Web からダウンロードしたりした電子書籍ファイルを閲覧する、という PDF ライクな運用が考えられている。また、Acrobat に相当する電子書籍ファイルの編集に対応した青典ビューワの上位版の発売も計画しているという。

この種のパッケージ化された検索・閲覧システムは、テキストの分析などの用途には適さないものの、最も一般的な用例・事例検索には十分であり、しかも他の Grep ツールなどと比べて技術的ハードルが低く手軽に使えるというメリットがある。それに加えて青典ビューワでは、従来の書同文製品に見られた異体字・同音字などの一括検索機能に加えて、検索語に指定した文字列と 1 ～ 2 文字が入れ替わった類似の文字列を検索するあいまい検索機能があらたに装備されるなど検索機能の実用性は向上しており、実用性が非常に高い。

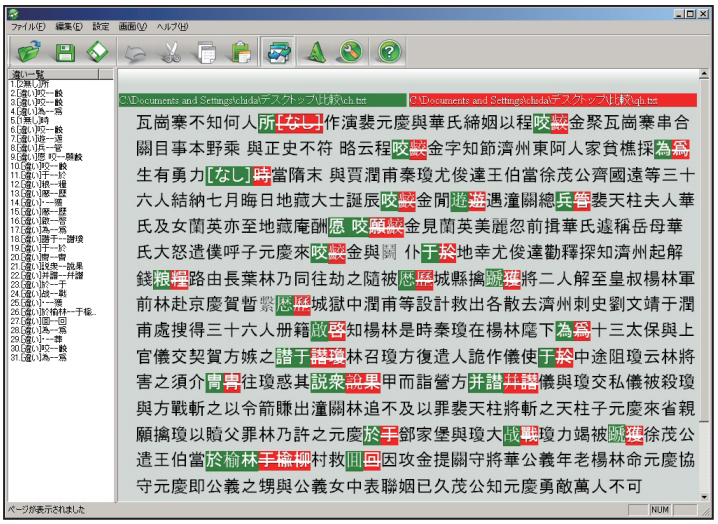
また、文献電子データ作成を委託する側からしても、テキストデータや XML を整理して Web に上げる作業は意外と面倒であり、はじめから Web 配布可能なフォーマットで作成してもらえればそれはそれで便利である。初・中級者にとっては、全文検索が簡単にできるのも大きなメリットだろう。インターフェイスや検索システムの開発に資金をかけられない小規模な個人レベルの古典文献デジタル化に、青典ビューワのシステムは大きな恩恵を及ぼすものと期待される。

できることならば、青典電子書籍ファイルのフォーマットもオープン化し、ユーザーが自由に電子書籍ファイルを作れるようにしてほしいものだが、まだ関連アプリケーションを開発中の段階であり、投資資金の回収という問題もあるため、当面は難しいとのことであった。

◈ 青典版本比較システム

創新力博は電子デジタル化のみならず、人文情報向けソフトウエアの販売も重視しているという。そのようなタイプのソフトとして、現在開発が進められているのが青典版本比較システムである。執筆時点では日本語版も作られており、機能限定試用版が用意されている。

青典版本比較システムは、二つのテキストデータを対校するためのツールである。一方のテキストをベースに、もう一方のテキストとの異動箇所を自動でチェックし、表示してくれるというものだが、チェックと同時にテキストの校正を行うこともできるようになっている。対象はテキストファイルで、文字コードは Unicode・UTF-8 を読み込めることを確認している。単純に文字列を比較するソフトであるから、Unicode で保存されてさえいれば中国語は勿論、日本語や英語、IPA などのテキストであっても、言語に関わりなく比較することができる。ただ、試した限りでは GB・Big5・JIS 等のローカルコードテキストは文字化けしてしまった。オンラインテキストにはローカルコードのものが多いことを考えると、これらの文字コードにも対応してもらいたいものである。



青典版本比較システム

図は、『伝奇彙考』と『曲海総目提要』の同一項目の文を比較したものである。異同箇所は、印刷ではわかりにくいと思われるが、比較元のテキストの字句が緑に白抜きで左に、比較対象テキストの字句が赤に白抜きで右に表示される。一方に字句が漏れている場合は［なし］と表示される。

また、対校にあたって字句の相違とは見なさない異体字のリストを、画面上で編集・保存することができる。

このように青典版本比較システムは、複数の電子テキストを比較し定本を作成する、テキストクリティーク支援ツールとして非常に有用である。また、複数のテキストや版本を比較対照するための支援ツールとしても使えよう。

ただし、複数の版本を比較して継承関係を樹形図化するというような用途に応用するのには少々力不足であると思われる。できうることであれば、今後、版本の異同の数量化と保存・分析、三種以上の版本の比較といった機能を開発・実装していってほしいものである。

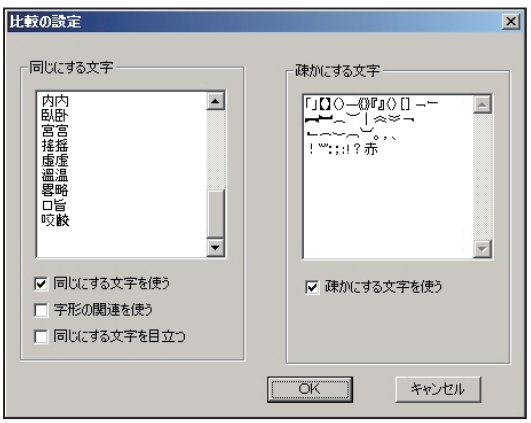
なお、青典版本比較システムは日本向け販売も計画されており、数千円程度の価格で提供できるようにしたい、とのことである。

❖ 文献デジタル化業務

文献資料のデジタル化受託は、当然のことながら創新力博の中心的業務ということになる。

さきにも触れたように、青典数字化系統は OCR によって文献をデジタル化する方法を採用している。基本的なシステムは書同文のものと大差なく、版本の文字を自動で区切り、人がそれを修正して文字を認識させ、同じ文字に認識された

青異体字同一視設定ダイアログ



箇所を抽出して文字毎に校正し、認識結果と原文を並べて対校する、というものである。木版本であっても OCR 認識できるのが強みである。当然のことながら、XML フォーマットや Ext.B、あるいは日本語文献のデジタル化にも対応している。

書同文では 21 元／ 1000 字という価格が一つの目安になっていたが、創新力博ではそれよりも若干安い値段になるとのことである。

回 おわりに

以上、中易中標と創新力博を中心に、中国の人文情報処理企業について紹介してきた。両社はともに文献デジタル化の実績とノウハウを持っており、また Unicode の最新規格に対応するなど、中国における人文情報処理企業の進取の精神とスピード感とを感じさせる。

文献デジタル化の品質については、今後実際に文献デジタル化を依頼して細かいクセなどの検証をすすめる必要があるものの、両社ともに錯誤率一万分の一以下のクオリティは保証されるので、実用的水準はクリアされているものと思われる。

ユーザー側にとっては、さまざまな関連ソフトウエアやツールの充実はもちろんのこと、文献デジタル化を委託できる高度な人文情報処理技術を有する企業の選択肢が広がり、デジタル化する文献の特性や作成するデータの形式に基づき、それぞれの企業の技術的特色やコストを比較検討できるようになったことも大いに歓迎される。

中国学のデジタル化は、工業生産的な大規模データベースによって主導され、もはや唐代以前の文献のデジタル化はほぼ完了している。そのメリットは言うまでもないことであるが、一方で、細かな専門研究ニーズを満たし、かつさまざまな手法による分析に利用可能なフリーテキストの充実が、中国学への情報処理応用の新たな課題として浮上してきている。これは以前から我々が主張しているように、個々の研究者や研究会組織が意識的に取り組んで行かねばならない課題である。その際に、本稿で紹介してきた中国の人文情報処理企業とのコラボレーションは、非常に有効な方法となろう。

側聞するところでは、このほか中国基本古籍庫の開発元である愛如生 [7]、雑誌・新聞のデジタル化実績の豊富な深圳の企業・点通 [8] なども、古典文献デジタル化受託業務に乗り出しているという。それらの企業の訪問調査および紹介は、今後に期したい。

本稿は、平成 17 年度科学研究費基盤研究 (C)「中国古典戯曲総合データベースの基礎的研究」（課題番号：17520237、研究代表者：千田大介）による成果の一部である。

注

[1] 中易社をインタビューした際に、なぜ「SimSong」でなく「SimSun」なのかを質問したところ、この名称は MS 側が指定してきたもので理由はわからない、との答えであった。

[2] 正式名称は北京中易中標電子信息技術有限公司。

[3] 『漢語大詞典』2.1 版は、上海数字世紀網絡社（**http://edu.ewen.cc/**）の委託により開発したもので、販売等は上海数字世紀網絡社が行っている。

[4] **http://www.china-e.com.cn/**

[5] これは、西夏文字処理システムや康熙字典でも同じ。

[6] 正式名称は、北京創新力博数碼科技有限公司（北京 ILIBO デジタルテクノロジー）。以下は、本年 8 月上旬に同社を訪問・インタビューした結果に基づいてまとめたものである。

[7] **http://www.cn-classics.com/**

[8] **http://www.datum.com.cn/**

電子書籍をめぐる状況

野村　英登（のむら　ひでと）

回 1.情報化の三つの位相

古典籍の情報化電子化をどのように実現していくか、人文学研究者の立場からこの問題を考えた場合、大まかには三つの位相を想定すると、議論の切り分けがうまくいくと思われる。

まず第一に、人文学の方法それ自体の再検討を目指す情報学的利用が想定される。第二には、伝統的人文学の方法をより強化するための工具書的利用が想定される。そして第三に、人文学の成果を広く世に知らしめるためのメディア的利用が想定される。

例えば中国学の場合、台湾中央研究院の「漢籍電子文献」[1] や中国で市販された「四庫全書」データベースなどの巨大全文データベースについていえば、まずこれまでの逐字索引の代替としての工具書的利用を目指して構築され（人文学的利用）、利用者の読みやすさを重視したインターフェースが実装されている（メディア的利用）。そして、膨大な用例の抽出が容易になることで、従来の文献読解を行うにあたりリソースの割り振りが変化する。これは量から質への転化がおこっているということで、ある種の情報学的利用の問題となるだろう。

人文学の現況においては、デジタルアーカイブをめぐる一連の動向の中で、いまだ電子化の行われていない文献資料の電子化が次々に計画され、工具書的利用の局面がいっそう増大している [2]。他方、作成された電子情報に情報処理技術の適用による、人文学の新たな方法の確立も目指されつつあり、情報学的利用の局面からの研究も行われるようになってきている。しかし、人文学研究者に意外に忘れられがちなのが、電子化情報化によるメディア的利用ではないか。

そもそも人文学の情報化が謳われる最大の理由は世の中全体が情報化しつつあるためである。従来は研究の成果は紀要や書籍などの公刊により社会に還元されていた。研究者ではないがその分野に興味がある者は、図書館で借りるなり書店で購入するなりして情報収集を行っていた。ところが現在では、インターネットによる情報収集が急速一般化し、情報をインターネットで公開することは社会的義務になりつつある。紙媒体でしか入手できない情報が差別される傾向が出てきたといえる。

ところが、日本の中国学研究についていえば、こうした状況にほとんど対応できていない。例えば「三国志」＋「中央研究院」で Google 検索を行うと、日本語サイトで 8,640 件がヒットし（2005 年 8 月 15 日現在）、趣味で中央研究院のデータベースを利用して『三国志』の現代語訳をすすめているページすらある。もちろんこうした一般利用者は日本人研究者の出版した研究書や翻訳を参考にするだろうし、研究者個々人でインターネットで活躍している例もあるが、学界としての貢献はなきに等しい。

もちろん、本を手にとって学ぶことを勧める教育活動が大切なのは当然だが、社会の情報化は避けようがない以上、データベースの構築などの情報化電子化は、単に研究のためのものだけではなく、より広く研究成果が利用されることを通じ、人文学のプレゼンスを高めるための情報発信メディアとして位置付けていくべきではないだろう

か。ワンリソースマルチユースが電子化のメリットである。学術研究に資するだけでなく、一般利用にも供することを想定した計画が望まれる。

回 2.電子書籍の可能性

メディア的利用としての情報化を考慮した場合、一般に注目されている電子出版について検討する価値は十分にあるだろう。広義の電子出版には、CD-ROM の販売、Web ページによる情報提供、電子メールによるニュース配信、IC 電子辞書などまで含まれるが、本稿では、従来の書籍出版を電子化したもの、いわゆる電子書籍をとりあげることとする [3]。

電子書籍は、市場規模では既存の出版にはるかに及ばないが、パソコンや PDA、携帯電話、専用の情報端末、電子辞書などで様々なかたちで利用が出来るため、出版業界の低調から新たな市場として期待されている。また注文を受けるたびに必要な部数を印刷するオンデマンド出版とあわせて絶版対策としても有効であるとされている [4]。

もっとも、電子書籍のフォーマットはたんにテキストを読むことまでしか想定していないので、それ自体では多くの情報を盛り込んだリッチな電子化を行うことはできない。したがって人文学からの利用については、まず研究用に十全なコンテンツを作って元のデータは保存しておいて、そのデータから自動的に電子書籍の形式にダウングレードできることが望ましい。実際、国文学分野については、すでに青空文庫 [5] のようなフリープロジェクトが電子書籍とのコラボレーションをすすめており、電子書店パピレス [6] の携帯電話サイトでは、青空文庫で公開されている著作権の切れた文学作品のデータを携帯用の電子書籍リーダーに対応させて無償配布サービスを行っている。

また中国学、それも古典学研究の立場からすると、電子化の最初にして最大の課題は漢字があるかどうかである。日本においても漢学の伝統がある以上、多漢字電子書籍の実現の可能性も重要な課題であろう。

レイアウトについては原典のレベルをそのまま再現することは画像でも用いないかぎり難しい。端末にあわせて可逆的にレイアウトが変化することで携帯性と視認性の両立を図れることが電子書籍のメリットなのであまり拘泥する必要はないだろう。むしろレイアウトまで考慮された元データの形式から容易にダウングレード可能かどうかを検討した方がよい。ただ、ある程度訓点が表現できることは考慮すべきではあろう。

回 3.電子書籍のフォーマット

さて、現在流通している電子書籍の形式は以下の通りである [7]。

- テキスト形式
- PDF 形式／ Adobe Reader
- Adobe eBook 形式／ Adobe Reader
- ドットブック形式／ T-Time
- XMDF 形式／ブンコビューア
- ebi.j Book Reader 形式
- デジブック形式／蔵衛門 2005 デジブック
- Kacis Book 形式／ Kacis リーダー
- シーモア形式／シーモアリーダー
- Hatch 形式／Hatch ビューア／Σ Book
- BBeB Book 形式／ LIBRIe

実に乱立といった感がある。このうちすでにあるデータからの変換ができ、かつテキスト検索を可能にするものが望ましい。専門的な書籍においては、単に内容を閲覧するだけでなく逐字検索など工具書的機能があることが望まれるであろうことは通常の出版事情からも明らかである。

まず、テキスト形式は単なるプレーンテキストであるため検討するまでもない。

Kacis Book 形式やシーモア形式は PC や PDA でしか利用できない。ハードが PC に限定されているなら、PDF で十分ではないかと思われる。

また最終的に画像データで配布する ebi.j Book Reader 形式や Hatch 形式、画像主体のデジブック形式も配布後に利用者が検索目途の利用ができず問題である。Hatch 形式／Σ Book はテキスト

の読み込みが可能とされるが、これはテキストデータを画像に変換した上で取り込んでいるため、テキスト形式のメリットは失われている。

したがって、人文学のニーズに応えうる電子書籍の候補としては、PDF 形式、ドットブック形式、XMDF 形式、BBeB 形式が残ることとなる。

回 4.電子書籍の多漢字処理

さてここでは、筆者が専門とする中国学でもっとも重要な課題である、どの程度漢字を使えるのかに注意して、PDF／Adobe eBook 形式、ドットブック形式、XMDF 形式、BBeB 形式の各電子書籍フォーマットを比較してみたい。

❖ 4.1 PDF／Adobe eBook 形式

まず、PDF ／ Adobe eBook 形式については、特に説明を要さないだろう。紙媒体での印刷物について、版下を PDF で作成する一方、そのデータをそのままインターネットでも公開することは、現在では情報公開の標準的な作業である。Adobe eBook 形式は、PDF に著作権保護処理をほどこしたものである。

PDF の作成については、Adobe Acrobat のような正規のオーサリングツール以外にも有償・無償を問わず多くのツールがある。

PDF は多言語対応に加えてフォントの埋め込みも可能なので多漢字環境としては問題がない。ただ電子書籍として考えた場合、PDF は本来他のソフトウエアで作成したコンテンツを電子印刷するものなので、端末上での可読性は必ずしも高くはない。

❖ 4.2 XMDF 形式

XMDF（モバイル・ドキュメント・フォーマット）はシャープが推進している電子書籍フォーマットで、WindowsPC やザウルスなどの PDA から携帯電話 [8] まで専用ビューワが用意されている、もっとも広汎な利用を期待できる形式である。XMDF は、中間フォーマットに XML を採用して、そこから閲覧用バイナリデータを生成している。このため同様に XML や類似の方法によりマークアップした元データがあれば、半自動的に電子書籍の作成を行うことが期待できる。また XMDF はフットプリントによりコンテンツの発行記録を残すことで、ゆるやか著作権保護機能を実現している。

XMDF のコンテンツ作成には専用のオーサリングツールが必要で、現在は商用を目的として出版社など業者向けの販売のみで、個人での購入は難しい。

XMDF では文字コードに Unicode（UTF-8）を採用しているので、多漢字電子書籍の作成が期待できる。しかし、シャープ株式会社 SST 推進センターに問い合わせたところ、フォーマットの仕様としては文字フォントさえあれば Unicode の文字は表現可能だが、現在は再生環境が存在しないため、オーサリングソフトサイドで JIS 外の文字が来たときはエラーを返す実装にしているとのことである。Unicode 対応のリーダーソフトの開発から行えば、多漢字電子書籍の可能性はある訳だが、その道のりは遠いといったところだろうか。

図 1　XMDF の表現力（SpaceTown ブックス－ XMDF の機能 [9] より）

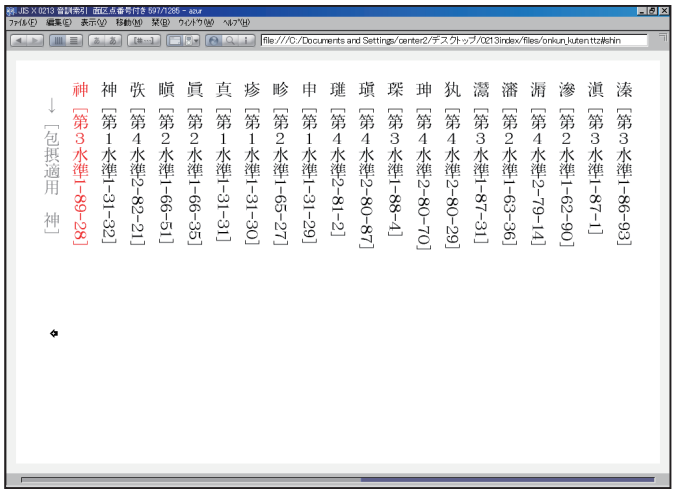


図２　azur の第三、四水準文字の表示例 [13]

なおレイアウトに関する表現力は図 1 の通りである。

※ 4.3 ドットブック／T-Time 形式

ドットブック形式 [10] はボイジャーの推進する電子書籍フォーマットで、基本的には同社のテキストビューワ T-Time の専用フォーマットである TTZ に著作権保護処理を行ったものである。

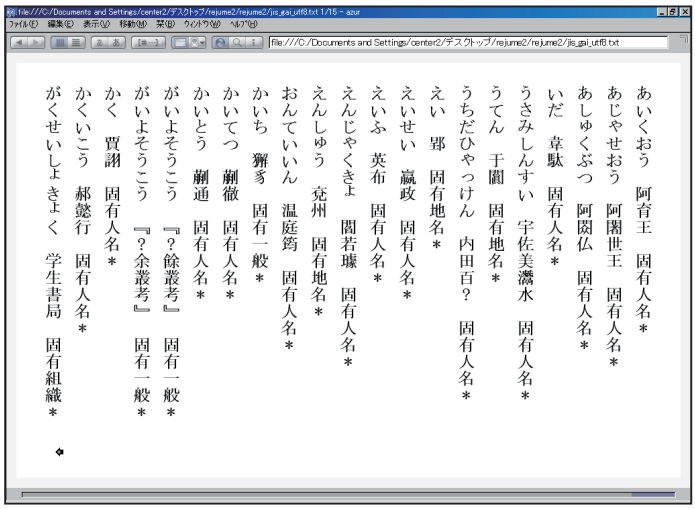
最新の T-Time5.5 からは電子書籍のデータを画像データに変換出力することで、携帯電話はおろかデジカメや iPod、PSP にいたるまで電子書籍端末にしてしまおうという機能が実装された。またドットブック形式は、当初画像データによるフォーマットのみを対象としていたΣ Book の標準フォーマットの一つとして採用もされ [11]、より幅広い利用形態が期待できる。

ドットブックは HTML を中間フォーマットとして利用できるが、現状ではオーサリングツールが市販されておらず、作成にあたってはボイジャーにコンテンツ単位で発注する必要がある。

さて、ドットブックで扱える漢字は、本来 JIS 第一、二水準までだが、変則的な方法で JIS 第三水準、第四水準対応をすすめており、多漢字電子書籍の作成は一応可能といえる。青空文庫とのコラボレーションによる電子書籍リーダー azur[12] は、ドットブック／T-Time 形式と HTML 双方の閲覧・全文検索ができ、対応フォントをインストールすることで、図２のように JIS 第三水準、第四水準の表示が可能である。ドットブック形式のコンテンツを作成するには専用オーサリングツールが必要だが、HTML 形式でも青空文庫の独自タグを追加することにより表現力の向上を図ることができる。

ドットブックで採用されている文字コードは Shift-JIS だが、azur については、どうも Unicode 対応を進めているらしく、UTF-8 テキストでも表示することができた。ただし、図３のように JIS 第三、第四水準以外の文字は文字化けしてしまう。こうした Unicode から距離をおいた立場は、漢籍の電子化を阻害することになるため非常に残念である。日本の伝統文化を漢籍や漢学の要素抜きに語ることの難しさを考えれば、漢籍の表現もできて、はじめて日本文化のための電子書籍フォーマットと言えまいか。

図３：azur の Unicode テキスト表示例 [14]



T-Time や azur の日本語組版を重視した縦書きの美しさはたいへん魅力的であるだけに、是非とも Unicode 対応をすすめていただきたいところだ。

※ 4.4 BBeB 形式

多漢字対応という点で、現在もっともすぐれているのは BBeB（Broad Band e-Book）Book 形式ではなかろうか。BBeB 規格はソニーが推進している電子書籍フォーマットで、書籍向けの Book フォーマットと辞書向けの Dictionary フォーマットに分かれる。ともに XML を中間フォーマットとしているため、マークアップした元データがあれば、半自動的に電子書籍の作成を行うことが期待できる。BBeB 形式のコンテンツにはソニーの著作権保護技術 OpenMG により著作権保護処理がほどこされている。

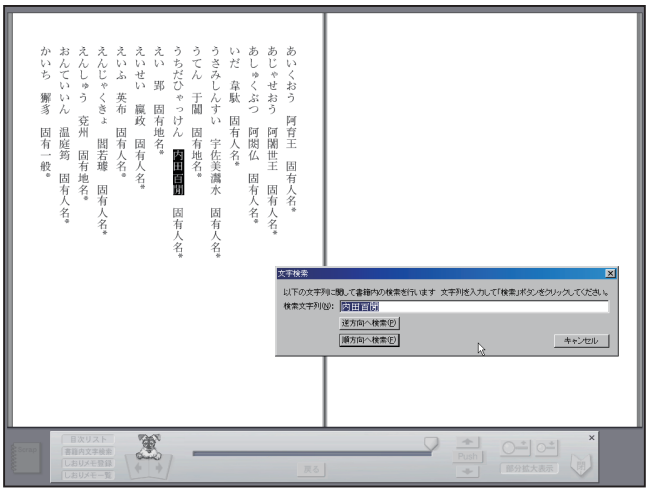


図 4　BBeB の表示例

BBeB Book 形式の電子書籍の利用にあたっては、電子書籍リーダーである LIBRIe[15] がソニーから、オーサリングツールの BookCreator[16] がキヤノンシステムソリューションズから販売されている。 BookCreator は高価すぎるが個人でも購入可能であり、電子書籍フォーマットの中では一応利用がしやすい部類に入る。

さて、BBeB Book 形式が採用している文字コードは Unicode（UTF-16）で、標準規格で 14,375 文字が制定されていて（標準添付のフォントのグリフ数は 14,631 字）、補助漢字レベルまでの漢字の表示にデフォルトで対応しており、図 4 のような JIS 外の漢字の表示・検索が可能である。LIBRIe と BookCreator の双方が当初から Unicode 対応しているものの、専用端末の LIBRIe ではキーボードがあっても辞書引きに使えるだけで、コンテンツ自体は閲覧だけ、全文検索ができるのは PC 用リーダー上のみといった問題がある。Mac 用リーダーもない。

BBeB Book 形式に期待できるのは、現時点の環境下で、正式な操作方法として推奨されていないものの、Unicode3.0 相当の漢字の表示が可能であり、古典籍や現代中国語の電子書籍を作成することが可能な点である。この点については別に論じた [17]。

回 5.おわりに

古典籍の情報化は、必要な情報が研究者の解釈から決定されるために、どうしても流動的にならざるを得ない。だからすべてを電子化することを目的として設定するのは、思考実験として必須であるが、現実的な選択ではない。

学問の本来のあり方からすれば、インターネットで自身の研究成果を誰でも利用できるかたちで公開していくことが、もっとも理想的だろう。しかし、現実にはすでに出版されている研究成果については著作権の壁が立ちはだかるし、これから研究成果を公開するにあたっても、とうぶんは学術成果を出版によってきちんと形にすることが、研究者自身の欲望と周囲からの評価の双方から望まれるだろう [18]。

電子書籍は、おそらくこうした人文学研究の電子化で目の前にある現実的な問題を解決する選択肢の一つになるのではないだろうか。電子書籍の著作権保護機能を利用すれば、これまでの学術成果を電子化し、Google Print[19] のようなサービスを行うことは可能であろう。また学術書の出版の

困難さを考えるとき、オンデマンド出版や電子書籍出版は、今後現実的な選択肢にならざるを得ないではないか。もちろん研究者個人で実現するのは難しく、学会・研究機関レベルでのプロジェクトの立ち上げが必要になるだろう。

出版文化の歴史を振り返れば､そもそも「古典」は口伝に始まり、書写され、印刷され、活版されていくにしたがい、それぞれの場面で何らかの情報が切り落とされ、代替として別の情報が付加されてきたといえよう。その基準は人文学内部の思想的な価値判断でなされたというよりも外部からの要請に拠っていたのではないだろうか。そもそも人文学研究者は現代社会における古典文化の継承者でもあるので、当事者として情報の取捨選択を古典籍に加えることは、必ずしも倫理的に問題があるわけではない。むしろ積極的であってもよいはずである。

参考文献

- 北村義弘・岩崎圭介・田中秀明「電子出版と XMDF 技術」､『シャープ技報』第 84 号、2002 年 12 月
- ソニー株式会社 e-book ビジネス推進室「「BBeB 規格」概要説明 Ver.1.00」
- 高木利弘『電子書籍ビジネス調査報告書 2004』、インプレス、2004 年
- デジタルアーカイブ推進協議会『デジタルアーカイブ白書 2005』、トランスアート、2005 年
- 名和小太郎『ディジタル著作権』、みすず書房、2004 年
- 野村英登「リブリエできるかな――多漢字電子書籍の試み――」、東洋学へのコンピュータ利用第 16 回研究セミナー、2005 年 3 月
- 野村英登「電子辞書の多漢字コンテンツ作成について」､『情報処理学会研究報告』Vol. 2005, No. 52 (2005-CH-67)、2005 年 7 月

注

[1] http://www.sinica.edu.tw/~tdbproj/handy1/

[2] デジタルアーカイブ推進協議会 2005 参照。

[3] IC 電子辞書については野村 2005.7 参照。

[4] 高木 2004 参照。

[5] http://www.aozora.gr.jp/

[6] http://www.papy.co.jp/

[7] 高木 2004 参照。

[8] au の携帯電話では標準フォーマットとなっている。http://www.sharp.co.jp/corporate/news/040722-a.html

[9] http://www.spacetown.ne.jp/dynamic/app/F101/book/use/beginner/xmdfmoji.jspl

[10] http://www.voyager.co.jp/dotbook/dotbook.html

[11] http://panasonic.co.jp/pss/news/jn2005/050512/

[12] http://www.voyager.co.jp/azur/index.html

[13] サンプルデータとして青空文庫の「新 JIS 漢字総合索引」を使用。http://www.sumomo.sakura.ne.jp/~aozora/jisx0213/0213tools/

[14] サンプルデータとして二階堂善弘氏の「JIS 外漢字辞書」を使用。http://www2.ipcku.kansai-u.ac.jp/~nikaido/jisho.html

[15] http://www.sony.jp/products/Consumer/LIBRIE/

[16] http://ps.canon-sol.jp/bc/index.html

[17] 野村 2005.3 参照。

[18] 名和 2004 参照。

[19] http://print.google.com/

偽古文尚書の「賢」と「官」

χ^2 値による語彙偏差の数量化を通して

齊藤　正高（さいとう　まさたか）

回 はじめに

本篇は『尚書』58 篇について、その語彙を計量的に分析し、とくに偽書と認められる 25 篇について、その独自の要素を一部抽出し、考察を加える試みである。

回『尚書』について

『尚書』は経書 [1] の一つである。その内容は、まず堯・舜・禹など、「神人」（神話上の人物）の事績をのべ、「武王克商」[2] など、夏殷周三代の王朝に関する記述があり、秦の穆公の言葉とされる「秦誓」で終わっている。

伝統的経学の見地から『尚書』は孔子の編修をへているとされる [3] が、苛烈な刑罰思想など、基本的に徳治を主張する孔子の思想とは異質な点もふくまれている。

『尚書』のテキストについては、唐の孔穎達らが編纂した『尚書正義』（653A.D.）によって、概ねその本文が定められており、「尚書序」につけられた疏によれば、『尚書』はつぎの二つの部分に分かれる。

第一の部分は、秦の博士伏生が前漢文帝期（180B.C.–157B.C.）に伝えた今文尚書（以下、今文）と共通する 33 篇 [4] である。

第二の部分は、東晋元帝期（317–322 A.D.）に豫章の内史、梅賾が献上したと伝えられる「孔安國伝古文尚書」[5] にだけ存在する 25 篇である。この部分は宋代以後その真偽が疑われ、清の閻若璩（1636–1704）に至って偽書であることが決定的になった。この部分を「偽古文尚書」（以下、偽古文）という。

それぞれの篇名をあげると以下になる。

今文　33 篇

堯典　舜典　皐陶謨　益稷　禹貢　甘誓
湯誓　盤庚上　盤庚中　盤庚下　高宗肜日　西伯戡黎　微子
牧誓　金縢　大誥　康誥　酒誥　梓材
召誥　洛誥　多士　無逸　君奭　洪範
多方　立政　顧命　康王之誥　呂刑　文侯之命　費誓　秦誓

偽古文　25 篇

大禹謨　五子之歌　胤征　仲虺之誥
湯誥　伊訓　太甲上　太甲中　太甲下
咸有一德　説命上　説命中　説命下　泰誓上　泰誓中　泰誓下
武成　旅獒　微子之命　蔡仲之命　周官
君陳　畢命　君牙　冏命

この分類については異論もあるが、先学の『尚書』訳注がほぼ従う所であり、本篇では尚書学の議論の土台としてうけいれ、以後の考察をすすめ

たい。

『尚書』という一つ書物のなかに存在する今文と偽古文の二つのテキストは、その性質が異なる。

まず、今文尚書は重層的に成立した文献であることが指摘できる。そのなかには、まず最古の部分として、金文との類似が指摘され、西周の作とされる「大誥」「康誥」などの周誥がある。また、秦にまで下る内容を指摘できる「堯典」があり、春秋期に実用化され、戦国期に普及した鉄の記述をふくむ「禹貢」がある[6]。つまり、これらの篇に代表される今文尚書の成立年代を、西周の末年（771B.C.）から戦国時代の末年（221B.C.）までと想定しても、およそ 500 年の幅があることになる。

一方、偽古文尚書は作者こそ定かでないものの、漢代から伝えられてきた古文尚書が散逸した永嘉の乱（307–313 A.D.）の後、数年で出現している点から考えれば、今文に認められるような重層性は弱いことが予測できる。語法的には偽古文尚書がつくられた時代は、上古から中古への過渡期にあたり[7]、この点も今文と異なる。また、偽古文の文章は全くの作りものではなく、先秦諸子や秦漢の書物に引用されて残った逸文尚書や、ほかの経書からの引用が核となっており、この出典についてはほぼ解明されている[8]。

回 本篇の扱う問題と方法

上に述べたように、今文尚書と偽古文尚書は、文献としての基本的性質が異なる。しかし、性質が異なるからといって、両者の比較が無意味だということにはならない。なにより偽古文が『尚書』の経文として作られている以上、そこには今文尚書との一体化が企図されているはずであり、今文尚書を模倣しようとした擬古の意図が働いていることが推察される。

つまり、偽古文尚書は文献学的に把握されてきたその今文との性質の差異とは裏腹に、その製作意図においては今文との一致を目指していたのであるから、その模倣の綻びを剔出することは、偽古文の特徴を明らかにすると同時に今文の特徴を明らかにすることにもなり、『尚書』全体の理解を深める助けになるはずである。

さらに、偽古文がさまざまな書物の引用から、何を構成しようとしたかという問題も示されるはずである。その構成しようとするものは、今文尚書に含まれる思想からの逸脱や、製作者の理想を反映している可能性がある。

本篇はこのような問題について、近年展開している計量的分析方法から解明を試みた。管見の及ぶところでは、同様の試みはないように見うけられる。

計量的分析方法には三つの問題がある。第一の問題は分析対象となるテキストデータの問題であり、第二の問題は文を分割し語彙ごとに集計するインデキシング（索引化）の問題であり、第三の問題は語彙の偏差を数量化する問題である。

※ テキストデータ

分析に用いたテキストデータは、『吉川幸次郎全集』（筑摩書房 1970）第 8 巻 –10 巻に収められた翻訳『尚書正義』の經文[9]によって作成した。基礎データは以下である。

	今文	偽古文
字数	16,932	7,578
字種	1,624	1,196
共通字数	15,268	7,128
共通字種	904	904
固有字数	1,164	450
固有字種	720	292

表 1　分析に用いたテキストデータの基礎統計（いずれも序及び篇名を含まず[10]）

偽古文は、今文の約 45% の長さしかなく、約 1200 の文字種からなりたち、そのうち約 900 種（75%）は今文と共通である。これらの文字種の使用頻度を合計すると 7,128 字になる。実に偽古文の全文 7,578 字の 94% が今文に共通する文字を使用している。これは偽古文がいかに今文を模倣しているかということのひとつの度合を示しているといえよう。

❖ インデキシング（索引化）

テキストデータの分割には N-gram[11] をつかい、『尚書』の全体を通じて頻度 5 以上の 1gram と 2gram を索引とした。

この頻度は後でのべる χ^2 検定について、筆者が設定した有意水準（0.1%）[12] とその棄却域（10.82）をから導いたものである。3gram 以上は、同じ用例が少なく、極端に全文に対する割合が低下する [13] ので、考察の対象から除外した。

gram 数	語種	頻度合計	長さ	全文比
1	775	22,381	22,381	0.913
2	545	5,067	10,134	0.413
3	77	596	1,788	0.073
4	16	119	476	0.019

表 2　『尚書』の見出し語

❖ 偏差の数量化

今文と偽古文の間における語彙偏差を評価するには、計量言語学で用いる χ^2 値（カイ二乗値）を用いた [14]。

具体的には、以下の例で紹介する χ^2 検定（独立性の検定）をへて、偏差の評価を行った [15]。

例 1：「我」

まず、対象とする語彙の頻度と、それ以外の語彙の頻度を計算し、以下の表のようにまとめる。このような分類の仕方を自由度 1 という。

	我	その他	合計
今文頻度	196	16,736	16,932
偽古文頻度	33	7,545	7,578
合計	229	24,281	24,510

表 3　「我」の分布

この頻度の表から、「今文と偽古文の文書集合の間に“我”という語彙の使用傾向に差がない」という仮説を立てる。そして、この場合の理論頻度は今文と偽古文の全文字数の比、つまり 16932:7578 に等しいはずだから、以下のように理論頻度を計算する。

今文の「我」の理論頻度
　229 × 16932 ÷ 24510 ＝ 158.20
偽古文の「我」の理論頻度
　229 × 7578 ÷ 24510 ＝ 70.80
今文の「その他」の理論頻度
　24281 × 16932 ÷ 24510 ＝ 16773.80
偽古文の「その他」の理論頻度
　24281 × 7578 ÷ 24510 ＝ 7507.20

つぎにこれらの値から、理論頻度からの偏差をあらわす χ^2 値を計算する。χ^2 値は次式でもとめる。

具体的には以下のように計算する。

$$\chi^2 = \sum \frac{(\text{実際の頻度} - \text{理論頻度})^2}{\text{理論頻度}}$$

$$\begin{aligned}\chi^2_{我} = &(196 - 158.20)^2 \div 158.20 \\ &+ (33 - 70.80)^2 \div 70.80 \\ &+ (16736 - 16773.80)^2 \div 16773.80 \\ &+ (7545 - 7507.20)^2 \div 7507.20 \\ &= 29.49\end{aligned}$$

つぎに、有意水準を 0.1% と設定し、自由度 1 の χ^2 分布から棄却域を求める。

p	χ^2
0.05	3.84
0.01	6.63
0.001	10.82

表 4　自由度 1 の χ^2 分布表　（p は有意水準）

「我」の χ^2 値 29.49 は、χ^2 分布の 0.1% の棄却域 10.82 より大きい。これは、「今文と偽古文の文書集合の間に“我”という語彙の使用傾向にちがいがない」という最初の仮説が棄却されることを意味している。したがって、仮説と対立するつぎのことが結論できる。

有意水準 0.1% で、今文尚書と偽古文尚書において、「我」の使用傾向に差がある。

例 2：「上帝」

「上帝」は 2gram（2 文字）である。長さ L の文章から 2gram を全て切り出すと、(L-1) 個だけ取り出すことができる。これによって、『尚書』各篇について、それぞれ 2gram が何個取り出せるかを計算し、それを今文・偽古文について合計すると、以下の表をつくることができる。

	上帝	その他	合計
今文頻度	22	16,877	16,899
偽古文頻度	10	7,543	7,553
合計	32	24,420	24,452

表 5 「上帝」の分布

「我」の場合と同様に「今文と偽古文の文書集合の間に“上帝”という語彙の使用傾向に差がない」という仮説をたて、理論値を計算すると、以下になる。

$$32 \times 16877 \div 24452 = \quad 22.11$$
$$32 \times 7543 \div 24452 = \quad 9.88$$
$$24420 \times 16877 \div 24452 = \quad 16876.88$$
$$24420 \times 7543 \div 24452 = \quad 7543.12$$

χ^2 値を計算すると 0.002 である。この値は $p < 0.05$ の棄却域 3.84 よりも小さく、有意水準 5% でも仮説は棄却されない。したがって次のことが結論できる。

有意水準 5% で、今文と偽古文の文書集合の間に“上帝”という語彙の使用傾向に差はない。

回 結果

※ 1gram の偏差

以下に、χ^2 検定を行い、有意水準 0.1% で差があると認められた 54 種の 1gram を χ^2 値の大きいものから示す。

官	6:31	(48.49)	后	24:44	(36.45)
萬	15:35	(35.83)	賢	1:17	(34.03)
我	196:33	(29.49)	德	115:105	(29.37)
曰	376:97	(24.47)	殷	81:6	(23.59)
以	124:104	(23.27)	道	12:24	(21.57)
世	9:21	(21.48)	兆	1:11	(20.74)
必	1:11	(20.74)	汝	148:27	(19.80)
神	7:18	(19.77)	戒	2:12	(19.69)
惟	396:251	(19.30)	良	3:13	(18.99)
五	89:11	(18.65)	克	85:74	(18.29)
治	6:16	(18.02)	聖	6:16	(18.02)
用	126:22	(17.97)	窮	0:8	(17.88)
物	4:13	(16.53)	終	18:26	(16.38)
善	5:14	(16.28)	商	17:25	(16.12)
越	63:6	(16.00)	學	0:7	(15.65)
忠	0:7	(15.65)	掌	0:7	(15.65)
至	64:7	(14.78)	説	4:12	(14.57)
足	0:6	(13.41)	謂	2:9	(13.35)
庶	75:11	(13.28)	訓	21:26	(13.13)
政	20:25	(12.81)	二	51:5	(12.71)
今	83:14	(12.39)	罔	87:68	(12.25)
允	13:19	(12.15)	師	29:31	(12.12)
多	52:6	(11.52)	罰	48:5	(11.48)
亡	2:8	(11.28)	令	2:8	(11.28)
寵	0:5	(11.17)	期	0:5	(11.17)
狎	0:5	(11.17)	莫	0:5	(11.17)
諫	0:5	(11.17)	與	8:14	(11.04)

表 6 『尚書』における 1gram の語彙偏差

結果の見方　例：官 6:31（48.49）

- ●官の今文頻度　6
- ●官の偽古文頻度　31
- ● χ^2 値　48.49

ほとんどの結果が、大きな使用頻度の差異を示している。しかし、「德」のように、今文と偽古文の間でほとんど使用頻度に差がない例もある。これは、もともと今文は偽古文より 2 倍以上長いので、偽古文における理論頻度は、今文における理論頻度の半分程度でもよいはずだからである。つまり、偽古文は今文の半分程度の長さであるのに、今文と同程度「德」を使用するのであるから、「德」は偽古文に偏って使われているのである。

今文に偏って使われている文字は、以下の文字

である。

今文尚書の代名詞や数詞の用例については、最近の研究では、銭宗武氏の著作 [16] に指摘がある。

我・曰・殷・汝・五・用・越・至・庶・今・多・罰

表 7　今文尚書に偏って使用される 1gram

偽古文尚書を扱う本篇では言及できないが、「殷」や「罰」といった文字の今文偏出については、別に用例を検討する余地があるだろう。

※ 2gram の偏差

1gram と同様に、有意水準 0.1% で差があると認められた 18 種の 2gram を χ^2 値の大きいものから示す。

「至于」と「公曰」は、例外的に今文に偏って出現している。「至于」61 例のうち、19 例が地理書の「禹貢」に出現する。「公曰」26 例のうち 12 例は「周公曰」であり、7 例が「無逸」に集中している。

兆民	1:10（18.57）	萬方	0:8（17.90）
至于	61:5（16.85）	伊尹	1:9（16.37）
萬邦	4:12（14.59）	一德	0:6（13.43）
或不	0:6（13.43）	掌邦	0:6（13.43）
爾萬	0:6（13.43）	百官	0:6（13.43）
衆士	0:6（13.43）	罔以	0:6（13.43）
厥德	6:13（12.55）	公曰	26:0（11.63）
天道	0:5（11.19）	惟賢	0:5（11.19）
戒哉	0:5（11.19）	明王	0:5（11.19）

表 8　『尚書』における 2gram の語彙偏差

回　考察

以下、偽古文に偏って使用されている文字の背景を原文に基づいて検討していきたい。

※「賢」について

「賢」は今文尚書には以下の 1 例しか使われておらず、偽古文によって『尚書』に導入された概念として最もはっきりした特徴をもっている。

- 在祖乙時、則有若巫賢（君奭）
  訓読：「祖乙の時に在りては、則ち巫賢の若き有り」

ここの「巫賢」は殷の祖乙の時代に王家を補佐した者の固有名である。「巫」の字を冠している点は、この人物が呪術的職能者であったことを示唆している。賢字の用例を今文尚書と同時期の成立とみなすことができる『毛詩』に探してみると「序賓以賢」（大雅「行葦」）と 1 例だけである。文献上、賢字が多く用いられるのは、『論語』に至ってからである。

偽古文尚書には賢字は次のように多く用いられている。

①野無遺賢、萬邦咸寧（大禹謨）
訓読：「野に遺賢無く、萬邦咸な寧し」
出典：「萬國咸寧」（『周易』乾彖傳）
②任賢勿貳、去邪勿疑（大禹謨）
訓読：「賢に任じて貳する勿く、邪を去りて疑う勿れ」
出典：『戦国策』趙策、書云
③成允成功、惟汝賢（大禹謨）
訓読：「允を成し功を成すは、惟れ汝の賢なり」
出典：「成允成功」（『左傳』襄公五年）
④不自滿假、惟汝賢（大禹謨）
訓読:「自ら滿假せざるは、惟れ汝の賢なり」
⑤侮慢自賢、反道敗德（大禹謨）
訓読：「侮慢自ら賢として、道に反し德を敗る」（有苗の君主の不徳をそしる）
⑥簡賢附勢、寔繁有徒（仲虺之誥）
訓読：「賢を簡にし勢に附く、寔に繁く徒有り」
出典：「寔繁有徒」（『左傳』昭公二十八年）
⑦佑賢輔德、顯忠遂良（仲虺之誥）
訓読：「賢を佑け德を輔け、忠を顯わし良を遂ぐ」

⑧任官惟賢材（咸有一德）
訓読：「官に任ずるは惟れ賢材」
⑨爵罔及惡德、惟其賢（説命中）
訓読：「爵は惡德に及ぶ罔く、惟れ其れ賢。」
出典：「兌命曰、爵無及惡德民」（『禮記』緇衣）
⑩惟后非賢不乂。惟賢非后不食（説命下）
訓読：「惟れ后は賢に非ざれば乂まらず、惟れ賢は后に非ざれば食まず」
⑪剖賢人之心（泰誓下）
訓読：「賢人の心を剖く」（紂王の非道をそしる）
⑫建官惟賢、位事惟能（武成）
訓読：「官に立てるは惟れ賢、事に位するは惟れ能」
⑬所寶惟賢、則邇人安（旅獒）
訓読：「寶とする所惟れ賢なれば、則ち邇人安し」
⑭惟稽古崇德象賢（微子之命）
訓読：「惟れ古を稽うるに德を崇び賢に象る」
⑮推賢讓能、庶官乃和（周官）
訓読：「賢を推し能に讓れば、庶官乃ち和す」
出典：「推賢讓能」（『荀子』仲尼篇 [17]）
⑯商俗靡靡、利口惟賢（畢命）
訓読：「商俗は靡靡として、利口惟れ賢とす」

これら偽古文に用いられた賢字の用例を総合すると、賢者を官に任用し、政治を補佐させ、「宝」として尊重すべきだということである。また、君主が「自らを賢」としたり、君主が「賢人」を弾圧したり、口の達者な者が「賢」とされることなどへの批判もみえる。つまり、偽古文の作者は今文尚書には存在しない「賢者を尊ぶ」という考え方、尚賢思想を自らの経書に投影しているのである。

偽古文尚書の賢字は、16 例のうち 10 例は出典がない。つまり、この 10 例は偽古文の作者が任意に書いた部分である。

残りの 6 例のなかで、3 例が『周易』や『左傳』の引用文と組合せて使用されている（①・③・⑥）。また、1 例⑨が『禮記』に引く逸書からの引用と組合せて用いている。これらは、偽古文の作者の地の文に使われている賢字が、ほかの経書の引用と連動して説得力を醸し出している例といえよう。

偽古文の「賢」の使用例のうち、出典があるものは 2 例しかない。そのなかで、『荀子』からの「推賢讓能」⑮が最も古いと思われる。

ここで問題となるのは、『荀子』には「推賢讓能」は 1 例しかないが、「尚賢使能」が 10 例もあるという点である。偽古文の作者が『荀子』のなかからわざわざ「推賢讓能」を選択した根拠をはっきり示すことは難しい。だが、一応の筋道を考えれば、『荀子』のなかでは、統一へむかう戦国後期にあって、「尚賢使能」（賢を尚び能を使う）が君主の政治手法、つまり能力主義として主張されているのに対し、「推賢讓能」（賢を推し能に譲る）は臣下として能力がない者の消極的処世術として叙述されている。この点を考慮すれば、すくなくとも、秦漢帝国における官僚制の成立以後に生きた偽古文の作者にとって、「尚賢使能」より「推賢讓能」の方が、官僚制における有能な者への機会拡大という面において、理想の政治としてシンパシーを感じる言葉であったのではないかと説明することは可能である。

偽古文以前の「尚書」に賢字が使用されたことを窺わせる用例は、『戦国策』に引く尚書逸文の「任賢」の例②がある。

ほかにも尚書逸文に賢字を用いる例が 2 例ある。

⑰往者不可及、来者不可待、賢明其世、謂之天子（『呂氏春秋』聴言引周書）
訓読：「往者は及ぶ可らず。来者は待つ可らず。賢は其の世を明かにす。これを天子と謂う」
⑱在上位而不能進賢者逐（『説苑』臣術引泰誓）
訓読：「上位に在りて賢を進むる能わざる者は逐う」

⑰の『呂氏春秋』に引く「周書」は現在の『尚書』にはなく、『呂氏春秋』成立時の伝本にしか存在しない文である。⑱の『説苑』に引く「泰誓」は前漢宣帝期に「河内女子」によって発見された書物であり、現行の「泰誓」とは別の偽書である。

『呂氏春秋』『説苑』『戦国策』は、みな秦から前漢末にかけての書物であり、これらの引用の内容はみな尚賢思想を述べたものである。したがって、秦漢の尚書には尚賢思想を述べるものがあったことが推察できる。

出土史料をみてみると、郭店楚簡『唐虞之道』に「堯舜之行、愛親尊賢」[18]とあり、戦国期に堯舜と尚賢思想とを連絡させる考えがあったことが分かる。秦漢の尚書逸文にみえる賢字の用例は、『唐虞之道』に見られるような考えが発展し、『尚書』の断片として伝えられた結果なのかもしれない。

このように賢者を尊ぶという尚賢思想は、戦国時代の出土文献や秦漢の書物に引用された逸書、そして何より『孟子』や『墨子』などに明白であり、偽古文の作者がこれらの記述を参照していたであろうことは他の引用の出所からみて、疑う余地がない。

偽古文の作者が三代の政治に投影した尚賢の理想は、歴史的にみれば、春秋時代末期における血縁的秩序の弱体化の後、戦国時代に盛んになる能力主義の考え方である。今文尚書のなかにもその使用語彙から成立が戦国時代に下る篇はいくつも指摘されているが、これらは賢字をつかって尚賢思想を語ることがない。

偽古文尚書の著作動機が、今文尚書との一体化にあったとするならば、作者は戦国諸子の著述を用いるなかで、今文尚書とは異質の要素を彼の『尚書』に書き込んだことになり、ここにその模倣のほころびの一つが見られるのである。

※「官」について

「官」は 1gram の偏差で最も高い χ^2 値を示し、偽古文に極端に偏って使用される言葉である。

今文尚書には官字が 6 例使用されている。

①鞭作官刑、扑作教刑（舜典）
訓読:「鞭は官刑を作し、扑は教刑を作す。」
②知人則哲、能官人。（皐陶謨）
訓読：「人を知るは則ち哲、能く人を官にす。」
③俊乂在官、百僚師師、百工惟時（皐陶謨）
訓読：「俊乂の官に在れば、百僚は師を師とし、百工は惟れ時（よ）し。」
④無曠庶官、天工人其代之。（皐陶謨）
訓読:「庶官を曠（むな）しくする無ければ、天工に人のこれに代わる」
⑤五過之疵、惟官、惟反、惟内、惟貨、惟來（呂刑）
訓読:「五過の疵は、惟れ官、惟れ反、惟れ内、惟れ貨、惟れ來」
⑥嗚呼、敬之哉、官伯族姓、朕言多懼（呂刑）
訓読：「嗚呼、之を敬せん哉。官伯・族姓、朕の言は懼れ多し」

今文尚書の官字の用例は、まず「舜典」や「皐陶謨」など、成立が戦国末から秦漢にまで下ると推定できる篇に使用されている点を指摘できる。

「呂刑」の２つの用例（⑤・⑥）については、通常の官僚組織の意味では使われておらず、特殊な用例といえるであろう。

「皐陶謨」の３例は、禹と皐陶の対話のなかで、皐陶の言葉として集中的に用いられている。そのなかには、官僚組織の意味で使われている例（③・④）を見出すことができる。

これらの例を総合すると、官字を官僚組織の意味で用いる例は、今文尚書のなかでも成立が戦国末以後に下る篇にしか見当たらず、非常に限定された範囲にしか存在しないことになる。

だが、具体的官名については、立政に「司徒」「司馬」「司空」などの記述がある。官僚組織の具体的記述は今文尚書にも備わっているとみるべきである。

偽古文尚書における官字の用例は 31 例である。全体の傾向として、まず、偽古文尚書 25 篇のうち 11 篇に用いられ、特定の朝代の文献にもちいられているのではないことを指摘できる。また、『周礼』の影響が濃厚である「周官」に 13 例が集中していることを指摘せねばならない。

一般的に「官」という文字は、具体的な官名ではなく、いわば抽象的な官僚機構をあらわす文字である。

	百官	厥官	庶官	有官	建官	ほか	総計
大禹謨	1					1	2
胤征	1	1				2	4
仲虺之誥						1	1
伊訓	1					1	2
咸有一德						1	1
説命上	1						1
説命中	1		1			1	3
泰誓上						1	1
武成					1		1
周官	1	3	1	2	1	5	13
冏命		1				1	2
総計	6	5	2	2	2	14	31

表９　偽古文尚書の「官」

この抽象的な官僚機構の総体を表す言葉として、「百官」がある。偽古文における官字の用例のうち、この「百官」は、今文にはないという点で偽古文独自の特殊な用例であり、かつ官字を含む２文字の用例では最も多く、すべての朝代に属する篇にいわば広く薄く用いられている点が特徴的である。

偽古文における「百官」の用例は以下になる。

⑦正月朔旦、受命于神宗、率百官、若帝之初（大禹謨）
訓読：「正月朔旦、命を神宗に受け、百官を率いること、帝の初の若し」

⑧百官修輔、厥后明明（胤征）
訓読：「百官修輔すれば、厥の后（きみ）は明明たり」

⑨百官總己、以聽冢宰（伊訓）
訓読：「百官己を總べて、以て冢宰に聽く」
出典：「子張曰、書云、高宗諒陰、三年不言、何謂也。子曰、何必高宗、古之人皆然。君薨、百官總己、以聽於冢宰三年」（『論語』憲問）

⑩天子君萬邦、百官承式（説命上）
訓読：「天子は萬邦に君として、百官は式を承く」

⑪惟説命總百官（説命中）
訓読：「惟れ説命ぜられて百官を總ぶ」

⑫冢宰掌邦治、統百官、均四海（周官）
訓読：「冢宰は邦治を掌り、百官を統べ、四海を均しくす」

「百官」の用例のうち、出典があるものは⑨の１例のみである。これは『論語』憲問の文をそのまま用いたものである。『論語』によれば、ここは子張と孔子が、当時の『尚書』について論じた内容である。

『論語』にはもう１例、「百官」の用例がある。

子貢曰、譬之宮牆、賜之牆也及肩、闚見室家之好。夫子之牆數仞、不得其門而入、不見宗廟之美、百官之富、得其門者或寡矣。（『論語』子張）

ここでは、子貢によって孔子が古代の（官制をふくめた）礼制の継承者として語られている。

『論語』におけるこの２例の「百官」は、孔子が『尚書』を「編次」[19]したとされる伝承を信ずる限り、偽古文の作者にとって、自らの経書に「百官」を用いる際の最も正当な根拠であったと思われる。あるいは、ここにみえるように『尚書』という経書に編纂者として伝えられる孔子の思想を徹底させるという理念こそが、偽古文尚書の作者のうちに経書を製作する動機としてはたらいていたのではないだろうか。

しかし、今文尚書との一体化という面では、「百官」は明らかに今文尚書とは異質の表現である。もし用いるとするなら、今文にある「百工」[20]などに新たな意味を与えるべきであった。

なお、この「百官」の用例は、さきにみた「賢」とも関連する。それは、「賢」としての「冢宰」、また、「野に遺賢無く」登用された「百官」たちとして、偽古文尚書がえがく理想の政治体制を反映しているのである。その理想政治は「賢才を挙げよ」[21]と述べた孔子の主張する理想とも共通するのである。

回 おわりに

以上、偽古文尚書に偏った語彙のなかで、最も大きな問題と思われる「賢」と「官」について考察した。このなかで、偽古文尚書が今文尚書には希薄であった尚賢思想を強調し、抽象化された官僚組織の記述を、『論語』に依拠しながら、「百官」として強調したことが明らかになった。

指標として用いた χ^2 値については、ほかの指標と同様に語彙に密接するので、抽象度を上げた場合の語彙の通用に難があり、語彙ごとに限定された大まかな傾向を示すのみにとどまることは否定できない。しかし、そうだとしても、これを参考に文献について語るべき問題の優先順位をさぐることはできるであろう。

『尚書』の問題についても、本篇で論じられなかった問題は多い。今文と偽古文の語彙偏差が示す問題としては、以下の点がのこるが、これらの問題については、別稿に譲りたい。

※『尚書』の語彙偏差が示す問題

⑴今文では「萬民」が使われ、偽古文では「兆民」が使われる傾向がある。今文で「兆民」を用いる唯一の例（呂刑）が、戦国時代の出土資料、上博楚簡「緇衣」の引用では「萬民」となっており、現行「呂刑」の書き換えが推察できる。

⑵今文では周書（君奭）に祖述されるのみである「伊尹」が、偽古文商書に集中的に使用されている。

⑶偽古文にのみ「天道」の用例がある。

（了）

注

[1] 『尚書』は歴史的記述を核としているが、純粋な史書ではなく、一種の経典化がなされている。この点、『尚書』の内容をそのまま歴史記述とみなすことは先学が戒める所である。本篇も『尚書』を経書とみなし、その思想内容を扱うものである。

[2] 武王克商の年代については、紀元前 1046 年が天文学上の暦算、出土金文の暦年、および『尚書』武成・召誥・洛誥、『国語』周語の文献記述と矛盾せず、もっとも精度が高いとされている。（夏商周断代工程専家組『夏商周断代工程 1996-2000 年階段成果报告』世界図書出版公司。2001 年。P.49）また、中島敏夫「歴史と神話への視座――疑古派禹天神論の検証からの再出発――（上）」（愛知大学現代中国学会『中国 21』Vol.15。 2003 年 3 月）にもこの点の指摘がある。

[3] 屈万里『尚書釈義』（中国文化大学出版部。1980 年重排版。）叙論。P.3-4

[4] 『尚書正義』序を参照。ただし、伏生が伝えた『尚書』は序を除いて 28 篇であり、堯典の後半から舜典ができ、皐陶謨の後半から益稷ができ、盤庚が三篇に分かれ、顧命の後半から康王之誥ができ、合計で 5 篇が増えて 33 篇となったとする。

[5] 『隋書』経籍志および『春秋左氏伝正義』襄公伝三十一年正義など

[6] 池田末利『尚書』(『全釈漢文大系』11)集英社。1976 年。による。

[7] 王力『漢語史稿』科学出版社。1956 年（中華書局中国文庫版。2004 年を参照、第六節「漢語史的分期」P.43）

[8] 前掲『尚書釈義』では「偽古文尚書」の部分に朱駿声『尚書古注便読』をひき出所を示している。

[9] 東方文化研究所（京都）『尚書正義定本』（第 1 冊。昭和 14 年刊行）による。

[10] 「舜典」は「堯典」の後半から作られたため、「舜典」の先頭には偽作の 28 文字がある。この部分については、今文尚書にも偽古文尚書の字数にも含めず、考察の対象外とした。

[11] 『漢字文献情報処理研究』第 2 号。2001 年。を参照。なお N-gram 分割ツールとしては、師茂樹氏が製作した morogram を用いた。

[12] 一般的なアンケート統計では、有意水準の最低値を 5% に設定するが、語彙統計では非常に多くの文字が説明要素となる可能性をもつので、5% では有意な変量を絞り込むことが難しい。よって、有意水準を高めの 0.1% に設定することにした。また、0.1% の有意水準で帰無仮説を棄却できる最小の頻度は『尚書』の場合、今文 0・偽古文 5 の場合である。

[13] この点は、山田崇仁氏の「中国戦国期の語彙量につ

いて」（『漢字文献情報処理研究』第 5 号。2004 年 P.100）にすでに指摘がある。

[14] χ^2 検定については、伊藤雅光『計量言語学入門』大修館書店。2002 年　P.104-111 を参照した。また、「入門講座Ⅲ――1 カイ二乗検定」（『計量国語学』1958.6 P.29-38）、西平重喜「『その他』の扱い方」（同前 P.28-29）も参考にした。

[15] χ^2 値は、Microsoft の Excel によって段階的に計算した。また、χ^2 値の基になる今文と偽古文の各頻度は、XML 化した『尚書』のテキストデータを自作のスクリプト（JavaScript）に読ませて計算した。これらは筆者のサイト（**http://taweb.aichi-u.ac.jp/saitom**）で公開する予定である。

[16] 銭宗武『今文尚書語法研究』商務印書館。2004 年。

[17] 前掲『尚書釋義』P.247 は、ここの出典を非十二子篇に誤っている。

[18] 劉釗『郭店楚簡校釈』福建人民出版社。2003 年。P.148 を参照した。

[19]「孔子序書、上紀唐虞之際、下至秦繆、編次其事」（『史記』孔子世家）

[20] 堯典、皐陶謨、益稷、康誥、洛誥にそれぞれ一例づつ使用されている。

[21]『論語』、子路

漢情研 2005 年公開講座報告

東洋学研究と著作権問題

ここ二年、本会では情報の発信や利用に際して他者の権利を犯さないノウハウを知るという目的で、「東洋学情報化と著作権問題」をテーマとして公開講座を開催してきた。

本年度は「東洋学研究と著作権問題」と若干タイトルを改めている。これは、自らの研究を権利者としてどう捉えるかについて問題意識がシフトしたためである。

そこで本年度は、校訂権を中心にして研究と著作権の問題を中心に採り上げた。校訂権を俎上にしたのは、文献研究において、資料の校訂には多くの時間を費やさねばならないにもかかわらず、この作業が著作権法上、どのように扱われるべきか示唆するものが少ないためである。

文献研究における校訂の意味を問い直し、これに権利が存在するとすればそれをどう行使することが学術の発展に寄与することなのかということについて、本年度は本会 BBS に専用会議室を設けて議論を行い、それを踏まえた形で法学者の石岡克俊氏によるご講演が行われ、更に会員諸氏による活発なディスカッションが行われた。

以下は、公開講座の講師による論攷、および報告である。

※漢情研 2005 年公開講座の日程・会場等の詳細については、彙報（P.220）を参照していただきたい。

Contents

2005 年公開講座報告
東洋学研究と著作権

校訂とはいかなる行為か？

秋山　陽一郎（あきやま　よういちろう）

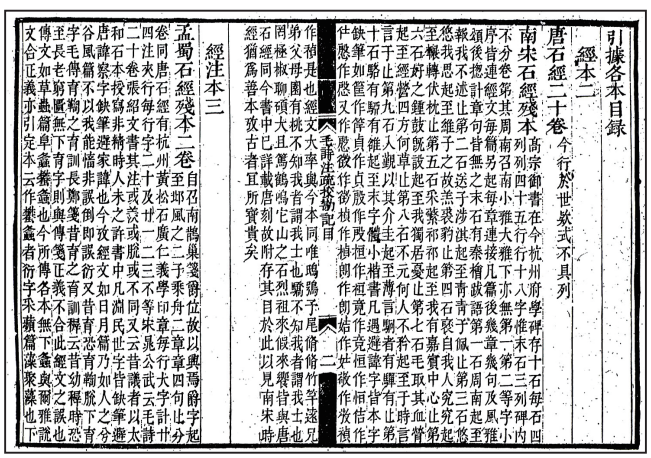
◇ はじめに

石岡克俊氏（慶應義塾大学産業研究所助教授）を講師としてお招きしての夏期公開著作権講座が行われるようになってから今年で３年目に当たる。今年度（2005 年度）は「東洋学研究と著作権問題」と題して、かねてから筆者を含む一部の参加者の関心を集めていた、いわゆる「校訂権」を軸として、次の点を軸として議論が進められた。

- 校訂作業を研究成果としてどのように位置づけるべきか
- 校訂権を著作権として認めるべきか

本稿ではこうした議論の前提となる「校訂」が、いかなる行為を指すのかを、公開講座前の BBS における議論も交えて概観する [1]。

図１：『十三經（毛詩）注疏校勘記』中で清朝中期の大儒 阮元が対校資料として参照した異本のリストとその説明



◇「校訂」の種類

ひとくちに「校訂」といっても、「校定」・「校勘」「校讐（讐校）」・「校訂」・「校正」・「対校（校対）」・「点校」・「標点」・「校注」と色々なバリエーションがあることは、（古典籍の整理を実際に行った経験者は別として）実は意外に知られていない。校訂という行為にはさまざまなレベルや種類があって、本来、行為の内容によって細かく区別されている。校訂という行為が指し示す範囲を弁別するためにも、それらがいかなる行為をそれぞれ意味しているのかを、まずはざっと追ってみたい。

⊙ 各種校訂の名義の問題点

個別の概念の検討に入る前に、一つ断っておかなければならないことがある。以上に挙げた「校定」・「校勘」・「校讐（讐校）」・「校訂」・「校正」・「対校（校対）」・「点校」・「標点」・「校注」といった術語は、よほど意識的に区別して用いられていない限り、しばしば混同されて用いられることが多く、区別して用いられている場合でも人や文脈によって名義が異なることがある。よって本稿で示した名義は、あくまで諸事例（就中、校讐家や校定に長じている近時の研究者の用法）を参考に筆者が一応穏当と思われる名義を挙げているに過ぎず、もとより異なる解釈もあって絶対的な定義ではないと

いうことをご了承おき願いたい[2]。

⊙ 対校（校対）

「校訂」・「対校」の「校」は「くらべる」（＝較）という意味。「対校」とは、複数の異本や字句を比較対照することを意味する校訂作業においてもっとも基本的な行為である。「校合」も同様で異本を較(くら)べ合わせることを意味する。

⊙ 校訂・校正

原本（底本）と校(くら)べて誤字や文章の明らかな誤りを訂正する行為。「校勘」や「校定」とは、本文を修正するのに絶対のよりどころ（底本）があるか、傍証や考証を必要とするかという点が異なる。次に紹介する倪其心氏の解説が参考となろう。

◆［参考］倪其心著、橋本秀美・鈴木かおり訳『校勘学講義 —— 中国古典の読み方』（ARCHIV、2003）

> 校勘とは、つまり古籍の校正のことではないか、と思われるかもしれませんが、校勘と校正には違いがあります。校正は書籍出版のために必要とされる作業の一環であり、校勘は古籍整理事業の中の一種の作業です。例えば中華書局出版の点校本《二十四史》は…学者たちが《二十四史》に対して校勘を行い、段落を分け、標点符号を付けたのは古籍整理に属する仕事であり、中華書局がそれを元に版を組んで印刷製本するに際して、組み上がった版の校正刷りを学者たちの整理した底本と何度も突き合わせて、両者の内容に出入りがないようにした作業は、書籍出版のための校正作業です。つまり、校正の場合には、すでに整理された底本が文字の正誤を判断する絶対の根拠となり、底本の文字の正誤の問題には関与しません。…これに対して校勘の場合には、各種の版本を蒐集(しゅうしゅう)して、それぞれの字句の異同を比較し、原本の文字を考証して、正誤の判断をします。直接の版本上の根拠がない場合でも、誤字や疑問の箇所について文章の原意に沿った判断をしなければなりません。もちろん、その際、古人の文章を書き改めてしまうような結果になってはいけません。これを肖像画に譬えて言えば、校正は、既に在る一枚の写真をそのまま模写するようなものであり、校勘は、さまざまな情報を利用して、本人の姿をより真に近く描き出すというようなものです。

現在の校訂・校正は、校正記号という業界固有の記号を使用したり、場合によっては漢字の使用字体や、送りがなの開き閉じ、表記揺れなどのチェックやルール設定から果ては版権処理など、素人では困難な専門的なスキルが必要になるが、一般に東洋古典学の世界では、校訂・校正は機械的に底本の誤植を修正する行為として、科学的な傍証を伴った学術的考証を必要とする校勘・校定と比べて往々一等低い地位を与えられがちである。これに対して校訂・校正を行う編集者の権利という観点から以下のような疑義が投じられた。

●2005/05/28 小島浩之氏コメント

> よく似たものに出版者の編集があります。本の出版の際に校正をしたり体例を整えたり、こういう行為はどうなるのでしょうね。
>
> 著名な作家や学者の全集や著作集には編集委員がいますが、編集委員に著作権はあり得るのでしょうか？
>
> 常識的に考えれば無さそうですが、もし法的に認められる場合があるとすれば、同様に校訂にも権利が認められる場合があるかもしれませんよね[3]。

●2005/05/29 千田大介氏コメント

> 秋山さんが校正とするものも、しかし現実には「機械的作業」ではあり得ませんよね。「明かな誤り」であることを認定するのに、やはり文脈を読んだり語彙の知識を動員したりという知的思考が必要であるのは、自分で書いた論文であろうと、古典の本文であろうと、本質的違いは無いと思い

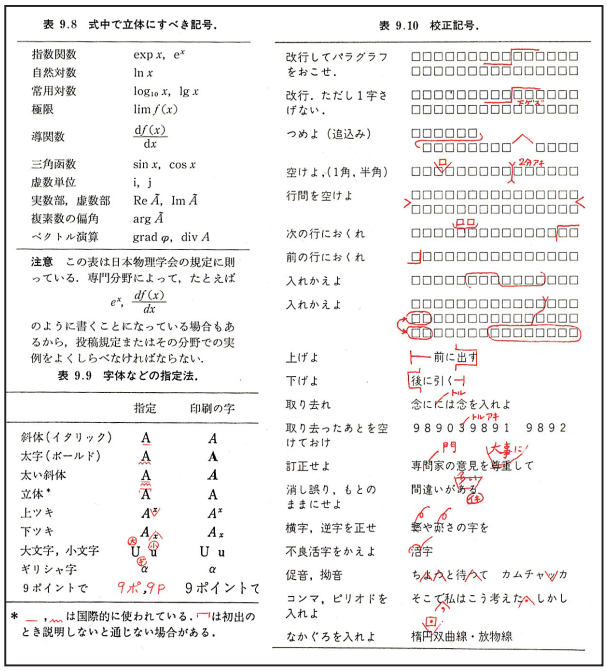


表 9.8　式中で立体にすべき記号.

指数関数	$\exp x$, e^x
自然対数	$\ln x$
常用対数	$\log_{10} x$, $\lg x$
極限	$\lim f(x)$
導関数	$\frac{df(x)}{dx}$
三角函数	$\sin x$, $\cos x$
虚数単位	i, j
実数部，虚数部	$\mathrm{Re}\,\dot{A}$, $\mathrm{Im}\,\dot{A}$
複素数の偏角	$\arg \dot{A}$
ベクトル演算	$\mathrm{grad}\,\varphi$, $\mathrm{div}\,A$

注意　この表は日本物理学会の規定に則っている．専門分野によって，たとえば e^x, $\frac{df(x)}{dx}$ のように書くことになっている場合もあるから，投稿規定またはその分野での実例をよくしらべなければならない．

表 9.9　字体などの指定法.

	指定	印刷の字
斜体（イタリック）	A	*A*
太字（ボールド）	**A**	**A**
太い斜体	***A***	***A***
立体*	A	A
上ツキ	A°	A°
下ツキ	$A_{\circ}$	$A_{\circ}$
大文字，小文字	U u	U u
ギリシャ字	α	α
9ポイントで	9ポ，9P	9ポイントで

＊ ‾，～は国際的に使われている．⌐は初出のとき説明しないと通じない場合がある．

表 9.10　校正記号.

改行してパラグラフをおこせ.	□□□□□□□□□□□□□□□□ □□□□□□□□□□□□□□□□
改行．ただし1字さげない．	□□□□□□□□□□□□□□□□ □□□□□□□□□□□□□□□□
つめよ（追込み）	□□□□□□□ □□□□□□□□ □□□□
空けよ，(1角，半角)	□□□□□□□□□□□□□□□□
行間を空けよ	□□□□□□□□□□□□□□□□ □□□□□□□□□□□□□□□□
次の行におくれ	□□□□□□□□□□□□□□□□
前の行におくれ	□□□□□□□□□□□□□□□□
入れかえよ	□□□□□□□□□□□□□□□□
入れかえよ	□□□□□□□□□□□□□□□□ □□□□□□□□□□□□□□□□
上げよ	前に出す
下げよ	後に引く
取り去れ	念には念を入れよ
取り去ったあとを空けておけ	9890399891　9892
訂正せよ	専門家の意見を尊重して
消し誤り，もとのままにせよ	間違いがある
横字，逆字を正せ	姿や形さの字を
不良活字をかえよ	活字
促音，拗音	ちょんと待って　カムチャツカ
コンマ，ピリオドを入れよ	そこで私はこう考えた　しかし
なかぐろを入れよ	楕円双曲線・放物線

図 2：校正記号（木下是雄『理科系の作文技術』中公新書より）

ます。

違いがあるとしたら、むしろ、行為の具体的過程の微差、それと対象なのかな [4]。

もとより「編集委員に著作権はあり得るのか」という問いへの解答は法律の素人である筆者がすべきものではないので、ここでは校訂・校正がいかなる行為を指すのかを示し、法律の専門家の判断材料となりうるものを提示するにとどめるが、編集者や出版社の権利という観点でも色々と議論すべきことは尽きない。

⊙ 校勘・校定

古典籍の本文を科学的な根拠にもとづいて考証し、復元しようとする行為を「校勘」という。またこの校勘によって新たにテキストを定める行為を「校定」といい、さらに校定によって作成されたテキストを「校定本」という。

これらの行為は、以下に掲げる参考資料にもあるように、（少なくともたてまえの上では）原則として著者の原本（オリジナル）に復元することを目的とする。「科学的な根拠」は、通常は異なる伝承経路をたどった複数の本（edition）との比較によって取得されるが、場合によっては全く別の古典籍から傍証を集めてきたり、同一本中から傍証を探すこともある [5]。

◆［参考］『校勘学講義 —— 中国古典の読み方』

古籍校勘の目的と任務は、本来の姿を保存・復元することに在り、原稿により近い善本を提供しようとするものです。ですから理論的には、校勘の根本原則とは本来の姿を保存・復元することであり、作者の本意に合わない校勘や、原作本来の姿を歪曲する校勘は、この根本原則に違反します。…しかし実際は同一古籍の各種校注本の間には、必ず異文や解釈の相違が存在しています。こうした状況は、本来の姿を完全に保存・復元するということは困難であり、過去の校勘実践の中に表現されている校注者たちの校勘の原則は必ずしも一致していないことを示しています。

この「本来の姿を保存・復元する」という校勘・校定の本分を墨守し、一定の成果をおさめているものとしては清朝考証学者の校勘がまず挙げられる。その一例として、以下に王念孫『読書雑志』の事例を紹介しておく。

◆金谷治『淮南子の思想』（講談社学術文庫、1992）より

さきにも述べた『秋萩帖』（国宝）の紙裏から発見された兵略間詁（かんこ）（後漢高誘注『淮南子』兵略篇。唐代の写本とされています）の古写をみると、その部分に相当する王氏の校語（王念孫のコメント）は十五ヵ条あるが、その十四条までが古写の内容とぴったり一致するのである。わずかに違った一条は、今本に十五字も脱字のある箇所で、さすがの王氏もこれには考えようがなかっ

たわけである。前後の文章や他書の用例によって訂正した点が、後出の資料によってそのまま証明されるというのは、何というすばらしさであろう。（括弧内は筆者による注記。）

しかしながら、洋の東西を問わず、現在では古くから多くの古典籍に多様な異本が存在したことが確認されており、亡佚してしまった異本のことも考慮に入れると、今日では単一原本を想定した復元は極めて困難（もしくは限りなく不可能に近い）と言わざるを得ない[6]。また校注者によっては銘々の思想や解釈を表現する手段として校定本や注釈書を作成することがしばしばあって、校勘・校定が必ずしも「真理の追究」になっていないことが少なくない。以下に掲げる宮崎市定氏の指摘は、中国古典学の一面としてそのような伝統があることを示唆するものである。

◆宮崎市定『現代語訳 論語』（岩波現代新書）後語より

注釈家は本文に対して従属的な態度をとりながら、根本においては注釈家としての権利、その固有の立場を留保している。これは歴史事実の上でも同様である。中国の忠臣というものは、日本でいうような滅私奉公を理想とはしていない。ちゃんと自己の個性を守り、個性を守る立場において忠義を尽くすのである。…だから例えば鄭玄は決して孔子に対する無条件な忠臣ではなく、強い自己主張を持っている[7]。既にある程度、経典に対して貢献する所があったならば、今度は同じ舞台を利用して、自分自身の学識をひけらかす権利があることを主張する。論語の本文以外にこれだけ余計なことを盛りこんでみせたぞ、というのが自慢の種であったようだ。鄭玄注という個人の名を出す以上、むしろそうすることが当然なのである。

なおこれと類似の問題を孕（はら）むものとして、師茂樹氏より親鸞の『教行信証』の事例をご紹介いただいた。

●2005/06/23 師茂樹氏コメント

仏教の場合、ある特殊な読みが「真理」そのものである、もしくは「真理」でなければならない、という例があります。

その一番の例は、親鸞の『教行信証』でしょう。これは、浄土教関連の経論を引用して浄土真宗の教義を叙述しようとしている浄土真宗の根本聖典のひとつですが、この引用に付された訓点は、親鸞の独自のかなり特殊な読みに基づいています。これはもちろん、親鸞独自の創造的読みと言うこともできますが、これこそが永遠普遍の「仏法」だ、とすると親鸞という個人に帰せられるのもまずいわけです。つまり、親鸞こそが「真理」を見出した人である、という考え方です。もちろん、これは、学界のメジャーな考え方ではありませんので、特殊な例のひとつとお考えください。

このように、テキストが特定の個人や国家、社会、宗教などの思想の影響を反映して改訂もしくは歪曲されることがある点もまた、まぎれもなく校定の一側面であるといえる。

しかしながら、このようにある時代や人物・集団・国家の影響を反映しているテキストそれ自体が研究対象になることもある。たとえば儒学には「今古文（きんこぶん）」問題と呼ばれる秦以前と漢以降のテキストにおける字句の相違を、経義の解釈にまで踏み込んで（時として漢代における政治的な都合や学閥争いの要素を含む）議論した歴史があるが、こうした漢代における今古文論争に焦点をあてて『尚書』（書経）や『春秋』といった儒教の経典を復元するのと、それらがそもそも制作された当時の姿の復元を試みるのとでは、自ずと復元される原文も復元するための方法も異なってくる。

⊙ 校讐

校讐には以下の２つの意味があるが、現在は

概ね「校勘」とほぼ同義で用いられることが多いようである。

①紀元前 1 世紀末の劉向（前 77－前 6）が当時、宮中の内外にあった書物を網羅的に整理する際に使用した術語。原義は「校」が一人で異本対校すること（目視チェック）で、「讐」が 2 人 1 組で異本対校すること（聴き取りチェック）。
②原義が転じて、後世「校勘」とほぼ同義で用いられるようになった。

⊙ 点校（校点）

校定作業からさらに一歩進めて（もしくはこの点校や標点も校定の内とすることもできる）、古典籍の本文に句読点をつけたり、固有名詞の範囲を示す標号をつけたりする行為をそれぞれ「点校」・「標点」という。

句読をつけるという行為は一見、校定とは無関係のようでいて、実はしばしば密接な関係がある。

子曰、加我數年、五十以學易、可以無大過矣。
（子曰く、「我れに数年を加え、五十にして以て『易』を学ばば、以て大過 無かるべし」と。）

上に例示したのは『論語』述而篇中の孔子（前 551－前 479）学易の章である。伝統的には、この章によって孔子は 50 歳前後で『易』を学んだとされているが、これに対して六朝末唐初の陸徳明（556－627）が異説を提示している。

《魯》は「易」を読みて「亦」と為す。今、《古》に従う。案ずるに《魯論》は「亦」に作りて下句に連ねて読む。（『経典釈文』論語）[8]

論語には『魯論語』や『古論語（古文論語）』という異なる出自のテキストが漢魏の頃にあった。陸徳明によれば、その中の『魯論語』というテキストでは「易」の字を「亦」に作っていて、これを下の句に掛けて読んでいるという。

五十以學、亦可以無大過矣。
（五十にして以て学ばば、亦た以て大過無かるべし。）

ご覧の通り、このように読んでしまうと、なんと孔子が 50 歳前後で『易』を学んだという事実が霧散してしまうことになる。

実は 1980 年代に河北省定州から出土した前漢末の『論語』の写本でもこの部分は「亦」字に作っているのだが[9]、「易」と「亦」は、字音が同じで漢代以前ではしばしば文脈自由に通用していたため、この竹簡本をもって孔子の学易がなかったことが確定したとは必ずしもいえない。しかし、陸徳明のいうように、「亦」を下の句に掛けるということになると事情は異なる。この場合、孔子は「50 歳になってなお学ぼうとした」のであって、明らかに「『易』を学ぼうとした」のではなくなる。このように句読点の置きどころ一つで古典の意味が大きく変わってしまうことがしばしばある。

この点校という行為については、千田氏より次のような指摘を頂戴した。

●2005/05/29 千田大介氏コメント

点を付けるというのは、古典籍について言えば、大抵の場合は原テクストには存在しないキャラクターを付け加えることになるので、果たして真理追究行為と言えるのか？ という疑問が生まれます。そもそも「？」「！」で疑問・反語の判定がなされていたりすることを考えると、点は確実に本文解釈の範疇に踏み込んでおり、点を付けた人の本文解釈を示す道具であるとも言えます。

まあ、孔子の意図を含めて真理を追求した、という言い方もできるでしょうけど、するとやはり、テクストとその解釈というのをどう定義するか、という文芸理論的問題になってしまいますね。

個人的には、秋山さんの挙げた例で言えば、

加我數年五十以學易可以無大過矣

とするのが真理の復元、点を付けるのはこのテクストを解釈するという行為の結果の刹那的産物であり、創作的な行為であると考えたいですね[10]。

⊙ 標点

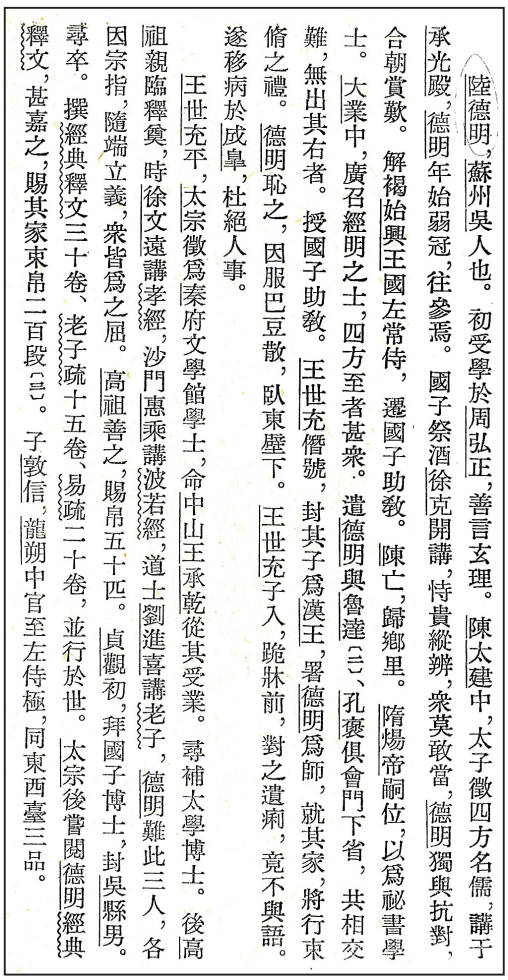
元来白文（句読点も何も振られていない漢字だけの文）の古典籍に、句読点や各種標号を加えた本のことを「標点本」といい、標点本中で加えられている各種の標号や句読点のことを「標点」もしくは「標点符号」などという。その主要なものは以下の通り。

◆『標點符號用法手册』（三聯書店、1996）より

｜ 〰	專名号（傍線・波線）人名・書名など
、	頓号（【日】読点は逗号に相当。）
。	句号（【日】句点）
・	間隔号（【日】中黒は頓号に相当。）
.	句号
,	逗号
：	冒号
；	分号
「 」	引用号（【日】鉤括弧）
？	問号
！	嘆号

このような標点は 20 世紀に入ってから行われたものだが、これとよく似た行為は前近代の日本や中国でも行われており、必ずしも 20 世紀に入ってから全く新しく創出された手法というわけではない。たとえば、松本市美術館や天理図書館が所蔵している狩谷棭齋（かりやえきさい）（1755－1835）旧蔵の『漢書』・『後漢書』（南宋慶元建安黄善夫・劉元起校刊本）を見ると、固有名詞に朱線が引かれているのが分かる。しかも仔細に見ると、人名には中央を貫くように、地名は右側、官職名は左側に傍線が引かれており、さらに書名の中央に二重線が、また年号の左にも二重線がそれぞれ引かれていて、人名・地名・年号を傍線、書名・篇名を波線としたている現在の中国の標点よりもさらに名詞が細かく区別されていることがわかる。この独特のマークアップについては、當山日出夫氏より、荻生徂徠門下の太宰春台（1680－1747）『倭読要領』に次のような朱引き歌があることをご教示いただいた[11]。

図 3：標点本二十四史『旧唐書』陸徳明伝


陸德明，蘇州吳人也。初受學於周弘正，善言玄理。陳太建中，太子徵四方名儒，講于承光殿，德明年始弱冠，往參焉。國子祭酒徐克開講，恃貴縱辨，衆莫敢當，德明獨與抗對，合朝賞歎。解褐始興王國左常侍，遷國子助教。陳亡，歸鄉里。隋煬帝嗣位，以爲祕書學士。大業中，廣召經明之士，四方至者甚衆。遣德明與魯達（三）、孔褒俱會門下省，共相交難，無出其右者。授國子助教。王世充僭號，封其子爲漢王，署德明爲師，就其家，將行束脩之禮。德明恥之，因服巴豆散，臥東壁下。王世充子入，跪牀前，對之遺痢，竟不與語。遂移病於成皐，杜絕人事。

王世充平，太宗徵爲秦府文學館學士，命中山王承乾從其受業。尋補太學博士。後高祖親臨釋奠，時徐文遠講孝經，沙門惠乘講波若經，道士劉進喜講老子，德明難此三人，各因宗指，隨端立義，衆皆爲之屈。高祖善之，賜帛五十匹。貞觀初，拜國子博士，封吳縣男。尋卒。撰經典釋文三十卷、老子疏十五卷、易疏二十卷，並行於世。太宗後嘗閱德明經典釋文，甚嘉之，賜其家束帛二百段（三）。子敦信，龍朔中官至左侍極，同東西臺三品。

みぎところ　なかはひとのな　ひだりかん　なかにはしょのな　さにはねんごう
（右所　中は人の名　左官　中二は書の名　左二は年号）

ただ、こうした朱引きは銘々の読み手（所蔵者）によって手ずから行われているもので、目下、印刷物としてこれだけ精緻なマークアップを施している刊本は管見の及ぶ限りでは見当たらない[12]。

⊙ 校注

一般的には校勘記＋注釈のことを「校注」という。研究者によって校勘のコメント（普通はこれを「校勘記」または「札記」などと呼ぶ）部分を「校注」と呼んだり、「校」（校勘記）と「注」（注釈）を区別した上でそれらを総称して「校注」と呼んだり用法にはしばしば個人差が見られる。校勘記・札記には、改訂前と改訂後の字句とその改訂した根拠を書き記す。校定本においてはこうした字句の修正をした科学的根拠を明示したり、改訂個所と改訂前の原文とを容易に参照し可逆的に復元できるようにするための注記をつけることが重視される。

●2005/06/14 千田大介氏コメント

校定本の学術業績としての認定については、校注が付いていることが必要条件であると、一般的にされているのではないかと思いますが、如何でしょう？要するに、作業の結果、なぜそれらの字句に同定したのか、その道筋が明らかでないモノはダメであると。

◆武内義雄『支那学研究法』（岩波書店、1949）

妄改を戒しむ。凡(およ)そ校讐にあたっては私意を以て本文を改めてはならない。底本に異同があるとき、本文はそのままにして別に札記を作って甲乙を論定するか、又は底本の誤を改正してその下に原本の字を注して改訂の理由を説明するか、二者いずれに従ってもよいが、ことわりなしに本文を改竄することは慎まなければならない。

◈ おわりに

以上、いわゆる校訂権を議論するための下ごしらえとして、東洋古典学において校訂やそれに類する一連の行為がどういう種類や範囲の行為を指しているのかを、具体的事例も交えて概観した。実際には本稿で列挙したようなさまざまな概念や行為に細分化され、たとえば「校訂」・「校正」と「校勘」・「校定」の関係のような知的行為の質的な相違のようなものもあったが、それ以上に重要なのは、一見しただけでは表面上に現れない、行為の目的の差なのかもしれない。

- あくまで原作者の作品に対して忠実であろうとする立場にたち、その原文の復元を最終目標とするもの。
- 校定者（もしくは注釈者）自身の思想や主張を表現する手段としているもの。

こうした様々な校訂およびその周辺行為の権利に関する法的な解釈は専門家に委ねることになるが、本稿で概観してきた諸概念がいくらかでもそれに資することがあれば幸いである。

注

[1] 本稿の執筆、および 2005 年度夏期公開著作権講座「東洋学研究と著作権問題」の予備説明において、筆者が特に参考にしたのは以下の諸書である。
- 武内義雄『支那学研究法』（岩波書店、1949。のち角川書店刊『武内義雄全集』 9 に再録。）
- 倪其心著、橋本秀美・鈴木かおり訳『校勘学講義 —— 中国古典の読み方』（ARCHIV、2003）
- 程千帆『校讐広義（校勘編）』（齊魯書社、1998）
- 兪樾『古書疑義挙例五種』（中華書局、1956）
- 池田亀鑑『古典学入門』（岩波文庫、1991）

[2] これまでの既見の限り、これらの術語を厳密に区別し

ている辞書や工具書類が見つからなかった。これは用法が曖昧で、よほど専門的な校讐家でもない限り、区別して使用している例の方が圧倒的に少ないからであろう。たとえば近藤春雄『中国学芸大事典』（大修館書店、1978）や長澤規矩也『図書学辞典』（三省堂、1979）のような邦人による専著にも詳述されていない。

[3] 石岡克俊氏の指摘によれば、こうした編集者による校訂・校正は著作権法上では「著作隣接権」の枠組みに該当する可能性があるという。また編集という行為については、著作権法 12 条に規定されている「編集著作物」という概念があるものの、「素材の選択又は配列に」ついてよほど積極的な「創作性」が認められない限り、編集委員の著作者性を肯定するのは難しいのではないかという。（石岡克俊氏 2005/06/07 コメント）

[4] 「編集者の改善意見」については、創作的寄与の程度により、著作権法 2 条 1 項 12 号に規定されている「共同著作物」としてとらえられる可能性がある点が指摘されている。（石岡克俊氏 2005/06/07 コメント）

[5] 陳垣『校勘學釋例』（中華書局、1959）。同一古典籍の複数の異本を突き合わせて比較する方法を「対校法」、同一古典籍・同一本の内部で比較する方法を「本校法」、他の古典籍における引文や用例から比較する方法を「他校法」、比較すべき資料がなく、あらゆる状況証拠や用例などの知見を総動員して本文字句のあるべき状態を推察する方法を「理校法」という。このように、校勘・校定は必ずしも異本対校に依らないことがある。

[6] 欧米における文献学でも作業仮説として原テキストの存在を想定しえなくなってきている点を師茂樹氏よりご指摘いただいた。（師茂樹氏 2005/05/29 コメント）
- ●明星聖子「『正統なテクスト』の終焉――ドイツ文献学史概説の試み――」（『埼玉大学紀要　教養学部』36 - 2、2000）
- ●明星聖子『新しいカフカ――「編集」が変えるテクスト』（慶応大学出版会、2002）

[7] 鄭玄（127－200）は『詩』・『書』・三礼（『周礼』・『儀礼』・『礼記』）・『論語』など幅広い儒教の経典に注解をつけた後漢末の大儒。

[8] 陸徳明『經典釋文』（通志堂経解所収本）

[9] いわゆる定州漢簡。河北省定州八角廊にある前漢の中山懐王の墓から出土した。河北省文物研究所・定州漢墓竹簡整理小組『定州漢墓竹簡《論語》』（文物出版社、1997）。

[10] 清の段玉裁（1735－1815）は校書の難しさは、底本（著者の書いた稿本）の是非の判断と立説（著者の述べる真意）の是非の判断とがあって、これらは区別されるべきものだという。（段玉裁「與諸同志論校書之難」）点校は後者の領域に属する行為といえるものの、段玉裁は立説の是非も校定のうちだとする。句読をつけるという行為も校定の一貫であるとした著名な儒者に『十三經注疏校勘記』の阮元（1764－1849）がいる。たとえば段玉裁の『説文解字注』（経韵楼本）の出版にあたって、阮元の子の阮長生が校勘を担当した巻六上には他巻にはない句読がつけられている。（補注：経韵楼本の影印は芸文印書館版を参照されたい。通行する上海古籍出版社版の影印は全巻に句読を追補されているなど、経韵楼本刊行時の姿そのままではない。）

[11] この後、さらに小島浩之氏より『実語教講釈』にもこの「書物朱引歌」が載っているとの指摘を受けた。往来物ということで、寺子屋における初等教育として教えられた歌のようである。

[12] 電子化された古典籍のうち、固有名詞を XML（Extensible Markup Language）でマークアップしたものなども、この標点本の流れに帰属すべき本といえるだろう。

2005 年公開講座報告
東洋学研究と著作権

東洋学情報化と法律問題――第 3 回

「校訂」の著作権法上の位置
――校訂権とその周辺（その一）

石岡　克俊（いしおか　かつとし）

◇ 1.はじめに

これから数回にわたり「校訂権とその周辺」と題して、「校訂」[1] とその法的取扱いについて検討を加えていくことにする。この論点は、昨年一昨年と 2 年にわたり、漢字文献情報処理研究会主催の講座や同会企画の研究会において、議論の対象となったところである [2]。

本稿は、これらの議論を踏まえ、校訂権をめぐる諸論点につき検討を加えるものである。ここで展開されるいくつかの議論は、これらの講座や研究会の成果である。殊に、本年夏期公開講座にあたって開設された漢字文献情報処理研究会の電子会議室（BBS）・著作権問題 [特設] において展開された議論やそこで紹介された文献は、人文学を専門としていない筆者にとって有意義であり、本稿の執筆に当たって大変有益であった。

このような機会と場を提供してくれた漢字文献情報処理研究会の幹事諸氏並びに議論に積極的な寄与をしてくれた会員諸氏に感謝申し上げたい。

◇ 2.古典文献のディジタル化

われわれが、現在において、洋の東西を問わず、さまざまな古典文献に接し、その豊穣たる人類の文化的所産に触れることができるとすれば、それは、多くの人々の有形・無形の努力、殊に、数多くの人文学者による文献学的研究の多大なる恩恵の下にあるといえる。このような古典文献をめぐる状況は、今も昔も変わらない。

もし、何らかの変化にわれわれが気づくとすれば、それは情報通信技術の革新 [3] に伴うさまざまな技術的な条件や基盤の差異にほかならない。

このようなコンピュータ及びネットワーク関連技術の発展・一般化に伴い、公有に帰した著作物をディジタル化（電子テキスト化）する動きが始まった。通俗的利用を目的としたものとしては、先駆けとしてあまりにも著名なプロジェクト・グーテンベルグ（**http://www.gutenberg.org/**）があり、わが国においては青空文庫（**http://www.aozora.gr.jp/**）がその代表例であろう [4]。これにより、われわれは必要となれば、いつでもどこでもさまざまな形態で先人たちの足跡に触れ、これらを享受することができるようになったのである。

他方、学術的利用を目的としたものとして、古典文献のディジタル化（電子テキスト化）、さらにこれらのデータベース化の動きがある。古典文献のディジタル化は、次世代への文化の継承、あるいはその保存、とりわけ質的劣化の可能性がより低い媒体（メディア）への移行という意義も認められるが、そればかりではない。近年のコーパス（corpus）を用いた言語学研究や自然言語処理の展開を受け、これらの成果が積極的に文献学研

究に取り入れられている。このように、古典文献のディジタル化はより高次の学術的利用の面においても新たな地平を拓きつつある。

古典文献のディジタル化は、通俗的利用の面においても、また学術的利用の面においても当該文献の本文の確定を必要とする。この本文の確定は古典研究の成果に大きく依存しており、本文確定の原理及びその方法論の体系は「文献学」として一つの学問分野を形成している。

後で述べるように「校訂」は、文献学において大きな位置を占める本文批判[5]と密接に関わっており、それ自体古い歴史を有している[6]。

そこで、本稿では、まず、準備作業として「校訂」が古典の文献学的研究においてどのように意義づけられているのかを、本文批判との関係で把握し、一旦「校訂」を最広義に捉えつつ、われわれが通常「校訂」と呼ぶいくつかの用例に触れ、「校訂」の諸相を素描してみることにしたい。その上で、「校訂」がいかなる幅を有する概念なのか、また、それと隣接する概念にはいかなるものがあるのか、さらに、これを法概念として捉えた場合、著作権法との関係でいかなる位置を占め得るものなのか、明らかにしていく。

◈ 3.古典研究における「校訂」

一般に、古典文献は著作者の自記又は口述の筆記（原手記・原本）として成立し、印刷技術が未発達の段階では、著作者又はそれ以外の者の書写によって展開する。しかし、転写本は、常に時の経過に伴う物理的損傷と誤読や場合によっては功名心といった書写者の精神的状況により、その本文は変化を来たし、異文を生ずる。諸写本の作成や多様化の中で、本文の誤謬と欠陥はその数を増やし、一方で、原手記・原本は失われる。現在、われわれが目にしているのは、多くの場合、誤謬と欠陥にあふれた写本である。

そこで、客観的な方法と規準とによって、また正当な根拠に基づいて、現存する写本群の相互関係を推知し、より原本に近い文献に復元し、本文を確定する必要が生ずる。そのための方法論とそれに基づく判断を、本文批判という[7]。

他方、文献学的研究には、大きく分けて２つのアプローチがある。ひとつは、今述べたところの文献学的批判であり、いまひとつは、文献学的解釈である。前者は、一定の原理に基づき伝来の諸文献を検討し、一定の価値判断を通じて、できるだけ形式、内容ともに文献自身を本来の姿に再建し、少なくともこれらをそのかつて存在した歴史上の正当な位置にまで復帰せしめようとする知的活動であり、後者は、それらの文献を正しく理解し、さらに他人に理解せしめ、かつその成立と生成とを説明しようとする知的活動である[8]。これらは、文献学的研究において、文献的事実の批判的確立とその説明という２つの学問的課題として捉えられ、これらがより厳密な文献学研究の成立に寄与するものと理解されている[9]。

とりわけ、文献学的事実の批判的確立の上に解釈が成り立っていることを考えると、本文批判が文献学研究におけるより重要な部分を占めていることがわかる。この枢要な位置を占めている文献学的批判ないし本文批判を、文献学研究においてこれまでしばしば「校訂」と呼んできたし、これらとほぼ同様の意義を有するものとして理解してきた。本文批判の成果を、しばしば校訂本といって出版するが、これも本文批判と「校訂」の意味の近接性を示す証拠である。

ところで、本文批判ないし「校訂」は、伝来の本文証跡（写本）を根拠とし、可及的に原文に近い本文を推定する活動である。この延長線上には本（校訂本・校注本）の出版があり、そこに報酬という形での金銭の授受が生まれる（今後はディジタル化・電子テキスト化もこうした取引の対象となるかもしれない）。ここに、純然たる学術的活動が、出版技術の一部としての役割を担い、経済活動としての出版事業、ひいては市場取引と結びつくことになる[10]。「校訂」の功利主義的な一面である[11]。この学術情報の市場化ともいえる特質が、後にさまざまな問題を提起することになる。

すでに述べたように、本文批判ないし「校訂」とは、できるだけ形式・内容ともにより原本に近

い文献に再建・復元し、本文を確定するための方法論とその実践であるといえる。それでは、本文批判・「校訂」の狙いとされる原本の再建・復元は、ここではいかなる意味を有するのであろうか。この問いかけは、本文批判・「校訂」の活動の性質を考える上で重要である。

これに対する一つの解答として、従来から原文（Original）／原型（Archetypus）両概念の厳密な区分が指摘されている。原文については既に述べた原手記・原本とほぼ同様に考えてよいが、原型（Archetypus）とは、われわれが持っている知識によって推知することの可能な伝来諸本の最初の分岐点とされる。これへの接近は、絶対唯一の著者自筆手記の再生を意味するのではなく、現在われわれが持ち得る無数の本文を、本来の系譜の中に還元し、位置づけることによって実現される[12]。つまり、現存の知識を前提に、学者や研究者の仮説・検証によって理論的にたどり着くところのものが原型ということになろう。したがって、原型は原本（自筆完成原稿）と同一である必然性はなく、それに限りなく近いものであるということになる[13]。

なお、昨今のドイツ文献学において、大きなパラダイム転換を経験したことが紹介されている。これは、未だ出版される前の草稿や、そもそも公表されることが企図されていない原稿が発見され、文学研究に不可欠と考えられるに至ったことに端を発している。近代以降現代に至る過程で、文献学研究の対象は、伝承された写本相互の語句の異同を問うものから、作者の創作過程で苦闘した跡であるところの草稿上の削除や訂正に関心が移っていった。これは、人々の文学作品に対する関心が、作品から作者の才能に向かい、彼の創作の秘密を明かす草稿、就中、草稿に残された創作過程に注目が集まったことによる。もちろん、この背景には印刷技術の発明以降、写本相互の違いは稀にしか見出されなくなったことも大きな原因である。ただこれらの議論はあくまで近代以降の文献において特徴的なものであり、中世以前の文献においてはこれまでの本文批判が十分に妥当することを忘れてはいけないだろう[14]。

◈ 4.実務における「校訂」の取扱い

学術研究、とりわけ文献学的研究において「校訂」がいかなる地位を与えられてきたのかを、その意義とともに瞥見してきた。こうした学術研究の成果としての「校訂」が、事業者による出版、ディジタル化ないしデータベース化等の実務において、どのような法的問題を孕むのか、また現実にどう対処しているのかを示す二つの興味深い例を紹介することにしよう。

⊙ 岩波文庫版『風姿花伝』を底本とする青空文庫での公開

青空文庫は、すでに公有に帰した文学作品などをボランティアの手によりディジタル化し、電子テキスト化したデータをさまざまな端末で利用可能なようテキストファイルを含めたいくつかのフォーマット（規格）に成型し、インターネット上で公開しているサイトである。

『風姿花伝』は室町時代に活躍した世阿弥の主著であるが、これをディジタル化（電子テキスト化）し、同サイトにて公開を希望する申し入れがなされた[15]。世阿弥は 1443 年に没し、すでに死後 50 年を経ているため、その著作物は公有に帰している。

ディジタル化にあたっては、岩波文庫版を底本とすることが企図されていたが、同版は校訂者として野上豊一郎と西尾実の名が明示してあった。なお、野上氏は 1950 年に、西尾氏は 1979 年に他界している。

青空文庫の代表者は、関係者及び出版社への問い合わせを行ない、版元の岩波書店からは以下のような回答を得た。

文庫版『風姿花伝』を発行する岩波書店は、校訂（者）・校注（者）の取扱いにつき、

- 古典作品の校訂、校注者に対して、翻訳者に相当する独立した「著作権」を認め、印税を支払っている。
- 校訂、校注者が他界した後も、死後 50 年

を経過するまでは、増刷に際し遺族に印税を支払いを行っている。『風姿花伝』に対しても、この方針に従って処理している。

ここで問題となっている校訂（・校注）は、前節において検討した古典研究、とりわけ文献学における本文批判の成果としての校訂である。既に示唆したように、純然たる学術研究が、ディジタル化（電子テキスト化）あるいはその前提たる校訂を通して、出版事業と利害関係を有するに到った象徴的事例だとも言えよう。

なお、この事案においては、最終的に、岩波文庫版を底本とした『風姿花伝』のディジタル化（電子テキスト化）、青空文庫への収録を行わないこととした[16]。

このように、岩波書店では、古典作品の校訂・校注者に対して、著作権による保護に準ずる保護を――少なくとも報酬の面で――行ってきたようである。これは、出版社と校訂・校注者の間で取り交わされる出版契約の中でなされているものであり、そこでの権利義務関係は契約法理に基づくものである。つまり、拘束されるのは契約の当事者のみである。二次的著作物の創作者として位置づけられる翻訳者の場合はひとまず措くとして、校訂・校注者の権利が、当事者以外の第三者に及ぶか否かは、後に続く本稿の検討を俟たなければならない。

⊙　丸善における著作権調査とその処理

かねてより丸善は、さまざまな機関が所蔵する資料のディジタル化業務を請け負っており、その際の著作権処理について数多くの実績を有している。本誌５号において小島浩之氏は、丸善の著作権処理業務担当者の報告に基づき、ディジタル化に伴う著作権の調査・処理の実情及びその問題点について述べ、その後、同社における校訂権の取扱いにつき、その要点を指摘している[17]。

- ●校訂者の問題は、厳密には「創作に相当する行為」（つまり「相当な加筆」の有無）があったかどうかに関わる。しかし編集内部の（ゲラ刷り校訂など）実際工程が分からないと判断できないことが多い。従って、図書本体中から解る校訂者について検討した。
- ●講義を筆記し、諸本により校訂したと思われるものは、校訂者を著作権者とみなした。
- ●名前貸しと分かるものは、著作権者として扱わなくても良いが、この場合、文化庁に名前貸しである具体的な証拠を示さねばならなかった。
- ●現在の出版社では、内部編集者の校訂が大量にあるので、これらを含め版面権（出版権）と呼んでいるようだ。ただし現行著作権法では、著作者、学術研究校訂者、内部編集者という関係がある著作物であれば、前２者のみが著作権関連者。ただし「△△出版社編集部編」と明記されている場合は、出版社を著作権者として見なす。

丸善の場合、著作権者を確定するということは、所蔵資料のディジタル化しその後の利用に関し許諾を受けるために不可欠な処理であり、そうした観点から「校訂」を行った者に対する取扱いのルールを定めている。ただし、ここでの校訂は、後に「校正」として取り上げる編集過程での作業も含まれており、この点には注意が必要であろう。何れにせよ、「校訂」概念の幅の広さと、いかなる場合に校訂者を著作権者とみなしているかは、これ以降において検討すべき興味深い実例である。これらは論を進めていく中で再び取り上げることになろう。

◈ 5.著作権法における「校訂」の位置づけ

	創作性・有	創作性・無
非公有	A：■＋●	B：■＋○
公　有	C：□＋●	D：□＋○

表　　著作物の公有・非公有の別と創作性の有無

⊙ 著作物の公有／非公有と創作性の有無のマトリクス

著作権法において「校訂」がいかなる位置に存するか、表を用いて説明していくことにしよう。まず、行にはある著作物が権利保護期間内にあるか（非公有 [■]）、それともすでに公有に帰したものか（公有 [□]）の別を、列には当該著作物に創作性が付加されたか否か（[● / ○]）の別をとる。すると、4 通りの組合わせが現れる。これらを表のとおり、それぞれ [A：■＋●]、[B：■＋○]、[C：□＋●]、[D：□＋●] と名付け、それぞれにつき具体例の指摘とそれに関連する著作権法の諸規定を触れることとする。

⊙ セル A：■＋●

これは、ある著作物につきそれが創作されるのと時を同じくして、もしくは創作がなされた後において [18]、何らかの創作性が付加される場合である。

著作権法上、前者の例として「共同著作物」（著作権法 2 条 1 項 12 号）[19] を指摘することができ、後者の例としては「二次的著作物」（著作権法 2 条 1 項 11 号）[20] をあげることができる。

まず、「共同著作物」として認められるためには、(i)二人以上の者の間で共に創作することにつき意思の連絡が存在していること（共同性）[21]、(ii)いずれの創作者においても何らかの創作的寄与が認められること（創作性）、(iii)各人の創作的寄与を分離して個別に利用することが不可能であること（分離不可能性）を要する。

なお、ここで問題となり得るのは、(ii)にあげた創作性である。なぜなら、(iii)は（共同）著作物の外形や態様を見れば、ある程度明白であり、また、(i)は「校訂」をなす場合にそもそも共同性を欠くものは想定し難いからである [22]。次に、「二次的著作物」について、である。これは、ある著作物（原著作物）の翻訳や編曲、あるいは二次元キャラクターの三次元化に代表される変形、また、脚色や映画化といった翻案など、ある著作物（原著作物）に何らかの創作性を付加したものをいう。二次的著作物の創作者は、当然著作権者としての地位を有するが、当該二次的著作物の原著作物の著作者は二次的著作物の著作者と同一の種類の権利をも有している（著作権法 27 条 [23]・28 条 [24]、なお同法 11 条 [25]）。

これら両方に関連して、かつて、わが国の古典文献の翻訳（英訳）作業に関与した外国人を共同著作権者と認めるか否かが法廷で争われたことがある。「英訳平家物語」事件として知られている事件である。後掲【6 参考：「英訳平家物語」事件】をご覧いただきたい。

⊙ セル B：■＋○

これは、著作権法の保護期間内にある著作物に何らの創作性も付加されない場合である。典型的には、出版物など印刷物を編集する工程における「校正」がこれに当たる。校正とは、印刷物と原稿とをひきあわせて、その文字などの誤謬をしらべただすこと、すなわち印刷物と原稿との校合である。通常、校正は原稿に基づき校正刷りの誤字・誤植を正す作業 [26] であるから、一般に、ここに創作性の入る余地はない。

著作権法は、出版契約及び出版権設定に関する規定を有しているが、ここにおいて出版権は著作物を原作のまま印刷その他の機械的または化学的方法により文書または図画として複製・頒布する専有権を内容とする旨定めている（著作権法 80 条 1 項）。複製に当たっては「原作のまま」再現することを要するから、著作物の内容に変更を加えることは許されない。最低限許容されるのは、誤字・脱字・仮名遣い等を補正するいわゆる校正の範囲内とされる [27]。このように、編集工程における校正者は、当然のことながら、共同著作者になることが想定されていないのは、かかる取扱いからも明らかである。なお、出版権を設定した著作権者の権利としてどの程度校正権が認められるかは、法の規定上からは明らかではなく、出版権設定行為又は慣行によって定まる [28]。

⊙ セル C：□＋●

これについては、3 通りの可能性を考えること

ができる。ひとつは、保護期間満了前のある原著作物に何らかの創作性を付加し、その後、原著作物が公有に帰した場合である。この場合、当該著作物は創作時において二次的著作物として現れ、原著作物の公有後は創作性付加部分につき通常の著作物として保護期間満了まで著作権が存続することとなる。

いまひとつは、そもそもすでに公有に帰した著作物に何らかの創作性を付加する場合である。例えば、古典作品の翻訳や編曲、又は脚色・映画化などの翻案がこれに当たる。さらに、そもそも著作物とは認められないもの[29]／著作物と認められても権利の目的とならないもの[30]に何らかの創作性を付加する場合もある。

⊙ セル D：□＋○

著作物が公有に帰しており（ないし、上記同様そもそも著作物とは認められないもの／著作物と認められても権利の目的とならないものであり）、かつ、何ら創作性が付加されていない場合である。著作権法上、このような場合を議論する実益は必ずしも大きくないようにも思われるが、古典作品のディジタル化や、校訂行為の隣接権的構成の可能性を含め考えるとき、検討を要する論点として再び姿を現すこととなるだろう。

⊙ 「校訂」の著作権法上の位置

表の各セルにおける検討を終えたところで、古典作品のディジタル化及びデータベース化、又は「校訂」が法律上いったいどのような位置にあるか整理・確認しておくことにしよう。言うまでもないことだが、ここで取り扱うのは、古典作品であり、これらは既に公有に帰した著作物である。したがって、表中、［領域 A - B］の組合わせ以外を見ておけば充分である。

● 領域 A-C

まず、A から C にまたがるものとして、「編集著作物」（著作権法 12 条）[31]と「データベースの著作物」（著作権法 2 条 1 項 10 号の 3）[32]とがある。いずれも素材ないし情報が著作物であろうとなかろうと、また、保護期間内にあろうとなかろうと、問題ではない。編集物においては「素材の選択・配列」につき創作性が付加された場合、データベースにおいては「情報の選択・体系的構成」につき創作性が付加された場合、著作権法上の著作物として認められる。

● 領域 B-D

つぎに、B から D にまたがるものには、一般に行われているさまざまな素材のディジタル化をあげることができよう。ディジタル化について、そもそも創作性を認める余地はなく、現行法において狭義の著作権（著作者人格権・著作財産権）を認めることは困難である。ただ、これに対し著作権法上の保護（ディジタル化権）を求める声もないわけではない。この場合、理論的には著作隣接権による対応も可能だと思われるが、従来から著作隣接権が認められているレコード製作者等に比べ、これによる保護の正当性を見出すことは困難と言わざるを得ない[33]。

● 領域 C-D

古典作品の「校訂」は、このカテゴリーに属する。これまで見てきたように、古典作品はすでに公有に帰しているので、ここでは「校訂」という行為が著作権法にいう創作性を有し得るのかどうかが問題となる。

しかし、この二項対立で議論は終りではない。仮に創作性があると認められるとしても、通常言われているように単なる模倣や盗用を排除するという意味ではないことは明らかである。一旦、公有に帰した古典作品に、再び校訂権という名の新たな権利を生じさせ、半世紀以上に及ぶ保護を安易に付与することが、著作権法の目的にいう「文化的所産の公正な利用」と整合的か否かという点については未だ検討の余地がある。

他方、仮に創作性が認められなくとも、ディジタル化（権）のところで指摘したように著作隣接権として校訂者の権利（校訂権）を考えることもできる。ドイツの著作権法はかかる考え方を採用している。

いずれにせよ、他国の状況等も踏まえ、校訂についてはさらに踏み込んだ検討が必要である（つづく）。

◇ 6.参考：「英訳平家物語」事件

● **事件の概要**

Y（被控訴人・被告、日本人）は、昭和 40 年 8 月から、岩波書店の覚一本平家物語の翻訳（英訳）を始めたが、その際、日本人的な表現になることを避けるため、知人の米国人 A（訴外）に英訳の点検を受けたり、文章のスタイルや言葉の用法につきアドバイスを得ながら、約 7–8 ヶ月にわたり 2 人で翻訳作業を続けていた。その間、平家物語 1 巻の 7 章ぐらいまで綿密な検討が加えられた。

しかし、A が昭和 41 年に帰国することになり、Y はその後任を求めていたところ、X（控訴人・原告、日系三世の米国人）を紹介された。A と同様、英訳の助力校訂を求めたところ、X はこれを承諾。X 及び Y は翻訳作業を続けることとなった。なお、X は「平家物語」の原典を読み、理解する能力はなかった。

翻訳作業は、Y が原典にしたがって英語訳を作成し、タイプの後、X が訳文の文法上の間違い・用語の訂正、きごちない英文、堅苦しい英文、退屈平板な英文を適当な英文に変更し、その訂正変更部分について、Y が原典を X に説明しながら共に再検討を加え、最終的に Y が原典と照合して訳文を決定するという手順で行った。

この作業は、昭和 41 年 2 月から X が帰国する昭和 42 年 7 月まで続けられた。X が帰国した時点で、「平家物語」の 6 巻 10 章ぐらいまでの翻訳は終わっていた。

Y は、X の協力が得られなくなったので、他の外国人数名から同様の協力を得て作業をすすめ、昭和 47 年 2 月ごろ、「英訳平家物語」（巻の一から巻の十二及び灌頂）の翻訳を完成させた。

その後、Y は、昭和 47 年 5 月 1 日、東京大学出版会と単独で本件「英訳平家物語」の出版契約を結び、単独の著作物として出版の運びとなった（なお、出版前に X からの異議をうけ、X および Y の間で交渉協議がもたれ、その結果、X を共同翻訳者として氏名を表示すること、そして出版による印税については両者折半との合意がなされている）。

また、「英訳平家物語」翻訳についての新聞記者のインタビューにおいて、Y は、あたかもほとんど独力で「平家物語」の翻訳をしたかのように発表し、その旨の記事が報道されるに及んだ。

本件は、かような事実関係を踏まえ、(i) Y が「英訳平家物語」を単独の著作物として出版しようとしたこと、また、(ii) Y が新聞記者に対し、独力で翻訳したかのような発言をしたことにつき、これらが共同著作権侵害行為である旨主張し、X が Y に対し損害賠償及び謝罪広告の掲載を請求した事案である。

● **【地裁判決】**

（判旨）請求棄却

「翻訳とは、「ある国語で表現された文書の内容を他の国の国語になおすこと（広辞苑）Webster に は rending into another language express the sense of in the words of another language interprete explain or recapitulate inother words. とある」をいうから翻訳者とは特定の国語で書かれた原典の意味を理解した上で、その原典を他の国語で表現できる者をいい、ある翻訳がなされた場合、その翻訳物の著作権は、特段の意思表示なき限り、そういうことをなし遂げた人に帰属することはいうまでもない。しかして、原典の翻訳作業に複数の者が関与した場合、誰が翻訳者であるのか問題となるが、翻訳作業に関与した者の中から翻訳者を決定するには、関与者が基本となる翻訳、校訂、再校訂、完訳と続く一連の翻訳作業の中で如何なる役割を担ったかという質的面と関与者が翻訳された書物の全体の如何なる分量の翻訳作業にたずさわったかという量的面とを相関的に評価して決定すべきである。特に関与者の翻訳作業の中での役割を評価するにあたっては、翻訳には、原典に対する正確な理解と移し換える国語への精通が必要であるから、右関与者

の原典の理解力、移し換える国語の精通性の程度が重要な要素となる」[34]。

「而して原告が被告に与えた援助は…被告の行った英語訳につき文法上の間違いを正し、用語の訂正、変更、リズムの調整を行い、英語を母国語とする人から見ると感ぜられるぎごちなさを正し、更にそれらの訂正、変更部分につき被告から原典の説明を受けて二人で再検討し、最終稿は被告が決定したものであるから原告の寄与は、被告には難しいぎこちなさの除去、リズムの調整という質的に高い部分を含んでいるがこれを以て翻訳とみることは相当ではない。このことは被告は原典を理解し、これを英語に訳し得る能力をもっているから作品のよしあしは別として単独でも翻訳をなし得るのに対し原告は原典を理解できないのであるからそもそもそうした翻訳ができないことを考えても明らかである」[35]。と述べ、X の請求を棄却した。これに対して X が控訴した。

- **【控訴審判決】**

（判旨）控訴棄却

「…右「英訳平家物語」は、著作権法上の翻訳著作物に該当するというべきところ、翻訳の定義はさて置き、右「英訳平家物語」作成の過程において控訴人が果たした役割およびその成果に着目するならば、右「英訳平家物語」の創作には、控訴人独自の、被控訴人と対等の立場よりする、創意工夫や精神的操作が存在する、というべく、しからば、この点において、同法上、控訴人は、右「英訳平家物語」につき、共同著作者としての地位を有する、と認めるのが相当である」[36]。

「そこで、…付言するに、著作権法上共同著作者となり得るためには、その要件の一つとして、創作の際の共同関係の存在を必要とするところ、右共同関係の存在は、客観的にみて、当事者間にお互に相手方の意思に反しないという程度の関係の存在、をもって、必要かつ十分とし、右共同関係の存在は、当事者間の経済的対価の支払いの有無とは関係ない、と解するのが相当であるから、右見地からするならば、前叙認定からして、本件においても、控訴人と被控訴人の本件英訳への関与につき、右共同関係の存在を認めるに十分というべきである」[37]。

「ただ、控訴人が本件英訳に関与した分量については前叙認定のとおりであって、右認定からすると、形式的には、控訴人が右英訳の全部にわたって関与していないこと、明らかである。／しかしながら、前叙認定にかかる、本件「英訳平家物語」の原典たる平家物語そのものが前叙各巻から成り、その各巻が独立の内容を持ちながら相互に密接に結びつき合って一貫した一つの物語を構成しているとの点、右「英訳平家物語」も又、原典たる平家物語の右構成に相応する構成を取っている点、したがって、右「英訳平家物語」の内控訴人の関与した部分を分離しては、右「英訳平家物語」が一貫した一つの物語として成立たない点、のみならず、控訴人が関与した部分自体についても、その性質上、控訴人と被控訴人の寄与度が明確に分離計量できない点、現に、本件「英訳平家物語」は、控訴人の関与した部分を含め一体として、一つの英文学作品としての評価を受けている点、特に、東大出版会が本件英訳原稿を審査した際の、同出版会担当者の、右原稿の内平家物語巻の一ないし巻の六に相当する分とそれ以後の巻に相当する分の間に内容的差異はない、と評定されている点、控訴人以後本件英訳に協力した人々が、当時の職業、専攻科目からみて控訴人と同等の英文学的素養および詩才を持っていたとは認め得ない点、を総合勘案すると、控訴人の本件英訳に関する創意工夫は、被控訴人の本件英訳における創造的精神的活動に作用し、それが、控訴人の関与なしに行われたその後の本件英訳にも引継がれ、あるいはこれに強い影響をおよぼした、と推認することができ、この点からすると、本件においては、控訴人が本件英訳に関与した部分を機械的形式的に分離し、その計量から、控訴人の右英訳に対する寄与度を評価することはできない、というのが相当である。／しからば、控訴人の本件英訳の関与量が形式的には全体の約 50 パーセント相当であっても、同人の本件英訳における創意とその精神的労力は、右関与部分を超え、残余の約 50 パーセントの部分にもおよんでいる、と

評価し、控訴人の本件英訳の関与量は、著作権法上、控訴人に本件「英訳平家物語」の共同著作者としての地位を認めるにつき、何等妨げとならない、というべきである」[38]。

「叙上の認定説示から、控訴人に本件「英訳平家物語」の共同著作者としての地位を是認する以上、同人は、右「英訳平家物語」につき、被控訴人と、その著作権を共有する（著作権法 65 条 1 項）、換言すれば、共同著作権を有する、というべきである」[39]。

なお、最終的に、本判決は、(i) X は、Y とともに共同著作者の地位にあるとしたものの、(ii) Y が単独で出版しようとした行為が不法行為に当たるか否かについては、単独著作権であるか共同著作権であるかを判断することは困難であるので、Y には過失がない（仮に損害賠償責任があるとしても、X、Y 間で、請求権を放棄する旨の和解が成立している。）、(iii)また、Y が記者に対して、独力で翻訳したかのような発言をした行為は、不法行為に当たらない（過失がない）、と判断して、X の請求を棄却した。

● **本件の「校訂」との関連性**

本稿との関わりで重要なポイントは、わが国の古典文献の翻訳（英訳）につき本件事実のような関与—「校訂」—を行った者が共同著作者となり得るのか、裏返していうと、本件「校訂」を行った者に創作的寄与を見出すことができるのか、またその際「校訂」に創作性を認めるにはいかなる要素を考慮しなければならないか、ということである。

地裁の判断によれば、2 つの視点を提示している。まずひとつは、関与者が基本となる翻訳、校訂、再校訂、完訳と続く一連の翻訳作業の中で如何なる役割を担ったか（質的面）、いまひとつは、関与者が書物の全体の如何なる分量の翻訳作業にたずさわったか（量的面）である。地裁の判決は、これらを相関的に評価し決定すべきであるとする。

判決は、前者（質的面）との関連で、関与者（校訂者）の原典理解力及び国語の精通性の程度が重要な要素となるところ、本件については関与者（校訂者）が原典を理解し得ないということが決定的に重視され、共同著作者としての地位を否定した。

しかし、判決は、翻訳が当然創作性を有し、それゆえ二次的著作物となるとの認識の下、本件関与者については校訂等、翻訳をするための過程の一部に関与したにとどまり、かつ Y の翻訳作業と同等の創作的寄与としては、X が原典理解能力薄弱であったために認められないとしたのである。つまり、本件は X の翻訳能力の欠如が創作的寄与を排除する根拠となっているに過ぎず、「校訂」それ自体の創作性の有無ないし要否には何ら触れてはいない。その意味では、本件「校訂」が創作性を有し得ることを完全に否定した判例であるとは必ずしもいえないことになる。

また、控訴審判決では、逆に関与者（校訂者）に共同著作者としての地位を認めている。確かに、形式的には関与者（校訂者）が英訳に関与した分量は全体にわたっていなかったものの（約 50 パーセント）、平家物語自体が一貫した一つの物語であり、関与者（校訂者）の寄与部分につき分離計算が不可能な上、関与者（校訂者）がなした創意工夫が全体の創造的精神活動にも作用していたことを認め、共同著作者としての地位を肯定した。

この高裁判決も、翻訳が二次的著作物としての性格を有することを前提として、本件関与者の翻訳作業において果たした役割を創作性を有するものとして肯定的に理解した。

この関与者の果たした役割を「校訂」と捉えるならば、この行為が創作性を有すると判断されたと見るべきであろう。しかし、地裁判決との比較において控訴審判決は、関与者の具体的な作業そのものではなく、その作業が古典作品の翻訳（英訳）にどのような影響を及ぼしたかについての検討がなされ、全体への影響を実質判断している点が注目される。つまり、本件「校訂」の寄与が形式的には一部であっても全体を覆うものであれば、創作性を認める余地があるが、本件のような「校訂」であれば即、創作性有りとするものではないのである。

参考文献

1. 池田亀鑑「第二部国文学に於ける文献批判の方法論」『古典の批判的処置に関する研究』(岩波書店、1941 年)
2. 長谷川鑛平『本と校正』(中央公論、1965 年)
3. 山本桂一『著作権法』(有斐閣、1973 年)
4. 中山信弘『マルチメディアと著作権』(岩波書店、1996 年)
5. 富田倫生「校訂者の権利に関する報告」[**http://www.aozora.gr.jp/houkokusyo/koteisha/koteisha.html**] (1997 年 12 月 17 日) (last visited 2005/09/01)
6. 文化庁長官官房著作権課内著作権法令研究会・通商産業省知的財産政策室編『デジタル・コンテンツの法的保護—著作権法・不正競争防止法改正解説』(有斐閣、1999 年)
7. 半田正夫 = 紋谷暢男 (編)『著作権のノウハウ』(有斐閣、第 6 版、2002 年)
8. 明星聖子「「正統なテクスト」の終焉—ドイツ文献学史概説の試み—」(埼玉大学紀要教養学部 36 巻 2 号、2002 年) 197-207 頁
9. 小島浩之「法理論と実務の狭間—『東洋学情報化と著作権問題Ⅱ』から」漢字文献情報処理研究 (漢字文献情報処理研究会、好文出版刊) 5 号 (2004 年) 52-56 頁

注

[1] 「校訂」の語は、本誌掲載の秋山論文 (「校訂とはいかなる行為か？」) において文献学研究、就中、東洋学研究の立場からその意義の説明がなされている。本稿においても、論を進めていく中で、諸々の文献や専門家の説明を素材に、法律的見地からの整理を試みていく。しかしながら、今の段階では、とりあえず「校訂」の意義を最広義のものとして捉えておく。

[2] 昨年 (2004 年) 7 月 19 日、早稲田大学において「東洋学情報化と著作権問題Ⅱ」と題して夏期公開講座が開催された。ここでの報告や議論の概要については、参考文献にあげた小島浩之氏の論稿を参照されたい。また、本年 (2005 年) 6 月 25 日には、花園大学にて夏期公開講座「東洋学研究と著作権問題」が催された。この講座は、従来の学術研究情報化をめぐる法的問題一般の検討から、学術研究、就中、文献学ないし東洋学研究がどのように法的に評価され、また、これらの諸活動がいかなる法的性格を有し得るものなのかという点に議論の中心を置いたものであった。このことは、自らの研究活動を改めて省み、その学問的方法論を相対化・客観化する作業でもある。その意味では、法律論を超え学術研究の在り様を問う、より根源的な議論がなされた。他方、昨年 (2004 年) 12 月 4 日には、関西大学において「漢籍の情報化—これからの出版文化—」と称するシンポジウムが行なわれた。これは、文部科学省科学研究費特定領域「東アジア出版文化の研究」G 班と漢字文献情報処理研究会との共同企画によるもので、筆者はここで本稿で取り上げることとなった校訂権について若干の問題点と考え方を提示した。

[3] 昨今の情報通信技術をめぐる革新の流れを「IT (Information Technology) 革命」と称することについては、すでに一般に受入れられている。しかし、この言葉そのものから、この技術革新がいかなる意味や内容を有しているかを見出すことは容易ではない。というのも、人類は、これまでに言語の使用や印刷技術の発明、電信電話技術の利用など、幾度となく情報通信技術の不連続な展開を経験しているからである。そこで、筆者としては、今般の情報通信技術の革新を(i)ディジタル化、(ii)コンピュータ化、(iii)ネットワーク化という 3 つ要因の複合として理解することにしたい。これらの要因は、今般の技術革新を意義づける顕著な特徴を示していると同時に、これらの要因を伴った変化が、社会の諸制度に大きな動揺をもたらし、法制度変更の背景となったとされているからである (たとえば、情報通信技術の革新に対応した著作権法改正 (平成 9 年・平成 11 年) についての解説が施されている文化庁長官官房著作権課内著作権法令研究会・通商産業省知的財産政策室編『デジタル・コンテンツの法的保護』(有斐閣、1999 年) 51 頁以下を参照)。

[4] なお、昨今の注目すべき動きとして Google Print Library Project をあげることができる。詳細は以下に掲げる文書を参照してほしい (**http://googleblog.blogspot.com/2005/08/making-books-easier-to-find.html**)。

[5] 文献批判・原典批判ともいう。Textkritik, critique

textuelle, text (or textual) criticism などの訳である。

[6] わが国では書写校正の任に当たる部署が奈良時代から存在していたと伝えられ、中国においては校讎学又は校勘学と称し周代に創始され前漢において確立したという。また、西洋ではギリシャ人によってこれが基礎づけられ、アレクサンドリア時代においてギリシャの全文献が蒐集・検査・整理され、その真偽・正否が批判されたと伝えられている。ただし、いずれも普遍的な方法論に支えられたものではなかった。「校訂」につき明確な方法論が導入されるのは、19 世紀に入ってからのことである。

[7] 池田亀鑑「第二部国文学に於ける文献批判の方法論」『古典の批判的処置に関する研究』(岩波書店、1941 年) 3-4 頁。

[8] 同上 10-11 頁及び 23 頁。

[9] 同上。

[10] 他方、池田・前掲注 [7]17 頁は、「学者は出版せんがために校訂するのではなく、本文研究の結果として、最良の本文が自ら出版される状態に導かれるのである。研究者は、必ず本文研究自体に没頭するのであろうが、出版には必ずしも関心をもたないであろう。文献学及びその方法としての本文批判が、出版のための技術であり、又手段であると考えられては本末転倒である」と述べる。

[11] 池田・前掲注 [7]16-17 頁参照。

[12] 同上 15 頁。

[13] 同上 16 頁は、「もし、…原作者の自筆本の「再建」が如何なる場合にも本文批判の目標とされるのであれば、実際にはそのような目標は殆ど達せられる機会はないのであるから、その結果、一切の本文批判は無力であるのみならず無用であるかの如き、誤れる見解を導き入れ、遂には恣意による本文の選択・改竄等の如き憎むべき行動に、尤もらしい口実を与えることになる」と述べる。

[14] 明星聖子「「正統なテクスト」の終焉—ドイツ文献学史概説の試み—」(埼玉大学紀要教養学部 36 巻 2 号、2002 年) 200-201 頁。なお、1997 年に刊行された「史的批判版カフカ全集」の第 1 巻目『審判』の出版については、同論文を参照。

[15] 経緯については、富田倫生「校訂者の権利に関する報告」を参照。

[16] 富田・前掲注 [15] によれば、著作物に対し補助的な形でかかわる者一般に、独立した著作権を認めることは妥当ではなく、また校訂が高い知的レベルを要するとしても、自由に利用できる方が望ましいと考えていたようだ。また、校訂者の知的努力に対しては敬意を表し尊重はするものの、それは道義的なものにとどまるということを指摘している。しかし、校訂者は岩波書店が提示した条件を踏まえ、自らが著作権または著作権と同等の保護が与えられるという期待を持って作業に当たっていたことを考慮し、最終的な判断をした模様である。

[17] 小島浩之「法理論と実務の狭間」漢字文献情報処理研究 5 号（2004 年）55 頁。

[18] 勿論、著作権の保護期間満了以前において、つまり当該著作物が公有に帰していない段階において、である。

[19] 法 2 条 1 項 12 号：共同著作物　二人以上の者が共同して創作した著作物であって、その各人の寄与を分離して個別的に利用することができないものをいう。

[20] 法 2 条 1 項 11 号：二次的著作物　著作物を翻訳し、編曲し、若しくは変形し、又は脚色し、映画化し、その他翻案することにより創作した著作物をいう。

[21] なお、後述・「英訳平家物語」事件控訴審判決、大阪高等裁判所昭和 55 年 6 月 26 日判決、昭和 52 年(ネ)第 1837 号損害賠償請求控訴事件、無体財産権関係民事・行政裁判例集 12 巻 1 号 266 頁〔280 頁〕参照。

[22] 少なくとも、これまでの著作物の創作にあってはそうであったという経験則に基づいて述べたに過ぎない。インターネット技術の進展はあらゆる態様の創作活動の可能性を否定しない。

[23] 法 27 条：著作者は、その著作物を翻訳し、編曲し、若しくは変形し、又は脚色し、映画化し、その他翻案する権利を専有する。

[24] 法 28 条：二次的著作物の原著作物の著作者は、当該二次的著作物の利用に関し、この款に規定する権利で当該二次的著作物の著作者が有するものと同一の種類の権利を専有する。

[25] 法 11 条：二次的著作物に対するこの法律による保護は、その原著作物の著作者の権利に影響を及ぼさない。

[26] 長谷川鑛平『本と校正』(中央公論、1965 年) によれば、校正の具体的作業として、「主として活版印刷の工程において、植字されたもの、つまり組版を適正に

するために、その校正刷りと原稿とを照合し、伏字・誤植・脱落・組誤り、さらに体裁上の不備、また明らかな原稿の誤りなどを直し改めること」と述べている（20 頁）。

[27] 半田正夫＝紋谷暢男（編）『著作権のノウハウ』（有斐閣、第 6 版、2002 年）188 頁〔半田執筆部分〕。但し、著作者から旧仮名遣いによることを指示された場合には、勿論これに従わなければならない（山本桂一『著作権法』（有斐閣、再版、1973 年）220 頁）。

[28] なお、山本・前掲注 [27]225 頁は、校正を何人が何回まで行い得るかなどは、勿論法定する必要もなく、判例または実務の集積によることを指摘する。

[29]「思想・感情の表現（著作権法 2 条 1 項 1 号）とはいえないもの。例えば、川のせせらぎや鳥や虫の鳴き声など。

[30] そもそも権利の目的とならない著作物（同法 13 条）。例えば、法令、公的機関が作成し発する文書。

[31] 法 12 条 1 項：編集物（データベースに該当するものを除く。以下同じ。）でその素材の選択又は配列によって創作性を有するものは、著作物として保護する。

[32] 法 2 条 1 項 10 号の 3：データベース　論文、数値、図形その他の情報の集合物であって、それらの情報を電子計算機を用いて検索することができるように体系的に構成したものをいう。なお、法 12 条の 2 第 1 項：データベースでその情報の選択又は体系的な構成によって創作性を有するものは、著作物として保護する。

[33] 中山信弘『マルチメディアと著作権』（岩波書店、1996 年）115-120 頁。

[34]「英訳平家物語」事件地裁判決、京都地方裁判所昭和 52 年 9 月 5 日判決、昭和 50 年(ワ)第 577 号損害賠償等請求事件、無体財産権関係民事・行政裁判例集 9 巻 2 号 583 頁〔593-594 頁〕。

[35] 同上〔594 頁〕。

[36]「英訳平家物語」事件控訴審判決、大阪高等裁判所昭和 55 年 6 月 26 日判決、昭和 52 年(ネ)第 1837 号損害賠償請求控訴事件、無体財産権関係民事・行政裁判例集 12 巻 1 号 266 頁〔280 頁〕。

[37] 同上〔280 頁〕。

[38] 同上〔281-282 頁〕。

[39] 同上〔282 頁〕。

2005 年公開講座報告
東洋学研究と著作権

「漢籍の情報化―これからの出版文化」
漢情研第七回大会から

小島　浩之（こじま　ひろゆき）

◇ 大会の概要

漢字文献情報処理研究会第七回大会は、2004 年 12 月 4 日㈯に文部科学省科学研究費特定領域「東アジア出版文化の研究」（**http://eapub.cneas.tohoku.ac.jp/**）Ｇ班とのジョイントセミナーとして、関西大学尚文館 1 階マルチメディアＡＶ大教室にて開催された。当日は季節はずれの台風の余波で、前線の活動が活発化したため途中からやや激しい雨に見舞われたが、60 名近くの参加者を得ることができた。主催者側の一員としてまずは参加された方々に御礼申し上げる。

このセミナーでは「漢籍の情報化－これからの出版文化－」という題目を掲げ山田崇仁（独立行政法人日本学術振興会特別研究員 PD）、秋山陽一郎（京都大学人文科学研究所 COE 技術補佐員）、野村英登（東洋大学講師）の各氏による研究発表と、パネルディスカッション（司会：二階堂善弘（関西大学助教授）パネラー：　相田満（国文学研究資料館助手）、石岡克俊（慶応大学産業研究所助教授）、守岡知彦（京都大学人文科学研究所付属漢字情報研究センター助手）、師茂樹（花園大学専任講師）【敬称略、職名は大会開催当時のもの】）が行われた。

◇ 今年度の本誌と内容的に関係の深い研究発表

前半の研究発表では、まず山田氏（「漢籍コーパスの歩みと現状」）が、これまでの漢籍大規模データベースの歩みを振り返りつつ、現在も有益であるデータベースの紹介と、その問題点の指摘をした。続く秋山氏（「漢籍テキスト処理の現状と展望」）は自然言語処理分野の研究技術を、どのように応用できるかということで、テキストマイニング、形態素解析、N-gram などの話題を初心者にも解りやすく解説した。野村氏（「漢籍と電子出版」）の発表は、電子出版についての総論的なもので、出版史上における電子出版の意義と問題点を論じた。電子出版というものを体系化する上で参考になるものであった。実際に電子書籍を持参するほどの熱の入れようで、参加者にも好評を博していた。

今回のセミナーの案内文には「漢籍の情報化の現状を明らかにすると同時に、『何ができるのか、これからどうすべきなのか』を広く討論したい」とあったが、三氏の発表はまさに「漢籍の情報化の現状を明らかにする」という部分に主眼が置かれたもののように感じた。山田氏の利用する側の視点からの話題、秋山氏の技術的な話題、野村氏の理論的な話題という構成は、この分野の現状と課題について広くカバーしており、大変良かったのではないかと思う。

またこれらの発表は、今年度の本誌の内容にも

深く関係している。山田氏の発表は、特集 1 のデータベースナビゲータの各企画記事にそのエッセンスが散りばめられ、野村氏の発表は、氏の寄稿論文 [1] に内容がまとめられている。秋山氏の発表内容は、特集 1 の秋山氏執筆部分や、特集 2「人文科学研究と自然言語処理」の各論攷に深く関わっている。本誌は特に大会を意識した誌面作りをした訳ではない。にもかかわらずこういった結果が生じるのは、三氏の研究発表が漢字文献の情報化にとって核心をついたものだったからに違いないだろう。

◈ 討論が短く残念だったパネルディスカッション

後半のパネルディスカッションは、司会の二階堂氏が苦笑しておられたように、パネラーに中国学プロパーが入っていないという、前代未聞というか斬新な構成であった。それぞれの専門を列挙すれば相田（国文学）、石岡（知的財産権）、守岡（電気・情報）、師（仏教学）【敬称略】で、中国学どころか出版史（学）や書誌学などの本の専門家もいないことになる。（本ということでは相田氏が最も近いのかもしれない。）

一方でこのように関連分野のスペシャリストが一堂に会することは、非常に有意義なことで、中国史や書誌学を専門とする筆者などは大いに参考になった。異なる分野の研究者からの示唆や意見は、内側から視るより核心をつくことも多く、今回のディスカッションは聞く者にとって良い刺激になったのではないだろうか。

簡単に司会およびパネリストの報告をまとめると次のようになる。

最初に二階堂氏（「青典閲読器による『封神演義』データ処理」）による電子テキスト作成の試みとして「青典閲読器」というソフトの紹介があった。これは簡単にいえば市販の『四部叢刊』と同形式のデータベースを作成できるものだという [2]。相田氏（「日本との関わり　オントロジから見た漢籍処理の問題」）は、日本の電子図書館や研究所等で公開されているデータの中の漢籍を題材とし、データ構造の定義（オントロジ）について、問題提起をおこなった。氏の報告における「日本もそれなりにあるのだ」という一言にはハッとさせられた。日本は遅れていると焦るばかりでなく、今あるものをいかに生かすかということも考えるべきだと反省した。石岡氏（「“創作性” は何処に？東洋古典研究とデジタル化の周辺」）からは、特に校訂権と呼ばれるものをどう考えるべきかについて、創造性という著作権の大原則と学術研究の成果とは何かということと絡めた報告があった。詳しくは次節で述べるが、結果的にこの発表内容は半年後の「東洋学研究と著作権問題」（2005 年度夏期公開講座）に発展的に吸収されている。内容の詳細は本誌の石岡、秋山両論考をご覧いただきたい。守岡氏（「門前の小僧から見た漢籍 s/ の / と / 情報化」）からは、京都大学人文科学研究所の東洋学文献類目、21 世紀 COE、CHISE Project といった各種の活動内容とその現状について報告があった。理系学者から見た人文科学研究への率直な意見が随所に伺われる楽しい（失礼）報告であった。最後に本会の代表でもある師氏（「大規模データベースの使い方」）から、データベースを単なる工具書として利用するだけでなく、データベースを使用してしかできない研究をとの問題提起がなされた。

このように非常に濃い内容のセミナーにも関わらず、質問や討論の時間が十分に確保できなかったことは大変残念であった。パネリストの方も参加者の方も消化不良気味だったのではないだろうか。主催者側としてこの点は深くお詫びしたい。ただ、このパネルディスカッションの内容も研究発表と同様に、直接ないしは間接に本誌の内容に反映されている。このことだけを見ても、第七回大会における一連の報告や討論が、今後の漢籍や出版文化の研究、またデータベースを利用した研究方法に一つの視座を与えたことは間違いないであろう。

◇ 2005 年度著作権講座の源流としての大会

例年本欄では当年度の著作権講座の参加報告を掲載してきた。従って読者諸氏の中には、「東洋学研究と著作権」という公開講座報告に、なぜ昨年度の大会の報告が掲載されているのか、疑問に感じる向きもあるだろう。当初、筆者に与えられたテーマは、今年度の著作権講座の報告であった。しかし今年度の著作権講座の内容は、石岡、秋山両論文に言い尽くされている。屋上屋を架す必要もないと判断し、編集部に我が儘を言って第七回大会の記録に代えてもらったのである。

実は第七回大会の記録は既に「漢字文献情報処理研究会メールマガジン」第 71 号に拙文を載せている。従って前節までの内容は、基本的にこれに加筆・訂正を施したものである。つまり筆者は編集部の方針に従わないばかりか、二重掲載の愚を犯したことにもなる。それでも第七回大会の記録を再び採り上げたのは、それなりの必然性があるからなのである。

古典籍の場合、撰者の著作権は切れているのが普通である。研究者が版本を利用してデジタル化やテキスト化を行うことに著作権法上の問題は無い（ただし版本所有者の所蔵権は別の問題として残っている）。しかし研究者にとって通行本の多くが、点校本である以上、デジタル化の原本として点校本を使用するのが普通だろうし、使用したいと願うはずである。

一方で自分が標点者や校訂者であったとしたらどうだろうか。汗水垂らして点を切り、校勘を施した労作が、何の断りも無いまま自由に扱われてゆく。これを不快に思う場合も当然あり得るだろう。「研究者として他人の点校本は自由に加工したいが、同時に研究者として自分の点校が他人に自由に使われるのは快く思わない。」極論だが、そういうジレンマがあっても可笑しくないだろう。

こなると校訂権を考えるに当たっては、校訂という行為が人文科学研究に持つ意味や、研究に占める位置から解きほぐすことが必要となってくる。その上で、学術発展のために校訂というものに権利を与えるべきか否か、学術情報の円滑な流通のために校訂の権利をどう位置づけるべきなのかを考えねばならない。欲望のままに権利を主張する前に、学術的な側面から権利の本質を議論しなければ、研究者は自分で自分の首を絞めることになりかねないのである。ここに至って「東洋学情報化と著作権問題」という講座名称を本年度より「東洋学研究と著作権問題」に改めた。これは上述の経緯から、自らの研究を権利者として如何にとらえるかを考える時に来ていると判断したからである。

漢情研第七回大会では漢籍の情報化や電子出版について議論する中で校訂権について採り上げた。これが起爆剤となり今年度の公開講座に繋がった。つまり今年度の著作権講座の源流は昨年度の大会にあると言える。今回、敢えて昨年度の大会記録を掲載したのは、昨年度の著作権講座と今年度の著作権講座を繋ぐ存在として第七回大会の意義を書き留めておくべきと考えたからなのである。

注

[1] 野村英登「電子書籍をめぐる状況」（本誌 19 ～ 24 頁を参照のこと）。

[2] 青典閲読器については同氏「中国古典文献における画像と電子テキスト処理」（『中国古典文献における画像及びテキストデータ処理の諸問題（平成 15・16 年度文部科学省科学研究費補助金特定領域研究⑵報告書）所収，2004.12』**http://www2.ipcku.kansai-u.ac.jp/~nikaido/gra_etext.html** も参照のこと。

特集 1

知っててお得！ 東洋学系電脳基礎教養

本会の設立目的は、「東洋学研究にコンピュータをどう活かすか」であり、本会の一般向けの教育・普及目的活動として、1999 年に『電脳中国学』を、2002 年に同 II をそれぞれ刊行している。

俗に電脳関係の時間は「通常の三倍速」とも言われるが、『電脳中国学』を刊行後既に三年もの歳月が経ち、執筆当時とは異なる環境が少なからず出てきており、新しい情報を一度まとめてきたいと常々考えていた。

そこで、現時点における漢字文献情報処理のための基本的な知識・テクニック・マナーなどを本特集では紹介することにした。

これらはいずれも、執筆陣が日頃から利用しているスキルの一端を開示するものである。読者諸賢におかれては、本特集を御一読いただき、参考として、あるいはネタもとしてぜひとも活用していただきたい。

Contents

Windows で多言語・多漢字を使う

二階堂　善弘（にかいどう　よしひろ）

◇ まだやってんの？ Shift-JIS オンリー

Windows や Mac OS などが Unicode を採用してから、もうかなりの時間が経っている。いまや、ワープロ文書はもとより、テキストデータやメールなどにおいても、多くの漢字を使うことができるようになった。

かつて何度も批判したように、Shift-JIS だけでは、墨翟の「翟」にしろ、八佾の「佾」にしろ、鄧小平の「鄧」にしろ、表示したり印刷したりするのに問題があった。むろん、中国語を打つこともできず、你好の「你」も表示できなかった。

中国でも旧 GB コード（GB2312）では文字数が少なくて「朱鎔基」の「鎔（镕）」が表記できなかったのは有名な話である。当時の首相の名前が打てないのは、さすがに問題だったと思う。

むろんこれらの問題はすでに大部分の OS やソフトウェアでは解決済みである。「墨翟」だろうが「八佾」だろうが「你好」だろうが、いまやそれほどソフト上の工夫を必要とせずに表記できるようになった。さらに日本語と中国語の混在もできるようになっている。

これはもちろん、JIS X 0208 に基づく Shift-JIS から、Unicode に技術がシフトしていった結果である。その Unicode も、はじめの頃は約 2 万字の漢字が表記できるだけだったのが、いまや領域が拡張されて、約 9 万の漢字が使用可能になっている。時代は変わったのである。

しかし、世間一般の認識はなかなか変わっていない。たとえば試みに、ニュース関連のサイトを見てみると、そこには相変わらず「深セン（深圳）」「草なぎ（草薙）」「トウ小平（鄧小平）」などの Shift-JIS 依存症の表記が目立つ。これも、多くのユーザが古いブラウザを使っていた時代であればやむを得ないところだが、さすがに現段階ではやめるべきだろう。

◇ まだやってんの？ 中国語独自コード

それから、いまだに中国語を表記する時に、古い Chinese Writer などが使っていた独自中国語コードを使っている人がいる。

中国語独自コードとは、かつて Unicode が使えなかった時期に、日本語と中国語だけを混在するために、苦肉の策として考え出されたものである。実は Shift-JIS の領域なのに、フォントだけを中国語に置き換えて、日本語と中国語を混在させているように「見せかけた」ものである。むろん、当時としてはこれしか方法がなかったので、仕方ないことではある。

しかしいまや、このような「なんちゃって中国語」ともいえる方法を使うことは、百害あって一利なしである。このような方法で作成した文書は、他のパソコンに送ると化けやすいし、メールで送ることもインターネットで表記することもできない。

よく学生が中国語でレポートを書くときに、意

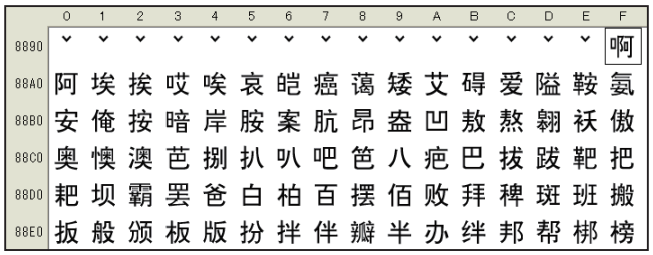
識しないでこの方法を使ってくることが多いが、見せかけの中国語表記にだまされてしまうわけである。中国語が「見えている」からといって、それは真に中国語が表記されていることにはならないのだが、情報教育の場でも中国語教育の場でも教えないためか、いまだに勘違いしている人が多い。この点もいいかげんに改めるべきだろう。

さて、実のところ JIS や GB といったローカルコードも発展してきている。例えば JIS であれば、JIS X 0208 から JIS X 0213 へと進化しているし、GB も GB2312 から GB18030 へと大幅に改訂されている。いずれも、収録される漢字の数を大幅増やしており、Unicode などとの連携も柔軟になってきている。

特に Windows XP は、多漢字や多言語については扱いやすいし、多くのソフトも揃っている。どんどん使わないと、もったいないのである。ここでは簡単ではあるが、Windows XP を中心にその多言語・多漢字の使い方を見てみたい。

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
889O	・	・	・	・	・	・	・	・	・	・	・	・	・	・	・	亜
88AO	唖	娃	阿	哀	愛	挨	姶	逢	葵	茜	穐	悪	握	渥	旭	葦
88BO	芦	鯵	梓	圧	斡	扱	宛	姐	虻	飴	絢	綾	鮎	或	粟	袷
88CO	安	庵	按	暗	案	闇	鞍	杏	以	伊	位	依	偉	囲	夷	委
88DO	威	尉	惟	意	慰	易	椅	為	畏	異	移	維	緯	胃	萎	衣

本当はこのような漢字コードが定義されている。



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
889O	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	啊
88AO	阿	埃	挨	哎	唉	哀	皑	癌	蔼	矮	艾	碍	爱	隘	鞍	氨
88BO	安	俺	按	暗	岸	胺	案	肮	昂	盎	凹	敖	熬	翱	袄	傲
88CO	奥	懊	澳	芭	捌	扒	叭	吧	笆	八	疤	巴	拔	跋	靶	把
88DO	耙	坝	霸	罢	爸	白	柏	百	摆	佰	败	拜	稗	斑	班	搬
88EO	扳	般	颁	板	版	扮	拌	伴	瓣	半	办	绊	邦	帮	梆	榜

しかしフォントだけを中国語に入れかえている。

◇ 各種入力・多漢字ツール

Windows XP は多言語 OS である。

だからアラビア語やタイ語を打ちたい場合、専用のソフトを入れなくても、そのまま対応できるようになっている。Word 2003 などは、左から書く言語と右から書く言語を混在することが可能である。

もっとも、われわれが一般的に使う場合は、そこまで要求されることはめったにないと思う。多いのは、多くの漢字を使うこと、さらに日本語と中国語の混在した文書を作成することであろう。

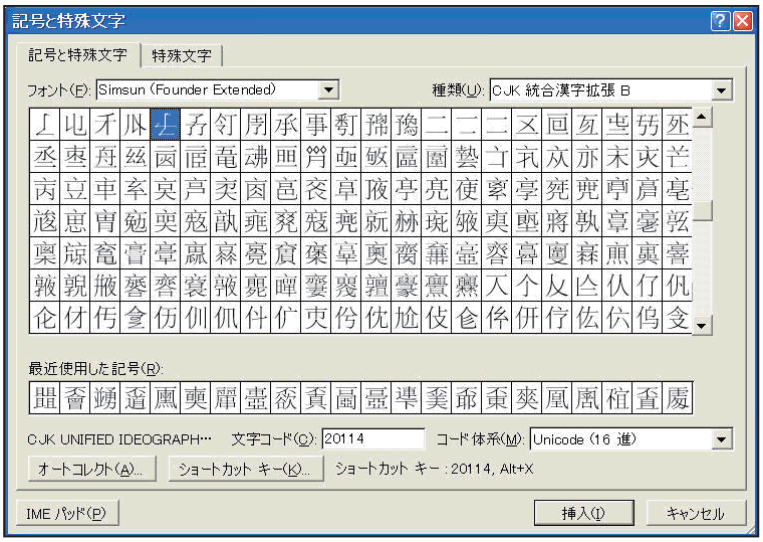
中国語を「表示する」だけなら、Windows XP であれば何の工夫も必要ない。かつては、そもそも中国語を表示させるだけで大変だった。中国語フォントをインストールしたり、表示設定のためのパッチを当てたりなど、いくつもの作業を行ったあとにようやく中国語の表示が可能となった。しかしいまの XP はそもそもデフォルトで中国語と韓国語のフォントが設定されており、インターネットの中国語サイトなどはすぐ見ることができるのだ。ただ、中国語を入力するためには、若干の設定が必要である。

ワープロなどで中国語を入力したい場合は、「コントロールパネル」から「地域と言語のオプション」を選んで、中国語の「MS Pinyin IME 2003」を追加すればそれで使用可能となる。他の設定はほとんど不要である。

しかしもし MS Office を使っているならば、その多言語拡張ツールである「MS Proofing Tools」を購入して使ってみるべきだろう。これは Office が XP であれば、その Proofing Tools XP を、Office が 2003 ならば Proofing Tools 2003 を使うようになっている。残念ながら日本語版はなく、英語版のみの提供であるが、大手のソフト量販店や大学生協ならば扱っているところがほとんどなので、購入するのは容易である。また Amazon のオンライン通販でも買うことができる。

むろん中国語のみに対応したソフトではないし、その機能も豊富であるが、特筆すべきは Unicode の拡張漢字のフォントが含まれることであろう。

Unicode においては、拡張漢字が定義されており、その使える漢字数も大幅に増えている。しかし、領域が定義されているだけで、フォントがな



拡張漢字 B を使う

ければ、ただの空き箱みたいなものである。その拡張漢字のフォントが、すべてではないが収められているのがこの Proofing Tools である。このフォントの名称は、「Simsun(Founder Extended)」という。むろん拡張漢字すべてのフォントには対応していないが、約 7 万の漢字が使えるようになる。

このように、特に別にソフトウェアを購入しなくとも、いまの Windows XP では中国語が使えるが、実際には MS の製品は日本人には使いづらい面もある。たとえば変換文字の候補を変えていく場合、スペースキーを押しても変換しなかったりする。日本語の IME に慣れていると、使いにくく感じるはずだ。

機能的に優れた中国語入力ツールとしては、オムロンソフトの「楽々中国語」に収められる「cWnn」や、高電社の「ChineseWriter」がある。もっとも、いずれも単なる中国語入力ツールというよりは、総合的な中国語パッケージソフトといったほうがよい。

MS Proofing Tools 2003

むろん、現在ではすべて Unicode に対応しており、独自コードは不必要になっている。逆に、かつて独自コードで作成された文書などをコンバートする機能も持っている。

特 に ChineseWriter は、バージョン 7 からは中国の GB18030 に対応し、その扱える漢字数などが格段に増えている。もちろん「朱鎔基」の「鎔（鎔）」だって打てる。辞書のボキャブラリも増えているので、中国語を必要とする現場にはあった方がいい。

このようなソフトは、中国語を打つ時だけに使用すると考えられがちだが、実際には日本語文書を打つ時もかなり使うものである。

Shift-JIS にない漢字を使った人名や地名があった場合、とっさに中国語入力ツールに切り替えた方が、漢字をいちいち探して入力するよりは早い。

たとえば、いま自分などは中国福建の「漳州」に関する報告を書いているが、「漳」の字は Shift-JIS 外であり、さらに辞書にも搭載されていないことが多く、表示させるのは面倒である。むろん、何度も出てくるものであれば、辞書登録をする必要があるが、数度しか使わないこともあるだろう。

そのような場合は、下の IME を中国語（繁体字）に切り替えて、「zhangzhou」と入れれば、すぐに「漳州」と出てくる。特に現代中国に関するボキャブラリであれば、いったん中国語 IME に切り替えて入力した方が便利な場合が多い。そういった場面でも ChineseWriter や cWnn は有用なソフトである。

MS AppLocale

Unicode が普及しても、依然として問題となったのは、中国の GB コードや台湾の Big5 コード

などの、「ローカルコード」をベースとして作られたソフトの動作であった。

いまや多くのデータベースソフトなどは、Unicode を基本として作られているので、日本語 Windows XP の上で動かすことは容易となった。しかしたとえば、国学のソフトなどは、いまだに GB を中心に作られており、文字化けすることが多かった。いや、文字化け程度ならよいが、そもそもインストール自体ができない場合もある。

そんなローカルコードのプログラムを動かすのに、必要なツールがマイクロソフトから提供された。「MS AppLocale Utility」である。このソフトを使えば、ローカルコードで作られたソフトを動かすことができる。使い方は簡単で、まずこの MS AppLocale を起動して、その中でさらにローカルコードで書かれたプログラムを起動する。それだけである。非常に便利なソフトである。

ただ、注意すべき点もある。このソフトはあくまで 1 つのプログラムをエミュレートするだけで、他のソフトとの整合性は必ずしもとれていない、ということである。たとえば MS AppLocale で国学のデータベースソフトを動かし、検索するとき、MS 以外の IME を使うと、入力できないことがある。使う時は若干注意が必要である。これを回避したい場合は、プログラムを動かす「基底の言語」を変更することにより可能である。

なお、MS AppLocale のダウンロードについては、**http://www.microsoft.com/downloads/details.aspx?displaylang=ja&FamilyID=8c4e8e0d-45d1-4d9b-b7c0-8430c1ac89ab** から行うことができる。より詳しい使い方は、本誌のレビュー（163 ページ）を参照していただきたい。

◇ CHISE IDS FIND と各種多漢字ツール

CHISE は一般ユーザにはまだそれほど知られていないかもしれないが、独自のコンセプトで多漢字処理に取り組んでいるプロジェクトで、京都大学人文科学研究所の守岡知彦氏を中心に運営されている（**http://kanji.zinbun.kyoto-u.ac.jp/projects/chise/**）。文字コードに依存せず、漢字を統合的に捉えて処理する仕組みを持つ。

ツールとしては XEmacs を主に使用するため、Windows ユーザでは使う機会が少なかったかもしれない。しかし現在では CHISE IDS FIND（**http://mousai.kanji.zinbun.kyoto-u.ac.jp/ids-find**）を Web サイトで使用できるようになった。これは漢字の部品、または『大漢和辞典』の番号などを入力して、検索の難しい漢字を簡便に使えるようにしたものである。出力結果をコピー＆ペーストすることによって、拡張漢字を入力することもできる。

いまや扱える漢字数が膨大なものとなり、その漢字が、JIS か、Unicode の BMP（基本面）に入っているのか、あるいは拡張漢字 A にあるのか B にあるのか、それとも文字コードに存在しないのか、われわれでも迷うことがしばしばある。その点非常に便利なツールである。しかも、音や構成の情報も含まれており、単なる漢字検索ツールと違っている。

ただ CHISE IDS FIND は、Internet Explorer で使うよりも、Mozilla Firefox などのブラウザで使う方が柔軟に扱えるように思われる。なおこれについても、詳しくは本誌レビュー（165 ページ）を参照して欲しい。

もちろん、他のツールとしてエー・アイ・ネットの「今昔文字鏡」（**http://www.mojikyo.org/**）も Windows でよく使われる多漢字ツールとして知られている。漢字の部品による検索については、非常に柔軟な機能を持っている。数多くの異体字フォントを持つことから、印刷時には威力を発揮する。

ただ、Unicode との連携が弱く、拡張漢字などを処理するのには向かない。さらに、表示フォントはあくまでも文字コードに定義されてるわけではなく、ある意味では先に見た中国語独自コードに近い。その点を踏まえて利用すべきであろう。

最近はあまり使われなくなったようだが、「GT 明朝」も多漢字のフォントとして有名なものである（**http://www.l.u-tokyo.ac.jp/GT/**）。ただ検索についてはあまり配慮がなされていない。またこれ

もフォントと文字コードとの乖離が生ずるので、その点についても注意が必要である。

別 OS の話になるが、パーソナルメディアの「超漢字」には、この GT 明朝が搭載されている。超漢字で使用するぶんには、検索ツールなどが充実しているので、問題はない。むろん、Windows との互換性は乏しいので、その点では注意が必要だ。

◇ 今後の展望

今後は、ローカルコードから Unicode へのシフトはますます進むものと予想される。OS も、ソフトウェアも、対応するのが「当たり前」という状況になっていくだろう。

実は、われわれはほとんどそれと意識しないままに、いつの間にか Unicode を使ってきている。例えば、Google のサイト（**http://www.google.co.jp/**）を見ていただきたい。ここでのデータは、検索・表示ともに Unicode が使われている。こうであってはじめて、世界中の情報が一気に検索できるのだ。

数種類の言語を使う、或いは多くの漢字を使用するならば、ワープロであれインターネットであれ、もう Unicode を避けて通れない。いやむしろ、Shift-JIS のようなローカルコードの使用を避けるのが「親切なサイト作り」になってくると思われる。

もっとも Unicode にも問題は多々ある。Unicode のこれまでの経緯や諸問題については、本誌の師茂樹・小林龍生両氏の論考（157 ページ）に詳しいので、そちらを参照していただきたい。

そのような問題を克服するための試みとして、CHISE のようなアプローチも行われている。Unicode を使いつつも、「その先」についての配慮も行っていくのがより望ましい態度であろう。

漢字文献情報処理研究会　会員制度変更のお知らせ

会員各位には既に BBS・メールマガジン等を通じてお知らせしておりますが、2003 年 8 月の臨時総会での議決に基づき、2004 年度より漢情研の会員制度が以下のように変更されました。

- **一般会員（BBS 利用＋『漢情研』購読）：年会費 3000 円**
- **BBS 会員（BBS 利用のみ）：年会費 1000 円**

従来からの会員の皆様は、自動的に BBS 会員となります。BBS 会員から一般会員への変更を希望される方のみ、会員資格変更の届け出をお願いします。

◇ 会員資格変更フォーム
http://jaet.gr.jp/JAET-BBS/change.html
※アクセスには漢情研 BBS の ID・パスワードが必要です。

データベースナビゲーター

山田　崇仁（やまだ　たかひと）・小島　浩之（こじま　ひろゆき）

◇ 本当に見つかりませんか？その情報

インターネットで情報検索。今日もはや当たり前となった情景である。「わからない」→とりあえず「ググる」[1]とばかり、Google でキーワードを入力して調べてみることなど珍しくもない。

今では、日常生活はおろか、専門のネタ探しや課題・レポートの作成にも使われていたりする。授業中に「何々について記せ」と小課題を出すと、パソコンで WWW ブラウザを起動し、Google や Yahoo から課題のキーワードを入力してそれらしい Web ページを見つけ、適当にコピー＆ペーストで一丁上がり！という受講生も（その是非はともかく）珍しくなかった[2]。もはや、一般的に「インターネット＝何かわからないことを調べてくれるツール」という認識になっているのである。

「探しものが見つからない」。別にこれはインターネットに限った話ではないが、本当に見つからないのだろうか？実は他人はそれをとっくに見つけているかもしれない。では何が悪いのか？あなたの探し方に問題があるのかもしれない。

ここでは、「インターネットでの情報の探し方」をテーマに、各種検索の方法について紹介する。

⊙ データベースの現況

まず始めに、ここでは、「データベース」を「各種調べ物をする際に利用する Web サイト」と定義する。実際には各 Web サイトで提供のデータベースやサーチエンジンが対象となるが、煩雑なので一括して「データベース」と呼ぶこととする。

では、インターネットではどんな「データベース」が公開されているのだろうか？ここでは、一般（いわゆるサーチエンジン系の Web サイト及びそこで公開されている各種検索サービス）・専門（OPAC やテキストデータベースなどの特定の目的に絞ったデータベース）の二つに分け、それぞれを紹介することにする。

ここで紹介しきれないデータベースも数多く存在する。また『四庫全書』や『四部叢刊』に代表される市販データベースも数多く存在する。それらについては、本誌や『電脳中国学』・同Ⅱのレビュー記事を参照していただきたい。

◇ サーチエンジンを使いこなせ

サーチエンジンとは、「インターネット上の情報を検索するサービス」である。

サーチエンジンには、サイト名＋紹介文を登録したデータベースを検索するカテゴリタイプと、インターネット上の情報を直接データベース化して登録するロボット型の大きく二つに分かれる。この分野の老舗である Yahoo! や本会で提供している漢風はカテゴリ型、今やサーチエンジンの代名詞ともなっている Google はロボット型である。

情報の量ではロボット型がカテゴリ型を圧倒するが、カテゴリタイプも小規模な Web サイトでも構築可能なこともあって、ある特定分野に絞った専門の検索サービスとしてよく使われている。

また、最近のサーチエンジン系 Web サイトでは、単なる Web 検索サービス以外に、さまざまなものを対象に検索するサービスを提供している。この分野では、Yahoo! や infoseek などの大

手検索サービスが、オークション・ショップ・地図・翻訳など Web サイトを利用してもらうためのポータル化をすすめている。しかし、ここにも巨人 Google が攻勢を強めているのが現状である。

ここでは、その巨人 Google を「もっと使い来なしてみよう」をテーマに、Google の多言語方面の使い方を紹介する。

⊙ Google を使いこなせ！

サーチエンジンの定番中の定番、それが Google である。俗に「Google 八分」という言い方もある。これは「Google で探せなければ、それは WWW 上に存在しない Web サイトである」という意味である。実際には、Google が検索しないようにすることも可能なのだが、一般的な感覚としては、あながち間違っていないだろう。

かくいう筆者も「とりあえずわからなければ Google で調べる」タイプである。しかし、目的の情報が見つからない場合も多い。その場合にどうすればよいか、それを先ず述べてみよう。

●キーワード選択と検索式の利用

まず、最初は、キーワードの選択である。

探し物が漠然としている場合、それに関連する単語を片端から組み合わせてキーワードに指定してみよう。特に、固有名詞や機能名、特定の表現などは有効なキーワードになる可能性がある。

キーワードを入れてみても思ったような検索結果が出なければ、類似の別なキーワードに変えてみよう。例えば、「おいしい京都の料理店」を調べたい場合、キーワードとして「おいしい」だけではなく「美味」「評判」「旨い」等に置き換ると、また違った検索結果が表示されるだろう。

更に、キーワードには「〜は」「〜に」のような助詞や助動詞などを入れない方が、より多くの検索結果が表示される（情報を絞り込みたい場合には、あえて含めるのも一つのテクニックである）。

これら複数のキーワードは、検索式を使って組み合わせることによって、より効率よく調べることができる（初めは and 検索だけで十分）。

探し物によっては、一回で目的の情報にたどり着けない場合もある。その時は以下の例を参考にして、少しずつ情報を絞り込む方法を使えばよい。

検索条件	検索式
or 検索（A または B）	A or B
and 検索（A かつ B）	A B
not 検索（B を含まない A）	A -B
空白も含めた文字列を一括検索（フレーズ一致）	“A B”

Google の検索式

例：中国正史の筆頭の書物の作者の名前をもじってつけられた作家の本名は？

この場合、いきなり［正史　筆頭　作家　本名］などをキーワードに入力して検索しても、まず見つからないだろう。そこで、見つけたキーワードをいくつかの固まりに分けて検索する方法を使う。

1. ［中国正史］の［筆頭］を検索→［史記］。
2. ［史記］の［作者］を検索→［司馬遷］
3. ［司馬遷］の［名前］をもじった［作家］→［司馬遼太郎］
4. ［司馬遼太郎］の［本名］→［福田定一］

●多言語検索を使いこなせ！

これら Google の検索サービスは、原則として多言語対応である。その為、キーワードによっては日本語と中国語など、複数の言語が検索結果に同時に表示されることも珍しくない。

しかし、検索対象の言語を指定した方がより有為な結果が得られる場合が多い。言語を指定しての検索は、検索時に「検索オプション」をクリックしてオプション設定画面を表示し、「言語」→「検索の対象にする言語」を適宜選択し、キーワードを入力して検索するか、「言語ツール」のページに移動して、「各言語、国に絞って検索」の所で検索すればよい。

●Web 以外の検索サービス

Google は、Web 検索以外にも様々な検索機能を持つ。例えば、画像専門のイメージ検索がある。

これはキーワードに関係すると思われる画像（そのページのテキストを見て判断するようだ）を検索するサービスである。その他に、ニュース専門の検索サービス Google News も便利だ。また、最近衛星画像と地図とをリンクさせた Google Map や地域情報と地図をリンクさせた Google ローカルが公開され、既存の地図サービスとは異なる利用法が注目されている。

それ以外にも、株価・路線・辞書・宅急便の調査など、Google を媒介とした各種データベースへのアクセスが可能である（実は計算機としても使えたりする）。

この分野に関係する試みとしては、google scholar（論文検索）や図書検索 google print（図書を検索）が物議を醸している。

また、各自のハードディスク内のデータを対象とする検索サービスである、Google Desctop Search も公開されて久しい。

更にはメールサービスの G-Mail、インスタントメッセンジャーの Google Talk など、検索以外のサービスの提供も試みられている。

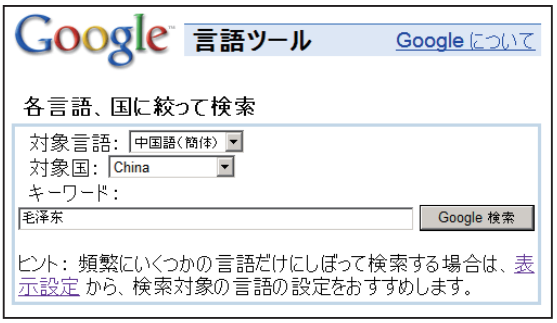
Google はもはや単純な Web 検索サービスではない。利用者が、「Google 無くしてはコンピュータ利用が成り立たない」という環境構築を志しているのである。

⊙ それでも見つからないときは？

上記のテクニックを使えば、これまでよりも効率よく情報を検索できるだろう。それでも見つかりにくい場合はどうするか？その場合以下の二つの方法を使ってみるとよい。

まず一つめが、「別なサーチエンジンを使う」方法である。

サーチエンジンにはそれぞれ癖があり、検索結果の数や上位に表示される Web サイトの種類が異なっている。従って、一つのサーチエンジンで見つからなくても、別なサーチエンジンを使えば見つかる場合もある。個別のサーチエンジンを渡り歩くのが面倒ならば、複数のサーチエンジンを同時に利用する「メタサーチ」サービスを提供する Web サイトを利用すればよい [3]。



「言語ツール」より、「各言語、国に絞って検索」

また、学術情報など、特定分野専門のサーチエンジンやデータベースもあるので、検索目的によっては、それらを利用した方が効率的だ（後述の「専門のデータベースを使いこなせ！」を参照。）。

二つめは「リンクを辿る」方法である。

Web ページの中には、関連情報へのリンクがまとまって掲載されていることがある。それを辿ることによって、場合によってはサーチエンジンよりも効率よく目的の情報に辿り着けるだろう。これは論文を読むときに脚注を読んで関連情報を知るというテクニックと同様なものといえる。

（以上、山田担当）

Google Print β版を使ってみた

◇ 東洋学系データベースを使いこなせ！

ここでは、東洋学研究者にとって有用な各種専門系データベースについて紹介する。

これら専門系データベースは、即時性という観点からはサーチエンジンに後れを取り、インターフェースもデータベース毎に異なるため、個々の使用方法や癖を把握する必要がある。しかし、目的の情報以外のノイズが混じり難いため、逆に必要な情報に手早くたどりつけるのである。

研究者たる者、これらを使いこなしてより効率よく研究を進めたいものである。

本文中の数値は基本的に 2005 年 8 月現在のものである。また適宜『電脳中国学 II』(「電中 2」) および本誌各号（JJ1 ～ JJ5 と略）の関連箇所を示しておいた。詳細はそちらを参照されたい。

⊙ 漢籍や論文を探す

●全国漢籍データベース

http://kanji.zinbun.kyoto-u.ac.jp/kanseki/

全国漢籍データベース協議会（**http://kanji.zinbun.kyoto-u.ac.jp/kansekikyogikai/**）が、日本における漢籍所在総合目録の作成を目指して構築しているものである。現時点で 34 機関の漢籍目録がデータベース化され、新たなデータも順次追加されている [4]。

システム面では UTF-8 ベースで異体字テーブルも用意してあるので、漢籍目録だからといって正字に縛られる必要はない。また詳細検索画面では、子目からの検索ができる。これはほぼ『漢籍分類目録』の機能に相当するものと言えよう。

惜しむらくは、使用法の説明など利用者への気配りが十分でない点が見受けられることである [5]。

●東洋学文献類目検索

http://kanji.zinbun.kyoto-u.ac.jp/db/CHINA3/index.html.ja

『東洋学文献類目』は京都大学人文科学研究所が、東洋学関係の文献目録として、東方文化学院京都研究所時代の 1935 年から発行している。国内外の関係書籍・論文が収録され、東洋学の研究者がまず紐解くべき基本工具書の一つである。この『東洋学文献類目』のデータベースは、以前 CHINA3 と呼ばれていた。現在では冊子体と同じ名称となり、6.0 α版まで公開されている。インターフェイスの改良も進められ、外字も減りタイトル以外に著者名からも検索できるようになった。現在の収録範囲は、第 4 版が 1981 年度版～ 2000 年度版、第 6 α版は第 4 版分に加えて 2001 年度版～ 2003 年度版（部分的に 2004 および 2005 年度版も含む）のデータだということである。

●CNKI

CNKI とは中国学術情報データベース（China National Knowledge Infrastructure）の略称であり、学術雑誌、新聞、博士論文、会議論文の 4 つのデータベースからなる。情報量は膨大であり、これまでは日本で入手できなかった雑誌の論文も簡単に手に入る。本文閲覧は有料だが一部書誌データの検索はフリーなので、論文索引や記事索引として利用できる。詳細は後掲のレビュー（「CNKI：中国最大の電子ジャーナル」）を参照。

⊙ 便利な電子図書館

一口に電子図書館と言っても、その範囲はかなり広いものになる。そこで本稿では、原則として「①図書館・研究機関・企業などが運営する大規模なもので、②書籍を画像化、もしくはテキスト化し公開していること」を条件として紹介する。

●画像系電子図書館

●京都大学電子図書館貴重資料画像

http://ddb.libnet.kulib.kyoto-u.ac.jp/minds.html

日本における電子図書館の老舗的存在。画像の閲覧には、上記 URL より「貴重資料画像のアイコンをクリック＞京都大学貴重資料画像のページ内の、“さがす”タブを選択」する。コレクションのうち、一般貴重書（和）、谷村文庫、清家文庫、

近衛文庫、中国清代民国公私文書コレクション（部局所蔵資料＞法学部図書室のプルダウンメニュー参照）は漢籍や中国の一次資料を多数含んでいる。

【参照】JJ2（p.158），JJ4（p.139-），JJ5（p.150）

●国立公文書館アジア歴史資料センター

http://www.jacar.go.jp/

日本における大規模デジタルアーカイブズの嚆矢。明治期より第二次大戦期終了までの、アジア関係公文書や重要な記録をデジタル化し（約730万コマ）公開している。原資料は国立公文書館、外務省外交資料館、防衛庁防衛研究所図書館が所蔵している。画像形式は高圧縮のDjVuのため、閲覧用プラグインのインストール（無料）が必要。

【参照】JJ4（p.154）

●国立国会図書館近代デジタルライブラリー

http://kindai.ndl.go.jp/

NDL所蔵の明治期刊行図書のうち著作権処理済の約39,000件（約59,000冊）が閲覧できる。元データは丸善刊行の『明治期刊行図書マイクロ版集成』である。一般的な書誌情報に加え、目次の検索が可能なものも多い。案外とアジア関係の書籍も含まれており、中には入手困難な研究書もあって重宝する。画像形式はGIFもしくは高圧縮のLindraで、後者の場合は閲覧用プラグインのインストール（無料）が必要となる。

【参照】JJ4（p. 138-139）

●超星数字図書館

http://www.ssreader.com.cn/

中国最大の電子図書館であり、数十万冊の書籍と300万編余りの論文を電子的に公開している。その数量は頁数にして4億、電子的な容量は3万GBに上るという。23万人の著作者と権利契約を結び、400の出版社と提携していると喧伝しているが、著作権法的にはグレーゾーンのものもある。

内容的には古典籍から新書まで、全てのジャンルに及ぶ。100元の「超星読書カード」[6]（上図参照）を購入すれば、1年間は自由に閲覧、ダウンロード、印刷ができる。

超星読書カード

画像は独自形式のPDG、閲覧には専用ブラウザ「超星閲覧器（SS Reader）」（日本語環境のWindowsで利用するには英語版のSS Reader）が必要となる。また別途付属OCRプラグインをダウンロードすることで、テキスト化することもできる。ただしこのソフトは非常に重く、ダウンロード途中でタイムアウトになってしまう。このため、筆者はまだこの機能を試せていない。

無料で閲覧できる書籍もあるので、まずは試してみて、日本にはない大規模バーチャル図書館の雰囲気を味わっていただきたい。

【参照】「電中2」（p.142-149），JJ3（166-169），JJ2（p.165-166）

●テキスト系電子図書館

日本の大規模テキストアーカイブズは、個人や任意の団体によるものが主流である。いずれも著作権切れの文献について、テキストデータを作成し、無料公開している。従ってこれらのサイトでは、先述のNDL近代デジタルライブラリーの原本画像に対応するテキストデータを入手できる場合もある。中でも最も知られているのは次のサイトだろう。

●青空文庫

http://www.aozora.gr.jp/

文学作品だけでなく、内藤湖南や桑原隲蔵といった東洋学の大家の論考も収められている。

⊙ OPAC も多言語の時代！

日本の図書館の OPAC は、UTF-8 を採用することで、多言語に対応するようになってきている。これにより OPAC で、アジアの特殊言語資料を検索できるようになった。

多言語用 OPAC の先鞭を付けたのは、国立情報学研究所（NII）を中心とする大学図書館だった。しかし現在では NDL や公共図書館にも広がりを見せている。いずれの OPAC でも、漢字の場合は、異体字テーブルの導入により、検索時に漢字の字体差を考慮しなくて良くなった。また中国語の場合はピンインからの検索もできる。

日本の代表的な多言語 OPAC は次の二つである。

●NACSIS Webcat　英語版
http://webcat.nii.ac.jp/webcat_eng.html

●国立国会図書館アジア言語 OPAC
http://asiaopac.ndl.go.jp/

Webcat では全国の大学図書館の所蔵状況が確認できる。ただし UTF-8 に対応しているのは英語版のみであるので注意が必要だ。近年 Webcat Plus（**http://webcatplus.nii.ac.jp/**）という連想検索を利用した次世代 OPAC が登場したが、2005 年 8 月現在ではまだアジア言語には対応していない。

NDL のものは、1986 年以降に受け入れられた中国語、朝鮮語、モンゴル語、ベトナム語、インドネシア語、マレーシア語の図書に対応している。

なお OPAC は検索エンジンのような全文検索ではない。このため OPAC の検索には多少コツが必要となる。このコツについての詳細は拙稿「大学図書館利用者のためのオンライン目録学」（JJ2, p31-39）をご覧いただきたい。

中国では 2002 年秋より国家図書館（**http://www.nlc.gov.cn/**）の OPAC が多言語に対応している。国家機関でありながら、GB に代わって UTF-8 を採用したのは、驚愕に値する。独自の異体字テーブルも用意しているらしく、日本の多言語 OPAC 同様、検索時に字体差を考慮しなくて良い。このため中国の OPAC でありながら、簡体字に不慣れな初学者でも十分活用できる。

【参照】「電中 2」（p.119），JJ2（p.156-157），JJ3（p.162），JJ4（p.138，140）

（以上、小島担当）

⊙ 漢籍電子文献を使いこなせ！

漢籍電子文献は、台湾中央研究院が公開する中国古典系データベースで、この分野の最古参かつ最大規模を誇る。本誌でも毎号どこかで採り上げているし、読者諸氏も日々使っておられるはずだ。まさに古典系データベースの定番中の定番である。

しかし、ここを利用していて「この用例絶対あるはずなのに。見つからない。」といった状況になったことはないだろうか？そう、それは本当ならあるはずなのだ。

では、何故にそれが見つからないのか。その理由は「キーワードの選択」「文字コード関係」の大きく二つに分けられる。次に、これらの問題について、解決方法を述べることにする。何もこれは漢籍電子文献だけではなく、寒泉など他の古典系データベースでも役立つテクニックである。

●検索式を理解しよう

「余りにも検索結果が余りにも多く目的の用例が見つからない。」という場合がある。キーワードによっては、数百〜数千〜数万といった単位で結果が出てくるため、目的の用例が発見できないのだ。そのような場合、キーワードを一工夫するだけでより効率よく目的の用例を発見することが出来るのである。

漢籍電子文献では、通常のキーワード検索の他に、and・or・not の検索式を組み合わせて使用できる。そのうち、and と or 検索はよく使われているが、not や括弧を含む複雑な検索式はそれほどでもない。しかし、これらを使いこなすことで、より用例の絞込が簡単に行えるのだ。使わない手はないだろう。

以下の表は、検索式の一覧と、その用例を挙げたものである。これを応用して使いこなしてもらいたい。

●キーワードの選択、間違っていませんか？

漢籍電子文献が入力元として選択しているテキストは、標点本が多い（標点本を選択した事情は、本誌第五号の陳弱水氏論考を参照）。従って、句点の切り方によっては、こちらの意図したキーワードが無視される場合がある。例えば、『論語』学而篇の冒頭「子曰學而」の「子曰學」をキーワードとして入力し、「十三經」の「斷句十三經經文」を選択して検索しても「找不到（探し出せない）」と表示される。これは、「十三經經文」の「斷句」が「子曰．學…（「．」は所謂「全角ピリオド」）」となっているためである。

そのような場合はどうするか。まずはキーワードを短くする。長い文字列だったら四字程度で入力してみる（これは古典漢文が四字句で構成されている場合が多いという傾向に基づいている）。上記の場合は「子曰」のみで検索すれば良かろう。その場合、当然多くの検索結果が標示されてしまう（この例では 1156 件！）。次にこれを絞り込もう。それには、「検索報表」を選択して検索結果を一覧表示し、そこから WWW ブラウザの検索機能で「子曰．學」とキーワードを入力して検索すればよい。その周辺のテキストが必要ならば、「子曰．學」周辺の適切な長さのキーワード（例えば「有朋自遠方來」）をコピーして、再検索すればよい。

漢籍電子文献の検索式とその用例

条件式（意味）	書式	具体例	意味
| （or）	A|B A もしくは B | は入れなくてもよい その場合は、各キーワード間に半角空白を入れる	劉備 | 關羽 | 張飛 劉備 關羽 張飛	［劉備］［関羽］［張飛］ どれか一つを含む
& （and）	A&B A と B との両方	劉備 & 關羽 & 張飛	［劉備］［関羽］［張飛］ 三者全てを含む
! （and not）	A!B B ではない A	關羽 ! 張飛	［関羽］を含むが ［張飛］は含まない
［&］と［!］との優先順位は同じ。［|］はそれらに次ぐ。 演算子は複数組み合わせて使う事も可能。その場合は、［(　)］で演算子の優先順位を決定。 検索例 (先主 | 後主) & (武帝 | 文帝) ※［&］演算子と［丸括弧（パーレン）］演算子との間には、必ず［半角空白 ］を挿入する事。 上記の例では、「先主と武帝」「先主と文帝」「後主と武帝」「後主と文帝」の四つの組み合わせのいずれかを意味する。			
{} （not）	A{BCD} A の後に BCD を含まない文字列 結びつく文字列は、大括弧の左右どちらでも可	關 { 平興 }	任意の［關某］。 但し［關平］と［關興］は含まない。 『三国志』では、關羽・關氏などが該当する。

●コードセパレート文字に注意

漢籍電子文献は、BIG5 コード＋外字で入力されている。外字については後述するが、BIG5 は台湾系文字コードであり、日本語用の JIS 漢字コードとは収録する文字に出入りがある。

更に、Unicode の「ソースセパレーションの原則（各ローカルな規格で分かれている文字は、Unicode でも別々の文字として扱う）」により、Big5・JIS それぞれ別個に収録されている文字（これを「コードセパレート文字」と呼ぶ）が、検索の際に問題となる。以下にいくつか挙げた。

BIG5	JIS	BIG5	JIS	BIG5	JIS	BIG5	JIS
產	産	巢	巣	銳	鋭	脫	脱
閱	閲	簞	箪	緣	縁	值	値
黃	黄	呑	呑	溫	温	內	内
虛	虚	囊	嚢	鄉	郷	剝	剥
俱	倶	晚	晩	啟	啓	姬	姫
揭	掲	步	歩	戶	戸	每	毎
吳	呉	戾	戻	娛	娯	歷	歴
查	査	錄	録	歲	歳	偷	偸
眾	衆	俞	兪	涉	渉	絕	絶
蔣	蒋	篡	簒	狀	状	喻	喩
稅	税	廄	廐	說	説	弒	弑

コードセパレート文字の例

これらの中には「説」「内」等の基本字まで含まれている。従って、どれがコードセパレート文字かを理解しておかないと、基本的な用語すら検索することすら出来ないのだ。

加えて、文字フォントのデザイン上、見た目だけ同じであっても実は異なった文字という場合がまま見られる。例として「爲」「為」両字が挙げられる（JIS 系フォントでは「為」であっても、台湾系フォントでは「爲」字デザインされている場合がある）。

また、このコードセパレート問題は、BIG5 と JIS の間のみならず、外字とも密接にリンクしているので更にやっかいである。

●独自外字に注意

上述のように、漢籍電子文献は BIG5 コードで構築されているが、BIG5 未収録の文字は外字で入力されている（外字未作成の文字は●で表記）。その数は実に 4000 字以上もあり、これをどう使いこなすかが効率的な検索のコツとも言える。

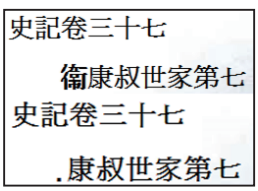
まずこの外字、繁体字版 Windows 専用である。従って日本語版 Windows98・Me では使用できない。当然、Windows 以外の OS でも使えない。一応、日本語版 Windows2000・Xp では使用可能だが、一部ハングルや GB18030 とコードポイントが重なる部分は、Windows の FontLink 機能の為にそれらの文字が優先して表示される（そのため、外字を含む文字を単純にコピーして貼り付けると、時々あり得ないハングル文字が表示される）。また、日本語 Windows では Internet Explorer でしか表示されない（特に FireFox を使っている方は要注意）。

外字である以上、珍しい文字が中心だと印象を持たれるかもしれないが、実は漢籍電子文献に限ってはそうではない。このデータベースは、基本的に入力元の本の活字デザインに忠実に入力するため、包摂で既存の字に置き換えられそうな物、或いは活字のかすれ・欠け・印刷字のインクのにじみによる、ちょっとした字形の違いすら外字として別字扱いにしているのだ（木版本ベースである『十三経注疏』等は、校勘記が附属していることもあって、外字＋●の数が段違いに多い）。

外字の一例を挙げよう。例えば「衞」「衆」。これが基本字であることは言うまでもないが、この字形は BIG5 にはない。従って、この両字は外字で処理されているのだ。

では、外字をどう効率よく入力するか？漢籍電子文献の外字は、入力中に見つかった外字相当部分を順に入力しており、字書のような部首画数などの整理は一切されていない。従って、IME の文字一覧から目的の文字を探すのは一苦労である。あらかじめその字が外字であると知っているのならば、IME の辞書に登録するのが一番だろう。し

上： Internet Explorer で表示。衞が表示されている。
下： FireFox で表示。衞が表示されていない。



かし、そもそも入力したいキーワードが外字を含むかどうか、それすら判別しがたいかもしれない。

その場合はどうするか？裏技だがこういう方法もある。まず、漢籍電子文献でキーワードを検索する。それで見つかれば問題ない。もしかして外字かな？と思ったら、寒泉などの他のデータベースやGoogleなどのサーチエンジンで該当の文字をキーワードとして指定して検索する。目的のテキストが表示されたらしめたもの。キーワード周辺の文字列をコピーして漢籍電子文献で検索すれば、外字を含む部分が表示されるだろう。

外字を含む文字を、Wordなどに貼り付ける場合も注意が必要だ。通常、IE環境ではWordに貼り付ける際、html形式で貼り付けられる。ところがその場合、外字を含む文章を貼り付けると、文字が化けるどころか、文章が一部入れ替わってしまう場合もある。そのため、Wordに貼り付ける際には、面倒でも「形式を選択して貼り付け→Unicodeテキスト形式」を選択して貼り付けよう。

●データが間違っている

漢籍電子文献のような大規模データベースでは、どうしても間違いは避けられない。それでも筆者の経験上、あからさまな入力ミスは殆ど遭遇していない（単に気がつかないだけだろうが）のは、対校に力を入れているためだろうと拝察する。

余り見かけない入力ミスのなかで筆者が一つ見つけたのは、「市（ふつ：ux5DFF）」「市（し・いち：ux5E02）」である。これは見た目が殆ど変わらないので「まあ、しょうがないかもなあ」という用例だが、実際の影響（特に「市（し・いち）」とすべき所を「市（ふつ）」にしていると、経済用語・或いは地方行政区画の検索に影響が出る）が多大であることが容易に想像できるだろう。

これが、これが外字を含む用例になるととたんにミスが多くなる。外字の問題は「外字の文字デザインが間違っている（間違った外字が入力されている）」に集約される。多少文字デザインの違いであれば、他の版本やテキストで作られた外字を流用しているのだろうと判断されるのだが、本来そこにはその文字が入力されるべきではないのに入力されている場合は、被害が大きい。これが、全てのテキストで同じ間違いをしているのであれば、単純に利用者が一括置換をすればよいだけの話である。しかし、時々「Aというテキストでは正しく入力されているのに、Bではその文字が誤って入力されている。」「本来Cという外字はA・B二つの異なる文字であるはずなのに、同じ外字で処理されている」という場合があるので問題なのだ。この様な場合は、逐次手元で訂正し、漢籍電子文献にその都度報告して訂正を依頼するしかないだろう。

◈ おわりに——情報の精度を求めて

以上、インターネット上のデータベースサービスを使いこなすためのコツについて述べてきた。これを読んでいただいて、これまでより効率的に目的の情報に達することが出来れば幸いである。

しかし、インターネットに必要な情報が全て流れているわけではない。また、その情報の真偽の程にも注意を払う必要がある。

事実、サーチエンジンで折角見つけた情報が「間違っている」事も珍しくない。そう、Webサイトに載っている情報は、全て正しいわけではない。単なる見解の相違や勘違いに始まる間違、更には悪意ある意図の下に誤った情報をわざと載せている場合もあるのだ。

その情報が正しいかどうか、他の同じような性格のサイトがないか、或いは過ちを検証しているところが無いかどうか、もう一度調べ直した方がよいのは当然である。

従って、これらの情報を利用する際には、そこに書かれてあるものが本当に信用に値するかどうかについて、引用者が判断する必要がある。例えば、そのWebページに参考文献や引用元が載っていればそっちでも確認するという行為である。そのような情報が載せられていない場合、たとえそれが真実であろうとも、情報元を確認できない以上、安易な引用は控えるべきだろう。

また、学術的な目的でデータベースを利用する

際は、比較的信頼出来る研究機関の作成にかかるものや、作成に当たって利用した情報源が妥当であると判断されるものを中心に利用すべきである。

インターネットで得られる情報は、玉石混合である（しかも限りなく石が多い）。そのため「データベースは本当に使えるのか？」という疑問を未だに持たれる方も多いようだ。確かに、入力元の版本が明示されていないデータベースも多い。しかし、その辺りは手元の良本で補えばよいのだ。それよりも、データベースを使って効率よく情報を集め、その結果の分析に時間を費やす方が遙か有意義であろう。

進士及第者ならともかく、凡人の我が身では『十三経注疏』や正史、はたまた『四部叢刊』や『四庫全書』をそらんじること、ましてやそこから数秒で用例を見つけ出すことなど逆立ちしたって不可能である。データベースを効率よく使う。これが今後の研究では当たり前の光景となるだろう。

しかし、現状での専門系データベースは、情報の濃淡が多いのも事実である。例えば古典漢文系では、経書・正史・諸子百家と言った基本的なものは収録されているが、地方誌・集部の著作・白話文学、それに『四庫全書』以降の著書については殆どデジタルテキスト化さえ未だされていない。また、思想分野でも仏教分野は早くからこの分野をリードしてきた物の、漢字文化圏の根本的アイデンティティたる儒学に関しては、上述の基本的なテキストしかデジタル化されていない。デジタル化＋データベースの恩恵に預かれる分野には限りがあり、それ以外の分野では相変わらず旧来の手法を用いざるを得ないのだ。

これは従来からの紙媒体での状況と変わりない。ある特定の書物・論文が見られるか否かで研究の進度・深度が決まった状況は、超星数字図書館による書物の閲覧環境の劇的な改善と、各種 OPAC や china3 による書物・論文の検索の容易さによって、ある程度平均化されてきたと言える。しかし、デジタル化の恩恵は何もそれがデジタル化されているかだけではなく、デジタルテキストやデータベースを利用可能か否かで、嘗ての書物を取り巻く状況よりも明確な差異を見せつけることとなるだろう。古くさい言い方かもしれないが、デジタルテキストとデータベースの恩恵に預かれる層とそうでない層とに階級文化を引き起こすのである。テキストを如何に読んで分析するかだけではなく、データベースを使いこなせる環境にあるか否かで、研究の進展に差異が現れる。出来るだけ各研究機関ではデータベースを導入して欲しいが、それが出来なければ、共同で購入或いは利用可能なライセンスを確保する等の努力を試みて貰いたいと願っている。

インターネットの文字や画像・映像等の情報にも著作権が存在する。引用には十分注意すべし（本誌 4 号以降に掲載の「東洋学と著作権」「東洋学研究と著作権」を熟読されたし。）。

（以上、山田担当）

注

[1] 「Google で検索する」という意味の造語。類似例に「ヤフる」などもあるらしい。

[2] 更に最近では、レポート・卒論のデータベース（中身付き！）や卒論の執筆代行サービスを提供する Web サイトまであるようだ。この手の代物は、昔から大学毎に存在していたが、ネットワークのおかげでこの方面も大規模化したものである。採点する方も一苦労だ。

[3] 検索デスク（**http://www.searchdesk.com/index.html**）は複数 Web 同時検索（メタサーチ）の一例。

[4] **http://kanji.zinbun.kyoto-u.ac.jp/kansekikyogikai/ichiran.htm**

[5] 全国漢籍データベースおよび東洋学文献類目検索については、山田崇仁氏のサイト睡人亭（**http://www.shuiren.org/**）に詳しい利用案内がある。

[6] 日本では中国書籍の専門店で購入可（3,000 円程度）。

手軽にできる情報分析

秋山　陽一郎（あきやま　よういちろう）

ローカル環境を活用せよ

漢籍電子文献や寒泉などのWeb検索サービスも確かに便利だが、それなりに使い込むようになってくると色々な問題点が見え隠れしてくる。

- 外字使用の問題
- 不適切なカテゴライズ
- 本文・注・校勘記を区別して検索できない
- 必要な文献がない
- もっと柔軟な検索をしたい
- 混雑時やメンテナンス時だと使えない

また四庫全書や四部叢刊といった市販のデータベースも個人で入手するには高価で、データの転用が困難であるなど使い勝手もそれほど良いわけではない。このような時､ふと「自分のコンピュータの中で検索環境が自由に構築できたらいいのに」などと考えることはないだろうか。実は自分のコンピュータの中（以下、ローカル環境）に検索環境を構築するメリットは思いのほか多い。

- オフラインでも作業が行える
- 自分の好みのテキストを思い通りに加工したり活用したりできる
- テキストデータを不特定多数に配信するわけではなく私的使用の範囲にとどまるため、原則として版権の制約を受けない。
- ネットワーク上では負荷のかかりすぎる複雑な処理も行える

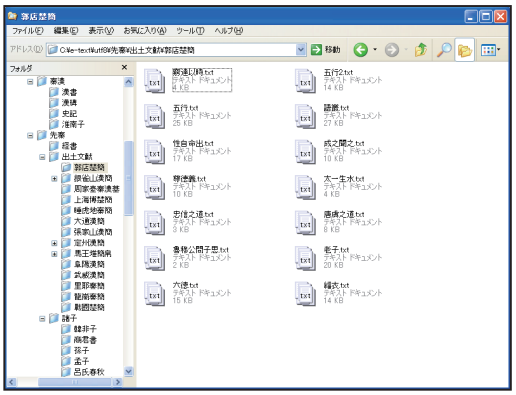
ここではローカル環境下でのデータ検索や管理法・分析法について紹介していく。

階層を掘ってテキスト文書を管理

いきなりローカルにデータベース環境を作るといっても特に小難しい知識やスキルがなくても、普通のテキスト文書だけでもかなりのことができる。といって、自作もしくはネット上にある電子テキストを一個所に保存して検索するだけでは芸がない。まずは手持ちの電子テキストをカテゴリーごとに分類してディレクトリ（＝フォルダ）階層ごとに管理することをおすすめしたい。

たとえば経書・史書・子書…、さらに子書の中に儒家・道家・法家というように、ジャンルごとに独自にディレクトリを立てつつ階層化すれば、「儒家系の文献だけを検索したい」という需要に容易に対応できるようになるし、先秦・秦漢・魏晋南北朝・唐・宋…のように時代ごとにディレク

図１　ディレクトリ階層例



トリを立てて管理すれば、たとえば同じ経学に関する文献でも漢代以前のものと宋代のものとを区別して扱える。

なおディレクトリの立て方については歴代の図書目録が参考になるが、基本的には各ユーザーの需要次第である。時代で区切ってからカテゴリーごとに分類するか、カテゴリーごとに分類してから時代で区切るかについても同様。

後述する正規表現を駆使して『史記』三家注を『史記』と銘々の注釈とに分離すれば、『史記』は漢代、裴駰の『史記集解』は魏晋南北朝、司馬貞『史記索隠』と張守節『史記正義』は唐代というように、これまた時代ごとに区別して内容を管理・検索できるようになる。こうした手法は特に輯佚学（佚文を収集して散佚してしまった古典籍の復元を試みる学問）の分野で効果を発揮するが、注釈が対応する本文の場所を同時に確認したい場合には、XML を使うなど別の手法を取る必要がある。

◈ GREP でパターンマッチング

⊙ GREP とは？

GREP と は、Global search for Regular Expression and Print の略で、正規表現（Regular Expression）を用いて単一もしくは複数のファイルの中から任意のパターンに一致する文字列を検索・出力することをいう。よく GREP 検索といいながら複数のファイルから任意の文字列を検索するだけで満足してしまっている人を目にするが、GREP の本領はむしろこの正規表現を利用したパターンマッチングをしてこそ発揮される。

⊙ 正規表現とは？

正規表現（Regular Expression）とは、”\”（エンサイン or バックスラッシュ）, “^”（キャレット）, “$”（ドル）, “*”（アスタリスク）, “+”（プラス）, “?”（クエスチョン）, “.”（ドット or ピリオド）, “|”（パイプ）, “[ ]”, “( )”, “{ }” といったメタ文字（Meta Character）と呼ばれる記号を利用して文字列のパターンマッチを行うための表記法のこと。ただし、この正規表現は、扱う言語によって若干の方言（表記法の差）がある [1]。以下に代表的な正規表現の記法と意味を少しだけ紹介しておこう。

^	行頭。［例］^ 子曰（行頭にある「子曰」）
$	行末。［例］乎？ $（行末にある「乎？」）
*	直前の文字 or パターンの 0 回以上のくりかえし。 ［例］顧 * 炎武（顧炎武 or 炎武）
+	直前の文字 or パターンの 1 回以上のくりかえし。(” *” は直前の文字 or パターンがあってもなくても OK だが、” +” は必ず 1 回は出現してなければならない。) ［例］\n+（1 つ or 連続する改行） ［例］[一二三四五六七八九十]+（漢数字）
\	エスケープ文字。\+, \[, \\ のようにメタ文字をただの文字として扱いたい時にメタ文字の直前に入力する。
\n	改行。
\t	タブ。
[abc]	[ ] で括られた文字のいずれか。a か b か c。 ［例］蘇 [洵軾轍]（蘇洵 or 蘇軾 or 蘇轍） ［例］[甲乙丙丁戊己庚辛壬癸][子丑寅卯辰巳午未申酉戌亥]（干支）
[^abc]	[ ] で括られた文字以外の 1 文字。左の場合は a, b, c 以外の 1 文字。 ［例］謝 [^ 靈運] 運（謝靈運・謝霊運以外の謝 x 運。謝惠運や謝恵運には一致する。）
(ab|xy)	“|”（パイプ）で区切られた文字列のいずれか。この場合は” ab” か” xy”。 ［例］史記 (集解 | 正義)（史記集解 or 史記正義）
(?!abc)	( ) で括られた文字列以外の文字列。左の場合は” abc” 以外の文字列。 ［例］『史記 (?! 索隠 | 索隠)』（『史記索隠』もしくは『史記索隠』以外の『史記～』という書名。）
.	任意の 1 文字。 ［例］謂 .. 曰（「謂」と「曰」の間に任意の 2 文字が入るパターン。「謂秦王曰」や「謂蘇秦曰」にはマッチするが、「謂武安君曰」や「謂起曰」にはマッチしない。「謂矣。曰…」にもマッチする。）
(abc)	任意の文字列パターン。 ［例］(惜哉。)+（「惜哉。」や「惜哉。惜哉。」にマッチする。）
\1 - \9	後方参照。 ［例］(.)\1 然（「蕩蕩然」や「莫莫然」のように任意の文字が「然」の直前に 2 字連続して現れるパターン。）

表 1　主なメタ文字や正規表現の意味

実際には以上に挙げた正規表現を複数組み合わせて利用することが多い。簡単なものでは、たとえば”王 *(念孫 | 引之)”で「王念孫」・「王引之」・「念孫」・「引之」にマッチさせたり、以下の正規表現で日付にマッチさせることができる。

[前中後]*[一二三四五六七八九十廿卅丗卌元]*[有又]*[一二三四五六七八九]*[年載][、，]*([春夏秋冬])*[、，]* 閏 *([一二三四五六七八九十正端]+ 月)*[、，]* 朔 *[、，]* ([甲乙丙丁戊己巳巳庚辛壬癸][子丑寅卯辰巳巳己午未申酉戌亥])*

注意しなくてはならないのは、たとえば「謂〜曰」（〜に謂ひて曰く）にマッチさせたいような場合に、「謂」と「曰」の間に任意の文字が 1 字以上入るパターンという意味で” 謂 .+ 曰” と検索してしまいがちだが、これだと「某者某之謂矣。曰某…」や、「某子謂之云、書曰…」のようなケースにもマッチしてしまう。こういう無駄なノイズを除くために、” 謂 [^、。•.,:;「」『』]*? 曰” (「謂」と「曰」の間に、読点「、」と句点「。」、中黒「•」、ピリオド「.」、カンマ「,」、コロン「：」、セミコロン「；」や、カギ括弧（「」『』）、全角・半角スペースを除く文字が 1 文字以上入るパターン）のように制限事項を考えながら正規表現を入力するようにしたい。

ex: “謂 [^、。•.,：；「」『』]+ 曰” の検索結果（抜粋）

C:\e-text\utf8\ 先秦 \ 諸子 \ 莊子 \ 雜篇 \ 徐无鬼 .txt(304): 子送葬、過惠子之墓、顧謂從者曰、「郢人堊慢其鼻端若蠅翼、使匠人斲之。石運斤成風、聽而斲之、盡堊而鼻不傷、郢人立不失容。宋元君聞之、召匠石曰、『嘗試爲寡人爲之。』匠石曰、『臣則嘗能斲之。雖然、臣之質死久矣。』自夫子之也、吾無以爲質矣、吾無與言之矣。」

C:\e-text\utf8\ 先秦 \ 諸子 \ 莊子 \ 雜篇 \ 徐无鬼 .txt(306): 吳王浮於江、登乎狙之山、衆狙見之、恂然棄而走、逃於深蓁。有一狙焉、委蛇搔、見巧乎王。王射之、敏給搏捷矣。王命相者趨射之、狙執死。王顧謂其友顏不曰、「之狙也、伐其巧・恃其便、以敖予、以至此殛也。戒之哉。嗟乎、無以汝色人哉。」顏不疑歸而師董梧、以鋤其色、去樂辭顯、三年而國人稱之。

C:\e-text\utf8\ 先秦 \ 諸子 \ 莊子 \ 雜篇 \ 外物 .txt(336): 子謂莊子曰、「子言無用。」莊子曰、「知無用而始可與言用矣。天地非不廣大也、人之所用容足耳、然則廁足而墊之致黃泉、人尚有用乎。」惠子曰、「無用。」莊子曰、「然則無用之爲用也亦明矣。」

⊙ Unicode 対応 GREP のできるエディタ

東洋学において GREP ツールを利用する場合、そのツールが Unicode に対応しているかどうかで利便性が大きく変わる。

GREP 機能を実装しているエディタは、サクラエディタや秀丸エディタをはじめとして幾つも選択肢があるが、Shift JIS にはない Unicode 文字列の GREP に対応しているエディタは少ない。

OS	Unicode 対応エディタ
Windows	EmEditor, Akira22+, Aprotool TM Editor
MacOS X	Jedit X, SubEthaEdit

表 2　GREP 可能な Unicode 対応エディタ [2]

⊙ Word でワイルドカード検索

MS Word でもワイルドカードと呼ばれる一種の正規表現を利用した検索が利用できる。テキストエディタほどの高度な検索はできないが、多言語対応が進んでいる点やリッチテキストを扱うワープロソフトならではの特殊な検索が行える [3]。

✥ KWIC で大量の文例比較

⊙ KWIC とは？

KWIC は、Key Word In Context の略で、検索語が文脈の中でどのように出現するのかを視覚的に比較・確認しやすいようにする検索のことをいう。GREP のように一行（≒一段落）まるごと出力するのではなく、あくまで検索するキーワードとその前後の文しか出力しないが、その分、大量の用例をいっぺんに比較したい時に便利な検索モデルといえる。

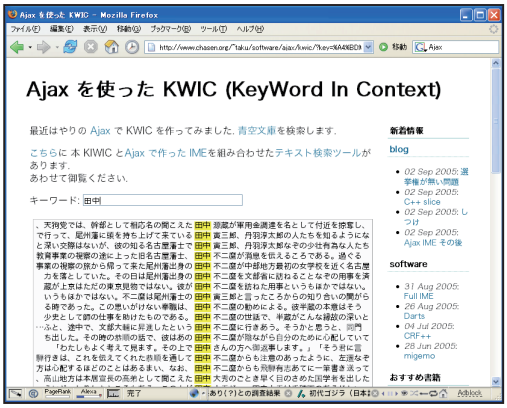


図 2　“Ajax を使った KWIC”で「田中」を検索

まずは KWIC 検索がどんなものかを Windows ユーザーにも Mac ユーザーにも特別なソフトをインストールすることなくお手軽に実感していただくために、工藤拓氏の手に成る Web 上に設置されている実験的な KWIC 検索システムを使ってみよう [4]。

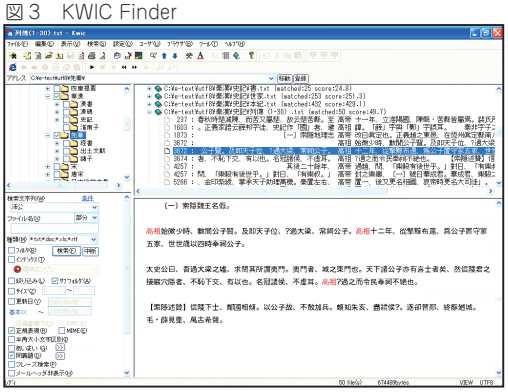
●Ajax[5] を使った KWIC

http://www.chasen.org/~taku/software/ajax/kwic/

このページの［キーワード］入力欄に「田中」と入力してみて欲しい。すると中央に入力した「田中」が黄色くハイライトされ、それを挟んで左右に前後の文章が出力される。これによって「田中」というキーワードがどんな文脈（Context）の中で出現するのかを一目で比較・確認できるようにしている。これが KWIC 検索である。

GREP だと通常はヒットしたファイルとその位置（出現行数）を一覧出力するか、ヒットした行（≒段落）を一つ一つすべて出力してしまうので、一画面あたりでチェックできる用例の数が限られてしまう。（せいぜい 2 ～ 3 例といったところだろう。）これに対して KWIC は一画面に表示可能な行数分だけの用例を出力できるので、一度に比較できる用例の数が GREP よりもはるかに多い。またキーワードが中央に来るように文を抽出するので、比較したい目的の場所を探すのも容易である。

図 3　KWIC Finder



●KWIC Finder

工藤氏の”Ajax を使った KWIC”は実験的な目的で設置している Web 検索ツールなので、青空文庫の一部のテキストを、それもオンラインでしか検索できないが、ローカル環境で利用できるより多機能なツールももちろんある。Windows 用の一般的なツールとしては KWIC Finder がある。

現在のところ、このソフトは Shift-JIS にない Unicode 文字列の検索・出力には対応していないが、正規表現やあいまい検索、出現頻度出力、簡易シソーラスの定義など柔軟で多彩な機能を実装している [6]。特に簡易シソーラス（KWIC Finder では「同義語」という）は、「捨」と「舎」と「舎」（「すてる」の正字と省文）、「（唐）太宗」と「（李）世民」と「秦王（世民）」（いずれも唐の太宗のこと）などを同一視させることができ、しかもこのシソーラスはスペース区切りのテキスト文書で管理できる。

◈ Ngram などの語彙分析手法

⊙ GREP や KWIC にも弱点はある

ここまで紹介してきた GREP や KWIC は検索対象がはっきりしている場合には便利だが、これらは必ずしも万能な情報分析の手段というわけではない。

［利点］

●手軽（検索対象がはっきりしている場合には便利）

［欠点］

- 大量のデータを扱うには不向き
- 検索・抽出するデータの選択が主観的・恣意的になりがち

たとえばGREPやKWICは、検索するキーワードを検索者が選定するという時点で、すでに恣意的な判断が加わっており、特に数百年から千年以上の研究の蓄積のある古典学にあって、この手法で未開拓の問題点を発見するのは容易でない。かといってテキストの一部を抽出するのではなく、その全体を分析対象とするとなると数千件・数万件にもおよぶ大量のデータを一度に扱うことになるが、少なくともGREPやKWICはこうした大量の情報を分析するのには必ずしも向いているとはいいがたい。

⊙ 未知の問題点を採掘せよ！

集めたデータの中から未発見の有用なデータを統計的手法によって発見することを、鉱脈を採掘（mining）することになぞらえて「データマイニング」というが、採掘するデータの対象（鉱脈）が自然文（text）である場合、これを特にテキストマイニングという。本誌の読者ならご承知の通り、本章で取り上げるN-gramは、このデータマイニングを行う上で、（原始的だが）非常に有効な手段のひとつであるといえる。

⊙ N-gramとは？

N-gramとは自然言語解析手法の一種で、2進数変換といったコンピュータにおける基礎理論を構築したクラウド・エルウッド・シャノン（Claude Elwood Shannon, 1916-2001）が編み出した文字列を機械的にn個単位で区切って解析する言語モデルで、近年、近藤泰弘・近藤みゆき両氏を嚆矢として東洋学人文情報学の業界でも注目されるようになってきた[7]。

表4のように『老子』第一章の冒頭部分であれば、「道可道非常道…」という文句を2 gramで解析する場合だと、「道可」「可道」「道非」「非常」「常道」…のように2文字ずつ本文を切り出していく。こうして切り出された字句の出現頻度や共起頻度（出現する字句の組み合わせの頻度）を下図のように取っていくことで、結び付きやすい文字や言葉が客観的な数値の形で現わされる。

なお、これまでに公開されているN-gramツールは以下の通り。導入法については山田崇仁氏の「N-gramモデルを利用したテキスト分析」（**http://www.shuiren.org/chuden/teach/N-gram/index-j.html**）にすでに詳細な解説があるので、そちらを参照されたい。

道	可	道	非	常	道	名	可	名	非	常	名
道	可										
	可	道									
		道	非								
			非	常							
				常	道						
					道	名					
						名	可				
							可	名			
								名	非		
									非	常	
										常	名
											名

表4　『老子』第一章を2 gram（bi-gram）で解析

- ngram（藤原滋氏作、元祖ngram解析ツール。）
  http://www.jaist.ac.jp/~shigeru/ngram-ja.html
- morogram.pl（師茂樹氏作。頻度1の抽出、（拡張漢字を含む）Unicode対応など。Perl 5.8に対応。）
  http://sourceforge.jp/projects/morogram/
- morogram.exe（極悪氏作。Windows用 実行ファイル版。Active PerlやUnix環境がなくても単体で稼働する。）
  http://hpcgi1.nifty.com/dune/gwiki.pl?morogram
- norogram（大谷由佳氏作、Java版N-gram。頻度1抽出にも対応。）
  http://sourceforge.jp/projects/norogram/
- ngmerge.pl（近藤泰弘氏作。NGSM用マージスクリプト。Unicode 3.2以降のCJK拡張漢字も問題なくマージしてくれる。）
  http://klab.ri.aoyama.ac.jp/tool/index.html

1gram（Uni-gram）	2gram（bi-gram）	3gram（tri-gram）
道　3	非常　2	道可道　1
名　3	道可　1	可道非　1
可　2	可道　1	道非常　1
非　2	道非　1	非常道　1
常　2	常道　1	常道名　1
	道名　1	道名可　1
	名可　1	名可名　1
	可名　1	可名非　1
	名非　1	名非常　1
	常名　1	非常名　1

表5　『老子』第一章1～3 gramの共起頻度

⊙ N-gramを利用した分析モデル

N-gram自体は一次的な解析モデルであって、テキストをN-gram解析した後に、どういう手段で解析したN-gramデータを評価するかがむしろ重要になる。以下、これまでに報告されている成果の中から参考になるものを幾つか抜粋して紹介するので、N-gramでどんなことができるのかの参考として欲しい。

●NGSM

NGSM（N-gram based System for Multiple document comparison and analysis）とは、石井公成氏の提唱にかかる、N-gramモデルを用いた複数の文献の間での共起頻度を比較する手法をいう[8]。要するに文献比較や異本比較であって、もっとも基本的な方法であるといえる（これ以下の項目でもこのNGSMを起点にしてそれを何らかの形で可視化したものが多い）。

表6　『老子』諸本NGSM[9]

	道藏傅奕本	明和王弼本	景龍河上公碑	敦煌河上公本	道藏河上公本	武内義雄本	四庫全書本	玄宗御注本	馬王堆甲本	馬王堆乙本
也	62	18	1	0	14	28	11	11	174	178
矣	27	11	2	1	10	22	9	11	25	25
焉	11	9	0	0	9	16	8	3	11	10
乎	14	11	1	0	11	12	6	11	12	12
邪	3	0	0	0	0	1	1	3	0	1
耶	0	3	1	1	2	2	2	1	0	0
歟	2	0	0	0	0	0	0	0	0	0

表6は『老子』諸本における句末の助字の出現頻度を比較したものである。同じ『老子』という古典文献でありながら、景龍河上公碑と敦煌出土鈔本（いずれも唐代の写本）における出現頻度が著しく低く、逆に馬王堆帛書本（漢代の写本）における出現頻度が著しく高いことがひと目で見て取れる。

●有意なデータを絞り込むには

ただし、単純に各テキストを統合（マージ）したデータを比較するとはいっても、統合したNGSMデータ自体が厖大な量になるため、その中から手探りで有意な部分を抽出するのは容易ではない。

これまでのところは、NGSMデータの中から各文献もしくは各本（edition）の間の標準偏差、分散、KWIC2乗値などを取って有意なデータの絞り込みを行った研究がある[10]。

なお、NGSMデータの整理は通常はMS ExcelやSPSSといった表計算ソフトを使うことが多いことと思うが、Excelは扱えるデータ量に65,536行という上限がある。扱う文献によってはこの上限を簡単に超えてしまうので注意が必要である。

●クラスター分析

NGSMデータを起点にして、その中に内在する傾向を可視化する手段は幾つかあるが、その代表的な手段の一つとしてクラスター分析がある。

クラスター分析は、異なる性質を持った集団（対象）の中から類似するもの同士を集めて（通常は階層的に）分類（クラスタリング）する方法の一種をいう。クラスター分析された結果はしばしば以下のような樹形図（テンドログラム）の形で表現される。

クラスター分けの線引きはサンプル（異本なり異文献）間の距離（≒類似度）によって変わってくるが、図4の樹形図の場合は、『景徳伝灯録』所収本・『瀑泉集』所収本・『禅門諸祖偈頌』所収本・『人天眼目』所収本を第1クラスタ（中国系諸本）、

『祖堂集』所収本を第２クラスタ（高麗系の本）、『参同契不能語』所収本・『参同契吹唱』所収本を第３クラスタ（日本系諸本）として３つのグループに綺麗に分類される。（さらにこの『般若心経』の異本データについていえば、概ね時代順にならんでいるのも興味深い点の一つといえるだろう。）

ただしクラスター分析は、各変数間の距離（類似度）の測定法と各クラスター間の距離（類似度）の測定法は、一般的にはユークリッド平方距離（もしくは標準化ユークリッド距離）による Ward 法で行われるのが普通だが、それぞれに多様な種類がある上、どの測定法を選択するかで樹形図が大きく変わってしまうのが難点である。

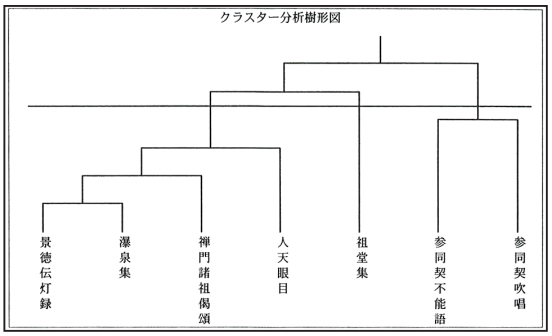


図４　『般若心経』NGSM 樹形図（標準化ユークリッド距離・Ward 法）[11]

1. 『景徳伝灯録』所収本（1004 年）
2. 『瀑泉集』所収本（～ 1052 年）
3. 『禅門諸祖偈頌』所収本（南宋末頃）
4. 『人天眼目』所収本（1188 年）
5. 『祖堂集』所収本（高麗版本）
6. 『参同契不能語』所収本（江戸、1736 年）
7. 『参同契吹唱』所収本（江戸、1767 年）

●主成分分析

主成分分析とは、相関のある多変数の情報を、無相関な少数個の総合特性値（主成分）に要約する分析方法で、要するにこれも似たもの同士で分類するための手法の一つである。

図５は、李賢平氏による、曹雪芹『紅楼夢』における虚詞使用率の主成分分析の事例で、第１～ 80 回は曹雪芹、第 81 ～ 120 回は高蘭墅の補作とする説を裏付けるかのような結果が出ているのが視覚的に見て取れる。

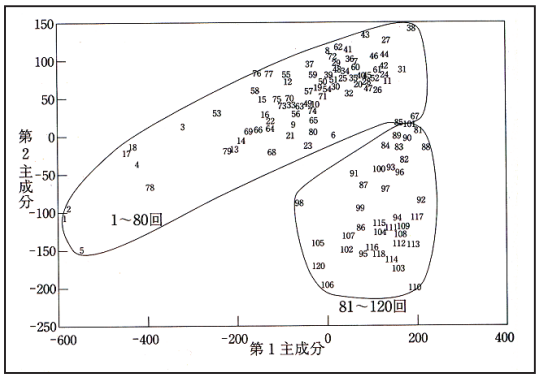


図５　『紅楼夢』虚詞使用率主成分分析 [12]

●K 特性値

K 特性値（K-Characteristic）とは、統計学者のユール（George Udny Yule,1871-1951）によって提唱された語彙量を表す統計指標で、語彙のバリエーションが豊富なほど値が小さくなり、バリエーションが少ない（≒文章が定型化されている）ほど値が大きくなる。

表９は、中国戦国時代の儒家・道家・法家・雑家の古文献における K 特性値による語彙量の差を一覧化したものである。

表９　主要先秦文献の K- 特性値 [13]

テキスト	2-gram	3-gram
『論語』学而・為政	19.841	1.577
『論語』述而	16.930	1.319
『礼記』坊記	17.505	3.600
『礼記』表記	13.892	3.520
『孟子』梁恵王	6.699	2.594
『孟子』公孫丑	6.857	2.099
『荀子』王霸	10.518	3.433
『荀子』正名	12.361	2.465
『墨子』尚同上	32.203	12.243
『墨子』尚同中	30.778	11.160
『荘子』逍遙遊	4.913	1.532
『荘子』斉物論	6.402	1.335
『呂氏春秋』有始覧	8.968	2.982
『呂氏春秋』孝行覧	7.351	2.655
『韓非子』顕学	10.080	3.413
『韓非子』説難	10.120	3.140

⊙ N-gram の利点と欠点

以上で紹介してきた N-gram モデルもやはり万能というわけではない。以下、簡単にその長所と短所を挙げておこう。

［利点］

- 基本的に検索漏れがない
- 辞書が必要ない（いかなる文に対しても適用できる）ため、古漢語は原則として 1 文字・1 音節・1 形態素の標語文字なので（少なくとも現段階では形態素解析よりも）漢籍のテキストマイニングに適している。

［欠点］

- ナンセンスな検索ノイズが多い（何らかの手段でフィルターをかける必要がある）。
- 0 頻度の問題（たまたま出現していない語句の取り扱いの問題）。
- 文脈や構文・語句の意味などは基本的にまったく無視。

◇ 展望

以上、ローカル環境で手軽にできる GREP・KWIC 検索や N-gram モデルを用いてできる情報分析方法の紹介を硬軟織り交ぜてしてきた。それぞれに利点と欠点のある方法なので、目的に応じて適切なモデルを選択したいところである。

また本稿で取り上げた情報分析法の数々は、ひとえにテキストの精度が分析結果の精度を左右する。古典文献の電子テキストは、古い時代のものほど網羅的に整備されてきているが、まだまだ研究者が手ずから校定した研究利用に堪えうる精善な電子テキストは少ない。この意味で今こそ邦人研究者による積極的な情報発信が求められる。

なお、テキスト処理について、関心を持たれた方は是非、本誌「特集 2　自然言語処理」を併読して欲しい。

注

[1] 秀丸エディタで利用されている HmJre.dll や 山田和夫氏作の jre32.dll のほか、EmEditor が採用している Perl 互換の C++ 用ライブラリ Boost Regex++、サクラエディタなどが採用する Ruby 系の鬼車などがある。ここでは EmEditor の Boost Regex++ の記法に従う。

[2] 詳細は山田崇仁氏のテキストエディタのレビュー（『漢字文献情報処理研究』4、2003）を参照されたい。なお EmEditor は次期バージョン（v4.20）からサロゲートペアをサポートし、CJK 統合漢字拡張 B の入出力と検索に完全対応する。

[3] 紙幅の都合もあり、本特集では詳述しないが、関心のある方は、山田崇仁氏の Web サイトに詳細な解説があるので、そちらを参照されたい。
http://www.shuiren.org/chuden/teach/word/

[4] キーワードとコンテキストの表現方法が若干異なるものの、実は Google の検索結果（キーワードにマッチしたページの見出しと、キーワードを含む文の一部を抜粋出力している）も KWIC である。

[5] Ajax というのは Asynchronous JavaScript And XML（非同期 JavaScript + XML）の略で、JavaScript を使って必要な時に必要な分だけ、サーバー上から XML（もしくはプレーンテキスト）データをリクエストする昨今流行中の Web 技術。

[6] このほか、XML と連携した KWIC 検索システムとして”全文検索システムひまわり”がある。
http://www.kokken.go.jp/lrc/index.php

[7] 近藤みゆき「n グラム統計処理を用いた文字列分析による日本古典文学の研究 ——古今和歌集のことばの型と性差——」（千葉大学『人文研究』第 29 号、2000）
近藤 泰弘・近藤 みゆき「平安時代古典語古典文学研究のための N-gram を用いた解析手法」（言語情報処理学会第 7 回年次大会『発表論文集』2001）など

[8] 石井公成「N-gram 利用の可能性 ——仏教文献における異本比較と訳者・作者判定」（『漢字文献情報処理研究』2、2001）
Ishii, Kosei. *“NGSM and Cluster Analysis: Its Usage in the Digitization of Variant Texts in the SAT (Taisho Daizokyo Text Database).”* Proceedings of PNC Annual Conference and Joint Meetings 2002.

[9] 拙稿「老子傅奕本来源考 ——テキスト処理による項羽妾本介在の検証」（『漢字文献情報処理研究』4、2003）

[10] 師茂樹「N グラムによる比較結果からの用例自動抽出――禅宗系の偽経を題材に」（『東洋学へのコンピュータ利用第 14 回研究セミナー』、2003）
齊藤正高「偽古文尚書の「賢」と「官」――χ^2 値による語彙偏差の数量化を通して」（『漢字文献情報処理研究』6、2005）
このほか、極悪氏が「頻度の合計・文字・文字数・各文献での出現頻度・出現頻度の分散」という順で出力する、mgsm.pl を公開されている。ソースは morogram メーリングリストの当該投稿を（http://www.google.co.jp/url?sa=t&ct=res&cd=1&url=http%3A//lists.sourceforge.jp/mailman/archives/morogram-users/2005-May/000048.html&ei=qiQfQ6eDGs-8YK3JgbIL）参照されたい。
[11] 師茂樹「N グラムモデルとクラスター分析を用いた漢文古典テキストの比較研究――『般若心経』の異訳の比較を例に」（『京都大学大型計算機センター第 69 回研究セミナー 東洋学へのコンピュータ利用』、2002）
[12] 李賢平「《紅樓夢》成書新説」（『復旦學報（社會科學）』1987-5）
[13] 山田崇仁「中国戦国期の語彙量について ――N-gram とユールの K- 特性値を利用した分析」（『漢字文献情報処理研究』5、2004）

情報発信のルール・マナー・スキル

小島　浩之（こじま　ひろゆき）

◇ 総論

⊙ バックナンバーから

本誌では、以前から情報発信のマナーやルール・スキルなどを題材として小特集を組んできた。下の表はこれまでの本誌の特集記事のうち、情報発信のルール・マナー・スキルに深く関わるものをまとめたものである。

第 3 号（2002 年 10 月）	
求められる学術研究情報の発信	他から評価してもらえる学術情研究情報の発信とは何か
第 4 号（2003 年 10 月）	
漢字処理技術の最新動向	最新のコンピューターでの漢字処理技術の紹介と展望。
第 5 号（2004 年 10 月）	
Wiki・Weblog と人文学	Wiki・Weblog を利用した、教育への情報発信。

これ以外にも、毎号、関連する論文や報告は数多い。また情報発信の先駆者を迎えた講演会や、著作権を主眼にした公開講座を開催しており、当然これらの内容は本誌に反映されている。つまり情報発信のルール・マナー・スキルについては、『漢字文献情報処理研究』のバックナンバーを紐解いていただくことが最も良い方法だと考えられる[1]。

ただバックナンバーには、様々なものが混在しており、記事や論文ごとの難易度に差がある。このため初心者や大雑把に知りたい方には取り付きにくい感じがある。そこで本稿では、バックナンバーへの橋渡しの意味も込めて、ルールとマナーの相違、著作権問題といったことについて、概説・紹介を試みることにした。

⊙ 最初の 1 冊とは？

インターネットの世界は技術の進展が早く、中身についても流行廃りが激しい。このためインターネットに関する書籍も回転が早く、優れた概説書を選ぶことは大変難しい。

その中で敢えて 1 冊を選ぶならば、インターネットで情報発信する際の根本精神を書いてあるかどうかに着目すべきだと思う。なぜならば、何事もその根底にある考え方を学ぶこと無くして、真の理解を得られるはずが無いからである。これは決してスキルの取得を軽んじているわけではない。本質理解があればこそ、有するスキルは威力を発揮できるはずなのだ。

筆者は、上述の点を考慮して、中村正三郎氏の『新版　インターネットを使いこなそう』（岩波ジュニア新書 391　2002.2）を最初の 1 冊として推薦する。以下に過去に筆者が書いたこの本の紹介文を再録しよう[2]。

> 本書が岩波ジュニア新書の一冊ということで違和感をいだく読者も少なからずあるだろう。しかし著者も自負しているように、本書は誰にでも読みやすい、非常に優れたインターネット世界への手引書だと言える。具体的にはインターネットのしくみ、歴史、技術、利用のポイント、そしてセキュリティーに至るまで、努めて平易に解説する。
>
> 著者の中村氏はオープンソースを推進する Ring Server プロジェクトの代表である。それ故、本書は“インターネットは全

ての人のもの”、“ボランティアと共有の精神”という考えに裏打ちされている。このためインターネット利用の際のエチケット（ネットエチケット＝ネッチケ）だけを採り上げた部分が無いにもかかわらず、一読すれば自然とこれが身に付くように配慮されている。エチケットを教えるというのは難しいもので、それだけを採り上げて書けば下手をすると説教じみてしまう。あまりに厳しく要求されれば、初心者の興味や意欲を失いかねない。この点、本書は読者にネット上のエチケットを自然に受け止めさせることに成功している。

ちなみに筆者のように必要に迫られ、よく解らないまま漠然とインターネットを始めた人には、よい復習教材となるだろう。

⊙ ルールとマナー（エチケット）

中村氏が述べるように、インターネットはボランティアと共有の精神によって発展してきた。自らが情報発信者となった場合にこのことを忘れてはならないだろう。特に研究・教育情報の発信は、それを共有することで、さらなる学問的発展や、教育効果を生むことが期待できる。情報の垂れ流しではなく、共有できる情報を流すことに意味があるのである。インターネット上の情報の質を高めるためにもこの点、留意すべきだろう。

ボランティアと共有の精神は、ルールとマナー（エチケット）に支えられてきた。ルールには拘束力を持つ法律、条例から、当事者間の取り決めまで様々な形態がある。しかしいずれの場合でも、これを破れば社会、もしくはコミュニティー単位で、制裁や罰則が加えられる。その最終判断は法律ならば司法に、特定のコミュニティーの規約ならば意志決定者（組織）に委ねられるのである。またルールの効力や制約範囲は、それぞれのルールにより異なっている。つまり法律であれば国内､条例であれば国内の一部地域､特定のコミュニティーの規約であれば、そのコミュニティー内というようにである。ここからルールは、管理や判断の責任者（組織）が決められており、効力や制約の範囲が限定されたものと言える。

これに対し、マナー（エチケット）の多くは礼儀や道義の問題であって、知らなければ恥をかくというべき性質のものである。管理すべき人や組織があるわけではなく、効力や制約の範囲もより普遍的だと言える。ただしマナー（エチケット）が過度に乱れるとルールをもって統制しなければ収拾がつかなくなる。またマナー（エチケット）が慣習法としてルールに転化する場合もある。

さて、自由なネットの世界では、多くの人がルールやマナー（エチケット）に意見を述べている。ただその内容（特にルールへの言及）に疑問や不安を感じる場合がある。もちろん既存のルールに対して意見を述べるのは大変良いことだと思う。しかし当否の判断資格の無い者が、勝手な解釈を押しつけたり、制約範囲外の人に自分たちのルールを強要したりするのはいかがなものであろうか。“私が法律”、“私がルール”というような態度をとっていないかどうか、ルールとマナー（エチケット）を混同していないか否か、今一度自分の行為を点検していただきたい。

◇ 憶えておくべし！　著作権

⊙ はじめに

総論でも触れたように、本会では毎年著作権講座を開催し、その成果を本誌で公開してきている。そこで以下、これまでの著作権講座で培った成果と、筆者の実際の経験を元に、インターネット上に他人の著作物を公開するために確認すべき点をまとめてみた。他人の著作物（今回は書籍など紙媒体のものに限定して論じる）をインターネットで公開する際の一例として読んでいただければと思う。なお紙幅の関係もあり一般的な用例を述べるに止まることご了解いただきたい。

⊙ 著作物か否か

まずこの同定をしなければならないが、これは非常に難しい問題でもある。著作物の定義は著作権法第２条、第１項、第１号には次のようにある。

思想又は感情を創作的に表現したものであつて、文芸、学術、美術又は音楽の範囲に属するものをいう

重要なのは、思想又は感情のオリジナリティーのある表現で、広く文化的な著作物ということである。ここから単なるデータや、単なる事実は著作物とは認められない。また技術的思想も著作権法の範囲外となる（これらに関しては、商標登録や特許といった形で保護される）。

一般的に出版された図書や論文の形態をとるものは、判断しやすいが、書状や手紙などは別途慎重に判断しなければならない[3]。

【著作権法→第 10 条～第 13 条】

⊙ 著作権者は誰か

著作物であれば著作権者がいるはずだ。論文であれば著作権者はより明白かもしれない。しかし掲載誌の編集部に著作権が移っている可能性もある。特に最近は後々の電子ジャーナル化を見据えて、論文掲載時に著者に著作権委譲の承諾書にサインを求める学会誌も多い。このように著者＝著作権者とは限らない場合もあるので注意が必要だ。

書籍だと、挿絵や本文著者以外の序・跋・解説などがあったりする。この場合、本文の著作権者とは別に複数の権利者が存在する。このように書籍の場合は著作権者が一人とは限らない。ただし最終的な公開手法により、同じ書籍でも対象となる権利者数は異なってくる[4]。また論文と同様に著作権が出版社に委譲されている場合もあるので気をつけなければならない[5]。

これ以外に、出版社が版面権と称して出版の権利を主張してくる場合もある。古典籍の校訂本では、校訂者の権利に関する問題[6]もある。いずれも現行の日本の著作権法で認められていない権利だが、対応に十分注意すべきである[7]。

【著作権法→第 14 条～第 16 条】

⊙ 保護期間の確認

次に該当著作が著作権保護期間にあるかどうかを確認しなければならない。著作物が保護期間内にある場合、複製や公開に著作権者の許諾が必要となる。

保護期間は、原則として著作者の死後 50 年、団体著作にあっては公表後 50 年である。ただし計算を簡略化するため、全ての期間の起点は、死亡、公表、創作などの翌年の 1 月 1 日となっている点、注意しなければならない。

このほか、現著作権法（昭和 46 年 1 月 1 日施行）以前の著作物は、旧法の影響を受ける場合がある。また戦勝国の著作物に対する戦時加算措置もある。このため昭和 46 年以前に公表された著作物や、戦前の海外の著作物は注意が必要である。

なお著作権が切れているにも拘わらず、掲載料と称して、法外に著作権料的な料金を請求する出版社や新聞社があると聞く。こういった場合は法的根拠をきちんと質し、毅然と対処すべきである。

【著作権法→第 51 条～ 58 条，附則第 2 条，附則第 7 条、連合国及び連合国民の著作権の特例に関する法律第 4 条】

⊙ データ加工と著作権

著作物をデータとして加工する場合は、著作人格権（公表権、氏名表示権、同一性保持権）に注意が必要となる。著作人格権は譲渡できず、著作者の死とともに消滅する。しかし著作人格権消滅後も、改竄や剽窃などこの権利を損ない、著作者の人格を傷つけるような行為をしてはならない。

【著作権法→第 18 ～ 20 条，第 59 ～ 60 条】

⊙ 複製権と公衆送信権

複製することとインターネット上で公開することは別の権利である。著作権保護期間内にある著作物を、著作権者の許諾を得てインターネット上に公開する場合、複製権とともに公衆送信権の許諾を得なければならない。

許諾のない場合、サーバーにアップロードした時点で、公衆送信権の侵害となる。

【著作権法→第 23 条】

⊙ 所有権と著作権[8]

所有権は民法に規定される権利で、著作権とは

別個の権利である。著作権は無体物を規律対象とする表現された発想への権利であるのに対し、所有権は有体物を規律対象とし、所有物を使用、収益および処分する権利だと解釈されている。図書館や資料館などが、著作権の切れた所蔵資料でも、複製等に許諾を求めるのは、民法上の所有権を行使していると言える。また書簡などは、所有権は受取人にあるが、著作権は差出人にあるので、許諾を得る相手を間違いないこと。

⊙ その他の権利との関係

著作権や所有権だけでなく、文献を公にする際には侵害に注意すべき権利がある。もっとも大きいものは人権である。近世～近現代の一次資料などは十分気をつけなければならない。公開することで、特定地域や関係者の子孫が差別等を受ける恐れのあるような資料は、たとえ著作権が切れていても電子化して Web 公開すべきではない。一次資料をインターネット上に公開する場合は、一点一点きちんと内容を精査しなければならない。

⊙ おわりに

以上、簡単ではあるが、他人の著作をインターネット上で公開する場合に、注意すべき最低限のポイントについて述べてみた。実際には個々様々な事例が多くあり、周辺の法的知識や判例知識をある程度有していないと対応できない場合もあるだろう。頭の痛い問題だが、他人の著作物への理解は自分の著作物への理解に繋がると信じて、自己研鑽を積むしかないのである。

◇『漢字文献情報処理研究』掲載関連文献目録

⊙ 再びバックナンバーから

総論で述べたように、情報発信のルール・マナー・スキルについて、さらに知識を得たい方は、本誌のバックナンバーをご覧いただきたい。そこで展望にかえて、本誌バックナンバーに掲載の関連論文・記事の目録を作成した。特に本稿ではスキルについて具体的に言及できなかったので、以下の目録を参照の上、関心のある分野を一読いただきたい。なお情報発信以前のもっと根本的なスキル（OS の特性、各種ワープロソフトや表計算ソフトの使用法、多言語表示など）は、これまでの本誌各レビュー、『電脳中国学』および『電脳中国学Ⅱ』、その他各種マニュアル類を参照していただきたい。

⊙ 第 1 号（2000.10）

- 二階堂善弘「『武王伐紂平話』データベースについて――電子テキストと画像データ――」
- 千田大介「漢字文献データの構築と公開をめぐって――中国古典戯曲文献データを例に――」
- 師茂樹「仏教学データベースにおける XML の活用：INBUDS における ID 検索の実現にむけて」
- 野村英登「漢－英術語データベースの構築へ向けて」
- 佐藤仁史「中国における近現代史史料のデジタル化の試み――上海デジタル図書館の場合――」

⊙ 第 2 号（2001.10）

- 千田大介「中国における古典文献データベースの構築：書同文公司へのインタビューを通じて」
- 師茂樹「GB18030 とは何か　大陸の戦略」
- 陳弱水（野村英登 訳）「中央研究院歴史語言研究所漢籍全文自動化計画の発展､現状､未来」

⊙ 第 3 号（2002.10）

- **特集　求められる学術情報研究の発信**
  - 小川利康「中国の場合」
  - 小島浩之「国内図書館における学術研究情報発信の現状」
  - 大内英範「日本文学の場合」
  - 岩本篤志「新潟大学人文・敦煌プロジェク

トについて：小規模な研究会による Web を使った情報発信の一例」
- 二階堂善弘「多漢字・多言語 Web サイト構築における諸問題」
- 千田大介「電子版学術雑誌をめぐる諸問題『中国都市芸能研究』創刊始末記」
- 小島浩之「人文情報処理および情報リテラシー関連書籍ガイド」

⊙ 第 4 号（2003.10）

- 師茂樹「『東洋学情報化と著作権問題』参加レポート」
- 石岡克俊「東洋学情報化と法律問題――第 1 回――所有権の行使と無体財産の法的保護：判例の分析と解説」
- 小島浩之「著作権についての知識を深めよう：東洋学のための著作権サイト・ページ指南」

●特集　漢字処理技術の最新動向

- 守岡知彦，江渡浩一郎，苫米地等流，宮崎泉，師茂樹「CHISE Project」
- 上地宏一「文字コード外フォント処理」
- Christian Wittern（師茂樹 訳）「Embedding Glyph Identifiers in XML Documents」
- 川幡太一「JIS X 0213 の改正と UCS との関係について」
- 師茂樹「Unicode 4.0」

⊙ 第 5 号（2004.10）

- Christian Wittern（秋山陽一郎 訳）「唐代ナリッジベースに向けて」
- 陳弱水（山下一夫 訳）「デジタルアーカイブと東洋学――中央研究院歴史語言研究所の経験から――」
- 師茂樹「春期公開講座レポート」
- 石岡克俊「東洋学情報化と法律問題――第 2 回――収蔵作品へのアクセスと法」
- 小島浩之「法理論と実務の狭間――「東洋学情報化と著作権問題Ⅱ」から――」

●特集　Wiki・Weblog と人文学

- 田邉鉄「プロジェクト研究における Wiki の活用」
- 千田大介「学術情報発信ツールとしての Wiki」
- 小川利康「授業に生かす Weblog と UniWiki――その特性と活用――」

注

[1] これら本誌掲載の関係論部については、後掲「『漢字文献情報処理研究』掲載関連文献目録」を参照のこと。

[2] 拙稿「人文情報処理および情報リテラシー関連書籍ガイド」（『漢字文献情報処理研究』第 3 号，2002.10）

[3] 手紙の著作物性については「三島由紀夫の手紙」事件（東京地裁 平成 10 年 (ワ)8761 号、平成 11 年 10 月 18 日判決、東京高裁 平成 11 年 (ネ)5631、平成 12 年 5 月 23 日上告棄却）が有名である。単なる時候の挨拶など事務的な内容でなく、思想又は感情を個性的に表現したものであれば手紙が著作物性を有すること、著作権が手紙の差し出し人にあること、著作人格権を著作者の死後も妄りに侵害してはならないことなどを判決の中で確認している。

[4] 例えば挿絵の著作権については、原本を画像化して公開したいならば配慮しなければならないが、本文のテキストのみ公開したいのならば配慮は必要なくなる。

[5] 著作権の委譲等が行われない限り、出版社は著作権を有していない。著作権法では複製権について、とくに出版社との関係を触れている【著作権法→第 80 条第 3 項】。また著作権法の規定する複製権以外の諸権利についてもほぼ同様である。翻訳権の許諾を著者ではなく原書の出版社に求めたため、著作者と日本の出版社の間で裁判となり、出版社側が敗訴した例もある。

[6] 本誌掲載の石岡論文および秋山論文を参照のこと。

[7] 法的に認められていない権利に対して従う必要はない。ただし、法的に認められていないことを説明できる法的知識を備えておかないと、対応しづらいだろう。

[8] 所有権と著作権の関係については、本誌第 3 号および第 4 号の石岡論文を参照。

データ入力下請けの使い方

千田　大介（ちだ　だいすけ）

❖ 手打ちデータか店屋ものデータか

❖ 自家製手打ちデータは難しい

ここにあなたの研究に欠かせない、非常に重要な分厚い本がある。あなたは繰り返し参照しなくてはならないその本を、自由に検索したり引用したりできるように、電子テキストに加工したくなった。さて、どうしようか？

紙に文字が印刷された「本」という物質的な存在を、「電子テキスト」という抽象的な電気的信号に置き換えるのは、それはそれで手間がかかる作業である。

電子テキストを作る一番単純な方法は、自ら手を動かすことである。本を傍らに、一生懸命日本語・中国語の IME を操りながら文字を手打ちしていく。自分の研究に欠かせない大切な本だけに、一字一字心を込めて、また本の内容を子細に読み直しながら入力していけば、必ずや研究を進める上でも大きな成果が得られるだろう。

しかし、高品質な手打ちデータを作るのには、それなりの技術と忍耐力が要求される。まず、IME などの漢字入力ツールを使いこなせなければならない。パソコンでどの漢字が使えてどの漢字が使えないのか、技術情報も知らなくてはならない。それになにより、入力には時間がかかる。分厚い本を自家製手打ちで美味しいデータに作り上げるのは、なにかと忙しい研究者にとっては至難のワザなのである。

❖ 忙しい研究者は店屋ものデータ

自家製手打ちデータの写経にも似た精神性へのこだわりを捨てる決心がついたあなたは、もっと手軽に電子テキストを作る方法を考えることになる。

手打ちが難しければ、機械で作ってはどうだろうか？確かに OCR（光学文字認識ソフト）という画像から文字を読み取ってくれるソフトもある。しかし、所詮道具は道具、それを操作して正確な電子テキストを認識させるのもそれなりに手間暇のかかる作業であり、また機械が見誤る形の似た文字は、人間も見逃しやすいというやっかいな問題もある。

ならばどうするか？餅は餅屋、プロに作ってもらえばよい。電子テキストを作るプロに電子テキストの作成を注文し、完成したデータを届けてもらうだけ、ただ自家製手打ちデータと違ってお金を払う必要はある。

❖ 電子テキスト発注の実際

❖ 業者の選定

店屋もの電子テキストを作る決心がついたら、こんどはどの店に注文するか考えなくてはならない。

中国学文献の電子テキスト作成をこなせる店は、日本国内にはほとんど見あたらない。そもそも山のような正字体を使った漢文文献のデータ入力、あるいは中国語簡体字文書の入力などというニーズは国内にほとんどなく、市場がないのだか

ら、店が生まれ育つはずもない。しかしよくしたもので、中国には漢文から現代中国語、さらには日本語まで、電子テキスト作成よろずお引き受け企業がいくつかある。

この分野では『四庫全書』『四部叢刊』全文検索版を作った書同文という会社が有名であったが、ここ一〜二年は技術的な立ち後れが目立つようになってきており、現在は「SimSun」フォントを作っている中易中標社と、書同文から枝分かれした創新力博社の二社が、技術面でも実績面でもおすすめである[1]。

❖ 価格の目安

電子テキスト作成のコストは、作成するデータの錯誤率と原本の複雑さによって変わってくる。

実は、中国の町中や大学の近辺には多くの「打字」屋さんがあり、電子テキスト入力から録音起こしまで、いろいろと引き受けてくれる。簡体字の原稿であれば 2 元／千字くらいなのだが、誤変換が多く見られ、歩留まりは一般に 90% 台後半程度である。つまり、百字に一文字以上、一頁に数字の誤字があることになり、研究につかう文献の入力には向かない。それに繁体字（旧字体）の入力はできないし、なにより大抵が零細個人経営で日本から注文して、送金して、というのが難しい。

中易中標・創新力博のような会社に委託する場合は、20 元／千字という価格が目安となる。電子テキスト化してもらう文献のコンディションや分量によって、この価格は上下することになる。

❖ 注文するときには

電子テキストを注文するには、まず、両方の会社のホームページに書いてあるメールアドレス宛にメールを送ろう。メールはできれば中国語で送ろう。ダメなら英語で、イザとなれば日本語でも社内に日本語のできる人がたぶん一人くらいは居ると思う。電子テキストの作成を注文する旨明記するほか、おおまかな分量、繁体字か簡体字か日本語かといった基本的な情報を書いておこう。

あとはメールで詰めてもよし、アポをとって北京の会社を直接訪問してもよし。筆者は、中国語で技術用語をやり取りするのがしんどいので、いつも直接担当者と相談して、価格や納期・データ形式などの詳細を決めるようにしている。

電子テキスト化する本は、原本やコピーを送ってもいいしスキャナで読み込んだ画像を CD-R などに焼いて送ってもいい。電子テキストの形式は、文字情報だけのテキストデータのほか、XML データにも対応しているし、両社ともに原本の版式を再現表示するようにデータを加工することもできるので、ニーズにあわせて判断しよう。相手はプロなので、こちらの好みにあわせて必要な電子テキストに調理してくれる。

❖ いろいろ使える店屋もの

下請けに出すのは、なにも研究対象の本ばかりに限ることはない。入れたいものは、コストが見合えば何でも委託してしまえばいいのだ。

例えば、個人の論文集を作るとき、原稿のデータはあるが校正したあとの最終決定稿のデータが無い、などというケースは多々ある。こんなときにも、中国の電子テキスト屋さんを使えば問題解決である。中国から送られてきた手書き原稿や発表の予稿を印刷するような場合にも、分量と重要度によっては電子テキスト屋さんに頼んでしまった方が手っ取り早いこともある。

使い方は皆さんのアイデア次第、リアルにあるいはバーチャルに日中を股にかけ、学術インフラ整備のグローバル化時代を存分に味わってみていただきたい。

注

[1] 中易中標：**http://www.china-e.com.cn/**
創新力博：**http://www.ilibo.com/**
両社についての詳細は、本誌掲載の拙稿「中国の人文学情報処理企業の最新動向」をご参照頂きたい。

特集2

人文科学研究と自然言語処理

東洋学においても、N-gram などの利用によりテキスト間の系統や、はては著作者の相違に及ぶまでを解析するような論考が出てきている。本誌はこれまでに 2 度 N-gram の特集を組んだところ、大変好評を博した。

しかし、東洋学において自然言語処理の方法論を理解している研究者はまだ少ないのが現状である。そこで東洋学だけ、N-gram だけといった枠を取り払い、人文科学研究と自然言語処理研究の関係を簡便に知り得るよう、少し裾野を広げた特集を企画してみた。執筆陣に自然言語処理の研究者も加わっていただいたため、各分野の研究史や現状、有効なシステムの紹介や検証、文理融合の研究の紹介など変化に富んだ論考が集まった。

この企画を通じ、読者諸氏にとって自然言語処理が少しでも身近なものになれば幸いである。

Contents

人文科学研究と自然言語処理

総論にかえて

小島　浩之（こじま　ひろゆき）

はじめに

我々は、コンピュータでテキスト中の文字を検索したり、置換したり、はてはgrep機能を利用したり、毎日の生活の中でいとも簡単にこれらをこなしている。コンピュータで文字が扱えることは、空気の存在と同じくらい当たり前のこととなっている。

しかしほんの数十年前までは、漢字は当然のこと、仮名であっても２バイト文字を扱うことに苦労した時代があったのである。そういった状況の下でも文献学へのコンピュータの応用は日々研究され続けてきた（この間の経緯は本特集内の近藤論文に詳細に綴られている）。当時､コンピュータを扱う人文科学研究者は一部の人間であったし､それでも良かった。

ところが現在では冒頭のような状況にあり、コンピュータを利用しない研究者の方が少ないだろう。個人レベルで相当なテキストの処理をこなすことも可能になり、以前ならば大型計算機を利用しなければならなかった分析が、パソコンレベルで可能になってきている。

読者の中には「自分はインターネットとメールの利用だけでしかコンピュータを利用しておらず、自然言語処理などは無関係だ」と言い切る方もおられるかもしれない。しかしインターネット上の多くの資源は自然言語処理の成果を利用している。検索エンジンも、図書館のOPACもみなそうである。我々は知ると知らずに関わらず自然言語処理の研究成果から大きな恩恵を被っているのである。

次節で述べるように、筆者は職務上の必要性から自然言語処理やその研究者と関わりをもつようになった。この成果や経過をまとめることで、少しでも人文科学研究に資することができればと思い、この特集を計画してみたのである。ただ、筆者の関わった範囲の話題だけでなく、もっと幅広い内容にしたいと考え、日本・中国・仏教学それぞれの分野における研究史と課題を寄稿していただいた。また一部のシステムについて言語学の立場から検証をお願いした。

先述のように、我々は多くの場面で、自然言語処理の成果の恩恵に浴している。本特集によって、人文科学研究と自然言語処理の関係史や現状、そして理科系の研究方法を知ることは、決して無駄ではないだろう。

前置きが長くなったが、以下本稿では、筆者と自然言語処理の関わりについて述べ、本特集への導入としたい。

Web情報へのキーワード付与

筆者の勤務する東京大学経済学部資料室では、従来からWeb上の重要資料の収集と、所蔵コレクションの電子化などが課題として挙がっていた。当時、省庁再編に伴う統廃合や財政難の影響

で、冊子で刊行されていた国の報告書・統計書類の一部が Web 上の公開へと移行した。結果的に多数には上らなかった [1] のだが、当時、紙媒体はいずれ無くなるといった憶測まで呼んだのである。この結果、筆者の職場では、まず Web 上の情報収集と適切なナビゲートについて検討することになった。方法としては、Web 上に氾濫する情報から適切なものを選び出し、それにメタデータ（書誌情報のようなもの）を付与し、ナビゲートすることとした。

当時の図書館界では、部門・分野に特化したメタデータ付与の検索システムの構築が盛んであり、東京大学経済学部もこれに準じたサブジェクトゲートウェイを構築しようと目論んだのであった。その際、付与するメタデータには、なるべく的確な情報やキーワードを盛り込みたいと考えた。特にキーワードは検索キーとなるため重要である。

直ぐに思いつくのは、人手でキーワードを付与する方法だが、人間の主観で判断した場合、キーワード付与の仕方に、自然とぶれが生じるのは避けられない。作業者が複数になればなおさらである。このため「人間の主観によらず、客観的かつ重要なキーワードを自動で付与できるシステムはないだろうか」こう考えを巡らせたのが、筆者が自然言語処理と関わるきっかけであった。

早速情報を収集したところ、東京大学情報基盤センターの中川教授と横浜国立大学の森助教授が“専門用語自動抽出システム” [2] というものを配付されていることを知った。すぐに当時同じ経済学部に勤務しておられた前田朗氏に相談し、二人で中川教授のところに出向いた。この結果、三人を中心にして小さな研究会を組織し、生みだされたものが、本特集内の前田論文で紹介されている言選 Web である。東京大学経済学部では、このシステムのおかげで、サブジェクトゲートウェイ Engel を公開することができた [3]。

中文版言選 Web への展開

筆者の目的は、Engel に利用できるキーワード付与ツールが完成したことで達成された。ところが、公開した言選 Web や関連ソフトが好評で、定期的に研究会を継続することになった。中川教授の理論と前田氏の技術により、言選 Web の性能は日進月歩の勢いで向上した。

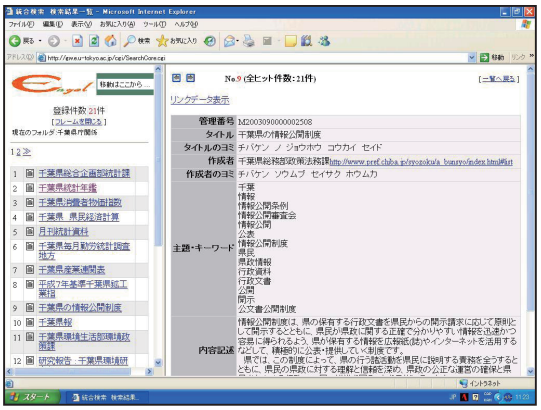


図　言選 Web によるキーワードを付与した Engel のメタデータ

日本語に関してある程度の目処がつくと、次は多言語に対応させたいということになった。西洋諸言語については、Kiwi（本特集の中川論文を参照）などで、中川研究室として既に実績があり難なく対応することができた。ところが中国語は予想外に難しかった。なぜなら言選 Web のシステムに組み込める、中国語用の形態素解析フリーソフトやフリーモジュールが見あたらなかったからである。

本特集の張・山本論文で述べられているように、形態素解析システムを一から組むのは至難であり、N-gram（本特集内の師論文参照）にしても言選 Web に組み込むには大部である。そこで考えついたのは、ストップワード（停止語）を指定して、そこを単語の区切りと見なし、言選 Web に判断させる方法であった。ストップワードの調整は筆者を中心に行ったが、一つの漢字が様々な品詞に成り得るという中国語の特性上、大変難航した。現在のストップワードリスト [4] は試行錯誤の結果だが、改良の余地が多分にある。例えば当初は Web 上の HTML1 ページ程度の、さほど文字量の多くないテキストを想定した。このため大量のテキストでの結果は芳しくない可能性があり、改良の必要がある。しかし、このストップワード

利用という思いつきが、無意味でなかったことは、本特集の山崎論文を見ていただければ明らかであろう。

なお言選 Web 中文版には、中国科学院の形態素解析システム ICTCLAS を利用して、品詞タグ付けを行った後、用語抽出を行う方法も用意している。ICTCLAS をシステムに組み込めないため、利用者には二度手間を強いているが、それなりの成果を挙げている（前田論文および山崎論文を参照）。

自然言語処理研究に関わって

このように、筆者が自然言語処理研究に関わることになったのは、ある意味偶然であった。三人の小規模な研究会では、多くの数式や専門用語が飛び出して、筆者はいつもとまどうばかりである。しかし理科系の研究者と一つのテーマに向かって研究を進めることは大変面白く、新しい発見の連続であった。研究の進め方、考え方から論文の書き方、業績の作り方など、筆者がこれまで馴染んできた歴史学や文献学的図書館学の世界とは異なり、非常に勉強になった。そしてこういった文理双方が関わる研究では、意思疎通とお互いの研究への理解が何より重要だと認識させられた。

私事になるが、筆者は学生時代、影絵劇の劇団サークルに所属していた。影絵劇は、光源からの光を、様々な素材を用いて作った人形や背景で遮り、シルエットを浮かび上がらせ、観客を魅了させる劇である。シルエットは基本的に平面しか表現できないので、空間を巧く演出することは難しい。シルエットをアニメーションのように自由自在に扱いたい一心で、いろいろな方法を試すのだが、今ひとつぎこちない。しかしずっと試行錯誤を続けていると、段々と巧く動いているように見えてくる。一生懸命そのことだけに没頭しているとつい錯覚してしまうのだ。当然のことながらアニメを見慣れた子供達には、ぎこちないおかしな動きににしか見えず、感動を呼ぶべき場面で大きな笑いが沸き起こるのである。

文理の双方が共同して行う研究の場合、各々が自分の認識の範囲内だけで進めると、良くも悪くも影絵劇の例と同じような弊害が起きてくるのではないだろうか。それぞれの思いの中では満足したように見えても、お互い見つめ合うと実はこっけいな結果になっているかもしれないのだ。

認識の違いは恐ろしいと感じた例を一つ挙げよう。本特集の各論文内で頻出する言葉に形態素解析がある。形態素とは意味を持つ最小限の単位のことで、「漢字文献情報処理研究」という語句であれば、次のように 10 個の形態素からなっている。

> 漢 / 字 / 文 / 献 / 情 / 報 / 処 / 理 / 研 / 究

形態素解析と言うからには、文章を上記のような形態素の単位で切り分けるシステムと思いがちだろう。ところが、これを形態素解析器にかけると、前田論文にあるように

> 漢字（名詞, 一般）、文献（名詞, 一般）、情報処理（名詞, 一般）、研究（名詞, サ変接続）

といった分割のされかたをする。中国語の形態素解析器 ICTCLAS にかけても

> 汉字/nz　文献/n　情报/n　处理/vn　研究/vn

という結果を得る。つまり形態素解析器による結果は形態素ではなく、単語単位で切り分けたものに、品詞を付与したものということになる。このことを自然言語処理の立場では次のように説明している。

> 言語学では、意味を担う最小の言語要素を**形態素**（morpheme）と呼ぶ。これに対して自然言語処理では、形態素を同定する処理、すなわち、入力文中の単語を同定し、その語形変化を解析する処理を**形態素解析**（morphological analysis）と呼ぶ [5]。

つまり、形態素解析とは形態素を抽出する作業ではなく、「形態素の概念を利用して単語分割と品詞タグ付けを行うこと」[6] だと言えるのである。したがって言語学的な形態素の理解だけで、自然言語処理における形態素解析を想像すると、とんでもないことになる。逆に自然言語処理の理解だけで、言語学の形態素の概念を想像しても、同様に見当はずれとなる。

おわりに

特集の総論的な文章として、このようなものが適当かどうか甚だ不安である。しかし、この機会に、人文科学を研究する者の率直な思いを書いておくのも良いだろうと考え、恥を忍んで書き散らしてみた。乏しい知識故の誤謬もあるかしれないがご海容いただきたい。

この特集によって、これまで以上に多くの人文科学研究者が、自然言語処理の世界に興味を抱いていただけたなら幸いである。

未筆ながら、お忙し中、本特集にご寄稿いただいた執筆者の方々に心より御礼申し上げます。

注

[1] この時期の政府刊行物の不安定さについては、拙稿「紙媒体資料と電子媒体資料」（『漢字文献情報処理研究』3，2002.10）を参照。

[2] **http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/resource/termext/atr.html**

[3] Engel 公開までの詳細な経緯は、拙稿「経済学部図書館サブジェクトゲートウェイサービス Engel について」（『図書館の窓：東京大学附属図書館報』43-1，2004.2）**http://www.lib.u-tokyo.ac.jp/koho/kanpo/vol43/vol43-1.pdf** を参照。

[4] **http://gensen.dl.itc.u-tokyo.ac.jp/doc/ChainesPlainTextGB#get_imp_word**

[5] 永田昌明「形態素解析」（『言語の科学３　単語と辞書』岩波書店，2004.6）54 頁。

[6] 東京大学情報基盤センター中川教授のご教示による。

この文書は『漢字文献情報処理研究』第6号の抜粋です（漢字文献情報処理研究会発行）。

自然言語処理と文献学研究

――日本語研究を中心に――

近藤　泰弘（こんどう　やすひろ）

1はじめに

自然言語処理を応用した日本語の研究は、当然のことながら、工学系の自然言語処理研究の中で、もっとも活発に行われてきている。しかしながら、本稿では、筆者の専門やまた本雑誌の性質から、もっぱら人文科学、特に日本語の語学的・文学的研究の立場からの自然言語処理ということについて述べてみたい。また中国の漢籍や漢訳仏典の情報処理研究については省略させていただくことを御了解されたい。

論述の順序としては、まずその研究史を追うことによって述べ、次に個別の問題点に順に触れてゆくこととする。紙幅の都合上重要な研究のすべてに触れることができないが、その分は参考文献などで補っていただきたい。

2研究史を追って

■ 2.1 日本語文献学研究と計算機の出会い

日本語研究の分野において最初にコンピューターを用いた自然言語処理的な研究が行われたのは、国立国語研究所（以下、国語研）の諸研究である。その初期の研究成果は『電子計算機による国語研究』（I ～ X）（1968 ～ 1980）、『電子計算機による新聞の語彙調査』（I ～ IV）（1970 ～ 1973）などによって公表されているが、後者の書名からわかるように最初は新聞データの漢字や単語の統計的分析からその研究は始まった。

この 1960 年代後半（昭和 40 年頃）は、コンピューター（当時は大型計算機）で漢字を扱うことがきわめて困難な時代であり、漢字入力や漢字印字をすること自体が自然言語処理研究の中心であったといっても過言ではない。1965 年に始まった国語研の研究では、『朝日新聞』『毎日新聞』『読売新聞』のそれぞれ 1966 年 1 年分の朝刊夕刊の全紙面から、ランダムに全体の 60 分の 1 の標本を抽出し（約 200 万語）、それを形態素解析したデータを作成した。それにより、異なり語彙表・使用頻度・その他品詞別の使用状況などを明らかにしたものである。そもそも、この時代の漢字印字出力に用いられた漢字テレタイプライタは印字スピードが秒速 2 字というようなものであり、その研究は困難を極めた。当然のことながら、JIS の漢字コードはまだ定められておらず、漢字キーボードディスプレイ（入出力装置）や、漢字テレタイプライタ（のメーカー）ごとに漢字コードが異なっている状態であったため各種の処理を行うごとに漢字コードの変換がつきまとう（当然のことながら、それぞれが覆う漢字集合の大きさも異なっている）のもひじょうに大きな問題となった。

この時代の研究としては、先にあげた新聞の調査があるが、国語研でこの同じシステムを用いて、試験的に行われた文学作品の研究が存在する。ひとつは、『志賀直哉『城の崎にて』用語索引――電子計算機による用例集作成の一実験――』（1971）である。これは文学作品を対象とした日本で最初の KWIC 索引である。なお元データは漢字であるが、プリンタが低速であったため、実際の印刷出力は半角カナのラインプリンタ出力によっている。また、同じ頃のものとして、国語研『森鴎外『寒山拾得』用語索引』（1974）がある。これは日本で最初の漢字プリンタを使った本格的な KWIC 索引である。これらの研究をもって日本の文献学における本格的自然言語処理が始まったといっても過言ではないであろう。国語研の用いた新聞データの一部（約 75 万語）は 1972 年に自然言語処理研究者に向けて公開され、多大な恩恵を及ぼしたことも時代に先んじるものとして特筆すべきことである。

この国語研のデータを用いた代表的研究としては、植村俊亮『電子計算機による自動索引の研究（上）（下）』（電子技術総合研究所・1974）がある。これは漢字による KWIC 索引の作成方法について理論的に追求したものであり、現在に至るまでこれを越える研究はない。その後、日本語研究者が KWIC を作る際に大きな影響を与えた研究である。また、同じデータを用いて品詞を用いた KWIC 状の索引を作った西村恕彦他編『日本語品詞列集成』（右順篇上下・左順篇上下・電子技術総合研究所・1977）も文法要素をキーとしてテキストを処理することができることを示した点で重要である。

■ 2.2 大型計算機の時代

さて 1980 年代前半まではその処理能力から、自然言語処理が可能なのは大型計算機（メインフレーム）に限られていた。1970 年代後半からはそれまでのバッチ処理（ジョブをパンチカードなどで投入して、まとめて処理されて後から結果がわかる方法）中心の用い方から、いわゆる TSS 処理（画面を見ながらキーボードで対話的にコンピューターにコマンドを発して処理を行う方法）を用いることが多くなり、使い勝手もかなりよくなっていた。また 1978 年に JIS 漢字規格（第一水準・第二水準）が定められたこと、漢字プリンタの能力も徐々に向上したこと、大型の磁気ディスク装置の発達でかなりの量のデータを磁気ディスク上でソートしたりすることができるようになった（古くは大量データのソートは磁気テープを交互に入れ替えて行った）ことなど、周辺の環境がよくなってきて、人文科学においても自然言語処理研究と言えるものが起こってきた。

この時代の代表的な研究をあげてみると、まず人文科学のための自然言語処理の教科書ができたということがあげられる。西村恕彦『人文科学の FORTRAN77』（東京大学出版会・1978）、この書は当時大型計算機で最もよく使われた言語である FORTRAN（当時最新の 77 規格）の教科書であるが、「本書は、文学部における電子計算機教育のための教科書である」と述べるように、その素材は文字・乱数・図形・数学・統計といった人文科学のすべての分野を網羅しており、その素晴らしい例題群とともに、日本語で書かれた人文科学者のための最高のプログラミング入門テキストであることは今に至るも変わらない。本稿の筆者は実は大学学部生の時に、この教科書（のもととなったプリント）で著者自身による授業を受けたのであるが、それはまことに幸運なことであったが、この教科書がその後広く使われずに絶版となっていることはまことに残念である。新しいプログラミング言語によって再生されることがぜひ望まれる。

ところで、大型コンピューターを用いて実際に文献学に応用した研究としては、国文学研究資料館の諸研究をあげることができる。国文学研究資料館では、『万葉集』をはじめ、多くの漢字を含むデータの扱いについて研究を行った。特に、漢字の異体字をどのようにコード化するかという問題、（現在の JIS 漢字の用語では「包摂基準」）について、日本で最初に取り組んだ研究である田嶋一夫による『データ処理システムのための漢字シソーラス［試作版］』（国文学研究資料館・科学研究費報告書・1980）はその集大成であり、その

後の JIS 漢字の改訂などに際して大きな影響を与えた。田嶋自身も JIS の補助漢字の規格策定に大きく関わった。またこの異体字システムを応用した文献学的研究として、国文学研究資料館編『連歌資料のコンピュータ処理の研究』（明治書院・1985）がある。これは国文学研究に自然言語処理的手法を用いて公刊された最初のものである。

この時代の計算機による文献学的研究のひとつの中心であった東大文学部言語学研究室から出版された『言語研究の中の計算機』（計算機利用言語学研究会・私家版・1982）には、荻野綱男・豊島正之・丸山徹等による諸研究が掲載されており、初期の研究の状況を知ることができる。同じ頃のものとして、金水敏『古文献の計算機処理（一）――東京大学国語研究室蔵恵果和上之碑文――』（神戸大学教養部・私家版・1984）は、計算機による訓点資料の処理として日本最初のものであり、また、山口明穂・近藤泰弘・金水敏・古田啓編『音訓篇立索引』（汲古書院・1985）は、古辞書の索引を対象とした最初の自然言語処理による研究である。

この時代には言語処理プログラムを研究者自らが作成して処理することが普通であり、言語としては FORTRAN,COBOL,PL/I などが主に用いられた。文献のデータの形式もこれらの言語に適した一行ごとの固定長フォーマットによることが多く、作品名やページ数といった情報を行の最初の固定された数十バイトに置き、その後に本文を置くといった形式がよく用いられた。

■ 2.3 パソコンの発達

1990 年代になりパーソナルコンピューターが高性能になり、またそれに付属する磁気ディスク装置（いわゆるハードディスク）やプリンタも徐々に高性能になり、かならずしも大型計算機を用いなくても、様々な言語処理ができるようになってきた。特に仮名漢字変換能力を備えたワードプロセッサーソフトウェアの登場によって漢字入力という点では大型計算機よりも使いやすくなったことがひじょうに大きいものがあった。またパソコン通信と言われるひとつのサーバーにアクセスするタイプのネットワークが生まれ、PC-VAN や NIFTY サーブといったそれぞれのサーバーに人文科学者の集まる掲示板が作成され、パソコンによる文献学研究の方法論を交換できるようになったこともパソコンによる研究の進展に大きな影響があった。パソコン上でのデータ圧縮ソフト（pkzip や lharc）、また圧縮ファイルを英数字にエンコードして掲示板などに掲載できるようにするソフト（ish や uuencode）などの登場は文献学的データを通信によって交換できる道を開いた。

この時代の自然言語処理的な文献学研究の特徴としては、データの汎用性およびその汎用的な処理方法についての議論が深まったことにあると思われる。これは文献学的データを誰もがパソコンで作ることができるようになったこと、また、それをネットワーク上で交換できるようになったこと、そしてパソコンの OS が MS-DOS か Mac に統一されたことなどにより、データの交換性といった問題について誰もが関心をいだくようになったことが要因としてあげられよう。

したがってこの時代のキーワードとしては「テキストファイル」あるいは「プレーンテキスト」があげられる。アプリケーションに依存しないもので、明示的なタグを持たないデータ形式が好まれ、またそれを grep の様な汎用ソフトウェアツールで文字列を検索したり、awk のような簡易でかつポータブルなプログラミング言語によって処理することが始まった。これらの UNIX 系のテキスト処理ツールが MS-DOS に移植されたことによりこれらのツールが利用しやすくなったことも特筆すべきである。

情報処理語学文学研究会（JALLC）は 1990 年代に活発に活動を行った研究会であり、フロッピーディスクで配布されたその会報には有益な情報が多い。豊島正之「TEI から見た SGML の話」（12 号・1992）同「perl で遊ぼう」（13 号・1992）、池田証寿「漢字字書データベースの作成とその利用」（15 号・1994）など現在でも見るべき論考が多い。いずれかのホームページに全巻の公開がされることを期待したい。

またこの時代の研究活動を知ることができる

ものとして 1990 年に刊行が始まった雑誌『人文学と情報処理』（勉誠社）がある。これは創刊号「特集　コンピュータ利用の現在」から始まり、2000 年の 29 号まで続いたもので、その 10 年間における人文科学のコンピューター利用の状況についてコンパクトに知ることができる資料である。

ほぼ同じ頃に、情報処理学会に、「人文科学とコンピュータ」研究会が発足し、年に 3 回ずつ研究会を行うことが始まった。第 1 回は 1989 年であり、その後現在まで継続している。この研究会の研究成果をまとめたものとして、及川昭文監修『講座　人文科学研究のための情報処理（全 5 巻）』（尚学社・1998）がある。これは 1995 年度から 4 カ年に渡り上記研究会を中心に行われた特定研究「人文科学とコンピュータ」の成果である。特に安永尚志による第 3 巻「テキスト処理編」は文献学的テキストのコンピューター処理についての概論となっておりきわめて有益な参考書である。

統計的な研究では村上征勝による諸研究が重要である。村上征勝による『文化を計る　文化計量学序説――』（朝倉書店・2002）は大型計算機の時代からの村上の統計的な文献学の理論をまとめてあってわかりやすい参考書となっている。

■ 2.4 インターネット時代の到来

1995 年頃からインターネットにおける WWW の利用が盛んになり、それまでごく一部の研究者のネットワークに留まっていたインターネットは爆発的な普及を見せるに至った。ホームページに作成した文書を置くことで世界中からアクセス可能になるというこの技術によって、使うことのできる日本語の文献学的データの量も飛躍的に増大した。またその種類も、上代から近代にいたる日本文学作品、古辞書、漢文に至るまでのものが揃うようになり、それ以前とは比べ物にならない研究環境が実現してきた。国文学研究資料館・国立国語研究所・国会図書館デジタルライブラリー・東京大学史料編纂所・京都大学付属図書館などには膨大な量のテキストや画像データが公開されている。

また、Windows パソコンが標準化され、誰でも自由にパソコンが使える環境になったこと、また、『国歌大観 CD-ROM 版』（角川書店・1996）のような市販のアプリケーションソフトが販売されることになったことにより、ごく一般の研究者レベルでも、自然言語処理的手法によって研究ができることになったこともこの時代の大きな変化であろう。その後、『源氏物語』『吾妻鏡』『群書類従』など種々の文献が CD-ROM 化された。

また、漢字コードが JIS の第一水準・第二水準に留まらず、ユニコードをはじめ、様々な大規模な漢字集合の提案があり、またそれが実用化されたこともこの時代のおおきな発展である。大規模漢字集合としては、JIS の補助漢字・第三四水準、GT 漢字・今昔文字鏡などそれぞれ種類の異なるものがあるが、ここではその詳細は省略する。漢字を含む文献の自然言語処理的手法については、漢字文献情報処理研究会の活動による発展が大きいものがある。機関誌の『漢字文献情報処理』（2000 ～・好文出版）や、研究会編による『電脳中国学』（1998・好文出版）でその研究活動の概観を知ることができる。漢字を含む多言語の文献学的テキストを処理するノウハウに関してはもっとも先進的な情報を提供していると言ってよいだろう。

また日本語の漢字文献についての扱いについては、池田証寿による研究がもっとも詳しい。氏の主導する研究会の雑誌『古辞書と JIS 漢字』（1 号～ 5 号）（北海道大学文学部言語情報学講座（国語）・1999 ～）はその方面の情報についてのもっとも詳しい情報源であり必読のものである。

もうひとつインターネットの普及によって変化した側面としては、データの構造化ということに改めて注目されてきたことがあげられる。日本で公開を前提に作られた本格的な文献学的なデータは『源氏物語』であり、1990 年に長瀬真理等によって作成され公開された。このデータは COCOA 形式でタグ付けされていたが、この後、国文学研究資料館の作成した諸データが KOKIN ルールというタグが付されていたのをはじめ、大

型計算機の時代にすでにデータの構造化については留意されていた。それがWWWの普及によりHTMLというテキストマークアップ言語が普及するにつれ、一般にもテキストの構造化の意義が十分にわかってきたということが重要である。そのため、近年作られる大規模な文献学的データは何らかの意味での構造化（タグ付け）がされていることが多い。代表的なものとしては、国語研の『太陽コーパス』（博文館新社・2005）がある。これはXMLによって近代語研究資料をマークアップしたものであり、XSLTによる各種データへの変換ツールも含まれていることもあり、今後の研究のひとつのモデルとなるべきものであると思われる。

■ 2.5 21世紀の研究

さて、今後の文献学における自然言語処理を応用した研究はどのようにあるべきであろうか。現在の最新の研究動向を見て将来を予測してみることにしよう。まず、国語研や国文研の研究に見られるように、工学系の自然言語研究処理との共同作業が今後ますます重要になるであろう。コンピューターの発達により自然言語処理の理論は急速に進歩しているが、人文系研究者にはそれにおいついてゆくだけの時間的余裕がないのが実情である。したがって、共同研究によってお互いに足りない部分を補いあって研究を進めることが必要であろう。また、工学系で作成された種々のツール（奈良先端大で開発された形態素解析ツールChasenやKWIC索引操作ソフトChakiなど）ももっと人文科学で使われてもいいと思われる。

また、従来の文献学における自然言語処理は基本的には単語の統計と検索につきていたといっても過言ではない。しかしコンピュータにできることはこれ以上のものがある。近年注目されているのは文字のN-gramの考え方に基づく文字の出現頻度統計をとることにより、特徴的な文字の並びを抽出できることがわかったことである。これにより、平安時代語の語彙の位相（近藤みゆきによる）、完訳仏典の諸本の特徴（石井公成・師茂樹による）などを調査する研究などが行われており、今後の発展が注目される。これらについてはいわゆるデータマイニングの手法のひとつであると思われ、近年自然言語処理で脚光を浴びている機械学習の応用も考えられるところである。

▌3分野別の問題点と参考文献

■ 3.1 テキストの構造化

文献学的研究のためのテキストは何らかの構造化が必要であることは現在の時点ですでに明らかになっていると思われる。テキストの構造化については、次の文献が詳しい。

これらを参考にしてXMLによる簡易なタグ付けをしておくことが今後の文献学のためのデータには必須の事項である。幸いに、マイクロソフトのオフィスアプリケーションの基本ファイル形式が次のバージョンからはXMLになるために、XML文書を作ることはきわめて容易となった。XMLについての基本的知識は今後の文献学研究者にとってますます欠かせないものになるだろう。

- 安永尚志『講座　人文科学のための情報処理　第3巻　テキスト処理編』（尚文社・1998）
- 豊島正之「TEIから見たSGMLの話」（『情報処理語学文学研究会会報』12号・1992・この趣旨を生かした発展版論文が豊島氏のWebページにあり）
- 神崎正英『ユニバーサルHTML/XHTML』（毎日コミュニケーションズ・2000）
- C. M. Sperberg-McQueen and Lou Burnard, *Guidelines for Electronic Text Encoding and Interchange (TEI P4)*, The TEI Consortium, 2002
- 近藤泰弘「古典語のコーパス」（『日本語学』4月臨時増刊号・コーパス言語学・2003）
- 国立国語研究所編『雑誌『太陽』による確立期現代語の研究』（博文館新社・2005）

■ 3.2 アプリケーション

文献学の立場からの各種アプリケーションソフトの使い方については、次の書物に詳しい。ソフトウェアを使う上では、漢字コードの問題を避けては通れないが、これについてはきわめて問題点が多いのでこれらの参考書に譲ることとする。

- 漢字文献情報処理研究会編『電脳中国学』（好文出版・1998）
- 文字鏡研究会編『パソコン悠悠漢字術 今昔文字鏡徹底活用』（紀伊国屋書店・1999・改訂版 2002）
- 漢字文献情報処理研究会編『電脳国文学』（好文出版・2000）
- 中村康夫『古典研究のためのデータベース』（臨川書店・2000）
- 伊藤雅光『計量言語学入門』（大修館書店・2002）

■ 3.3 プログラム言語とアルゴリズム

自分自身でプログラムを作って文献処理をすることを考えている人にとっては現在は適切な参考書が少なくきびしい時代である。しかし、とりあえず、次のようなものがあげられる。

最後のものは文献学対象ではないが、文献学と似たテキスト処理の必要がある DNA 分析のためのプログラミング入門でありきわめて有用な書物である。

- 西村恕彦『人文科学の FORTRAN77』（東京大学出版会・1978）
- 田中章夫他編『朝倉新日本語講座　運用 II・人文系研究のための言語データ処理入門』（朝倉書店・1983）
- 水谷静夫『次第立て言語　小朱唇の手引き』（東京女子大学日本文学科・1986）
- 影浦峡『計量情報学　図書館／言語研究への応用』（丸善・2000）
- 中尾浩他編『コーパス言語学の技法 I　テキスト処理入門』（夏目書房・2002）
- 佐野洋『WindowsPC による日本語研究法』（共立出版・2003）
- James Tisdall『バイオインフォマティクスのための Perl 入門』（オライリージャパン・2002）

この文書は『漢字文献情報処理研究』第6号の抜粋です（漢字文献情報処理研究会発行）。

中国語のコンピュータ処理について

コンピュータによる中国語処理の発展と課題

張 玉 潔（ちょう　ぎょくけつ）・山本　和英（やまもと　かずひで）

はじめに

中国語は世界の 20% の人口が使う言語であり、近年の中国経済の発展に伴い、中国語で発信される重要な情報もますます増加してきた。そのため、中国語への関心は近年ますます高まってきており、これに伴ってコンピュータによる中国語の自動処理も重要視され始めてきた。最近、いくつかの著名な研究機関や企業が、中国語と英語や日本語などとの機械翻訳の研究開発に本格的に力を入れているのはその証拠である。

当然、中国国内でも中国語の言語研究及び中国語の情報処理は盛んに行われている。本稿では、中国国内における中国語処理の発展状況と最新技術及び課題を概観する。

中国では中国語に関する研究の歴史は長い。文字と韻律に関する学問は昔から盛んで、多くの輝かしい業績を挙げてきた。20 世紀になると、中国語研究は西欧言語学理論の影響を受け、現代中国語の音韻学・語彙論・文法及び意味論の分野が開拓され、その後の現代中国語研究は、西欧言語学の思想を吸収しながら発展してきた。特に 1980 年以降、中国の改革開放政策を背景に、コンピュータ技術の導入と計算言語学の成果を導入し、中国語の言語研究は急速に発展してきている。

中国語の言語研究は以下の二つの面から推進されている。すなわち、一つは中国語を第二言語とする言語教育であり、もう一つは中国語の情報処理である。中国語の情報処理の研究については、主に三分類することができる。すなわち、⑴西欧の計算言語学の成果を中国語に適応し、中国語の特徴に適合するような中国語の計算言語学の理論を立てること、⑵機械翻訳・情報抽出・情報検索・自動要約などの他に応用可能なシステムを開発すること、⑶中国語の言語知識のデータベースとタグ（情報）付けされたコーパスを構築して、情報処理のための様々な法則を発見することである。

コンピュータによる漢字の入出力

コンピュータによって中国語を処理できるようになったのは、英語などよりも遙かに遅れていた。1980 年以降コンピュータでの漢字の入出力について取り組み、実現までには約 10 年を要した。

1983 年当時は、コンピュータは主に外国からの輸入品であったため、漢字を表示することができなかった。当時筆者（張）は、漢字のフォントをデザインし、コンピュータの文字表示機能を使ってコンピュータに漢字を表示させるというシステムの開発に携わっていた。

当時のコンピュータの記憶容量がごく小規模だったため、綺麗にデザインされた漢字フォントを圧縮しなければならず、また出力時には元の綺

麗な字形に復元しなければならなかった。このような制限の下で、研究者は漢字フォントの圧縮技術をめぐって様々な知恵を生み出したのである。

漢字を入力する方法は研究者と一般人の双方により数多く考案され、また実践された。その中で、発音表記（拼音）に従う方法が一般の人に最も利用されている。他方、字形に基づく方法もある。これは一つの漢字を上下、左右、内外の部分に分解し、一定のルールに従って入力する方法である。この方法は入力速度が速いため、大量の入力作業が必要な人によく使われている。

漢字の入力における重要な技術の飛躍は、連想の概念を導入したことであった。文字と文字、あるいは単語と単語の連想により、コンピュータは自動的に後続する文字の候補を絞り込んで提示することが可能となり、これによって入力スピードが格段に速くなったのである。例えば、「中」を入力するだけで後続文字候補として「国」、「华」などの文字が提示され、入力したい文字の番号を押せばその文字が入力される。これは後に携帯電話なで使用される予測入力技術（または予測変換技術）と原理的には同一技術である。

キーボードによる漢字の入力問題の解決後、漢字の手書き入力や音声入力技術も開発され、それぞれ実用のレベルに達している。現在では、多くの OS が中国語の入力や表示を実現している。

中国語の形態素解析

中国語は日本語と同様分かち書きをしない。つまり単語間に空白（スペース）がないため、言語の自動処理の第一歩として「文を単語に分割」「単語ごとに品詞の情報を付与」の二つが必要である。これらの処理は日本語の形態素解析に該当する。この段階は 1990 年からおよそ 10 年間を要した。

■ 単語の定義

文を単語に分割するには、まず「単語」（中国語文法で言うところの「詞」）がどのような単位かを定める必要がある。中国語では、単語の定義について多種多様な意見がある。中国語を母語とする人に調査したところ、テスト用文章に対して単語の認定は約 70％しか一致しなかった。

漢字の情報処理を推進するために、1992 年に中国政府が中国語情報処理用単語分割基準（中国国家基準GB13715）を制定した。これによると「結びつきが強く、安定した状態で使われる二つあるいは三つの文字列を単語の分割単位とする」と定義している。ここでの「強く」と「安定」という用語は工学的にはあまり厳密ではなく、解釈に問題が残るものの、この基準によって単語分割用の単語リストを作成することが可能となり、単語分割ツールの開発が一歩先に進むこととなった。

■ 単語分割タスクの難しさ

日本語では、用言は時制とアスペクトによる活用（語形変化）があり、また使用文字として漢字以外にひらがな、カタカナ、アルファベットと多様なため、これらがいずれも単語分割の大きな手掛かりになる（ただし文字種の区切りと単語の区切りとは完全には一致しない）。しかし、中国語には語形変化がなく、文字種も原則として漢字のみであり、日本語と比べ単語分割は困難である。

■ 単語分割手法

主な分割手法は、語彙構造知識に基づくルールベース手法（規則に基づく手法）と大量の用例に基づく統計的手法、及び両方を統合した手法とに分かれる。以下に、語彙構造知識の例を挙げる（Gao 他 2003）。

⑴単語の重複：干干净净，研究研究
⑵接頭語＋単語：副部长，非党员
⑶単語＋接尾語：全面性，朋友们
⑷動詞間に助詞・副詞挿入：看得出，看不出
⑸動詞＋時制助詞：克服了，蚕食着
⑹動詞＋方向補語：走进，走出来，
⑺動詞の分離形式：洗了澡，洗个澡

例えば、⑴の「干净」は単語辞書に登録されているが、「干干净净」は未登録である。⑴の規則により、単語を構成する二つの文字がそれぞれ重

複してできた四文字列も単語になる。よって、文「房子打扫得干干净净」に対し、「房子 打扫 得 干干净净」のように分割される（空白は単語の区切りを表す）。この規則がないと、誤った分割結果「房子 打扫 得 干 干净 净」となってしまう。

規則に基づく手法の中で最も平易な手法は、文頭から文末に向かって文中の単語を単語リストと照合し、その中の最長文字列を単語とする方法（最長一致法）である。

例①はこのようにして分割した結果を示す。二番目の「将」と三番目の「来」は照合する単語リスト中に「将来」があるため、「将来」の一語となる。同様に六番目の「外」と七番目の「交」は「外交」の一語と認識される。さらに、上の語彙構造知識⑶「単語＋接尾語」を用いることで、単語「外交」と後続の接尾語「家」が一単語として認識される。

例① 他 将来 想 当 外交家。

しかし、この手法では例②の文が例③に示すような分割結果になり、正しく解析できていない。

例② 他将来日本留学。
例③ 他 将来 日本 留学。

文字列「将来」は二つの分割の可能性がある。すなわち例①の「将来」一語か、例③の正しい分割「将」と「来」の二語である。これをコンピュータに正しく分割させるには、必要な文法と意味の知識を教える必要がある。例えば、例③では主語「他」の後ろで時間の意味を表す「将来」は「日本」を修飾することができない、というような文法の知識をコンピュータに教える必要があるだろう。

■ 単語分割の難点

単語の自動分割でよく間違いが起こる原因は、単語リストを利用するとき生じる曖昧さにあることが明らかとなった。文自体は唯一の単語列を持つが、局所的に文字列を単語リストと照合すると、一つ以上の分割結果が起こり得る。これは中国語で「偽曖昧」とも呼ばれ、主に以下の二種類がある。

⑴包含型：前後の二つの文字は、単語リストにより、一つの単語になることもできるし、二つの単語になることもできる。

例②はこのような例である。本来二つの単語に分割されるべきものであっても例③のように一つの単語として認定される。

⑵交差型：ある文字は、単語リストにより、直前の文字と結合して一つの単語になることもできるし、後ろの文字と結合して一つの単語になることもできる。

例えば、例④（黄他 2003）の中の文字「流」に関して、「信息流」と「流入」は共に単語である。正しい分割結果を例⑤に示す。

例④ 防止有害信息流入和传播
例⑤ 防止 有害 信息 流入 和 传播

このような問題が生じるのは、中国語の文字が他の文字と結合して単語になるという結合力の強さに起因する。文献によると、中国語では 1 文字あたりのエントロピー（情報量）は 9.71 ビットであり（劉他 1987）、日本語の場合（1,500 個のかな・漢字に限定した場合）では 1 文字あたりのエントロピーは約 3 ビット（中川 1988）と報告されている。中国語で 1 文字のエントロピーが大きいことは、文字と文字とが結合して単語となる多様性が高いことを意味する。

単語分割問題を解決する方策として構文解析と意味解析についての研究が行われ、実用システムも開発された。他方、このような方向性に対して疑問を持つ研究者もいる。すなわち、単語分割の問題がそれよりもより難しい構文解析と意味解析によって解決されるのは本末転倒ではないか、との疑問である。実際に、上で紹介したような曖昧な例文を対象とする実用システムを用いた実験では、良い結果が得られなかったと報告されている。

■ 単語分割のための言語情報

単語分割に使われる単語リストあるいは単語辞書には、単語の品詞や意味などの静的情報が記載

されるが、これらが実際のテキスト分割に際してどのような問題を生じさせるか、またそれをどのように解決するかに関する情報は記載されていない。そのため、自動単語分割に必要な情報がまだ不足していると意識されるようになった。

そこで、まず包含型と交差型の問題に対して、大量のデータを用いて調査が行われた。事前に決められた単語リストにより、問題が生じる文字列が抽出、分類された。そして出現頻度が高い順に、文字列の分類ごとにどのような場合でどのように分割するかの情報がデータベースに蓄積された。このようないくつかの研究成果により、単語自動分割には詳細かつ膨大な情報を必要とすることが明らかとなった（孫他 1999・黄他 2003)。同時に、単語という分割単位がより明確に定義されるようになった。すなわち、次のいずれかに該当するものを単語と定義したのである（Gao 他 2003)。

㈠分割用単語リストの中の一つの単語と照合できる文字列
㈡語彙構造知識により単語として生成できる文字列
㈢人名、地名または組織名
㈣期日、時間、時間の長さ、貨幣、温度、長さ、面積、体積、重さ、住所、電話・ファックス番号、メールアドレス

㈡の情報は規則として表現し、これを適用することで実現できる。㈢の固有名詞に対しては専用のツールを作る。㈣の単語の構造は生成規則で表現し、有限オートマトンで実現することができる。

■ 品詞同定処理

もう一つ、分割された単語ごとに品詞を決めるという処理方法がある。中国語には、一単語が多数の品詞を持ち、また 1 つの品詞が多数の文法機能を持つという特徴がある。例えば「把」は以下に示すように 3 つの異なる品詞を持つ（俞)。

例⑥ 他 把 茶壶 打 碎 了。(介詞)
例⑦ 他 买 了 一 把 茶壶。(量詞)
例⑧ 今天 由 他 把 球门。(動詞)

品詞分類体系は長期間議論されているが、広く受容された分類体系は未だ存在しない。また、仮に分類体系が決まっても、これに従って数万個の単語に品詞を付与するのは膨大な作業である。更に、実際のテキストに対して単語の品詞付けを行うのは、コンピュータにとって容易ではない。

■ 品詞同定の手法

現在の技術では、単語の品詞付けを単語分割と同時に行うのが一般的であり、音声認識分野で発展してきた隠れマルコフモデルと呼ばれる統計的手法を単語分割と品詞付けに応用することで、良好な結果が得られることが知られている（中川 1988)。

具体的には、まず品詞を付与した大量のデータから隠れマルコフモデルのパラメーターを推定する（この作業をデータからの「学習」と呼ぶ)。次に未知文に対し、学習したモデルを利用して単語分割を行う。統計手法は学習データがあれば効率的にモデルを構成できるが、学習したモデルがどこまで本来の言語規則に合致するかは、用意した学習データに依存することになる。そこで実際のシステムでは、得られた結果に対してさらに精密化を行っている。具体的には、頻繁に誤る品詞付け部分に対し、これを正しい品詞に直す規則を作っておいて修正を自動的に行う、といった処理を行うことで処理性能を改善させるのである。

■ 未知語の問題

未知語とは、「事前に用意した単語辞書に掲載されていない単語」のことで、新語などがこれに該当する。新語は時代と共に増え続けるため、単語辞書が常に全ての単語を網羅しておくことは原理的に不可能であり、従って未知語をどう処理するかは単語分割の品詞付けにおいて大きな問題になる。一般的に言えば、文の中に未知語が含まれると、影響がその前後の文字列にも波及し、間違って分割されることが少なくないからである。

これに対し、文中における単語の境界を測定し、

その品詞も推定することを目的とする研究が最近増えてきた。一つの解決方法として、品詞分類に未知語のための分類を設け、隠れマルコフモデルを学習させるやり方が挙げられる。このモデルにより、未知語の分類の単語を認識できるようになり、また、各文字が単語を構成する際の情報を収集し、これによって未知語を認識することが可能となる。例えば、各文字が単語の先頭、中央、末尾になるそれぞれの場合について頻度を測定することで単語境界の妥当性を判断するのである。

■ 研究グループの紹介

ここでは、中国語の形態素解析に関する研究とシステム開発を行っているグループを紹介する。

中国国内では、北京大学・清華大学・山西大学・中国科学院計算技術研究所が挙げられる。

北京大学は、10 年以上をかけて中国語の単語の定義、単語の分割基準を研究してきた。その研究成果として「現代漢語文法情報辞書」（約 7 万語）（俞 1997）や「人民日報のタグ付きコーパス」が作成され（俞他 2000）、自動単語分割と品詞タグ付けツールが開発された（Zhou 他 1994）。これらの研究成果は中国語処理のデファクトスタンダード（事実上の標準）として広く用いられている。また、上記コーパスは大規模であるため、中国語情報処理研究への影響も大きい。北京大学と清華大学は、中国・西洋・日本の人名などの人名辞書、中国国内と外国の地名などの地名辞書、及び中国国内と外国の組織名などの組織名辞書を広く整備した。これらの辞書はシステムの高性能化に貢献している。単語品詞分類は、幾つかの体系が各グループにより提示され、各グループ開発の形態素解析システムとコーパスとに使われている。中国科学院計算技術研究所は複数の品詞分類体系間を自動的に変換するツールを開発した。

上に紹介した研究グループはそれぞれ各自で形態素解析のソフトを開発し、単語分割と品詞付けの正解率が 90％以上に達したことが各グループにより報告されている。単語の定義及び単語分割において多岐にわたる問題がまだ残されているが、中国語を単語に分割する技術は実用のレベルに入ったと言ってよい。

■ 中国語の構文解析と意味解析

機械翻訳・文生成・要約など中国語情報処理では、中国語文の構文解析と意味解析が必要である。中国語には文法構造を表す表層的情報が少なく、語順もより自由であり、構文構造を把握するための手掛かりが少ない。中国語の意味は単語の意味から直接構成されたとさえ言われることもある。

■ 文法研究

古来、中国語文法の研究は、形式・構造より意味を重視していた。19 世紀後半から、西洋の言語構造と言語形式に関する理論の影響を受け、中国語文法を体系的に記述する研究が始まった（代表的な著作に『馬氏文通』『新著国語文法』『中国文法要略』『中国現代語法』などがある）。

近年、中国語の特徴を踏まえ、意味重視の中国伝統的言語研究手法と言語形式重視の西洋言語理論とを融合する方向に研究が進んでいる。近年の中国語の言語現象を記述する文法理論として、文字本位文法理論、単語・句本位文法理論、小句本位文法理論などが提案されている（詹 2000 年）。

■ 構文解析の発展

1980 年の後半よりコンピュータによる中国語情報処理が始まった当時、西洋言語の様々な文法理論と構文解析の方法が導入された。例えば、チョムスキーの変形生成文法・格文法・機能文法などが中国語構文解析の自動化に応用され、中国語の文法に関する知見が人手により蓄積されてきた。これら西洋言語を記述する仕組みは、中国語の言語現象に適用できない点が多い。しかし、言語現象を規則で表現する考え方と記述手段及びコンピュータ処理のためのアルゴリズムは、中国国内の研究者に有益な影響を与えてきた。

1990 年以来、上述した異なった複数の中国語文法体系について、研究グループはそれらの文法をコンピュータにより実現し、検証実験を試みた。この分野の研究は始まったばかりであり、どの理

論が中国語をより正しく記述できるか、またどの文法がコンピュータにおいてより効率的に実現できるか、といった結論を出すのは時期尚早である。

■ 構文解析の手法

構文解析の処理とは、文中の文法成分を認定し、文法成分の階層構造を識別し表現することであり、この手法も形態素解析の場合と同じく、主に規則に基づく処理手法とコーパスに基づく統計手法とがある。前者を補強するために、複雑な特徴を取り入れる手法もある。また、前者と後者との手法を融合すべきとも提唱されている。

構文解析の一例として、例⑨の文が例⑩のような構文構造を持つことを説明する。例⑩では、階層構造は括弧で表現される。まず括弧 1 で単語「是」と「主席」が結合して一つの文法成分（仮に成分 1 と呼ぶ）となる。次に、括弧 2 で単語「他」と成分 1 が結合して一つの文法成分（仮に成分 2 と呼ぶ）となる。同様に括弧 3 で単語「认为」と成分 2 が結合して一つの文法成分（成分 3）となり、最後に括弧 4 で単語「我们」と成分 3 が結合して一つの文法成分（成分 4）となる。

例⑨ 我们 认为 他 是 主席

例⑩ (4我们（3认为（2他（1是 主席))))

中国語文の構文解析研究は、当初チョムスキーの変形生成文法を取り入れて始められた。しかし、上述のように、中国語は英語と違い構文構造上の制約がより少ないため、西欧言語を対象とする解析手法を直接中国語に使用するのは問題があり、中国語に適応した方法が模索されている。

統計的手法とは、構文解析済みのコーパスから文法を学習して構文解析ツールを開発することである。この研究は最近活発であり、ペンシルバニア大学の研究グループによって開発・公開された Penn Chinese Treebank は、世界初の中国語構文解析済みコーパス（treebank と呼ぶ）である（LDC）。これは、英語の treebank の仕組みに基づいて作られたものであるが、中国語の文法と構文解析（Xiong 他 2005・吉田他 2003）の進展に大いに貢献した。清華大学は 1998 年から 5 年間をかけて、100 万字の中国語構文解析済みコーパスを構築した（周 2004）。このコーパスは新聞・小説・教科書などからの文章により構成され、現代中国語の言語状況をよくカバーしている。

文の構文構造は階層構造木で表現されるが、付与される情報の特徴として、構文単位の内部における構成成分相互の意味関係、構文単位の相互の結合関係、及び構文単位の中で焦点となる単語などが考えられた。例えば構造木の任意の非終端節点には「文法成分ラベル」「成分関係ラベル」の二種類の情報が付与される。文法成分ラベルは全部で 16 個あり、文字・単語→句（チャンク）→文→段落の各文法成分を表現できる。例えば、名詞句のラベル np、時間を表す句のラベル tp、数量を表す句のラベル mp などである。成分関係ラベルは全部で 27 個あり、例えば節の中の主語と述語の関係、述語と補語の関係、修飾関係の種類（「定中」と「状中」）、概念と機能語との間の構成関係、といった多様な関係を表現できる。

清華大学はこのコーパスから構文解析システムと文法を抽出するツールを開発した。また、このコーパスのデータをほかの文法体系、例えば依存文法に変換するツールも開発している。

中国語文の構文構造全体を解析する技術は、まだ実用的なレベルには至っていないため、文よりも小さい部分、例えば句（チャンク）を解析する研究（チャンク解析と呼ばれる）が多く行われている。名詞句・動詞句などのチャンクの認識は、文全体の構文構造を解析する際の基礎的技術であり、情報処理の応用的課題の中には、基礎レベルの解析によって問題を解決できるものがあるため、チャンク解析も必要とされているのである。

■ 構文解析の課題

一方、中国語の一部の言語現象に限定して観察し、それを記述する文法研究もある。その対象は、主にコンピュータではうまく解決できない個別の言語現象である。そのため、コンピュータに必要な知識を教えるために、その言語現象を解析し、文法を記述し、さらにコンピュータが実現できる表現方法を考案する必要がある。

●名詞＋動詞のような単語列から生じた曖昧な構造を解消する研究

例⑪ 维护大局积极进取
例⑫ 维护大局的稳定

例⑪は、维护(v)と大局(n)とは述語と目的語のVPになるが、例⑫は、维护(v)と大局(n)とは直接関係しない。

●"介詞「被」＋動詞句1＋動詞句2" のような単語列から生じた曖昧な構造を解消する研究。

例⑬ 被警察抓住 vp1 罚了款 vp2
例⑭ 被老师批评 vp1 写了检查 vp2

同じ "p「被」＋ vp1 ＋ vp2" の構文単位列でも、例⑬は介詞「被」の修飾範囲がvp2までだから、「警察」は「抓住」と「罚」の動作の主体となる。例⑭は介詞「被」の修飾範囲がvp1までだから、「老师」は「批评」の動作の主体となり、「检查」の動作の主体ではない。

構文構造を正しく解析できないと正しい意味は得られない。先行して構文上の曖昧な問題を研究しない限り、構文解析の実現は困難である。

以下に、中国語構文上の主な問題点を挙げた(陸1998)。

●単語の品詞列が同じで構文構造が異なる

上の例⑬と例⑭はこのようなケースである。前述したように、中国語は一品詞がいくつかの構文成分あるいは構文の役割となるが、表層的な識別情報がない。例⑮は、いずれも動詞＋動詞の列であるが、それぞれ異なった構文成分になる。

例⑮ 打算 回家（述語　目的語）
　　研究 结束（主語　述語）
　　挖掘 出来（述語　補語）
　　唱歌 跳舞（並列）

●単語と品詞列が同じで構文構造が異なる

例⑯ 不　适当地管教孩子，对孩子不利。
例⑰ 不　适当地管教孩子，对孩子不利。

例⑯では、「不」は「适当地管教孩子」を修飾する。例⑰では、「不」は「适当地」を修飾する。すなわち同文だが、意味解釈が複数存在するのである。

●単語列が同じでも、文法上で正しい構文成分になる場合とならない場合がある
●文中の代名詞の照合問題

■ 意味解析の研究

一般に、文の構文構造が分かれば文の意味を解析できるとされる。文の各構文成分や成分間の修飾関係から、動詞の施事（動作の主体）・受事（受動者）・与事（与えられた主体）・手段・時間などの情報が得られる。しかし、構文構造が文の意味により違ってくることが度々あり、構文解析の問題は意味解析と結合して解決すべきとされた。

例⑱ 鸡不吃了。

例⑱の「鸡」は「吃」の施事であるか、受事であるかをまず決めないといけない。多くの場合人間は文の前後の情報により判断できるが、これをコンピュータに如何に判断させるかは、情報処理の実用的な問題となる。

また、意味解析のために人間が持つ常識をコンピュータに教える必要がある。中国語のHowNetは、コンピュータ上の知識データベースを実践したものである（董 2004）。HowNetでは「個体」「個体の属性」「個体と個体の関係」「時間」「空間」などを中心に据え、意味を表現するために意味素の体系が設けられ、中国語単語から共通の文字を抽出し、これを意味素として使用する（意味素は全体で800程度ある）。個体と個体の関係は全体で16種類あり、例えば「上下」「同義」「反義語」「部分と全体」などの関係がある。HowNetは6千個の中国語単語を意味素と個体の関係により定義した。例⑲と例⑳に定義された単語の一例を示す。

例⑲
　単語＝打
　品詞＝V
　例＝～酱油，～张票，～饭，去～瓶酒，

醋～来了
英語訳＝ buy
英語訳の品詞＝ V
英語訳の例＝
意味素による定義＝ buy｜买

例⑳
単語＝打
品詞＝ V
例＝～毛衣，～毛裤，～双毛袜子，～草鞋，～一条围巾，～麻绳，～条辫子
英語訳＝ knit
英語訳の品詞＝ V
英語訳の例＝
意味素による定義＝ weave｜编织

HowNet を利用して様々な知識を抽出できる。例えば、破壊の意味を持つ単語を全て探すとか、二単語間の意味上の類似度を計算する、などといった処理が可能である。最近、HowNet を利用して、情報検索と機械翻訳に応用する研究が増えてきており、研究成果もいくつか報告されている。

終わりに

以上、中国語言語処理の発展と課題について紹介した。この分野の内容は多岐に亘っており、本稿で紹介したのはこれらのごく一部である。また、我々の知る限りにおける概要であるため、中国語言語処理における重要な研究テーマ・技術・研究グループが網羅できていないかもしれない。

中国語言語処理の理論と技術については、今後より一層研究を進める必要があり、我々もこの分野でより多くの研鑽を積む必要がある。機会があれば、改めて新しい情報を紹介したい。

参考文献

- 陸俭明・郭鋭「漢語語法研究所面臨的挑戦」『世界漢語教学』1998 年第 4 集。
- 詹衛東「80 年代以来漢語信息処理研究述評」『当代語言学』2000 年第 2 集。
- 俞士汶「民族特点的文化要求――漢字和漢語民族語言進入信息系統」『科学概覧之信息巻』第十七章。
- 俞士汶・朱学峰・王恵・張芸芸『現代漢語信息辞典』清華大学出版社。1997。
- 俞士汶・朱学峰・段慧明「大規模現代漢語標註語料庫的加工規範」多語言信息処理国際会議 2000'ICMIP 論文集。2000 年。
- 劉源等「漢語字詞的概率分布，shang 及冗余度」『中文信息処理国際論文集』1987 年。
- 中川聖一『確率モデルによる音声認識』第 5 章。1988 年。
- 孫茂松・左正平・鄒嘉彦「高頻最大交集型 qi 義切分字段在漢語自動分詞中的応用」『中文信息学報』1999 年第 1 期。P.27-34.
- 黄昌寧・高剣峰・李沐「対自動分詞的反思」『語言計算与基於内容的文本処理』p.26-38。2000 年。
- 周強「漢語句法樹庫標註体系」『中文信息学報』2004 年第 18 巻第 4 期。P.1-8.
- 吉田辰巳・大竹清敬・山本和英「サポートベクトルマシンを用いた中国語解析実験」『自然言語処理』2003 年 Vol.10 No.1 P.110-131.
- 董振東
  "HowNet Tutorial" *The First International Joint Conference on Natural Language Processing (IJCNLP-04).* **http://www.keenage.com/**
- LDC. Chinese Treebank Project. **http://www.ldc.upenn.edu/ctb/**
- Deyi XIONG, Shuanglong LI, Qun LIU, Shouxun LIN, and Yueliang QIAN. "Parsing the Penn Chinese Treebank with Semantic Knowledge." *The Second International Joint Conference on Natural Language Processing (IJCNLP-05).* 2005
- Jianfeng Gao, Mu Li and Chang-Ning Huang. "Improved Source-channel models for Chinese word segmentation." *In Proceedings of 41th Annual Meeting Computational Linguistics.* 2003
- Qiang Zhou, Shiwen Yu. "Blending Segmentation with Tagging in Chinese Language Corpus Processing." *In Proc. of COLING-94*, P.1274-1278. 1994.

この文書は『漢字文献情報処理研究』第6号の抜粋です（漢字文献情報処理研究会発行）。

仏教学における自然言語処理

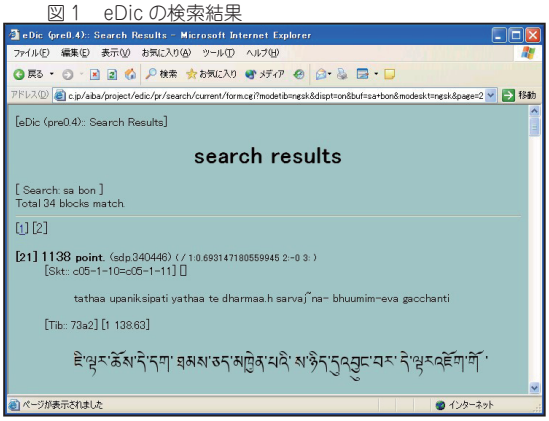
師　茂樹（もろ　しげき）

はじめに

仏教学においては、扱うテキストの多くがサンスクリット語、チベット語、古典中国語など複数の言語にまたがっていること、密接な関係にある隣接分野のインド学において言語研究が盛んなこと、偽作・偽訳などの問題が古くから大きな関心になっていること、等々の理由から、言語の問題に大きな関心が払われてきた。一方、コンピュータに関しても、その登場のころから強い関心が向けられ、現在に至るまで様々な分野で積極的に利用されている。

本稿では、仏教学の中で行われてきたコンピュータを利用した研究のうち、特に複数言語間の比較研究と確率・統計的な研究に焦点を絞り、簡単に紹介していきたい。

図 1　eDic の検索結果



複数言語間の比較研究

インドで創始されアジア全域に広がった仏教のテキストは、伝播の過程で様々な言語に翻訳されている。インドにおいてパーリ語やサンスクリット語などで書かれた仏典は、残念ながら多くが失われてしまったため、チベット語や中国語などに翻訳されたテキストを比較し、原語を復原、もしくは復原的に考慮しながら読解するという研究方法がとられてきた。またそれだけでなく、この方法は、例えばサンスクリット原典と漢訳の音写語の比較を通じた六朝音や唐代音の復原研究、あるいは西夏語訳された『華厳経』『六祖壇経』などと漢語原典との比較を通じた西夏語の復原研究など、言語学の分野でも活用されている。

横山紘一・廣澤隆之『漢梵蔵対照瑜伽師地論総索引』（山喜房仏書林、1996）はまさにそのような研究のための工具書として作られた労作のひとつであるが、両氏が本書を出版するために作成した漢訳語・サンスクリット語・チベット語の対照テーブルを 1999 年末に公開 [1] したことは、当時大きな話題となった。このテーブルは現在、Charles Muller 氏の Digital Dictionary of Buddhism[2] をはじめ、さまざまなデータベースに組み込まれ、活用されている。

また、漢字仏典とは直接関係ないが、鈴木隆泰氏をリーダーとするチベット語・サンスクリット語間構文対照電子辞書プロジェクト eDic[3] は、まさに自然言語処理的な手法による研究を目指すものとして、今後の展開が注目される。

確率・統計的な方法による文献比較

■ 計量文献学による研究

1851 年にド・モルガンが、単語の長さの平均値に著者の特徴がでるのではないかというアイデアを出して以来、欧米では文章の数量的性質からその特徴を見出そうとする研究が行われてきた。この方法は、実際に文章を読んで内容面から検討する従来の方法ではわかりにくい偽作（贋作）などの著者推定で特に用いられており、著者がはっきりとしない場合が少なくない宗教文献の研究においては大きな効果が期待できる。

この方法は、偽作の判定に用いられたり、記述内容ではなく形式的な特徴に注目することが多かったりすることから、文章の「指紋」を見つけるための方法と比喩的に説明されることが多い。

> 文体に指紋があるとすれば、それはどのようなものだろうか？それはおそらく、ある著者の文体的な特徴――例えば ‘such as’ の生起度数といった、まったく取るに足りないと言ってもよいような特徴を組み合わせたもの――であって、指紋と同様にその人に特有のものであろう。文体上些細で取るに足りぬ特徴だからといって、文体分析に利用しない理由にはならない。指先にある渦巻や輪が我々の容姿においては大切でも目につくわけでもないが、指紋が一生涯変わらないように、そういったものこそが著者の叙述において変化することのない特徴となるはずであり、他の書き手には見られないその人だけのものとなるはずであろう [4]。

この方法を漢字文献や日本語文献に対して適用しようという試みは、単語の分かち書きをしないという言語的な性質をはじめ、漢字処理技術の遅れ、そして日本の文献学界の保守性などもあり、欧米よりもかなり遅れることになったが、最近では国文学、仏教学、中国古典などでも徐々に活用されるようになってきている。

仏典に対して本格的に確率・統計的な分析を行った研究としては、まずは伊藤瑞叡氏・村上征勝氏らによる共同研究 [5] をあげなければならないだろう。この研究では、従来、日蓮の著作とされながらも偽作の疑いが強かった『三大秘法禀承事』の真贋を判定するために、日蓮の著作において特徴が出やすい「所」「これ」「是」などの語の使用率を比較し、クラスタ分析を行うことで、真作である可能性が高いと結論している。この結論をめぐっては論争もあったようだが [6]、いずれにせよ現在では村上氏らの方法が「計量文献学」という名前で定着した感がある。

その後、次に述べる N グラムモデルに基づいた統計分析や後藤義乗氏による研究 [7] など、この方法による研究は着実に実績が積み重ねられていると言っていいだろう。

■ N グラムモデルによる研究

上でも少し述べたが、仏教学における自然言語処理を利用した研究として注目されるのは、N グラムモデルによる仏典の比較研究である。この方法は、前の計量文献学による研究と共通する部分が多いが、本誌を中心として広がった方法であることから、ここでは独立して紹介したいと思う。

N グラムとは、確率・統計的自然言語処理の分野で広く用いられている言語モデルである。N グラムとは「n 個の文字列または単語列」のことであり、テキスト中に出現する文字列の頻度によってテキストの特徴を割り出そうという方法である。例えば、「摩訶般若波羅蜜多心經」という文章を 3 グラムで（3 文字ずつ）分解すると、

摩	訶	般							
	訶	般	若						
		般	若	波					
			若	波	羅				
				波	羅	蜜			
					羅	蜜	多		
						蜜	多	心	
							多	心	經

という８つの文字列がそれぞれ一回ずつあることがわかるが、それはつまり「摩訶般」も「羅蜜多」も同じ確率で発生し、「摩訶經」というような組み合わせの文字列が発生する確率はゼロ、ということを意味する。同様に日本語の文章で「ばらき」という音の列は「茨城」「原木中山」などの地名、「薔薇・木」「バラキ（マフィア映画のタイトル）」など、単語レベルでもそれなりに高い確率で発生するが、順番を並び替えただけの「らきば」という音の列は、「これから木場へ行く」など複合的な例はあるものの、「ばらき」と比べて発生する確率はかなり低いと言える。この確率の差や分布によって、文章や言語の特徴を記述しようというのがNグラムモデルである。すなわち、単語や文字（「アイテム」と総称）の生起が直前のアイテムのみに依存するという一方向的・線的性質なものとして言語をモデル化し、確率・統計的な処理の俎上にのせたものである。Nグラムモデルによるテキストの比較分析は、北研二氏による多言語コーパスの分類[8]をはじめ研究が積み重ねられており、古典文献の分析においても、近藤みゆき・近藤泰弘夫妻による国文学・国語学における応用を皮切りに、仏教学、中国古典などにも広がっている。

先に紹介した計量文献学的方法との違いをあえて述べるならば、計量文献学においては文章の特徴が出やすいと思われる部分（先の日蓮遺文研究では「所」「これ」「是」など）を研究者の側で指定することが多いのに対して、Nグラムモデルによる研究においてはあり得る文字列の組み合わせを網羅的に扱おうとする点が異なる。この点について、Nグラムモデルによる古典研究の先駆者の一人である近藤泰弘氏が、コンピュータを利用したテキスト研究の期待されることとして、次の2点をあげているのが示唆に富む。

1.徹底的に網羅的な研究（すべての単語・すべての文字の単位にまで網羅性を及ぼすことが可能になる）。
2.それによって現代人には通常認知できないデータの構造性や規則性を探り出す。それは、現代人の古典語に対する「内省」（introspection）（語感）の欠如を補うことができ、文学研究に貢献する。なぜなら、古典文学の正しい読みにとって、「内省」（文法的直観と言語外知識など）の欠如は大きな障害のひとつだからである。

これは別の言葉でいえば、計量文献学が仮説検証型の研究方法であるのに対して、Nグラムモデルによる分析は仮説形成型の研究[9]とも言えるだろう。

さて、仏典研究におけるNグラムモデルの応用は、石井公成氏の提唱するNGSM（N-Gram based system for Multiple document comparison and analysis[10]）によってツールや研究方法の開発が促進されたと言ってよい。石井公成氏[11]や角田泰隆氏[12]、道元徹心氏[13]らによって研究が蓄積され、筆者（師）もクラスタ分析やツールの開発などを行った[14]。特に筆者は、先に述べたような仮説形成の面を重視して、すべての玄奘訳仏典（一部外典も含む）のような大規模な文献群に対するクラスタ分析にも取り組んでいる（図2）[15]。

■ 文字知識データベースを用いた分析

沖本克己氏は禅宗の成立に大きな影響を与えた仏典の多くが中国撰述もしくはその疑いがもたれていることから、文字単位（先のNグラムモデルで言えば1グラム）での統計分析による比較・分類を行った[16]。この研究では、表語文字である漢字の特性に鑑み、文字単位での頻度分析から用語の特徴を探ろうとしており、先駆的な研究だと言える。

しかしながら、形・音・義を備える漢字を表現する方法として文字コードはあまりにも貧弱すぎる。したがって、従来の文字コードに依存した確率・統計的分析には様々な問題点や限界が指摘されてきており、その解決のためには余計な手間（例えば、音韻解析のために、ローマ字で発音を表記するテキストデータベースを作成するなど）をする必要があった。

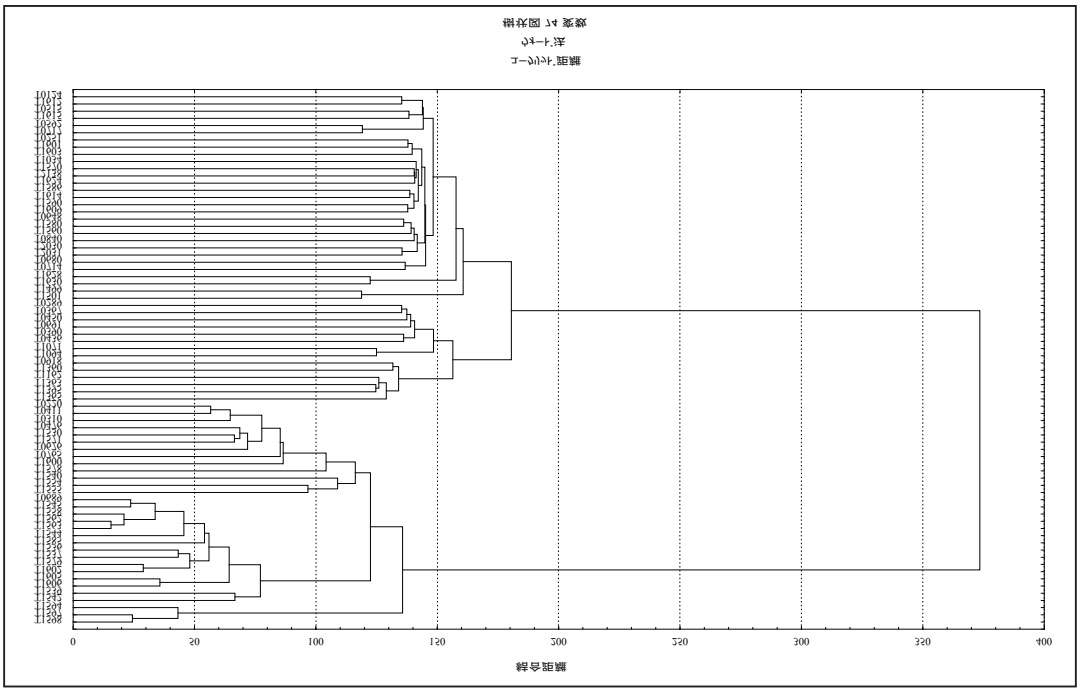


図２　大規模文献群のクラスタ分析による樹状図

筆者もまた、先に紹介したＮグラムモデルによる研究において、文字コードの限界を痛切に感じていたが、守岡知彦氏が中心となって研究・開発が進められている CHISE プロジェクト [17] に参加するようになったことで、文字コードではなく文字知識データベース（文字オントロジ）に基づいた文字処理および文献分析の可能性が開けてきた [18]。これによって、例えば従来認識されてこなかった仏典中の音韻構造など、様々な応用が期待できるのではないかと考えている。

終わりに

以上、仏教学における自然言語処理技術を利用した研究についてごく簡単に紹介した。現在は、知識データベースや確率・統計的モデルによる分析が主流であるが、今後は構文解析や文脈把握、より複雑な確率モデルを用いた研究が必要となってこよう。矢野環氏がバイオ・インフォマティクスで使用されるアルゴリズムを利用した古典文献（写本）の系統分析を行っており注目されるが [19]、そのような新しい方法による分析方法も検討されなければならないだろう。

確率・統計的な分析方法は、適当なモデルとツールがあれば、どのような種類の文献であれ、分け隔てなく適用出来てしまう。従来、学問領域はその方法と密接に結びついていたが、数理的な研究方法はその領域性を解体する傾向があり [20]、文献を仏典などに限定するのは研究史的な要請もしくは研究者の恣意でしかない。コンピュータを積極的に導入しつつも、人文諸科学はその領域を存立しうるような独自の問いを発することができるだろうか。この点についても、今後は議論されなければならないのではないだろうか。

注

[1] このデータベースは当初、**http://buddhist-term.org** で公開されていたが、現在このサイトは閉鎖されている。しかしながら、Richard Mahoney 氏が UTF-8 に変換した上で公開しているものがダウンロードできる（**http://indica-et-buddhica.org/homepages/lexica/**）。

[2] http://www.acmuller.net/ddb/

[3] http://suzuki.ypu.jp/edic/

[4] Kenny, Anthony. *Computation of Style: An Introduction to Statistics for Students of Literature and Humanities.* 1982; 吉岡健一訳『文章の計量　文学研究のための計量文体学入門』（南雲堂、1996）、p. 24。

[5] 藤本熙・村上征勝・伊藤瑞叡・春日正三『統計的決定理論の立場からの文献学的判別問題に対する研究——日蓮の三大秘法禀承事の真偽判別解析——』（文部省科研費一般研究研究報告、1981）、村上征勝・伊藤瑞叡「日蓮遺文の数理研究」（『東洋の思想と宗教』8、1991）、伊藤瑞叡・村上征勝「三大秘法禀承事の計量文献学的新研究」（『大崎学報』148、1992）、村上征勝『真贋の科学　計量文献学入門』（朝倉書店、1994）、村上征勝『文化を計る—文化計量学序説』（朝倉書店、2002）、村上征勝『シェークスピアは誰ですか？——計量文献学の世界』（文芸春秋、2004）等。

[6] 冠賢一「文部省統計数理研究所の「三大秘法禀承事」真作説に対する疑義」（『大崎学報』148、1992）、伊藤瑞叡「三大秘法禀承事の計量文献学的新研究　クラスター分析による真偽判定—本研究に対する批判疑義をも消通する」（『大崎学報』148、1992）

[7] 後藤義乘「計量文献学による漢訳者推定」（『印度哲学仏教学』12、1997）、後藤義乘「計量文献学による漢訳者推定」（『印度学仏教学研究』100、2002）、後藤義乘「計量文献学による漢訳者推定　『無量寿経』の漢訳者」（『印度学仏教学研究』104、2004）等。

[8] 北研二「確率的言語モデルに基づく多言語コーパスからの言語系統樹の再構築」（『自然言語処理』Vol. 4, No.3、1997）

[9] N グラムモデルを用いた文献分析による仮説形成については、師茂樹「大規模仏教文献群に対する確率統計的分析の試み」（『中國宗教文獻研究國際シンポジウム報告書』、2005 年 3 月）参照。

[10] Ishii, Kosei. "NGSM and Cluster Analysis: Its Usage in the Digitization of Variant Texts in the SAT (Taisho Daizokyo Text Database)." *Proceedings of PNC Annual Conference and Joint Meetings 2002.* 2002.

[11] 石井公成「N-gram 利用の可能性　——仏教文献における異本比較と訳者・作者判定——」（『漢字文献情報処理研究』2、2001）、石井公成「仏教学における N-Gram の活用」（東京大学東洋文化研究所附属東洋学研究情報センター編『明日の東洋学』、2002）、Ishii, Kosei. "Classifying the Genealogies of Variant Editions in the Chinese Buddhist Corpus."（『電子佛典』第 3 輯、東國大學校 EBTI、2001）、石井公成「敦煌發現之地論宗諸文獻與電腦自動異本處理」（『戒幢佛學』2, 2003）、石井公成「『大乗起信論』の用語と語法の傾向 —NGSM による比較分析—」（『印度学仏教学研究』52-1、2003）など。

[12] 角田泰隆「異本処理システムによる道元禅師関係文献の書誌学的研究（序）——真字『正法眼蔵』による試み——」（『駒澤短期大学研究紀要』31、2003）

[13] 道元徹心「『観心略要集』撰述者の再検討——N グラムの研究方法を通して」（『行信学報』17、2004）

[14] 師茂樹「XML と NGSM によるテキスト内部の比較分析実験　——『守護国界章』研究の一環として——」（『漢字文献情報処理研究』2、2001）、師茂樹「N グラムモデルとクラスター分析を用いた漢文古典テキストの比較研究——『般若心経』の異訳の比較を例に」（京都大学大型計算機センター第 69 回研究セミナー「東洋学へのコンピュータ利用」予稿集、2002）、師茂樹「N グラムによる比較結果からの用例自動抽出 ——禅宗系の偽経を題材に」（『東洋学へのコンピュータ利用第 14 回研究セミナー』予稿集、2003）、師茂樹「NGSM 結果のばねモデルによる視覚化」（『漢字文献情報処理研究』5、2004）、師茂樹「楞厳経惟慤疏の逸文をめぐる二、三の問題」（『禪學研究』特別号、2005）など。筆者が開発中の N グラムツールである morogram については、http://sourceforge.jp/projects/morogram/ を参照。

[15] 師前掲「大規模仏教文献群に対する確率統計的分析の試み」参照。

[16] 沖本 克己「MENSURA ZOILI　禅文献の計量語彙的研究の試み」（『禅文化研究所紀要』19、1993）。なお、師前掲「N グラムによる比較結果からの用例自動抽出 ——禅宗系の偽経を題材に」では、沖本氏の研究に対して N グラムモデルを用いた検証と再評価を行っている。

[17] http://kanji.zinbun.kyoto-u.ac.jp/projects/chise/。守岡知彦・師茂樹「文字素性に基づく文字処理」（『情報処理学会研究報告』2004-CH-62、2004）等も参照。

[18] 師茂樹「Perl/CHISE による正規表現の拡張の試み ――文字素性による後方参照の実装実験と課題――」（『Linux Conference 抄録集 : 第 1 巻 (2003 年)』、http://lc.linux.or.jp/paper/lc2003/CP-10.pdf）、 師茂樹「N グラムと文字データベースによる漢字仏教文献の分析」（『情報処理学会研究報告』2004-CH-61、2004）

[19] 矢野環「芸道伝書の発展経過の数理文献学的考察 ― Spectronet, Split decomposition ―」（『情報処理学会研究報告』2005-CH-65、2005）など。

[20] 柄谷行人氏は、このような状況を、哲学の問題としてすでに論じている。「フッサールは、哲学者がまだ自分のものだと思いこんでいる領域が錯覚にすぎないことを知っている。解析幾何学をモデルにする諸科学に反撥して、それではとらえられないようなものを見ようとする「文化科学」あるいは「精神科学」なるものが、「あたかも数学的なものの本質は数と量にあるかのように思う一般的先入見にもとづいていることを、彼は知っている。彼のいう「危機」は、自然科学と文化科学、諸学問（科学）と哲学といった区別を無効にするような形式数学を前提としていたがゆえに、そして、それが哲学固有の領域をなくしてしまうことを知っていたが故に、生じたのだ。彼の現象学には、最初から哲学（者）には何が残されているのかという問いが重なっている。（中略）フッサールは、二〇世紀の形式主義が、数学だけでなく、あらゆる領域に浸透せざるをえないことを察知していたといってよい。それは今日ではコンピュータ科学や分子生物学に典型的にあらわれる。つまり、一九世紀の人たちが、最後の牙城として残しておいた、精神、生命、詩といったものにそれが浸透するのである」（柄谷行人『隠喩としての建築』（定本柄谷行人集 2 、岩波書店、2004、pp. 38-39）。

Kiwi: 多言語用例検索システム

中川　裕志（なかがわ　ひろし）

1. はじめに

Kiwi[1] というテキストからの用例検索ツールの名前の由来を説明することから始めたい。古くから KWIC（Key Word In Context）と呼ばれるテキスト解析の方法およびそれを実現するソフトウェアツールがあった。KWIC は与えられたキーワードが対象であるテキストにおいてどのような文脈で出現するかを一覧として表示する。例えば、ここまでの記述をテキストと見做して、「テキスト」という単語で KWIC の表示を行うと次のようになる。

Kiwi という	テキスト	からの用例
と呼ばれる	テキスト	解析の方法
が対象である	テキスト	においてどの
までの記述を	テキスト	と見做して、

「テキスト」という単語の使用例が一目瞭然であり、テキスト解析ツールとしては有用であることが窺がえる。

KWIC では自分のパソコンにテキストコーパスを搭載し、検索する方法で使うことが多い。たしかに特定のテキストの性質を調べるためには、それでもよかった。しかし、世の中一般の言語使用について調べたいとなると、世界中で日々テキストデータが蓄積され続けている World Wide Web（以下では Web と呼ぶ）がなんと言っても魅力的である。そこで、KWIC を Web 対応化するという意味で Keyword In Web という概念を考えることにした。この頭文字をとると KIW となるが、なんとも読みにくい。

KWIC は、あまり大きくないテキストを解析するのであれば、指定した語句の前後に出現する全部の文字列を表示してくれる。また、それがかなり多数であっても、じっくり読み込む楽しみがあるというものである。ところが、Web のテキストデータというのは全く様相を異にする。例えば、前例の「テキスト」を代表的な Web 検索エンジン Google で検索してみると、瞬時に〔0.33 秒〕355 万件の検索結果があるという検索結果が得られる。いくらなんでも 355 万の Web ページを読むことはできない。実際にも、結果として取り出せるのは 1000 ページ程度である。しかし、1000 ページ全てに目を通すことも、時間的には無理であろう。もっと簡明に検索結果の全容やら、傾向やらが分かる賢い（Intelligence のある）方法、すなわち、Keyword In Web Intelligence のような概念は実現できないものだろうか。このように考えてきて、目指すシステムの名前は、Keyword In Web Intelligence の略称 Kiwi[2] と決まった。

Kiwi では、検索の対象が Web 全体だから、特定のテキストを対象にする場合とは結果がだいぶ違ってくる。現状の Kiwi に「花より」というフレーズで質問し、「花より」に後接する表現を求めると、図 1 のようになる。実際の Web 上のデータでは「花よりだんご」ではなくて、「男子」である [3]。これは世相を映していて面白いとも言えるが、Kiwi ではもう少し役に立つこともできるので後に紹介する。なお、現状の Kiwi の URL は脚注 [4] に示す。

以下の節では、Kiwi の仕掛け、使用例などに

ついて説明する。

2. Kiwi の仕掛け

■ 2.1 概要

Kiwi は Web 検索エンジンの検索結果として得られる snippet を統計処理して、1 章に述べた目的を達成しようとするシステムであり、およそ図２のような情報の流れで処理が進む。

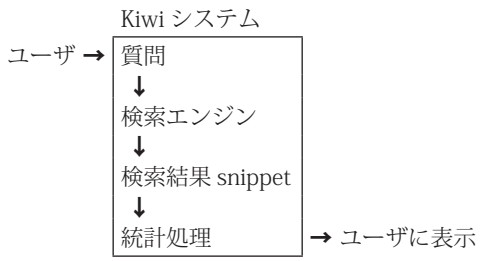


図２　Kiwi の情報の流れ

snippet の一例を図 3 に示す。ある質問文で検索すると、その質問文を含む多数の snippet が得られる。一度に表示されるのは 10 件あるいは 20 件程度だが、ブラウザ画面において良く見かける「次の 10 件」あるいは「次の 20 件」というボタンを繰り返してクリックすれば、表示を繰り返せば、1000 件程度まで snippet を得ることができる。Kiwi では、この最大 1000 件程度の snippet を使う。逆に言えば、Kiwi では、この 1000 件の Web ページ本体から見ればごくわずかな部分である snippet しか使っていない。それでもかなり面白い結果が得られる。テキストを隅々まで解析しようというのであれば、いちいち

図３　snippet の例

Hiroshi Nakagawa
中川裕志．My English page is here. 研究活動 . 研究テーマ ; 発表文献、等；中川研究室へのリンク；我々が開発した主要な Web 上のシステム．Web ページからの用語抽出システム（日英中仏独伊西瑞典語に対応）：言選 Web；多言語用例指南ツール：Kiwi　…

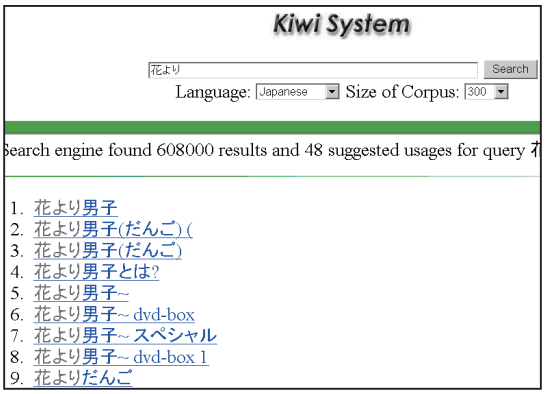


図１　Kiwi の検索例：「花より」に後接する表現

元の Web ページを取りにいく必要がある。それは理想的かもしれないが、1000 件の Web ページあるいは 100 のページでも実際にダウンロードすると時間がかかりすぎる。snippet を使うというアイデアによって、質を若干犠牲にはしたが速度を早めることに成功した。

sinppet が得られると、次は役に立つ部分を取り出して表示するための統計処理である。Kiwi 独自のアイデアは、この統計処理の部分にあるので、次節以下で少し詳しく説明する。

■ 2.2 適切な長さの文字列の切り出し

snippet は質問文の周辺を取り出しているとはいえ、それを全部 KWIC のような形式で表示するのは、Kiwi の標榜する Intelligence ある方法ではないし、そもそも 1000 件の KWIC 表示を読むのは大変である。そこで、まず適当な長さの文字列にして切り出す作業をする。適当な長さだからといって、10 文字とか 20 文字とかいうように固定した長さにしてしまうと、重要な表現あるいは単語が途中で切れたりしかねない。また、同じ表現が繰り返し表示されると読むための時間がかかる。そうなると、Kiwi の出力結果を見ただけでは、質の良い情報が得られない。統計的に有意な頻出する表現を取り出して表示したい。「適当な長さの文字列」とは、意味的にまとまりが良く、頻出する文字列ということになる。

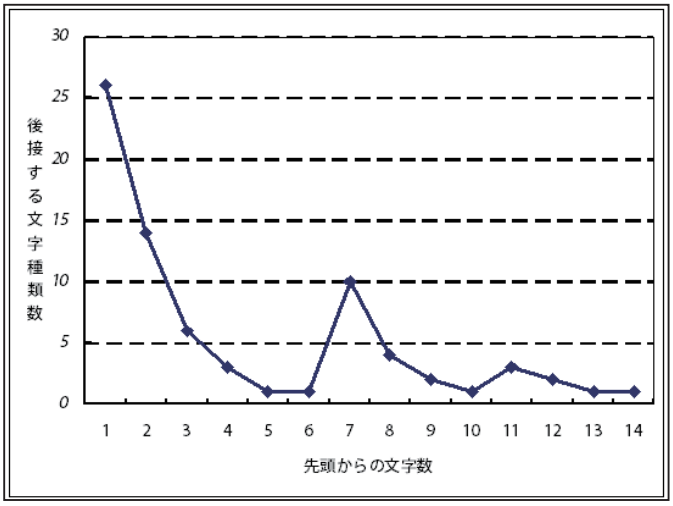
この問題に対処するために次のような観察を行ってみる。多数の言語には単語という概念があ

り、意味のまとまりを表す。単語は文字の連接である。文字の種類数は単語の種類数よりも多い。例えば、英語では大文字小文字を区別し、数字や記号を加えても 100 種類程度の文字である。一方、単語は一般的な語彙でも数十万（10^5）、専門用語を入れれば 100 万は下らないだろう。すると、当然複数の文字を連接して単語を形成する。ところで、仮に文字種類数を 100 とすると、n 文字の単語は、10^{2n} 種類の可能性がある。例えば 5 文字の単語の種類は、10^{10} である。だから、実際の単語は可能な文字列のうちごく少数を使っていることが分かる。そこで、英語の単語を先頭から 1 文字ずつ見ていくとしよう。1 文字目が決まっても、英語なら 26 種類のうちの 1 個に絞ったというだけなので、2 文字目に来る文字種類はアルファベット 26 文字全部とは言わないまでも相当多い。ところが 2 文字まで決まると、26 の 2 乗（すなわち 676）通り 1 種類が選ばれたので、数十万の語彙のうち可能なものは数千程度に限定されてきている。よって次の 3 文字目の文字種類数はだいぶ限定されてくる。4 文字目まで進むと、原理的には 26 の 3 乗（すなわち 17576）のうちの 1 個が選ばれたので、4 文字目の文字種類数はいっそう限定されてくる。このように考えれば、先頭から進むにつれて後接する文字種類数は減少する傾向は、一般的に成り立つといえよう。

実際、日本語のカタカナ単語を調べると、先頭から文字が進むにつれて後接する文字種類数は単調に減衰していく。概念的には図 4 のようになる。

先頭から進むにつれて上に述べた理由で後接する文字種類数は単調に減衰していくのだが、単語が終了すると（英語なら空白が入ると）、次には多種類の単語が後続して現れる可能性があるので、図の横軸 7 の場所のおけるように急激に増加する。その後はまた単語の内部に入るので、後接する文字種類数は単調に減衰していく。このようなことを繰り返すわけだが、単語という意味表現の単位を持つ言語であれば、この傾向は現れると考えてよい。また、単語の区切りという形式的に明確な境界でなくても、ある種に固定的な言い回しの内部では、同じように後接する文字種類数は単調に減衰する傾向があることは予想されるところである。そして、その固定的表現が終われば、後接する文字種類数は増加する。このような考察から、意味的にまとまりの良い文字列を切り出すには、後接する文字種類数を調べ、それが減少から増加に転ずる直前までで切り出せば良い候補が得られることが予想できる。以上の説明からお分かりいただけるように、この方法は主に言語が有限個のアルファベットからなる文字の列であることを利用しているだけなので、言語を問わず有効である。

図 4　先頭からの文字数 vs 後接可能文字種類数



上記のアイデアを具体的に実現する方法は、以下の各 step に沿って行う。

step1: Web 検索エンジンに利用者の入力した質問の文字列を送り、検索結果の snippet を得る。得られる snippet の数は 300 から 1000 の間で利用者が指定するようにした。なお、現在は検索エンジンとして AltaVista[5] を用いている。

step2: 各 snippet のテキスト

を質問文字列に合致した直後から走査し、Trie 構造化する。Trie 構造は上記の後接する文字種類数を計算するのに都合の良いデータ構造であり、後で説明する。

step3: Trie 構造を先頭から調べ、後接する文字種類数が増加に転ずるところまでの文字列を用例候補として切り出す。

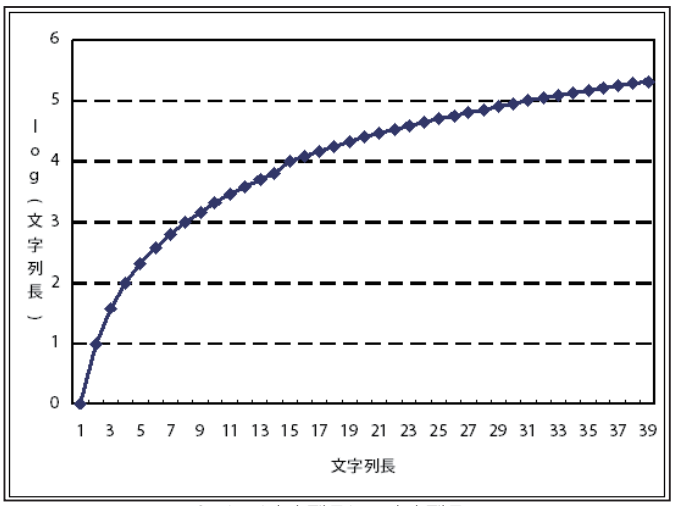


図 6　log(文字列長) vs 文字列長

さて、Trie 構造およびその作成方法について例を用いて説明する。仮に「犬も」という質問に対して、以下の(a) – (d)の 4 本の文字列が検索結果の snippet 群に含まれていたとしよう。また、1 文字目は、「歩」のほかに 4 種類の文字から始まる文字列があったとする。

(a)　歩けば棒にあたる
(b)　歩けばなんとやら
(c)　歩くとなんだっけ
(d)　歩くと飼い主も歩く

すると、図 5 のような木構造でこの sinppet を表現できる。

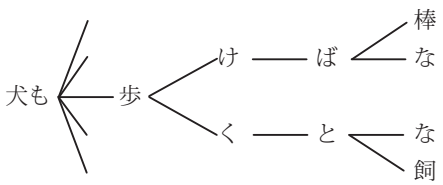


図 5　Trie 構造の例

このような木構造を Trie と呼ぶ。この構造を作るモジュールそのものが備わっているプログラム言語もあるので、snippet の集合が準備できれば、Trie は比較的容易に自動構築できる。Trie が構築できれば、後接する文字種類数は簡単に知ることができるので、それが増加に転ずるところまで文字列を切り出すことも実現できる。図 5 の例だと、「歩けば」と「歩くと」が切り出される。

■ 2.3 順位付け

2.2 節に述べた方法でまとまった意味の文字列が切り出せたのだが、snippet の数が多いので、多数の文字列が切り出される。それらを無作為に表示するだけでは、知的でない。良いものから順に表示したいところである。そこで、良く使われるものは重要であること、また長い表現のほうが利用者には理解しやすい、という 2 点を考慮し、実験的に調べた結果、候補文字列（ここでは S と書くことにする。）を次の式 K(S) の値の大きい順に表示することにした。

結果の snippet 群における出現頻度を freq(S)、文字列 S の長さ（文字数）を length(S)、としたとき

$$K(S)=freq(S) \times \log(length(S)) \quad (1)$$

実験的に式 (1) を決めたので、後知恵になってしまうが、この式について少し考察しておきたい。2.2 節で説明したように、アルファベットのような文字セットが決まったとき、その文字セッ

トから生成されうる文字列は文字列長に対して指数関数的に増大する。しかし、元々の snippet 数は一定だから、枝分かれのたびに出現頻度は何分の 1 かになる。よって出現頻度は指数関数的に減少する。つまり、長さ n の文字列が与えられると、それに後接する文字列の数は n に対して指数関数的に減少する。単語や固定的な言い回しが終了したところで増加に転ずるもの、その絶対数は減少している。一方、図 6 は文字列長の log である。これは、増加するものの、増加率は逓減していく。

そこで、式 (1) のように出現頻度と log(文字列長) を掛け合わせた結果はおよそ次のような傾向をもつであろう。すなわち、出現頻度が指数関数的に減少するなら、それが支配的になり、全体としては現象傾向だが、log(文字列長) が増加関数なので、減少の度合いが緩和される。このような傾向の式 (1) において、出現頻度が指数関数的な減少よりは減少の度合いが少ない文字列があれば、式 (1) の値は大きくなる。つまり一般的傾向としては、(a) 短くて頻度が高い文字列でも log(文字列長) のせいで値が抑えられ、(b) 長くても絶対的頻度は小さくとも、ある程度の頻度がある文字列は log(文字列長) のおかげでかなり大きな値を持つ。この性質からみて、式 (1) による順位付けは、質問文に後続する意味的にまとまりのよく、頻繁にかつ安定して使われる文字列であり、適度な長さのもの、すなわち用例というべき文字列を取り出すのに適したものといえよう。

■ 2.4 ワイルドカードの使い方

これまで説明してきたのは、質問文字列に後続する文字列を検索して統計処理して表示するものであった（前方一致検索）。しかし、質問文字列の前方に連接する文字列も検索したい（後方一致検索）。さらには、二つの質問文字列で挟まれた部分を検索したい（前後一致、中間検索）、という要求もある。

後方一致検索は、2.3 節までで説明してきた方法を、前後逆転して適用すれば良い。統語構造を利用しているわけではないので、文字列の構造、すなわち文字の連鎖の仕方の統計的性質は前後反転しても、かなりの程度に成り立つから、前後反転しての適用でもうまく動作するであろう。

前後一致の中間検索は、若干異なる。まず、前方、後方の質問文字列を検索エンジンに投入し AND 検索を行う。その結果に対して、前方の質問文字列と後方の質問文字列がこの順序で出現する snippet を集めて、前方質問文字列と後方質問文字列の中間に現れる文字列を頻度の大きい順に表示している。

Kiwi では、これらの検索は、ワイルドカード「*」を使って、次のように与えることになっている。ただし、abc, xyz は質問文字列を表すとする。

前方一致検索 :	abc*
後方一致検索 :	*xyz
前後一致の中間検索 :	abc*xyz

このように Kiwi は必要最低限の質問方法を備えているが、より高度な質問をする場合には、ワイルカードが 2 箇所指定できる方法も検討しなければならない。（藤本 2005）

▌3. Kiwi の使い方と評価

■ 3.1 多言語の用例検索

Kiwi の用例文字列切り出しは、言語に依存しないので、基本的にはいかなる言語にも対応する[6]。そのような応用例として Kiwi は英語で論文を書くとき、自信のない用例を確認する、あるいはある単語や言い回しの部分を与えて、適切な用例表現を効率良く探し出す場合がある。いくつか例を示そう。

例 1 ：keep と touch の間に入る前置詞を知りたい。

Kiwi で検索すると以下のような結果となる。正解である in のほかに 'n という省略表現も見出せるし、ハイフンでつないで 1 語のようにする用例もかなり使われていることが分かった。

文法的というよりは、むしろ現実の言語使用における例を調べたい場合の例を示そう。「動詞 it seriously」という表現において良く用いられる動詞を調べるために「* it seriously」という検索を行った結果を図 8 に示す。

面白いことに上位は take という動詞で独占される。つまり、少なくとも Web 上では、take it seriously という表現はほとんど熟語と言えるほどに安定して使用されていることが分かる。

中文の例を示そう。アメリカについてどのような記述が多いかを調べるために「美国 *」（美国とは中文でアメリカ合衆国のこと）という検索をした結果を図 9 に示す。

留学が第 1 位にくるのは、中国人にとってアメリカ留学が人気の高いことの表れであろうか。

次に少し文法的用例を検索してみたい。「很喜欢看」すなわち「ＸＸを見るのが好き」という動詞句の直前にはどのような表現が使われるかを調べてみた。結果が図 10 である。

いきなり人称代詞が使えること（1 位、2 位、7 位の例）、「不是」で否定できること（3 位の例）、程度を表すためには「都」「真的」「一直」などが使えること（4 位、8 位、10 位）、「以前」「也」の使用できうる位置（5 位、6 位、9 位）、などいろいろな文法的知識が得られた。これらの例は丁寧に Web 検索エンジンの結果を読めば分かることだろうが、一瞬にして、これらの頻出表現が得られるのは Kiwi の威力であろう。また、金庸 [7] の小説は現在でも人気があることが分かって興味深い。一方、中国語圏の人たちは何を見たいのか、という問いかけに答えるために「很喜欢看 *」という質問をしてみた結果を図 11 に示す。

2 番目の答えは率直ながら面白い。映画（5 位、8 位）や動画（4 位。たぶんアニメのことであろうか。）のほか、小説（3 位）、しかも武侠小説というジャンルに人気があることは面白い発見である（9 位）。12 位は球技の試合はやはりどこの国でも人気があるということか。11 位は意味深な感じがする。こういうときには原文にあたれるとありがたい。このために Kiwi は、この結果を行をクリックすると原文を表示できる。その結果、

1. keep in touch
2. keep-in-touch
3. keeping in touch
4. keep'n touch
5. keep 'n touch
6. keep_in_touch
7. keep you connecte
8. keeping-in-touch
9. keep 'n' touch
10. keep'n'touch
11. keep.touch

図 7　keep*touch の検索結果　上位 11 位

1. take it seriously
2. don't take it seriously
3. taking it seriously
4. to take it seriously
5. if you don't take it seriously
6. please don't take it seriously
7. ever take it seriously
8. never take it seriously
9. took it seriously
10. takes it seriously
11. not take it seriously
12. we take it seriously

図 8　* it seriously の検索結果　上位 12 位

図 9　美国 *　による検索結果　上位 10 件

1. 美国留学
2. 美国公司注册
3. 美国公司注册网注册美国公司,
4. 美国总统大选
5. 美国总统
6. 美国华人
7. 美国概况 news.sohu.com 200
8. 美国留学,美国签证,学分评估,留学课堂,
9. 美国概况
10. 美国,出国留学 美国,美国留学费用

図 10　* 很喜欢看　の検索結果　上位 10 位

1. 我很喜欢看
2. .* 我很喜欢看
3. 不是很喜欢看
4. ... 金庸:不错,侦探小说我一向都很喜欢看
5. ... mimi:其实看小说没有太多的时间,不过我以前很喜欢看
6. ... 很巧唉,我也叫小玉, 而且我以前的男朋友也很喜欢看
7. ... 个深深的酒窝,甜甜地,似乎溢满着蜜。他很喜欢看
8. 我真的很喜欢看
9. 我也很喜欢看
10. 。我一直很喜欢看

図 11 很喜欢看＊　の検索結果　上位 12 位

以下のような例からこの結果が得られたことが分かった。

很喜欢看别人的 msn spaces
我很喜欢看别人写的。
很喜欢看别人的照片。
我很喜欢看别人的文字

つまり「别人的ＸＸ」という形の修飾句がよく使われた結果であった。詳しく調べると面白い例があるようだが、筆者の中国語能力では説明できそうもないので割愛する。興味がある方は試してみていただきたい。

図 12「日本で二番目に高い山は」に対する Kiwi の検索結果

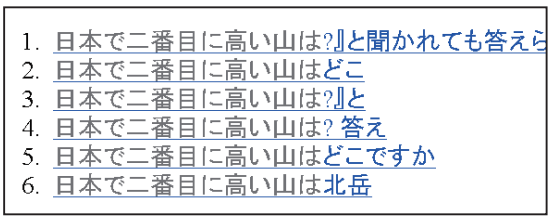
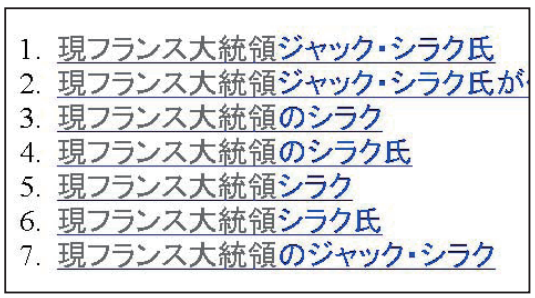


図 13「現フランス大統領」の検索結果



中文を対象に使い込んだわけではないが、文法的知識から中文における広い意味の語法、中国における社会常識まで垣間見ることができるので、Kiwi は単なる用例検索を超えて面白い情報を提供してくれる。

■ 3.2 More than search engine, Less than QA

文献（*Tanaka Nakagawa 2005*）では、現状の Kiwi を “More than search engine, Less than QA” と位置づけている。第 1 章で述べたように、Kiwi は単語やフレーズを質問として与えて検索する Google などの検索エンジンよりは、検索された内容そのもののテキストとしての特徴をうまく表示してくれる。また、Web 全体を対象データとする検索エンジンの snippet を用いているから、図 1 に例を示したように、結果は Web の現状を反映したものになっている。だから、“More than search engine” と喧伝しても恥かしくはない。

さて、情報検索や自然言語処理の分野でさかんに研究されている質問応答（Question Answering: 略して QA）という技術がある。QA は、テキストコーパス中における知識を問うものである。例えば、「日本で二番目に高い山は？」（百科事典的知識）とか「フランスの歴代大統領は？」（歴史的事柄）、さらには「最近、社長が交代した企業名と新旧社長の名前を知りたい」（現代社会における出来事）というような質問の答えをテキストコーパスから検索する技術である。QA をテキストコーパスではなく、Web に対して行うことを WebQA というが、なかなか実現の困難な技術である。Kiwi でこのような QA の問いに答えられるかというのは興味ある問題である。

では、実際に上記の質問を Kiwi にしてみよう。

結果を図 12 に示す。

6 位に正解が現れている。検索エンジンを直接使うよりは楽だが、高機能の QA システムなら、この答えを 1 位に欲しいところである。一方、「フランスの歴代大統領は？」という質問には答えが得られなかった。つまり、Web ページとしてはこの質問文を含むものがないということである。

このような問題は工夫が必要で、まず「現フランス大統領 *」を質問し、図 13 のように答えを得る。

次に「前フランス大統領 *」を質問したが、これには結果がなかった。しかし、「* 前フランス大統領」としてミッテランという正解を得た。歴代となるとさらにやっかいである。例えば、単に「フランス大統領 *」「* フランス大統領」で質問し、出てきた名前（仮に X とする）から「X 大統領」という質問をして確認していくようなことになろう。実際にはこの質問で、ドゴール・ジスカールデスタン・ポンピドウ・ミッテラン・シラクが得られた。このように、Kiwi は QA システムとして使うにはあまりに原始的なツールである。よって、“Less than QA” という位置づけになる。

4. おわりに

用例検索システム Kiwi について、その背景、仕掛け、使用例について述べてきた。Kiwi は文字列を対象にしているだけで、辞書などの言語固有の資源は使わないので、いかなる言語にも直接適用できる。しかし、その一方で限界があることも分かり、現状では 3.2 節で述べた “More than search engine, Less than QA” というのが的を射た位置づけであろう。実際に Kiwi を使っていると、検索エンジンとのやり取りに時間がかかり、そのほかにも問題が多い。ひとつの方向としては、テキストコーパスを手元に持ってしまうことが考えられるが、そうなると Kiwi とは何か、とう本質を考え直す必要があるようである。

謝辞

Kiwi の開発において、Web を使うという基本アイデアを提案され、その後の開発にもご尽力された共同研究者の田中久美子助教授（現東京大学大学院情報理工学系研究科創造情報学専攻。開発当事は東京大学情報基盤センター）、および Kiwi システムの開発に努力された中川研究室の学生である山本真人君、河内崇君、藤本宏涼君に深謝いたします。本研究は文科省科学研究費補助金　特定領域研究「情報学」課題番号 16016215、および科学技術振興機構 CREST「情報のモビリティを高めるための基盤技術」の補助を受けた。

参考文献

- Kumiko Tanaka-Ishii, Hiroshi Nakagawa, "A Multilingual Usage Consultation Tool based on Internet Searching ——More than search engine, Less than QA", *The 14th International World Wide Web Conference*（*WWW2005*）pp.363-371. 2005 May, Chiba, Japan
- 藤本宏涼、吉田稔 , 中川裕志 , “ローカルコーパスからのテキストマイニングツール : PortableKiwi” , 言語処理学会第 11 回年次大会 C1-8, 2005

注

[1] Kiwi は「キウイ」と読む。

[2] KiwI なら発音も「キウイ」としやすいし、覚えやすい。

[3] これを「だんご」と読むらしい。『花より男子』。神尾葉子作。全 36 巻（1992 ～ 2004）。集英社マーガレットコミックスからの引用だろう。

[4] **http://kiwi.r.dl.itc.u-tokyo.ac.jp/**
予告なく変更することもあるので、ご了承ください。

[5] AltaVista 以外の検索エンジンも考えたが、例えば Google だと安定して動くのは API だが、1 日 1000 件までの検索しか許さないという制限により使えなかった。また、Web の情報を利用する研究では AltaVista がよく使われている。

[6] AltaVista は現在のところ、検索対象の言語指定が直接できない仕様になっている。したがって、実際のところは、質問文で使用した文字コードによって検索対象の範囲が規定されてくるという、あなた任せな方法であり、問題となっている。

[7] 金庸は有名な武侠小説家。筆者はＮＨＫのラジオ中国語講座でさわりを聞いた。

キーワード自動抽出システム「言選 web」

前田　朗（まえだ　あきら）

はじめに ～「言選 Web」へようこそ～

文章中の重要なキーワードをあらかじめ示してくれれば、概要をすぐにつかめるのにと思ったことはないであろうか。また、複数の文章を情報工学的に比較したいと思ったことは？「言選 Web」は文章からその概要をつかめるキーワードを取り出す Web 上のサービスである。これは、日本語のみならず、中文・西ヨーロッパの各言語（英語など）にも対応している。

「言選 Web」自体は Web アプリケーションであるが、そのエンジン部分や、Windows アプリケーションをフリーソフトとして公開・配布している。本フリーソフトは Perl モジュール “TermExtract”、Windows 用システム “termex”、テキストマイニングツール “termmi” からなる、まさに「言選 Web」ファミリーともいうべき一群のソフトウェアである。

本稿では「言選 Web」をはじめとするこれらの専門用語自動抽出システムについて、基礎理論から活用法までを紹介する。

第一章と第二章は、基礎理論である。第三章と第四章では、専門用語自動抽出システムを実際に使いこなすための情報を提供する。もし、「言選 Web」の活用法だけを知りたければ、第三章からお読みいただくこともできる。

本稿の読後はこの「言選 Web」（**http://gensen.dl.itc.u-tokyo.ac.jp/**）にアクセスし、その有効性をぜひ確かめていただきたい。

1. 用語抽出手法あれこれ

「言選 Web」では、⑴文章から用語を抽出し、⑵重要性の高い順に並べかえる、という２ステップの処理を行っている。ここでは、最初のステップである用語抽出をみていくことにしよう。

「言選 Web」における用語の抽出手法は大きく２種類にわかれる。ひとつは文章をいったん形態素解析により単語まで分割して、その上で断片をつなぐように用語を組み立てる手法。もうひとつは用語の要素になりえない単語（もしくは文字）を消去し、残ったものを取り出す手法である。

■ 形態素解析ソフトと用語の「まとめあげ」

まずは、単語から用語にまとめあげる方法を説明しよう。日本語では文章を単語に分割するソフトとして、茶筅、和布蕪といった形態素解析ソフトがある。形態素解析ソフトは、テキストを形態素（意味を持つ語の最小単位、漢字の場合は一文字にほぼ相当）に分割するソフトというより、形態素の概念を利用して単語分割と品詞タグ付けを行うソフトと理解していただきたい。なお、わかち書きパッチ」を当てた案山子は文章の単語分割を高速に行えるが、品詞情報が付与されない。

試しに「漢字文献情報処理研究」という語を形態素解析ソフトにかけてみよう。和布蕪では次の

４つの単語に分解される。

漢字（名詞，一般）、文献（名詞，一般）、情報処理 （名詞，一般）、研究（名詞，サ変接続）

用語として「漢字文献情報処理研究」が一語で出て欲しいのに、これでは用語の単位として小さすぎる。形態素解析ソフトは、複合語に対応したものではないといえる。

そこで「言選 Web」では個々の単語を用語レベルにまとめあげる。その基本ルールは、名詞となる最小の単位（単名詞）が連続した場合に、それらをまとめて複合名詞とみなすことである。

上記の例では、どの単語も全て名詞である。よって全ての単語を連結でき、「漢字文献情報処理研究」と、一語にまとめあげることができる。

英文の場合は、既に単語ごとに分かち書きされているので、品詞のタグ付けがされていればよい。この品詞のタグ付けを行うのが、POS Tagger という種類のソフトである。英語の場合は、Brill's Tagger という POS Tagger をフリーで入手できる。まとめあげのルールは日本語と比べて複雑になるが、基本的な考えは同じである。

■ 中文の形態素解析には“ICTCLAS“を使う

中文では、中国科学技術院の Windows ソフト “ICTCLAS” で、単語分割と品詞タグ付けを行える。「计算所汉语词法分析系统」とあるが、詞法とは形態論、形態素の意味であるので、これは中文の形態素解析ソフトといった意味になる。なお、Web 版もあり、以下の URL にて試すことができる。

http://mtgroup.ict.ac.cn/~zhp/ICTCLAS.htm

実際に、「ICTCLAS 的介绍及说明」を処理した結果は次のとおり（/nx などは品詞情報）である。

ICTCLAS/nx 的/u 介绍/vn 及/c 说明/v

さて、中文の場合は各単語を次のとおりまとめあげる。例外の少ないシンプルなルールであるが、人民日報の記事で試したところ、人手で指定した重要語のうち 5 割強を取り出すことができた。

- ●名詞に類する語（ng, n, nr, ns, nt, nz, nx, vn, an, i, j）* 以後「名詞」
  →名詞・形容詞・助詞・後接成分・連詞（和・与）に結合する。複合語の先頭と末尾になる。
- ●形容詞（ag, a）
  →形容詞・助詞・後接成分・連詞（和・与）に結合する。複合語の先頭になる
- ●助詞（u）・後接成分（k）
  →名詞・形容詞に結合する
- ●連詞（c）
  →和・与の場合のみ。名詞に結合する。
- ●区別詞（b）
  →名詞・助詞・連詞（和・与）に結合する。複合語の先頭になる

このルールで「ICTCLAS/nx 的/u 介绍/vn 及/c 说明/v」を試していただきたい。「ICTCLAS 的介绍」と用語抽出されるはずである。

さて、上記のルールをみて分かるとおり、動詞となる単語は基本的に複合語から除外している。例外は、名動詞（名詞的にも使われる動詞）で、これだけは名詞扱いにしている。このことは中文の用語抽出にどのように影響するのであろうか。

中国語は品詞種別や活用、時制が見た目には区別できないことが多い。実際の文中では［述語］［目的語］という並びでも、抽出されたものが十分に用語として通用する。つまり英文の場合” write a letter” という語が用語としては不適格で” writing a letter” となれば用語として適格となるが、中文ではどちらも “写信” ということである。文章の内容が手紙を書くことに関してであれば文章中の品詞に左右されず、“写信” という用語はキーワードとして適格といえる [1]

しかし ITCCLAS では、“写” は動詞、“信” は名詞と判定されるので、“写信” を用語抽出するこ

とができない。なお、後述するストップワード方式では、この弊害は起こりにくい。

■ カタカナと漢字熟語のみ抽出すると

ここからは用語の要素になりえない単語（もしくは文字）を消去し用語を取り出す方法を示そう。

まず、日本語の場合はカタカナと漢字熟語の並びを抽出するだけで用語候補になる。別の見方をすれば、用語は主に名詞句であることから、名詞句候補を抽出したとも考えられる。

たしかに、カタカナは外来の名詞に使われることが多い。また、漢字熟語は文章中の使われ方によらず、熟語部分だけ抽出すると名詞としても扱えてしまうことが多い。次の例1を見てみよう。

特に動詞「抽出する」が「抽出」の部分だけみると名詞扱いできることを確認いただきたい。

『日本語の場合、カタカナと漢字熟語を抽出するだけでキーワード候補になり得る。』
|
（カタカナと漢字熟語のみ抽出）
↓
『日本語　場合　カタカナ　漢字熟語　抽出　キーワード候補』

例1　カタカナ・漢字に着目した用語抽出

上記の例では「場合」を除き用語を抽出できていると思う。

■ 文章を特定の語で分けてみる

情報検索システムでは、キーワードとして登録しない語を、事前にストップワードとしてシステム設定する。もし、ストップワードを文章からはずしていけば、残った語はキーワードである可能性が高いといえないであろうか。もちろん、キーワード以外の用語も多いが、それは後述の重要度順の用語並べかえにより、判別できる。

例えば、“I have a chinese journal”という英文であるが、情報検索システムで使われるストップワードを除いていくと、“chinese journal”だけが残る。シンプルだが有効に働くことがわかるかと思う。

中文に関しては、「言選Web」オリジナルのストップワードリストを作成した。用語辞書から「動詞」となりえる単語を全て抜き出し、それらをストップワードにするなどの方法も試したが、前述の“写信”の例でも示したとおり動詞を除くとよい結果が出ない。用語辞書から品詞で選定するより、人手により1語づつ選定したストップワードのほうが効果的であった。

中文では、文章中の品詞の用語抽出に与える影響が、日本語や西欧言語に比べて弱い。逆に言えばストップワードによる用語抽出の特性を生かせる言語であるといえる。

2. 大事な用語から並べてみよう

■ オーソドックスなTF-IDF法

重要な用語をランク付けするオーソドックスな手法が、TF-IDF法である。TFはTerm Frequency、IDFはInverted Document Frequencyの略である。これは、ある文献中に多く出てくる用語は重要だが、一般的な用語は除外するという考えによる。文献中に多く出てくる語は出現回数を数えれば求められるが、一般的な用語を低くランクづけるのはどのように行っているのであろうか。

たとえば用語「漢字情報」が全8文献中、3文献に使われていたとする。この情報を元に確率の考えを使えば、「漢字情報」がどのくらい一般的な用語かを示せる。前述の例の場合は、3/8である。この確率が高い、つまり大きい値ほど一般的な用語であり重要性が低いと判断できる。しかし、用語の出現頻度では大きい値ほど重要性が高いため、このままでは両者を掛け合わせることができない。

そこで、確率の逆数を用いる。先の例では確率が3/8だが、これを反転させて8/3にするということである。しかし、このままでも十分ではなく、100万件中1件も、100万件中2件も、一般的ではない用語として大差がない。しかし、重要度の数値としては2倍になってしまう。その調整として対数（log）の計算を行う。また、全

ての文献に含まれる用語は、対数の計算を行うと0になる（log1=0）ため、その調整のために1を加える。

最後に式にまとめると以下のとおりである。なお、後述する termmi では TF-IDF 法もサポートしている。

TF-IDF の重要度 =
用語の出現頻度×（log（総文献数 / 該当の用語を含む文献数）＋ 1）

■ FLR が「言選 Web」の基本

次に「言選 Web」がメインで用いている FLR を紹介する [2]。用語を名詞として扱える語の最小単位に分割したものを単名詞と呼ぶ。つまり用語は単名詞そのものか単名詞の組み合わせで構成されるといえる。この理論では、他の単名詞と連接して複合名詞をなすことが多い単名詞ほど、文書中で重要な概念を示すと考える。

簡単な例で、「漢字文献情報処理」を考える。この語を単名詞に分割する。そして分割した単名詞が他とどれだけ連接したか文章中の用語から統計をとり、次の表1のとおりわかったものとする。

単語	前の語に連接	後の語に連接
漢字	2	3
文献	3	4
情報	4	5
処理	2	0
研究	0	3

表1　「漢字文献情報処理」の単名詞連接

用語の重要度はこれらの10（単名詞 x 2）の数値の平均から求める。用語の重要度には、相乗平均が相和平均より効果的なため、「言選 Web」では相乗平均を用いている。なお、連接0回の単名詞を含む語は相乗平均が0になってしまうため、実際には各連接回数に1を加えた値を用いている。

こうして得られた単名詞の連接情報による重要度に、用語の出現頻度をかけたものが「言選 Web」の重要度である。出現頻度（Frequency）に左（Left）と右（Right）の語の連接情報を組み合わせて使うため、これを FLR と呼ぶ。

■ 中文は「文字」も使える

漢字は一字一字が意味を持つ「表意文字」として使われることが多い。FLR は他の単名詞と連接する単名詞ほど重要度を高くする。これは他の単語と結びついて複合語をなすような単語ほど、まさに文中で核となる意味を表す単語であると考えられるからである。一字一字が意味を持つ漢字であれば、FLR の「単名詞の連接」を「文字の連接」に替えるという選択肢もありえる。

例えば、「漢字文献」を次の表2のように重要度計算できるのではないであろうか。（数値は仮の値）

文字	前の字に連接	後の字に連接
漢	6	7
字	3	6
文	2	5
献	4	3

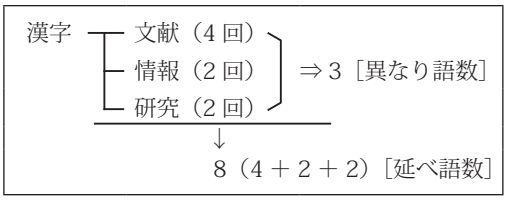
表2　「漢字文献情報処理」の文字連接

中文版「言選 Web」では、文字と単語、それぞれに着目した重要度計算のモードを用意している（停止語方式版と ICTCLAS 版）。両者がどのような傾向を持つかは今後の研究課題である。

■「多様性」（パープレキシティ）でランクづける

単名詞が他の単名詞に連接した回数のカウントにはいくつかの方法がある。直感的にわかりやすいところでは連接の延べ数と種類数である。これに換えて情報理論的な回数、すなわちパープレキシティを使うのが、東京大学情報基盤センター中川研究室における最近の研究理論である。

例えば、次の「漢字」の連接の例（例2）で考えてみる。出現回数（延べ語数）なら8回（4＋2＋2）、種類数（異なり語数）なら3回（「文献」



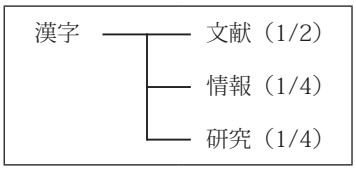
例2　異なり語数と述べ語数による連接回数

「情報」「研究」の3種）である。

この回数のカウントにパープレキシティを使うと、たとえ連結する語の種類数が多くとも、特定の語とばかり結びつけば、カウントが少なくなる。逆に語の連接が多くの語に分散していれば、カウントが多くなる性質を持つ。

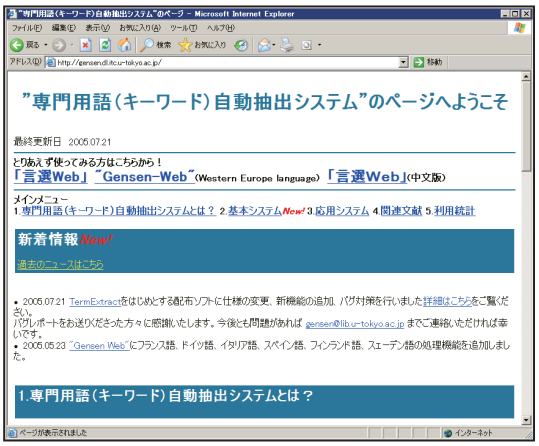
パープレキシティが初耳でもエントロピーならご存知のかたも多いかと思う。情報理論でいうエントロピーは、「情報を平均して何バイトで示せるか」、つまり情報の多様性を示す指標である。パープレキシティはエントロピーを2のべき乗した数値であり、同じく情報の多様性を示している。

パープレキシティは確率の考え方を使う。例えば先のケースを割合で示してみると次（例3）になる。

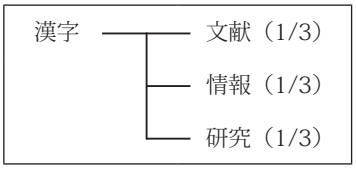


例3　連接の確率1

図1　専門用語自動抽出のページ



この確率の分布が平均しているほど、パープレキシティが増大する。例えば、以下の例4は例3よりも多様性があるということである。



例4　連接の確率2（パープレキシティ最大）

■ 文章をどんどん「学習」させる

FLRは、単名詞同士の連接についての統計情報を必要とする。この統計情報が正しければ正しいほど、計算の精度は上がるといえる。

文章を新規に読み込むたびに、この統計情報を蓄積し、次回以降のFLRの計算に生かす機能を与えれば、統計情報ひいてはFLRの精度が良くなっていくのではなかろうか。その考えから「言選Web」のエンジン部分、"TermExtract"では学習機能をオプションとして実装している。

ただし、雑多な分野の文章を学習させると、あまりに一般的な用語の重要度が高くなるという弊害もある。あくまで特定の分野に限って「学習」させていただきたい。なお、ユーザが不特定多数の「言選Web」では採用を見送っている。

3. "言選Web"を使ってみよう

■ まずは"専門用語自動抽出のページ"にアクセス

まずは"専門用語自動抽出のページ"（以下のURL）にアクセスしてみよう。

http://gensen.dl.itc.u-tokyo.ac.jp/

このページでは、「言選Web」をはじめとする専門用語（キーワード）抽出システム全体について、案内を行っている。はじめての方は、このペ

ージから必要な情報やソフトを探して欲しい。

また。新機能の追加や、バグレポートの掲示も頻繁に行っている。「言選 Web」を既にお使いのかたも、たまにアクセスしていただきたい。

■「言選 Web」は準備不用

「言選 Web」は文章中から専門用語（キーワード）を自動抽出する Web 上のサービスである。使い方は次のとおりいたって簡単である。

1.URL 入力欄に解析を行う Web ページの URL を入れるか、テキストボックスに解析対象の文書を貼り付ける
2.専門用語（キーワード）自動抽出」ボタンをクリックする、
3.文章中の用語が重要な順に表示される。

「言選 Web」には日本語版のほかに、中文、西ヨーロッパ言語版も用意している。これは上部のタブメニューで切り替えられる。インターネットが利用できる環境なら、準備は不要である。「言選 Web」の機能をぜひお試しいただきたい。

■ termex なら自分専用

「言選 Web」は不特定多数のユーザにサービスするため、いくつかの制約を設けてある。この制約なしで使いたいという要望にこたえるのが、Windows 用ソフト“termex”である。これには、通常の日本語版と中文版、また日本語の簡易版である“termex lite”がある。

使い方は処理対象のテキストファイルを termex の実行用アイコンにドラッグするだけでよい。Windows の「メモ帳」が起動し、用語抽出の結果が表示される。

Web ページを主に扱うなら、Internet Explorer との連携機能を試してみよう。Web サイト内の複数のページを蓄積した上で、用語を抽出できる。

また、“termex”は重要度計算パラメータの設定や、学習機能を利用できるなど、個人ユースに特化してある。

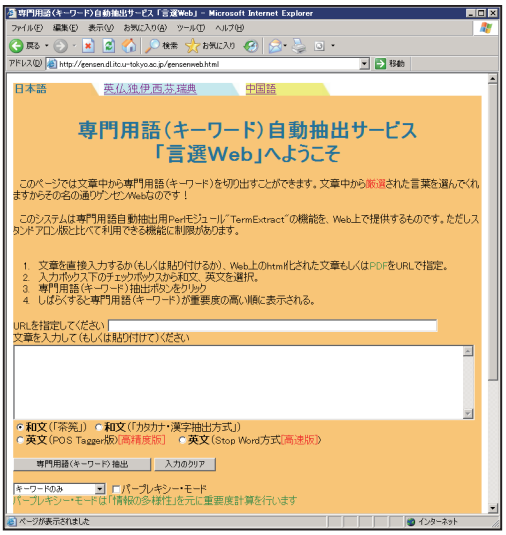


図 2 「言選 Web」

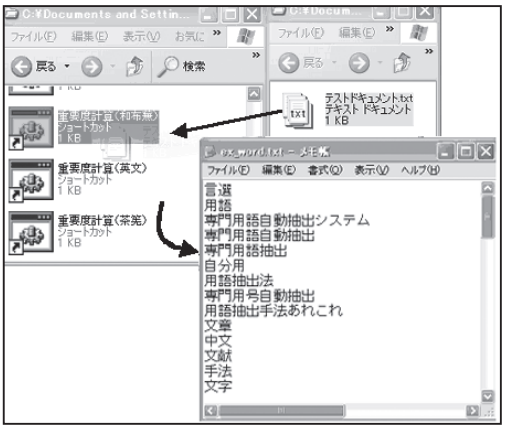
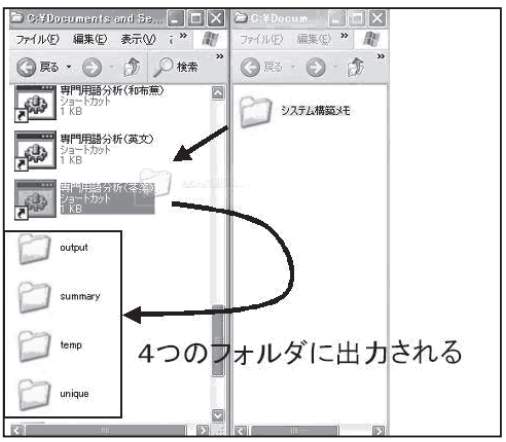


図 3 termex の動作例

図 4 termmi の動作例



■ termmi で文章をまとめて解析

「言選 Web」のキーワード抽出機能をテキスト・マイニングに応用したのが、“termmi” である。テキスト・マイニングとは、大量のテキストデータから隠れた知識を発見するための手法の総称である。termmi はその中でも「手軽」「シンプル」「フリーソフト」であることに特色がある。

Windows のフォルダの中にテキストファイルを複数置き、“termmi” のアイコンにドラッグすることでフォルダ内の全テキストファイルの解析が行われる。全テキストで共通な用語、個々のテキストのユニークな用語、個々のテキストの用語抽出結果、形態素解析結果を見ることができる。

また、別に用意したベクトル空間法による類似度計算スクリプト（vector_space.pl）を使うことで、文献群の中で特徴的な文献を数値的に判断できる。

termmi は、統計的な手法を駆使する学術的なテキスト・マイニング研究とは異なる方向性である、しかし、手軽でシンプルなだけに一般のユーザにはむしろよいのではないかと考えている。

4 “言選 Web”を使いこなす

■「言選 Web」（Web によるサービス）

◆いくつかの制限を理解しよう

「言選 Web」は不特定多数のユーザが利用する。そのため、やむなく機能を制限している部分がある。不便を感じるかたもいるかもいるかと思うが、どうかご理解いただきたい。

A. 一定時間過ぎると処理を中断
B. データ量が大きい場合は、処理を中断。
C. 動的 Web ページへのアクセスを禁止
D. 同時アクセスの禁止

◆アクセス先 Web ページの文字コードは大丈夫？

「言選 Web」では、アクセス先の Web ページの文字コードによらず動作する。これは、文字コードの自動認識結果と文字コード変換により実現している。

例えば、日本語で書かれた Web ページであれば、日本語版「言選 Web」は Shift-JIS, EUC, UTF-8 のいずれでも動作する。西欧言語版「言選 Web」では、Latin-1, UTF-8 の２つに対応している。

中文版「言選 Web」に関してだけは、アクセス先 Web ページの文字コードを自動認識しない。中文版「言選 Web」のデフォルト文字コードは GB であるが、UTF-8 で書かれた中文 Web ページにアクセスするのであれば、「中文（UTF-8）」チェックボックスにチェックの上、お使いいただきたい。

◆目的にあった用語抽出法を選ぼう

「言選 Web」の利用法は簡単だが、オプションがいくつかあり、どれを選んだらよいか悩むところかと思う。以下は開発者からのお薦めである。

A. 日本語であれば、まずは「茶筅」版を選び、よい結果が出なければ、「漢字・カタカナ抽出方式」を試す。
B. 英語の場合は、“POS Tagger version” がデフォルトだが、多量の文章では「制限時間オーバー」を起こしがちである。大量のデータの場合は、“Stop Word version” を選びたい。
C. 中文の場合は、停止語方式版の Web ページからリンクを張っている ICTCLAS 版もある。通常の停止語方式版で満足できない場合は、下準備の手数がかかるが、ICTCLAS 版もお試しいただきたい。

◆下準備が必要な中文 ICTCLAS 版

「言選 Web」はサーバ内部に形態素解析ソフトを組み込んでいる。そのため、［形態素解析］→［用語まとめあげ］→［重要度ランク付け］の一連の処理ステップをユーザが意識することはない。

しかし、中文の形態素解析ソフト ICTCLAS だけは「言選 Web」サーバ内に組み込めなかった。そのため、ICTCLAS 版に限っては、ICTCLAS で事前にタグ付け済みのテキストを「言選 Web」のテキストボックスに貼り付けて使う必要がある。

◆パープレキシティモードはどこが違う？

パープレキシティモードは「情報の多様性」をもとに用語の重要度ランクづけを行う。抽出される用語自体は同じだが、その並び順や重要度の値が異なるということである。

東京大学情報基盤センター中川研究室による最近の研究では、パープレキシティモードが通常の FLR よりも優れた結果を示した。「言選 Web」ではオプション扱いであるが、かなり優れた性能を示すはずである。ぜひ、お試しいただきたい [3]。

■ TermExtract（Perl モジュール）

◆サンプルスクリプトから自作スクリプトへ

“TermExtract”は「言選 Web」のエンジン部分をなす Perl モジュールである。そのまま使えるサンプルスクリプトも付属している。

このサンプルスクリプトを参考に、自分専用のスクリプトを作成してみよう。自作の Perl スクリプトの中に「専門用語（キーワード）自動抽出機能」を組み込むことすら容易に実現できるはずである。

◆自分用の追加モジュールを作る

TermExtract では、「茶筅」解析結果や英文プレーンテキストなど、さまざまな入力データ形式に容易に対応させるため、入力データ形式依存部分と重要度計算部分でモジュールを切り離して作っている。

この入力データ形式依存の部分と、重要度計算部分（Calc_Imp.pm）のデータ入出力仕様さえご理解いただければ、TermExtract の追加モジュールを自作できる。例えば、「アラビア語対応モジュール」など現在は、“TermExtract”がサポートしていない言語への対応も可能であろう。また、第一章で述べた用語切り出し方法を自分の好みに変更することもできる。詳細な仕様を Web でも公開しているので、Perl を使い慣れたかたなら、自在に改良ができると思う。

■ termex（Windows 専門用語抽出）

◆各種パラメータを変更してみよう

termex の Perl スクリプトをエディタ（Windows の「メモ帳」など）で開き、パラメータを変更できる。その説明はスクリプト中のコメント欄に詳しいが、以下にまとめてみた。

A. 重要度計算で、連接情報の重要度計算のモードを選択できる。可能なモードは、連接語の“延べ数”・“異なり数”・“パープレキシティ“・“連接情報を使わない”のいずれかである。

B. 連接情報と組み合わせる頻度を Frequency, と TF（Term Frequency）のいずれかから選ぶことができる。「言選 Web」において、TF は用語が他の用語の一部に含まれていた場合もカウントするが、Frequency はカウントとしない。例えば「情報システムと情報」の場合、TF，Frequency のカウントは次の例 5 のとおりである。

```
「情報システムと情報」
  TF          →
      「情報」2 回、「情報システム」1 回
  Frequency   →
      「情報」1 回、「情報システム」1 回
```

例 5　TF と Frequency の違い

上記 A の“連接情報を使わない”設定がなされていれば、頻度情報のみ出力できる。

C. 重要度計算で、「ドキュメント中の用語の頻度」と「語の連接の重要度」のどちらに比重をおくかを設定する（デフォルト値は 1。値が大きいほど「ドキュメント中の用語の頻度」の比重が高まる。

D.オプションの学習機能を使用できる。デフォルトは、“使用しない”である。

◆学習機能をうまく使うには

学習機能は雑多なテーマの文献を学習させてしまうと、一般的すぎる語が上位にきてしまう。常に同じテーマの文献に限って使用していただきたい。

また、学習機能を使い続けると、重要度中の頻度と連接情報の重みが変わっていってしまう。そのため、重み付けの割合を「ドキュメント中の用語の頻度」側に戻すなどの手当てが必要になる。

■ termmi（テキスト・マイニングツール）

◆どの文献が似ているかを調べてみる

“termmi”には、「ベクトル空間法」による類似度判定用スクリプト vector_space.pl が付属している。「ベクトル空間法」では TF-IDF の重要度を使う方法がオーソドックスであるが、termmi では FLR の重要度を採用している。なお、オプションとしてオーソドックスな TF-IDF の処理モードも用意してある。

vector_space.pl は [文献群全体] と [個々の文献] との類似度計算を行う。文献群全体ではなく特定の文献と他との類似度計算には、⑴ OUTPUT フォルダ中の比較したいファイルを SUMMARY フォルダ中の total.txt に上書きし、⑵その上で vector_space.pl を実行いただきたい。

FLR の重要度と、ベクトル空間法の組み合わせは、研究で実証されていない機能である。読者のかたにも評価いただければ幸いである。

◆どの単語が隣り合うかを確かめる

“termmi”では単名詞の連接情報を、次回の termmi の実行まで学習用データベースに保存している。この連接情報を付属スクリプト get_stat.pl で見ることができる。単語の連接は単語バイグラムモデルにも使われる統計情報である。termmi の場合は用語中の単名詞の連接に限定されるが、研究目的にもご活用いただけるかと考えている。

▌おわりに ～さらなる展開に向けて～

「言選 Web」は平成 15 年 4 月 23 日から一般に公開を始め、月間アクセス件数を数百から多いときには数千まで伸ばしてきた。原稿執筆時（2005 年 9 月 1 日現在）でインターネット・エクスプローラーの「お気に入り」に登録された件数は、3,000 をも超えている。インターネット上での評判もおおむね好評といえそうである。

「言選 Web」の機能は文章から用語を重要度つきで抽出するだけというシンプルなものである。だからこそ、その応用範囲は広いといえる。開発当初は、メタデータベースにおけるキーワード選定、言語学の研究、Web サイトの解析を想定していた。公開後にわかったところでは翻訳補助ツールとしての需要もあるようである。今後ともさまざまな用途に活用いただければ喜ばしい限りである。

「言選 Web」は発展していくサービスである。公開後 2 年あまりの間にも、新たな研究理論の導入や、テキスト・マイニングへの応用、多言語対応など、さまざまな手当てを行ってきた。今後とも、研究から一般向けサービスへの橋渡しとして、開発・運用を進めていていきたい。

謝辞

東京大学情報基盤センター図書館電子化部門の中川裕志教授には、当システムの構築にあたり、専門用語抽出理論の実装、システムの仕様策定などにおいて多大なる助力をいただきました。

また、東京大学経済学部資料室の小島浩之助手によるメタデータベース（東京大学経済学部サブジェクトゲートウエイ Engel）の「キーワード半自動付与」構想がなければ「言選 Web」は生まれなかったといえます。さらに中文版「言選 Web」においても、小島浩之助手を抜きにしては語ることができません。

お二人に深く感謝いたします。

参考文献

- 前田朗・小島浩之・中川裕志「言選 Web」の世界『図書館の窓』vol.43 No.3, 2004 年。pp.61-65
- 小島浩之・前田朗「キーワード（専門用語）自動抽出システムの構想とその展開」第 51 回日本図書館情報学会研究発表要綱。2003 年。p.17-20.
- 佐良木昌・新田義彦『正規表現とテキスト・マイニング：情報発見のツール・キット』東京　明石書店。2003 年。

注

[1] Hiroshi Nakagawa, Hiroyuki Kojima, Akira Maeda. "Chinese Term Extraction from Web Pages Based on Compound word Productivity" *.42nd Annual Meeting of the Association for Computational Linguistics (ACL2004), Third SIGHAN Workshop on Chinese Language Processing*, Barcelona, Spain, July, 2004, p.79-85.

[2] Hiroshi Nakagawa, Tatsunori Mori. Automatic Term Recognition based on Statistics of Compound Nouns and their Components. *Terminology, Vol.9 No.2*, 2003, p.201-209.

[3] 森山聡・吉田稔・中川裕志「複合語のパープレキシティに基づく重要語抽出法の研究」言語処理学会第 11 回年次大会発表。2005 年。

連絡先変更および会費支払いのお願い

会員各位へは既に BBS、メールマガジン等でお知らせしておりますが、連絡先変更届が無いため、事務局から連絡不能となっておられる会員の方が多数おられます。住所、メールアドレス等に変更があった場合は下記 URL より事務局宛にご一報ください。

会員資格変更フォーム
http://jaet.gr.jp/JAET-BBS/change.html
※アクセスには漢情研 BBS の ID・パスワードが必要です。

また本会の運営は会費収入に依存しております。定期的な会費納入にご協力いただきますようお願い申し上げます。

この文書は『漢字文献情報処理研究』第6号の抜粋です（漢字文献情報処理研究会発行）。

キーワード自動抽出システム「言選 web」（中国語バージョン）を検証する

山崎　直樹（やまざき　なおき）

1. この文章の目的

この文章の目的は、「キーワード自動抽出システム「言選 web」（中国語バージョン）」で幾つかのテキストを分析し、その処理結果を検証することである。この中国語バージョンには、次の 2 種類の方式による版が存在する。

(a)「言選 web」（中国語バージョン）
　停止語方式版
(b)「言選 web」（中国語バージョン）
　ICTCLAS 版

この文章では、この 2 種類の方式の処理結果の特徴を比べながら、話を進めていきたい。

この 2 種の版が比べられることは、あるいは、製作者たちにとって本意でないかもしれない。しかし、筆者の試行の結果、一手間よけいにかけて下ごしらえをした ICTCLAS 版のほうが、必ずしも常に「良い」結果（この「良い」は主観的なものであるが）を出すわけではない、ということがわかったので、この比較にもそれなりの意味があると考え、ここに提出する次第である。

2.「言選 web」について

「キーワード自動抽出システム「言選 web」（中国語バージョン）」の停止語方式を用いた版と、ICTCLAS（北京大学計算言語学研究所）の自動分節システムを用いた版については、前田 2005（本誌所収）に解説があるので、そちらを参照していただきたい。

3. ノイズと重複

ICTCLAS 版のほうが、必ず、少ない量のキーワードを抽出して終わる。だいたい、停止語方式版の 50% ～ 60% 前後（多くても 70%）の量になる。停止語方式版が、なぜ、キーワードの量が多くなるのかについての専門的分析は擱いておいて、筆者のナイーブな感想を言わせてもらえば、これは、ノイズと重複が多すぎるせいではないかと思われる。以下に、幾つかのテキストの処理結果のうち、それぞれの方式で上位にランクされたキーワードを挙げる。

【テキスト⑴の処理結果】

停止語方式	ICTCLAS
体态语	体态
体态语的	有声语言
的体态语	人类
与体态语	社会
语言	自然语言
大部分体态语	动作
(全 269 語)	（全 165 語）

【テキスト⑵の処理結果】

停止語方式	ICTCLAS
茶艺馆	茶艺
茶艺	茶室
茶艺馆中	魅力的中国茶文化
见茶艺馆	悦耳的中国古典音乐
茶艺室内	茶和茶点
五福茶艺馆	文化氛围
（全 43 語）	（全 30 語）

【テキスト⑶の処理結果】

停止語方式	ICTCLAS
名词	汉语
现代汉语	名词
古代汉语	古汉语
现代汉语中	汉语普通话的事实
古代汉语中	古汉语的事实
古代汉语的某	汉语的一般规律
古汉语的事实	词的语法功能的时候
宾语	习惯说法
成语	金戴银
古汉语的干扰	主语
（全 28 語）	（全 18 語）

【テキスト⑷の処理結果】

停止語方式	ICTCLAS
同意	对方
的介绍	条件
人的介绍	面的问题
介绍	面的信
介绍的信	问题
介绍男	对方的条件
人介绍男	男女婚姻
（全 45 語）	（全 21 語）

例えば、テキスト⑴の処理結果を見ると、停止語方式のほうは、上位 6 語のうち実に 5 語までが“体态语”という語を含むフレーズである（“体态语”を含む、更に大きい複合語ではない。それであれば意味があるのだが）。このような重複は、何ら新しい情報を受け取り手に与えない。ICTCLAS 版には、そのような重複はない。また、テキスト⑵の処理結果で、停止語方式版が“茶艺”と“茶艺馆”を別にリストアップしているのは意味があるのだが、“茶艺馆、茶艺馆中、见茶艺馆”という重複はむだであろう。

また、テキスト⑷に著しいが、停止語方式版は分節の不適切さが目立つ。よって、抽出された語が意味がわからない文字列になっていることがある。これは、ストップワードのリスト（「言選 web」は、手作業でピックアップした独自のストップワードリストをもつ。この作業自体は敬意を払うに値する）が、まだ不完全なせいなのか、それとも方法論的な限界なのかは、筆者には判断できない。ただ、ICTCLAS 版に分節の誤りがないわけではない [1]。

4. 派生語の処理

■ 4.1 停止語方式版の明るい面

派生語の処理に関して、停止語方式版と ICTCLAS 版では妙な差がある。

テキスト⑴は、トピックが非常に明確な文章であり、「この文章に題をつけなさい」という課題を与えれば、誰もが“体态语”（身振り、手振りによるメッセージ）という語を選ぶだろうと思われる文章である。

停止語方式版は、その“体态语”をキーワードの筆頭に選んでいるが、ICTCLAS 版では、筆頭は“体态”であり、そして全 165 語において、つ

いに“体态语”は挙げられない。因みに、この文章では“体态”（身振り、手振り）そのものは、重要な語ではない。

同じく、テキスト⑵において、重要なのは“茶艺”そのものよりもむしろ“茶艺馆”なのだが、停止語方式版はそれを筆頭にしているのに対し、ICTCLAS 版は“茶艺”は挙げるが、全 30 語のうちに“茶艺馆”は挙がっていない。

結果としては、停止語方式版のほうが「良い成績」を残しているようだが、これは ICTCLAS 版の派生語の処理に問題があるためであろう。

つまり、ICTCLAS 版は、［自由形式の語］＋［付属形式の形態素］という構造を、1 つの概念をもつより大きな派生語としてとらえないのだと思われる。上述のキーワードもみなこの構造である。次のとおり。

[[体态 F] 语 B]，[[茶艺 F] 馆 B]
（F= 自由形式の語，B= 付属形式の形態素）

因みに、ICTCLAS の自動分節システムは、この 2 つの語を、次のように分節する。

体态/n 语/Ng，茶艺/n 馆/Ng
（n＝名詞，Ng＝名詞性形態素）

なお、別のテキストでは、ICTCLAS 版が、“医学界”“死亡率”というキーワードを切り出している。これなども、上述の [[_ F] _ B] という構造のように思えるが、ICTCLAS の自動分節システムは、次のように分節している。

医学界/n，死亡率/n （n＝名詞）

また、別のテキストの処理結果で、ICTCLAS 版が、同様の構造と思われる“姓名权”というキーワードを切り出している例が見られたが、これは ICTCLAS の自動分節システムが、この文字列を次のように分節しているからであろう（“权”を独立した名詞とするのは、どうかと思うが）。

姓名/n 权/n （n＝名詞）

要するに、ICTCLAS の自動分節システムの分析の不統一さが、キーワード抽出の結果の整合性に影響を与えているのである。

■ 4.2 停止語方式版の暗い面

§ 4.1 で述べたことだけを見ると、停止語方式版のほうが成功しているように見えるが、他の処理結果を見ると、そうとは言い切れない。[[_ F] _ B] というキーワードを認定できることが、多量の無意味な重複に繋がっているからである。【テキスト⑶の処理結果】の“现代汉语、古代汉语”と“现代汉语中、古代汉语中”という重複を見れば明らかである。【テキスト⑴の処理結果】の“体态语的、与体态语”などの不適切な分節による抽出も、おそらくこれと関係があろう。

■ 5. 品詞で振り分けるやりかたについて

前田 2005 によれば，ICTCLAS 版では，この自動分節システムにより動詞と判断された語をキーワードから除外する（ただし，名動詞と判断された語は除外しない）方式を採用している。

いっぽう，停止語方式版では、機械的に動詞を除くやり方はよい結果を生まないとの判断から、手作業による停止語のリストを作成し、動詞を一律に除外する方法とは別のアプローチを採っている。

具体的に言うと、“写信”（手紙を書く／手紙を書くこと）は、述語として使われていれば、キーワードとして挙げる必要はないが、動名詞的に使われていれば、キーワード候補になりうる。しかし、中国語のばあい、語形からのみではどちらの用法なのか判別がつかない [2]。よって、「品詞で判断」は現実的ではないとのことである。

ではさて、この 2 種類のアプローチがどのような結果をもたらしたか、以下で検証したい。

【テキスト⑸の処理結果】

停止語方式	ICTCLAS
安乐死	问题
问题	资源
资源	社会
需要慎重	医生
的问题	卫生资源
医生	时候
病人	病人
（全 186 語）	（全 91 語）

このテキストは、トピックが非常に明確な文章であり、「この文章に題をつけなさい」という課題を与えれば、誰もが、“安乐死”（安楽死）という語を選ぶだろうと思われる文章である。

停止語方式版は、その語を筆頭に挙げている。しかし、ICTCLAS 版は、全 91 語の中に “安乐死” がいっさい出てこない。これは、ICTCLAS の自動分節システムが次のように、“安乐死” を動詞として分節しているからである。

⑷“安乐死” 的问题
⑸“/w 安乐死/v ”/w 的/u 问题/n
（w＝句読記号, v＝動詞, u＝助詞, n＝名詞）

また、§３で挙げた【テキスト⑷の処理結果】を見ていただきたい。このキーワードのリストだけからでは、およそ文章の内容が察しにくいのであるが、要は、「紹介状の書きかた」がトピックである。よって、不適切な分節が多いものの、停止語方式版が “介绍”（紹介／紹介する）を含む語をキーワードリストの上位に持ってきたのは適切である。

しかし、ICTCLAS 版では、“介绍” のかけらもない。これは、ICTCLAS の自動分節システムが、次の⑷のように明らかに名詞的に使われているばあいでも、⑸のように分節してしまっているからである。

⑷作一简单的介绍（簡単な紹介をする），
　別人的介紹（他の人の紹介）
⑸作/v 一/d 简单/a 的/u 介绍/v，
　別人/r 的/u 介绍/v
（v= 動詞，d= 副詞，a= 形容詞，u= 助詞，r= 代名詞）

同じ例をもう１つ挙げる。§３の【テキスト⑶の処理結果】を見ていただきたい。停止語方式版のキーワードの終わりに、“古汉语的干扰”（古代漢語の干渉）というフレーズが挙がっている。実は、このフレーズはけっこう重要なのであるが、ICTCLAS 版では挙がってこない。やはり、下記の⑸のように分節がなされていることが原因である。

⑷ 古汉语的干扰
⑸ 古/a 汉语/n 的/u 干扰/v
（a＝形容詞，n＝名詞，u＝助詞，v＝動詞）

余談だが、下記の構造をもつフレーズは、重要な概念を表していることが多いのではないか（中国語のばあい、下記の構造の [名詞] は、動詞の対象であるより主体であることが多い）。

[名詞] 的 [動詞]#　（#= 統語論的境界）

▌6. 重要語が抽出できない例

以下に、テストに使ったあるテキスト（テキスト⑹とする）を掲げる。

伊斯兰教赋予男子有休妻和遗弃妻子的绝对权力。跟据伊斯兰教规，一个男人可以拥有四位妻子，于是离婚成了男子一生中娶多个妻子的可供选择的手段。他们往往以离婚次数多而自豪。对女人来说，离婚并不像汉人所想像的那么严重以至成为妇女的耻辱并导致严重的后果，这是因为女方娘家并不会强迫她在不幸福的婚姻里苦度光阴，女方在离婚后可以回到娘家，娘家具有从道义上，财力上支持她们取得社会经济地位的责无旁贷的义务，并会重新为她物色对象以期再嫁。

离婚后的妇女可以带走陪嫁财产，丈夫还给一定数量的钱财和物品。重新进入婚姻市场时，她们还能与年轻姑娘一争高低。离婚后的孩子的抚养也不成向题，孩子的祖父母或外祖父母可作为孩子的监护人。这样在伊斯兰教文化中，人们不以离婚为耻辱，男子有离婚的自由，妇女与孩子可以获得各种形式的社会和家庭的支持，因而伊斯兰教社会中的离婚水平很高。

このテキストの処理結果は次のとおりである。

【テキスト(6)の処理結果】

停止語方式	ICTCLAS
男子	孩子
娘家	伊斯兰
妻子的	社会
伊斯兰教社会中的	妇女
女人	妻子
妇女与孩子	男子
伊斯兰教赋	娘家
会重新	一定数量的钱财和物品
四位妻子	婚姻市场
伊斯兰教文化中	幸福的婚姻
（全 48 語）	（全 38 語）

この文章にタイトルをつけるとしたら、最もシンプルなのは、“伊斯兰教文化与离婚”（イスラム教文化と離婚）といったところであろうか。停止語方式版が“伊斯兰教社会中的、伊斯兰教文化中”というキーワードを切り出し、ICTCLAS 版が“伊斯兰”どまりなのは、おそらく§4で述べた理由によるものであろう。

しかし、問題は、“离婚”という重要語が、テキスト中に全部で9回も出現しているのにかかわらず、どちらのリストにも最後まで現れないことである。このあたりの処理は、将来の課題のようである。

注

[1] 例えば、“这方面的信”という文字列から“面的信”を切り出したりしている。

[2] このあたりの問題は、専門家の間でも意見の分かれる、やっかいなところである。詳しくは、三宅2005を参照。

【参考文献】

前田朗 2005
「キーワード自動抽出システム「言選 web」」。『漢字文献情報処理研究』第 6 号，pp.124-133。

三宅登之 2005
「動詞と名詞の区分をめぐって―品詞表示の比較のモデルケースとして―」。山崎・遠藤 2005，pp.43 - 73。

山崎直樹・遠藤雅裕編 2005
『辞書のチカラ―中国語紙辞書電子辞書の現在』。東京：好文出版。

中国語CAI 実践レポート

電子辞書の普及によって、私たちはあらためて「いい辞書とは何か」という問いと向き合うことになった。紙の辞書の重要性はいささかも低下していないし、やみくもに新しいものを追い求めても仕方ないが、「いい辞書」の条件をもう少し自由に発想してもいいのではないだろうか。今回は辞書の電子化によって、紙の辞書では考えられなかった形態や利用法を実現しようとする試みを取り上げる。

日本中国語 CAI 研究会について

本会（会長：田邉鉄北海道大学助教授）は「コンピュータ援用の授業方法を中心とした中国語教授法の研究・開発・普及を推進し、同時に教員・研究者・ソフトウェア開発者の交流をはかる」ことを目的とし、1996 年 11 月に発足した。

会員による研究発表・実践報告の場として、例会（年 1 ～ 2 回）、総会（秋、年 1 回）を開催するほか、常時メーリングリストで情報意見交換を行っている。

参加を希望される方は、中国語 CAI 研究会 Web **http://moli.cims.hokudai.ac.jp/~ccai/** を参照していただきたい。

Contents

オンライン中国語辞書『北辞郎』

清原　文代（きよはら　ふみよ）

✎ 1.多様化する辞書の形態

紙の辞書、携帯に便利な IC 電子辞書・パソコンで使う CD-ROM の電子辞書・パソコンのハードディスクにインストールするタイプの電子辞書、そして Web で検索するオンライン辞書と、従来からある辞書をわざわざ「紙の」辞書と言わなければならないほど、辞書の形態は多様化している。しかし、紙の辞書以外の辞書であっても、そこで提供される内容は紙の辞書を電子化したものがほとんどである。しかし、本稿が紹介するオンライン中国語辞書『北辞郎』[1] は元になる紙の辞書は存在せず、メンバーが単語を登録していくオンライン中国語辞書である。

『北辞郎』

http://www.ctrans.org/cjdic/index.php

『北辞郎』に単語を登録するには「辞書幇」のメンバーになる必要があるが [2]、『北辞郎』を辞書として利用するだけのユーザーであれば、登録等は不要である。

図 1 『北辞郎』トップページ

✎ 2.『北辞郎』の検索方法

2.1.Web ページ

『北辞郎』の Web ページでは以下の検索方法が提供されている。

- 中国語前方一致
- ピンイン検索
- 日本語検索
- 中国語完全一致検索
- 中国語全文検索
- 部首画数検索

各検索方法の詳細については『北辞郎』のヘルプ [3] に掲載されているが、ここではいくつか注意すべき点や気づいた点を述べておきたい。

中国語前方一致・中国語完全一致検索・中国語全文検索はいずれも中国語を入力して検索するが、『北辞郎』のヘルプによれば、単語登録は簡体字にも繁体字にも対応しているものの、現時点でのデータは簡体字のものがほとんどのようで、検索は簡体字を使って行なった方がよい。

ピンイン検索の拼音字母は半角英数字を使ってアルファベット＋声調を表す数字という形で入力する。声調の入力は必須である。例えば、“什么”は shen2me と入力する。ü は v で代用する。

ピンイン検索ではワイルドカードが使用でき、? は任意の 1 文字を、* は任意の文字列を表す。?

の 1 文字とはアルファベットまたは数字の 1 文字を指す。例えば b?? と入力すると、bu4 にも bao にもヒットし、更にはピンイン検索は前方一致検索のため bang 等の 4 文字以上のものにもヒットしてしまう。また、ヘルプによればピンイン検索は前方一致検索とあったが、ワイルドカードが使えるということは後方一致検索もできるのではと思い試してみたが、期待した検索結果は出てこなかった。

図 2　検索結果画面

日本語検索は語釈の中の日本語を検索することにより、本来中日辞典である『北辞郎』を日中辞典としても使えるようにする機能である。この日本語検索は前方一致検索である。

2.2. FireFox プラグイン

Web ブラウザ Firefox[4] の検索バーから『北辞郎』を検索するプラグインをタケウチ氏が提供しておられる [5]。

http://www.ctrans.org/software/index.php

このプラグインをインストールすれば『北辞郎』の Web ページを開けなくとも、Firefox の検索バーから『北辞郎』を検索することができる。中国語を検索バーに入れれば中国語前方一致で検索結果が表示されるが、タケウチ氏の Web ページによれば、py: を前に付ければ拼音字母での検索が、jp: を付ければ日本語検索が可能である。更に筆者が試したところ、ex: を付ければ中国語完全一致検索が、ch: を付ければ中国語全文検索ができた。

プラグインのインストール方法

- Windows 版 Firefox

  ダウンロードしたファイルを解凍し、Mozilla Firefox フォルダ内にある searchplugins フォルダに入れる。

- Mac OS X 版 Firefox

  Firefox のアプリケーションアイコンを control キーを押しながらクリックして「パッケージの内容を表示」を選び、Contents フォルダ→ MacOS フォルダ→ searchplugins の中に解凍したファイルを入れる。

✎ 3.検索結果の表示

単語が見つかれば、Web ページに下半分に表示される。辞書の本文の横には編集というリンクが表示され、メンバーであれば内容を編集することが可能だ。

見つからなかった場合は、『北辞郎』に登録するためのリンク（メンバー向け）、当該の単語を

図 3　検索語が見つからなかった場合

Google で検索するためのリンク、中国のオンライン辞書 Web サイト『词霸搜索』で検索するためのリンクが示される。Google と『词霸搜索』へのリンクはそれぞれのサイトの検索ボックスに当該単語を自動的に入力して検索した結果が表示され、非常に便利である。

✎ 4.『北辞郎』の収録語彙

2005 年 7 月現在で『北辞郎』には 10 万 2000 語余りが登録されているが、登録によって今後も収録語数は増えていくであろう。

紙の辞書の場合、重要度や使用頻度の高い語から順に収録されるが、『北辞郎』はメンバーによる登録という形を取っているため、紙の辞書のようなバランスのとれた語彙の収録にはなっていない。『北辞郎』が設置されている Web サイト Ctrans.org[6] の「このサイトについて」[7] によると、管理人タケウチ氏は、契約書・特許文書・ビジネスレター等の中国語の実務翻訳に従事し、主にソフトウェア・ハードウェア・通信・機械・医薬といった分野の文書を取り扱っておられるとのことである。そのためこれらの分野の語彙が多く登録されている。この収録語彙のアンバランスというのは、一見短所のようにも見えるが、逆に言えば紙の辞書ではなかなか見つからない専門用語に強く、紙の辞書では収録が間に合わない新語や流行語に対応しやすいという長所でもある。例を挙げると、最近日本でも注目されているブログ（Weblog）は中国語では“博客”と言う。中国の代表的ブログサイト博客中国 [8] が設立されたのが 2002 年、本格的に発展しだしたのは 2004 年からであるが、『北辞郎』にはすでに収録されている。

辞書の記述については、単語にもよるが拼音字母（アルファベットと数字による表記）と語釈のみである場合が多く、例文や用法の解説はほとんどない。したがって、初級～中級レベルの学習用辞書としてはあまり適当ではなく、中国語を一定レベル以上マスターした人が使う辞書であると言えよう。但し、この辞書はメンバーによる単語登録で日々成長していく辞書であり、これはあくまで現時点においての状況である。現在の収録語彙や意味記述に不満があれば、自ら「辞書幇」のメンバーとなり辞書の作成に参画すればよいのである。『北辞郎』は中国語に関わる人々が智慧と知識を持ち寄る場として今後大いに期待できる。

✎ 5.タケウチ氏開発の辞書関連ソフトウェア

タケウチ氏の Web サイト Ctrans.org では、『北辞郎』以外にも様々な中国語関連のソフトやパソコンで中国語を扱うためのノウハウを提供している。その中から辞書に関わるソフトを 2 点紹介する [9]。

5.1. PDIC 用中国語辞書 [10]

5 万 5000 語が収録された PDIC 形式の中国語辞書データファイル（シェアウェア）である。閲覧するためには別途辞書検索ソフトが必要である。Windows2000 / XP 用の辞書検索ソフトとしては、TaN 氏制作の PDIC / Unicode[11]（開発途上のβ版）があり、Mac OS 9 / X 用には Naoya TOZUKA 氏制作の PDIC Viewer[12]（シェアウェア）がある。

辞書の内容は単語・拼音字母・語釈という構成で例文はないが、PDIC 形式の辞書ファイルは自分で作ることができ、タケウチ氏はその制作方法を Web ページで紹介している [13]。

5.2. 中国語変換ツール PinConv+[14]

中国語（簡体字）の文章を単語に区切った上で、拼音字母に変換する Windows2000/XP 用ソフト（フリーウェア）であり、中国語辞書を内蔵している。簡体字と日本漢字を相互に変換する機能もある。

拼音字母への変換は、声調符号付き拼音字母、及びアルファベット＋数字の簡易表記形式での出力の 2 種類である。拼音字母付きの中国語教材を簡便に作成するという点から見ると、PinConv+ はフリーウェアである上に、操作が比較的わかり

やすく、お勧めである[15]。

PinConv+ には 5 万 5000 語が収録された辞書が付属しており、単語をダブルクリックすることによって辞書引きができる。辞書の記述は拼音字母と日本語の語釈のみのシンプルなもので、例文はない。内蔵の辞書は、PinConv+ の「現在の単語を編集」ボタンをクリックすることによって、自分で単語登録をしたり、辞書の内容を編集することが可能である。辞書データのインポート（タブ区切りテキスト）及びエクスポート（タブ区切りテキストと PDIC 形式）にも対応している。

注

[1] 『北辞郎』という名称は恐らく『英辞郎』（**http://www.eijiro.jp/**）の影響を受けたものだと思われる。『英辞郎』はプロの翻訳者・通訳者のグループである EDP による英語辞書である。

[2] 『北辞郎』利用規約
http://www.ctrans.org/cjdic/rule.php
「辞書幇」メンバー登録申請
http://www.ctrans.org/cjdic/member.php?mode=regist

[3] **http://www.ctrans.org/cjdic/help.php**

[4] **http://www.mozilla-japan.org/**

[5] タケウチ氏は同 Web ページにおいて中国のオンライン辞書『词霸搜索』用のプラグインも提供されている。

[6] **http://www.ctrans.org/**

[7] **http://www.ctrans.org/about.htm**

[8] **http://www.bokee.com/**
2005 年 7 月に博客中国から博客网に改称した。

[9] 辞書関連以外に、多言語に対応した Windows 用の検索ツール「朧」（シェアウェア）がある。「朧」は Unicode・GB2312・UTF-8・Big5・Shift-JIS で書かれたテキストファイルを正規表現で検索できる。
http://www.ctrans.org/entry.php/1121316568
多言語、特に中国語に対応し、且つ正規表現も使える検索ツールは少なく、「朧」は用例の抽出等に役立つであろう。

[10] **http://www.ctrans.org/entry.php/1120452496**

[11] **http://homepage3.nifty.com/TaN/pdic-unicode.html**

[12] **http://pdicviewer.naochan.com/**

[13] **http://www.ctrans.org/entry.php/1120453288**

[14] **http://www.ctrans.org/entry.php/1120984246**

[15] 同様のことは MS Office に別売の Proofing Tools を導入すれば可能である。Proofing Tools にはその他にも中国語 OCR 機能や追加の中国語フォント等が含まれている。

『北辞郎』に単語を追加する

田邉　鉄

『北辞郎』に単語を追加したり既存の単語を編集するには、辞書編集グループ「辞書幇」に参加する必要がある。参加資格は特になく、「中国や中国語に興味があれば誰でも」参加できる。参加方法も簡単で、初期画面から「メンバー登録」を選び、メールアドレスを登録すればよい。追って参加確認メールが届くので、その指示に従って URL をクリックして名前とパスワードを登録すれば、メンバーとして活動できる。

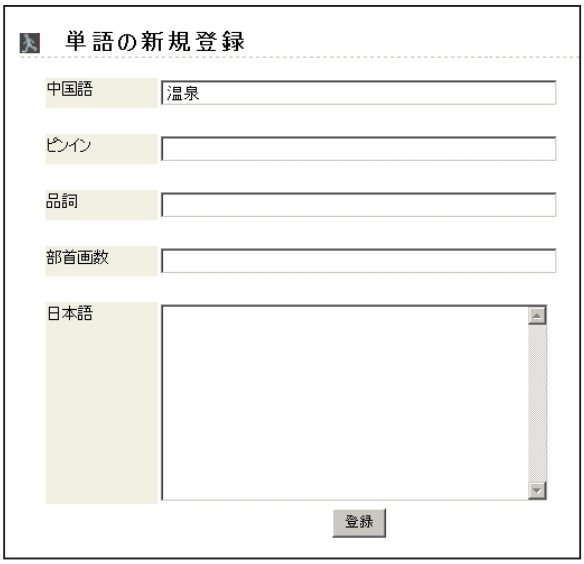


図 1　単語の新規登録

メンバーは単語の登録や修正のほか、運営やシステムについて意見を述べること、また『北辞郎』で利益が生じた場合に、その配分に与ることができる。

辞書には、中国語・ピンイン・品詞・部首画数（単漢字のみ）・日本語が登録できる。このうち中国語と日本語は必須で、他は任意である。登録方法は、単語一つずつの登録と、一括登録が選べる。

⑴ 単語一つずつの登録

『北辞郎』で単語を検索した際、まだ登録されていない単語の場合、「○○を辞書に登録する」というメッセージが表示される。これをクリックすると単語編集画面が開くので、各項目を入力する。『北辞郎』を使って翻訳などの作業をしている時に新語に出くわした、というような場合、この方法で手軽に登録できる。『北辞郎』はコンピュータ関係の技術用語など、結構充実しているので、この「登録する」メッセージが出るとなんだか嬉しい。

⑵ 一括登録

中国語に関わる仕事をしていて覚え書きとして単語帳を作っているとか、学生向けに特定分野の glossary を作っている、といった場合は、ぜひ『北辞郎』に登録してほしい。一括登録するには、1 行に中国語、品詞、ピンイン、日本語、部首画数の 5 項目をタブで区切って並べたファイルを作る。入力しない項目は？（半角）を入れておく。Excel などで作って、テキストファイルとして出力してもよい。先頭行には「北辞郎用ファイル」と書いておく。ファイルは UTF-8（BOM なし）で保存し、『北辞郎』の「一括登録用ファイルのチェック」機能で、登録済の単語ではないか、書式は正しいかなどのチェックを行う。修正したファイルを「ファイルから単語一括登録」機能で読み込ませると、辞書に登録される。自分で登録した単語は、いつでも「編集状況の確認」で一覧でき、削除したり編集したりできる。

例文や備考などは日本語の項目に、〈例〉、〈備考〉などのタグを添えて記述する。その他いくつか記述ルールが決められているが、ユニークなのは〈修正歓迎〉タグである。訳語などに自信がない時は、このタグをつけて書き込んでおくと、他のメンバーは気兼ねなく修正できる。日本語の項目は基本的にフリーフォーマットなので、用語や記法の厳密な統一は難しいだろうが、〈修正歓迎〉タグのような、「気づいた人が直せばいい」的な姿勢がエントリーバリアを下げ、参加者を増やし、結果として辞書の内容充実に繋がっている。これは本誌 5 号で特集した WikiWikiWeb の中国語辞書版と言え、外国語分野だけでなく、ネットワークコミュニケーション分野においても興味深い取り組みと言えよう。

手のひらに中国語を

小川　利康（おがわ　としやす）

✎ 1.PDA で中国語を使うには

Windows2000 以降、中国語が使えるのはもはや至極当たり前のことになってしまった。だから、PDA で中国語が使えないと逆にストレスになってしまう。小文ではなるべくストレスがたまらぬ Tips をご紹介する。

1) Palm 系 PDA の場合

Palm OS 系の PDA（Sony CLIE など）でも CJK OS Ver4.621（シェアウェア、杜氏工作室）[1] によって中国語の利用は可能である。ただし、これは本来的には英語版 OS を中国語化するソフトウェアであり、その前提となるのは、中国語版 Windows の母艦とシンクさせながら使う環境だ。Palm OS 自体が対応しない Unicode は当然サポート外。もし日本語を完全に棄て、中国語専用 PDA とするなら快適に利用できるが、日中混在を考えるとなると、日本語とのコード衝突は避けようがない。例えば、メモ帳を経由して中国語を読んだり、書いたりすることは可能だ。（図 1 は Sony CLIE-NX70V で中国語を表示させたところ）だが、Windows 上で Unicode に自動変換されぬようにネィティブコードのまま（文字化け状態のまま）転送しなければならない [2]。この作業は別段不可能ではないが、ややトリッキーな方法だ。更に Documents to go 7（シェアウェア、XLsoft）[3] があれば、Word や Excel のファイルを PDA で読めるが、編集すると母艦の Windows マシンに戻した段階で文字化けするので、結局 Palm では中国語ニュースを読むくらいが精々だろう。

2) Pocket PC2002, 2003 の場合

Windows CE の世代では単一言語しか扱えなかったが、Pocket PC 2002、2003 では内部的には Unicode で動作し、複数言語を扱えるようになった。Windows3.1 から何とか Windows95 ないしは 98 に近づいたレベルであり、とても完全なものとはいえないが、Word や Excel で作ったファイルなら、あれこれ考えなくても中国語を簡単に扱える。Palm 系に比べれば格段の進歩である。

とはいえ、OS に対応するソフトがなければ絵に描いた餅であることは、かつての Symbian OS

図 1　Sony CLIE-NX70V で中国語を表示

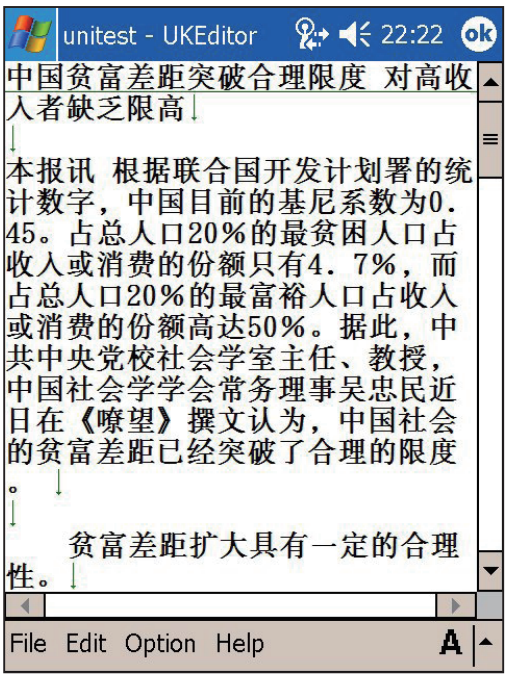


図２　UKeditor で中国語文書を表示

を搭載した PSION（サイオン）シリーズが良い例である[4]。当初から Unicode を搭載しながら対応する中国語フォントや入力環境がなかったために、結局日中マルチリンガル環境は実現しなかった。

幸い Pocket PC の場合、Gnu の Unicode フォント[5]が利用できるので、Unicode もしくは GB,Big5 にエンコードされたテキストファイル・Word ファイルを Pocket PC に転送してやれば、比較的簡単に読むことが出来る。Windows の母艦でも利用できる TrueType フォントなので、双方に同一のフォントを入れておけば文字化けをいちいち直す手間も省ける。標準でインストールされている Pocket Word だけでも充分だが、エディタとしては UKeditor（フリーウェア、UKtaniyama 氏、図２は UTF-8 で中国語を読み込んだところ）が秀逸だ。また、扱えるファイルが 64KB 以下に制限されるものの、読書用なら TextViewer2002（フリーウェア、楳田敏之氏作）も使いやすい[6]。ただし、いずれにしてもフォントは複数指定できないので、利用する Unicode フォントによっては読めない文字が出てしまう。これはソフトウェアというよりも文字フォントの問題である。

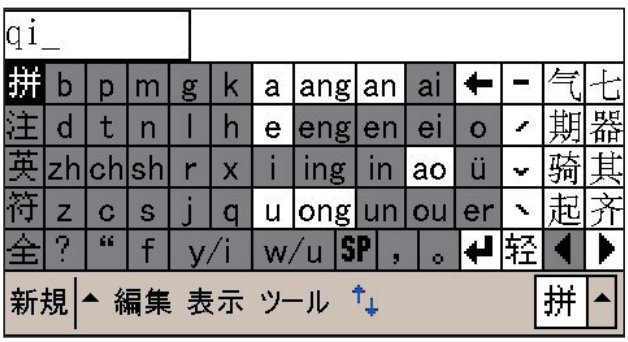
このようにフリーウェアだけでも中国語を読むだけなら十分可能だ。だが、入力まではできない。今のところ、その問題を解決する最も有力な選択肢が中国語入力 IME と日中・中日辞書を備えた「翻訳ウォーカー j・北京 V2」である。

✎ 2.翻訳ウォーカー j・北京 V2

1) 入力の実際

「翻訳ウォーカー j・北京 V2」は翻訳ソフトとして発売されたソフトであるが、同時に日中電子辞書を内蔵し、中国語入力 IME を備えた統合的中国語環境を提供してくれるソフトウェアである。添付された SD カードを挿入するだけでインストールが始まり、自動的に「中国語ピンイン入力」「中国語手書き検索」「中国語手書き入力」の３種類の IME が追加される。

図３　翻訳ウォーカーでピンイン入力



ピンイン入力の場合、キーボード画面からピンインを逐次選択して行くと、自動的に後に続く韻母が絞り込まれる設計となっており、最初はとまどうが、使い慣れれば便利である。図２には出ていないものの、注音字母入力も選択可能だ。オプションで繁体字も選択できる。

また、PDA 用の手書き入力も認識率がかなり高く、書き順を間違えてもある程度までは許容される。日本漢字の「骨」

で書いても、きちんと「骨」に変換してくれる。いちど「手書き入力」に慣れてしまうと、ピンイン入力が煩わしくなるほどだ。文章の推敲、短い文章書きは全く苦にならないだろう。ただ、単漢字変換だから大量に文章を書くのは余り現実的ではない。近年は中国の携帯電話も予測変換で語彙を提示してくれるぐらいだから、そうした入力支援機能があると更に快適だっただろう。

2）辞書の新機能

中国語辞書自体は近年の電子辞書に収録されている小学館版の日中、中日辞典であり、余り新味はない。だが、例文のなかから該当する語彙を網羅的に全文検索できる機能は従来なかったもので、学習者にとっては大きなアドバンテージになると思われる（図３は例文からネコを含む用例を引いたもの）。日本語からも、中国語からも引くことが出来るので、作文学習などには大いに役立つだろう。是非とも本家の Windows 版でも利用できるようにしてもらいたいものである。

3）翻訳機能

旅先でのコミュニケーションツールを意識して翻訳ウォーカーに定型文を五百例ほど収録している。この文例を使えば、機械音声による朗読機能で、相手にそのまま聞かせることも出来る。機械音声ながら、十分聞いて理解できるレベルに達している。ただ、文例は一文ごとに別ファイルなので、たった一言の表現を呼び出すために毎回ファイルを開かねばならない。この操作性は改善が望まれる所だ。

肝心の日本語、中国語の双方向翻訳機能については、短文ならば訳せるというレベルには達しているものの、複雑な文章を訳せるレベルではない。単語を並べれば、一括して辞書が引ける位のつもり使った方が賢明だろう。筆者の場合、Windows マシンの母艦で作成した Word ファイルの中国語から一部をコピーして、翻訳ウォーカーに貼り付けて、単語の意味を調べたり、ピンインを表示させてみたりするような使い方が中心だが、これなら十分実用に耐える。いちいちピンイン入力しなければならない電子辞書と違って、長い文章でも一気に貼り付けて引けるのは便利で、大きなアドバンテージといえる。

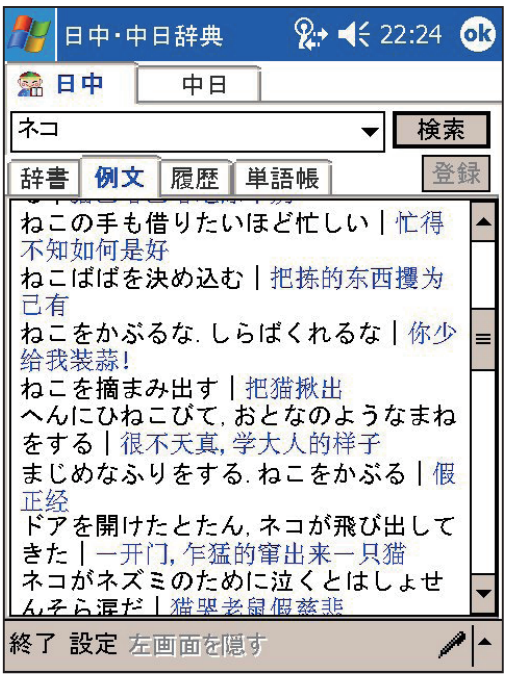


図４　例文からネコを含む用例を引いたもの

同様の製品として Windows 版には「j・北京 V5」があり、細かい使い勝手では見劣りするものの、基本的機能においては遜色ない。これだけ豊富な機能を Pocket PC に追加してくれるというのは驚異的でもある。とはいえ、豊富な機能ゆえに高額（SD カード付属で 41,790 円）であり、Pocket PC で中国語環境を整えたいと思っているユーザの誰もが気軽に買える値段ではない。電子辞書との競合も考えると、入力 IME 単体、もしくは辞書のみをつけた廉価版を要望しておきたい。

4）ブラウザとの連携

中国語環境を手に入れたユーザがまず考えるのは、出先でネットにアクセスして中国語のメールを読み書きしたり、中国語のサイトをブラウズすることだ。だが、翻訳ウォーカーをインストールしても、この機能は実現できない[7]。入力 IME とフォント周りの問題だけでなく、メーラー、ブラウザの多言語対応も必要だからである。これは

翻訳ウォーカーの仕様を越える部分だが、今後ぜひ対応して欲しい部分である。

現時点でどうしても中国語でメールやブラウジングをしたい場合は、Time Space System 社[8]の Effy-JC 3.1（日 / 中 / 英）for Windows CE をインストールする方法がある。ただし、システムフォントを書き換えてしまうソフトなので、万人に勧められるものではない。

つい先日、Windows Mobile(TM) 5.0 日本語版のリリースが発表された。Windows Mobile 2003SE からのバージョンアップで多言語関連はどれくらい改善されたのかはまだ全く分からないが、携帯端末との融合を視野に今後も進化は続くだろう。

注

[1] 杜氏工作室（**http://www.dyts.com/gb/products.html**、シェアウェア）。日本語による情報としては谷野氏による解説があるが、やや古いバージョンに関してのものになる。**http://www.rikkyo.ne.jp/univ/tanino/palm/**

[2] XLsoft（**http://www.xlsoft.com/jp/products/togo/index.html**）サポートされるのは表示までで、入力は出来ない。

[3] Becky! のエディタ部分、xyzzy（**http://www.mars.dti.ne.jp/~t-kamei/xyzzy/**）などを利用する。

[4] 昨秋、Symbian OS の開発元 Psion 社は Nokia 社に買収され、同社の PDA、PSION シリーズの命運も尽きた。この買収劇は PDA が携帯端末に飲み込まれる構図を鮮明に描き出すものであった。Sony の CLIE 開発中止決定に続き、東芝が Pocket PC 陣営から撤退する噂が流れているが、これもまた携帯端末が PDA を越える日が近いことを物語るものであろう。

[5] KaoriYa.net BDF UM+（**http://www.kaoriya.net/**） はビットマップフォントを固定サイズ TTF に変換したもの。コンパクトながら Unicode を実装する。GNU Software の中国語簡体字フォント（**ftp://ftp.gnu.org/gnu/non-gnu/chinese-fonts-truetype/** から）もあるが、4MB と大きい。もしメインメモリに入れる余裕がない場合は、Font On Storage（**http://www.geocities.co.jp/SiliconValley-Cupertino/2039/**）を利用して、外部メモリにフォントを待避するとよい。

[6] UKeditor は「UK-taniyama's Homepage」（**http://homepage3.nifty.com/UK-taniyama/**）で公開。このソフトウェアの存在は山田崇仁氏からのご教示による。TextViewer2002 は「うめぷのベルギー日記」（**http://www.tele.ucl.ac.be/PEOPLE/UMEDA/TextViewer2002/**）で公開。本来は英語版 Pocket PC を日本語対応させるためであったようだ。

[7] 著者の環境は iPAQ hx400（WindowsMobile2003SE）に標準で付属する Internet Explorer で UTF-8 による中国語ページを作成して試したが表示できなかった（GB, BIG5 のエンコードには非対応）。完全な UTF-8 対応ではないと考えられる。また、Access 社（**http://www.access.co.jp/top.html**）のブラウザ NetFront（ver3.2 および 3.3Technical Preview 版、シェアウェア）には UTF-8 のほか、GB, BIG5 も用意されているが、同じく実際には表示できなかった。ただ、今後改善される可能性はある。

[8] Time Space System 社（**http://www.tsshop.com/jp/**）でダウンロード販売している。価格は 4,790 円。本来は英語版 OS をローカライズするソフトであり、その特性を理解したうえで利用した方がよい。詳細は不明だが、通常の日本語のシステムフォントを日中韓の文字セットを含む Unicode フォントに置き換えることによって、日中韓の三言語を表示できるようにしているようだ。このため日本語版 PPC であっても日本語サポートも含めて購入する必要しないと、日本語だけ文字化けしてしまう。類似した試みが「★大陸諜報活動新聞★」（**http://asukal.seesaa.net/article/5519071.html**）でも展開されている。この点も山田崇仁氏からのご教示による。