【特別寄稿】

・坂内　栄夫　「人文情報学いまむかし」

筆者は、昨年九月に開催された東アジア人文情報学サマーセミナー「インターネット時代の人文学の技術（スキル）」を聴講させて頂く機会を得た。それに関して自らが専攻する中国哲学史（中国古典学）とコンピューターの関わりについて、文章を求められた。そこで、「人文情報学いまむかし」と題して、自らの過去・現在・未来について書いてみたい。

筆者がコンピューターを利用し始めたのは、八十年代後半の頃である。普通のワープロとして論文製作に利用していたほかに、麦谷邦夫助教授（当時）が MS-DOS 上で動く一字検索システムを TurboPascal で構築されたので、それを利用させて頂くようになった。そこで、検索用元データを作成するため、大型計算機センター（当時）の二階にあった OCR 専用機を使ってプレーンテキストを幾つか作成した。データの種類は、花園大学内の禅文化研究所で禅語録研究会の末席に連なっていた関係上、馬祖・百丈・黄檗・臨済の所謂四家のデータを作成する事にした。当時、入矢義高先生の手によって馬祖・黄檗・臨済は訳注が公開されていたので、この三人に関してはその訳注によってデータを作成すれば、底本についてとりあえずの問題はなかった。ところが、百丈については訳注も研究成果も公開されていなかったため、一般に利用されていた宇井伯寿氏校訂の『百丈広語』（『第二禅宗史研究』所収）を底本にした。これは、『古尊宿語録』や『四家語録』から百丈の説法・問答を集め、そこに『宗鏡録』や『伝灯録』等の引用文との校勘を加えたものである。

しかし、その当時には既に柳田聖山先生の研究により、『百丈語録』の最古のテキストとして、福州東禅寺版や開元寺版『天聖広灯録』（巻九）の存在が知られていた。更に、『百丈広語』の最後に付されている「大乗入道頓悟法門」に関しても、引用元である『祖堂集』と『伝灯録』の間に夥しい文字の異同が存在する。この様な、『古尊宿語録』『四家語録』と『天聖広灯録』との間の文字・文章の異同、また「大乗入道頓悟法門」間相互の異同などの情報は、プレーンテキストの形では保存の方法がなく、異同をカッコに括って書き加えておくしかなかった。更に、個人的に検出していた『百丈語録』の引用文・典拠などについての情報に至っては、有効な電子化の方法を考えつかなかった。「電子データは、プレーンテキストの形で一字検索に利用する以外に有効な活用法はないのか」と考え始めていた当時の筆者はこの段階で早くも挫折し、以後十年近くこの問題については放置し

たままであった。

ところが、近年家人が漢字情報研究センターの武田時昌教授の元に内地留学に行き、C.Wittern 助教授や守岡知彦助手の指導を受け「『説文解字』データベース作成の研究」を始めた事から、XML の古典文献用タグセットである TEI（Text Encoding Initiative）や、漢字を文字コードを離れて扱うことのできる CHISE プロジェクトの存在を知った。とりわけ、TEI によるマークアップを行なえば、プレーンテキストでは保存できなかった文字の異同情報や引用文・典拠という付加情報も有効に保存する事が可能である事。更に、XML から HTML に変換してブラウザ上で表示したりデータ処理する事も可能であり、また XEmacs（Emacs）上に限定されるのではあるが、XML テキストを Elisp により様々にデータ処理することのできるプログラムも開発中である事も知った。

その結果、長い間お蔵入りしていた『百丈語録』の電子データも、コンピューター上で有効に利用する事ができそうだと考えるに至った。そこで、個人的に Wittern 助教授や守岡助手の指導を受けに行き、PC-UNIX 上で動作する「XEmacs-TEI」や「XEmacs-CHISE」の使用法について習熟に努めるようにした。そして、今回自らの人文情報学研修の一貫として「東アジア人文情報学サマーセミナー」を聴講させて頂いた次第である。

今後は、上記『百丈語録』のマークアップを行ない、テキストの異同や典拠等についてすべて電子化し、コンピューター上で利用できるようにしてみたいと考えている。そのための準備として、現在手始めに読書会で読了した南朝梁顔延之の『庭誥』[1]について、中華書局版評点本を底本にして、宋版『冊府元亀』と明版『冊府元亀』の文字の異同を中心に個人的にマークアップを行なっており、それはほぼ完成している。マークアップ完成後、XSL を用いてどのように文字の異同処理や他の文字データ処理できるのか、現状では XSL について殆ど知識がないので、CBATA のソースを参照するなどして色々と勉強し、異同の処理だけでも最低限実現してみたいと考えている。サマーセミナーで岩井茂樹教授の研究成果を見せて頂いた感じでは、かなり柔軟に複雑な処理も行なえるようなので、XSL の処理能力については非常に期待している。その次の段階として、訳注についてもどのようにリンクさせて処理できるのか、検討してみる予定である。また、Elisp によるデータ処理プログラムについては、その開発状況を睨んで色々と利用法を探ってみるつもりである。

なお、この『庭誥』に関しては、諸々のマークアップが完成した暁には宇佐美文理研究代表と相談して、適当な WEB 上で公開できればと考えている。そして、『庭誥』で基本的な問題点・疑問点を解決したのち、本来の目標である『百丈語録』のマークアップに取り掛かりたいと考えている。

現在は、先に少し名称を出したように、TEI のマークアップを行なうのに Knoppix[2]上で「XEmacs-TEI」を使用している。周知のように、PC-UNIX 系 OS はそのインストールから環境設定まで非常に煩雑な作業が多く、UNIX システムに対しての基本的な知識がな

いと、実際問題として作業を行なうにも困難な場合が多い。更に、Emacs 系のエディターを扱うにも、フォントまわりを始めとして様々な設定が全く複雑怪奇で、誰でもが簡単に使用するにはまだまだ敷居が高いのが現状である。しかし、守岡助手がサマーセミナーで配布された「Knoppix-CHISE」[3]は、Knoppix に「XEmacs-CHISE」やフォント等を付け加えて独自に再構成したもので、「XEmacs-CHISE」やフォントの設定及びその他必要な環境設定の類いが予め施されている。その結果、「Knoppix-CHISE」導入後直ちに「XEmacs-CHISE」が使用可能になっており、「XEmacs-CHISE」や「XEmacs-TEI」[4]の普及と利用に多大の福音を齎す可能性を秘めている。今後「Knoppix-CHISE」がよりいっそう進化して、使い勝手のよくなる事を願ってやまない。

[1]「六朝隋唐精神史研究」（代表：京都大学文学研究科　宇佐美文理助教授）の科研報告書の一つとして訳注を公刊するため、現在原稿の整理を行なっている。

[2]　Debian ベースの Linux ディストリビューションの一つ。ドイツの K. Knopper 氏が開発している。それに基づいて、独立行政法人産業技術総合研究所の須崎有康氏が、日本語の使えるように改良した日本語版を配布をしている。URL は以下の通りである。「http://unit.aist.go.jp/itri/knoppix」

[3]　以下の URL で「Knoppix-CHISE」の ISO イメージとルートイメージが公開されている。「http://kanji.zinbun.kyoto-u.ac.jp/projects/chise/dist/KNOPPIX/」

[4]　残念な事に、C.Wittern 助教授が独自に開発された、TEI の設定は取り込まれていない様である。この点については、是非今後の改良をお願いしたい。

7　反省と今後の展望

2004 年度人文情報学サマーセミナーは、期間は１週間で、受講生 10 名の小規模な企画であったが、初めての試みということもあり、実施にあたっては思いのほか大変であった。大きな支障がなく、何とか無事終えられたことは大きな喜びであるが、振り返ってみて至らぬ点、反省すべき点は多々あるように思われる。

最も大きな問題は、やはり講義、演習のどのようなカリキュラムを組めばいいのか、そしてどの程度のレベルで教えればいいのか、ということであった。正規の授業ではなく、一週間で集中的に行うセミナーであるだけに、学問的な基本概念を教えることなく、パソコン操作の実習に終始すると、ますますと初心者向けのパソコン教室的なものになってしまう。また、学生の関心を向けるために、データベースの横断検索や n-gram 処理の応用例を紹介するのも悪くはないが、プログラミングやコンピュータ処理の基礎知識が理解できていないと安直な考え方を植え付けてしまいかねない。そうした危惧を回避するために、TeX と XML にターゲットを絞り、概論と実習をワンセットとし、１日目に入門、２日目に応用を行うことにした。そして、TeX から XML へという流れを理解させようとしたのであるが、果たして狙い通りにいったかどうかは疑問である。

セミナー受講生の感想によれば、初日の TeX の講義と演習は、比較的やさしかったために理解しやすかったらしく、組版ソフトである TeX、LaTeX によって論文を作成することに強い関心を示し、実習後も勉強を重ね、TeX によるレポートを提出してきた者も複数いたくらいである。ところが、その応用となると全体像がつかめずに中途半端な理解に終わってしまった印象が強い。実習では、インストールや基本操作の説明で時間的なロスがあって話の途中で終わってしまい、論理構造の記述するという特性を把握したり、多言語処理の手法という側面は、十分に教えることができなかった点は残念であった。

３日目以降の XML に関しては、XML 自体に馴染みがないために、さらに面食らったようである。受講生の意欲として、テキストデータよりも高度な資料データベースを学びたいという気持ちはあったが、基本的な知識が乏しいために、どのようなパソコン処理をしようとしているのかがわかりにくかったようである。マークアップ概念をもう少しわかりやすく概説したり、受講生が普段取り扱っている文献によってダグ付けの具体的な実習を取り入れてもよかったかもしれない。実習では、XML をパソコン上で走らせることで、親しみを持たせようとしたが、初心者向けの操作法を教える工夫がやや不足していたので、難解であるという印象を持たせてしまったかもしれない。今後の課題としてマークアップ手法の簡便な教材作りは、欠かせないように思われる。

しかしながら、レベルの高い話にまったく興味を示さなかったかというとそうではない。とりわけ師茂樹氏の講義で正規表現を用いた検索の実例を示すことで、XML 文書の構造的な本質を論じたこと、岩井茂樹氏の講義でホームページを閲覧しながら、共同研究会の会読テキストとして『元典章』のマークアップテキストを具体的に示しながら、プログラ

ミングの解説がなされたことに対しては、受講生は知的な驚きを伴って強い関心を抱いたようである。一週間のサマーセミナーでは、中上級の技法を学ぶまでには至らないから、プログラミングはとてもできそうにないと及び腰の受講生もいたが、これからの人文学にはデータベース作成が不可欠になってくることを痛感したにちがいない。だから、そのようなスキルを身につけることができれば是非とも活用してみたいと参加者全員が感じたことも確かである。そこに、サマーセミナーの成果の一つを見いだすことができるだろう。

全体的に振り返ってみて、情報量が多すぎて消化不良であった印象は拭えない。それは、授業のテーマやレベルの問題だけではなく、教え方や準備段階の工夫によって、ある程度は解消できるものがあるように思われる。とりわけ少人数とはいえコンピュータへの習熟度にかなりのばらつきがある受講生を相手に、TeX や XML の実習を行うのは、思った以上に難しかった。基本的な教え方として、インストールからすべて自力で行えるようにするということにしたので、インストールや基本操作に馴れるのに少し時間がかかるのはいたしかたがないだろう。しかし、パソコン操作に気を取られすぎると、何をしようとしているかの理解が疎かになってしまう。したがって、教える内容と手順を十分に吟味しておく必要があることは言うまでもない。１日目の TeX の演習には、比較的馴染みのある EmEditor を使ったのでさほど問題はなかったが、２日目の実習で用いた XEmacs CHISE や３日目の oXygen は、特殊なのでかなり難しかったにちがいない。今回は初めての実習ということなのでいたしかたがない面もあるが、もっとスムースな導入を工夫する必要があるだろう。例えば、基本操作について口頭で説明するだけでなく、すぐに理解できない受講生のために、基本操作のマニュアルがあるだけでも、かなり違ってくるはずである。また、事前に参考書を配布しておいたが、初歩的な基礎知識は、あらかじめ読むように具体的な指示を出したほうがよかったかもしれない。

新しい学問領域である人文情報学の人材育成を究極の目的としているのであるから、短期間のサマーセミナーだけでそれが達成できるわけではないし、本格的な教育の場を設けるべく取り組む必要性がある。しかしながら、たとえイメージトレーニング的なものに終わってしまったのだとしても、人文情報学の構築に向けての試行的な試みとして、大きな意義があったように思われるる。以上の諸々も反省を踏まえて、来年度以降も人文情報学サマーセミナーを継続して実施していきたいと考えている。セミナー受講生には、今回の実習体験を通して得られた技能と知見をさらに深め、今後の研究活動に大いに活用してもらいたいし、アフターサポートの協力も惜しまないつもりである。

８　セミナー資料集

・資料１　「セミナー開会式挨拶」　高田　時雄

我々の COE プログラムと言うのは漢字とコンピュータの心地よい関係を築きたいと言うのが眼目であります。

ご存知の通り、皆さんはほとんど気が付いた時からコンピュータを使っていたと言える世代だと思います。が、我々はずいぶん年を取ってからコンピュータと言うのが入ってきまして、それと格闘した経験があります。

これが 21 世紀の東アジアと言うものを考えると、現在非常に困難な状態にあると言えます。我々東アジアの中国、日本、韓国は漢字を使って文化を形成した歴史がありますが、コンピュータが入ってきてから非常に難しい問題がいろいろ起こってきていました。ご存知の通り、第一にコンピュータで使っている、文字コードと言うものがそれぞれ全部違っていました。中国は GB コード、台湾は BIG ５、日本は JIS コードと言うのがあって、昔は大変困りました。最近は Unicode というのが出現しまして、少しその困難は取り除かれつつありますが、東アジア全体を考えると、まだまだ漢字絡みの問題が沢山あります。現在グローバルな観点からみると、世界的にヨーロッパというのは一つになっていますが、東アジアと言うのは全然なっていません。政治的な問題はともかく、文化的には少なくとも共同でやっていくべきだと思っております。それを推進するためには、漢字の問題と言うものを我々が主体的に取り組んで、解決していくという姿勢がなければなりません。そのためには、これをコンピュータの専門家だけに任しておくというわけにはいけません。特に人文科学に取り組んでいる研究者の立場から、様々な問題を捉え直して、具体的に解決していくと言うことを考えなければなりません。今回、こういうサマーセミナーを開催したのも、まさにそういう意図がありました。皆さんには将来、これを自分の問題として解決してもらいたいと考えております。その出発点が実は今日だと思っております。

京都大学にはコンピュータの入門授業やセミナーがいくつかあると思います。でもそれはコンピュータ一般についての話であって、漢字関係、漢字文化の継承と発展と言う観点からこういうセミナーを開催するのは、おそらく日本でも最初で、東アジア、ラテン地域を考えても初だと思います。中国はおそらく中国国内のことしか考えていませんし、韓国もそうだと思います。日本も基本的にはそうですが、日本ではもう少し、国際的な視野で物事を考えようと言う気運が昔からあったと思います。今回、受講生の皆さんがたまたま東アジアのいろんな国から来ていますので、皆さんが今日を出発点として 21 世紀の東アジア人文情報学の戦士になっていただけたらと考えております。一週間という短い時間ですが、講師の先生方はまさにこの問題の専門家ですので、細かい問題、基礎的な問題でも遠慮なく聞いて、小さな問題から解決してスキールを高めていけたらと思います。このセミナーが終わった後でも、先生方にはメンテナンスをしていただけると聞いておりますので、COE プロジェクトの拠点の方に足を運んでいただけたらと思っております。よろしくお願いします。

資料２　セミナー教材テキスト（附録参照）

※　編集後記

本報告集は、2004 年 9 月 6 日（月）－ 10 日（金）に実施した２００４年度東アジア人文情報学サマーセミナーの実施報告書である。編集は、教育部門リーダーの武田時昌が担当した。セミナーの講義、実習記録に関しては、実習指導員の一人であった COE 研究員の山本一登氏が作成したレポートに加筆したものである。

なお、セミナーの会場については、金坂清則教授にお世話いただいた。また、セミナーのビデオ撮影、記録の整理には、ノートルダム女子大学３年生車愛順さんの協力を得た。ここに感謝申し上げる。（武田記）

セミナー教材テキスト