文学部光学的欧文文字読取システムについて

文学部文学科ドイツ語学　竹内義晴

始めに

前年度の臨時経費の要求を認めていただいて，OCR（Optical Character Reader）を買うことができた。「なんだ，ＯＣＲか？」と拍子抜けされる向きも多いかと思う。私だって，大型機につけられたＯＣＲを見て「すげぇ！」と思ったのは，学生時代のことだから，サバを読まずに２０年ぐらい前のことだ。日進月歩の世の中で，ＯＣＲごときで驚いていては笑われてしまう。たまたま，話を洩れ聞いたパソコンメーカー（の営業）は，「桁が違うのではないか？」と言った。確かに今時ＯＣＲは数十万円の代物なのだ。それを八百万円いただいて購入したのだから，何がしかの説明が必要なのだろう・・・と思っているところへ，原稿の依頼が飛び込んだ。渡りに舟である。

以下，若干のスペースを借りて，文学部で購入した光学的欧文文字読取システムの紹介をさせていただく。（数十万円にしても八百万円にしても，そういう価格の物が経常の予算で買えないし，買って貰ったと浮かれる所に，文系の研究条件の貧困があらわれていると思ってくださって結構です。）

１．　ＯＣＲって何だ？

ＯＣＲを「Optical Character Reader」と綴って見せても，漢字で「光学的文字読取機」と書いて見せても，わかる人にはわかるし，わからない人にはわからない（なんだこりゃ？）。禅問答（かけあい漫才？）のようになってしまったが，要するに，人間が文字を読むことの不思議と偉大さ（大変さ）を，受けとめていただけるかどうか，が問題なのかもしれない。「計算機はそのままでは文字が読めない」という作文が理解された結果，私の要求を認めていただけたのだろう（と私は思っている）。

郵便番号の機械読み取りの歴史もすでに長くなるし，今時の人はコンピュータが文字を読むぐらいはあたり前と思っているかもしれない。もちろん，パソコンにつけるＯＣＲが数十万円の時代なのだから，コンピュータが文字を読むのが「あたり前」であることはその通りなのである。この頃の郵便局では，手紙の宛名そのものを機械に読み取らせてさえいるのだから。しかし，この「あたり前」の構造は，すごく単純化しても，段落一つ分の記述は必要とするだろう。

例えばこのページ面の全体には，文字が印刷されて並んでいる（だろう）。これは，即物的に言えば，インクの染みの散らばりが，光を反射する部分と吸収する部分との模様を形づくっているだけで

ある。計算機はまず，この模様を画像として取り込み，その模様の全体を，小さい部分模様の並びとして整理する。並びの順に，一つづつ部分模様のマトマリを切り取り，画像データとしての文字のパターンの照合をおこない，特定の文字であると同定する文字の同定ができたマトマリに文字のコードを与え，文字データの並びに加える。このようにして，始めて，画像データは文字データとして計算機に取り込まれたことになる。

以上のプロセスには，文字のカタマリの切り取り方の問題，活字面の汚れ，活字の太さ，大きさ，傾き，曲がりや折れのクセ，にじみ，かすれなど，さまざまな要因がからんでいるから，実際にはこの照合作業は単純ではない。しかし，いずれにしても，このステップを乗り越えれば，コンピュータは文字を読むのだし，そして，そのこと自体は「当たり前」の時代になってきてはいる。

ところで，何が面白くて，コンピュータに文字なんぞを読ませるのだろう？別に面白い訳ではないが，郵便番号を機械に読ませれば，仕分けの人手が省ける。コンピュータに文字・テキストを読ませれば，キーボードから入力する手間が省ける。

コンピュータに入力する内容は，すべて自分の脳のプロセスを経たものに限る。印刷物を書き写すなどという非生産的なことは一切しない。という人には，コンピュータに文字を読ませる必要はないのかもしれない（私なども実は，そんな「脳様」の信奉者の一人である）。しかし，「脳様」に充分に働いていただくためには，やはり，「脳様」を機械的な作業・無駄な作業から解きはなってさしあげなくてはいけない。また，「脳様」に充分に働いていただくためには，効率よく栄養を補給してさしあげなくてはいけない。そして「脳様」の栄養は知識，とりわけ，書物からの知識なのである。私は，思考，実験，対話，体験，夢想，放心などの大切さを軽視するわけでは決してない。それらが大切だからこそ，栄養は，時間と労力を節約して，効率よく与えなくてはならないと思うのである。

人工知能などということばが先走りしたところで，現代の実用レベルのコンピュータに期待できることといったら，実際の問題としては，大きな規模のデータの蓄積・処理にすぎない（間違っていたらごめんなさい）。ワープロ・データベースなどによる，小回りのきくデータ処理もまた，結局の所，この範疇に属することがらなのだろう。とするなら，書物のデータを，キーボードから入力する必要がある場合に，コンピュータに直接読んでもらったらどうだろう。さらには，自分の必要に応じて自前のデータベースを作り上げておけば，必要なデータをコンピュータによって簡単に取り出せる。そのために，印刷されたデータをコンピュータに直接読ませて活用したらどうだろう。

「学際」ということが叫ばれる今日，自分が深くかかわりを持たなかった分野の知識にアクセスする必要に迫られるということも増えている。他方，諸科学の細分化もまた，止むところを知らない。自分の目的にかなったデータベースがないとか，既存のデータベースがその時々の目的にとってキメが細かすぎるとか，荒すぎるとか，そりが合わないということがあるだろう。そんな時，そんな必要に応じて，書物をコンピュータに直接読ませて，自前のデータベースを作りたいと思ったことのあるのは，私ばかりではないだろう。

「コンピュータに直接読ませて」というのはなぜか。それは，私たち自身が文字・テキストを読み

取り，キーボードから入力する作業が，たいへん肩のこる，厄介な仕事だからに他ならない。私たちの目と手は，そしてそれを統括する脳は，この大変な仕事にいつもくたびれ果てている。私は私の「脳様」や，それにつながっている色々な神経を早く楽にしてさし上げたい。それは私自身を楽に快適にすることなのだから。

以上，人間が文字を読むことの不思議と偉大さ（大変さ）について，だらだらと紙面を費やした。これでＯＣＲの概念について，およその「ＡＨ！」に到達していただけただろうか。ＯＣＲは私たちの視神経や脳のかわりに活字を読み取って，コンピュータのファイルに書きこんでくれる。

2．　どんな内容のシステムを揃えたのか。

今日では「当たり前」の，数十万円で買えるＯＣＲがどうして八百万円したのか。それを語るには，日本語・英語以外の言語を用いる文化圏を対象をする研究者のうらみ・つらみの物語から始めなくてはならないが，ばかばかしいので省略する。「国際化」の日本とか言っちゃっても，ＮＥＣ　なんかの日本のコンピュータは，英語と日本語しかできない，それ以外の言語は，特別な処理をしないと扱えないのである。だから，英語や日本語を読むＯＣＲだったら，数十万円で買えたのでけれど，そうでないものが欲しかったからそうなった　。経験のない人は「そんな馬鹿な！」と疑うかもしれないが，これはこれまでの所，どうしようもない現実である。（ＩＢＭコンパチ機の普及で，この辺の事情は変ってきたようでもある。）

システムの構成はどうなっているのかというと，サン・マイクロシステムズのワークステーション Sparc station 2GX に，ゼロックス・イメージング・システムズの OCR，Scan-worx が付いているだけである。ソフトは，ＯＳと欧文ＯＣＲのソフトの他には，ＤＴＰソフトを付けることができた。予算の関係からプリンタを買うことはできなかった。このシステムを学内ＬＡＮに接続してある。

卓上コピー機のような外見のスキャナーがＯＣＲの機械で，机の上にどんと威張っている。この機械は，紙の上の画像情報を画像情報として読み取るのだから，コピー機と似ているのも当然だが，コピー機と違って，出力はコンピュータに送られる。スキャナーの操作はＯＣＲのソフトウエア，icr（intelligent character reader）によってワークステーションから行われる。オートシートフィーダによって，オリジナルのシートを自動的に取り込むこともできるが，本をめくったりするのはやはり人間がしなければいけない。この点はコピー機と同じだ。（このスキャナーは　400dpi のすごくよい解像性能を備えているのだそうです。）

ｉｃｒのソフトウエアは，更に画像情報として取り込んだ紙の上の「模様」を，文字として認識し，文字情報としてファイルに書き込む。このソフトウエアの特徴は，辞書と学習機能などがついていることだ。読み取り言語として，例えばドイツ語を指定すると，ドイツ語の文字を読むばかりではなく，単語や簡単なコンテクストについての知識を参照しながら文字の認識を行ってくれる。また，

ある雑誌なら雑誌の文字のクセを学習させて，文字認識に利用・再利用することができる。

コンピュータに文書を読ませる場合に，文字の読み間違えは必ず起きてしまう。このことは困ったことだが，例えばコンピュータのプログラムだったら，ソフトウエアから文法のチェックをかけて，ある程度の間違いを探し出すことができるだろう。なにしろプログラム言語は形式言語なのだから。ところで，今回ＯＣＲを購入するについて，コンピュータに文書を読ませて文科系の教育・研究に役立てるという目的を設定した。この目的設定においては，読み取らせるテキストは，ほとんどの場合は自然言語によって書かれているだろう。私の知る限り，自然言語については，実用に耐えるような文法チェックのシステムは，私たちの「脳様」にしかインストールされていない。しかし，コンピュータに読ませたテキストを自分で読み直していては，大変な手間がかかってしまい，ＯＣＲシステムの導入が省力化・効率化ということにはなかなかならない。

この問題は，基本的には解決できない問題ではある。しかし，多少なりとも，この文字の読み間違えの問題に対処するために，ＯＣＲの辞書機能に加えて，さらにＤＴＰソフトウエアのスペルチェック機能を利用することを考えた。英語の他にドイツ語・フランス語など欧文１３言語を扱うことのできる ＤＴＰ－ソフトウエア，Interleaf ５が，システムに組み込まれていて，読み取り結果にスペルチェックをかけることができる。

Interleaf ５を使って，読み込んだファイルについて，スペルチェック以外にも様々な加工が可能である。しかし，予算の関係でプリンタを購入できなかったので，残念ながらInterleaf ５の備えている高度な印刷機能はシステム内部では使えない（ＬＡＮを通じて，他所のプリンタを使うことはできると思うのだけれど，まだ研究していない。ご近所で ｓｕｎ系のプリンタを使わせてくださる方がいないだろうか。けちな話を更につけ加えると，日本語ＯＣＲソフトも購入できなかった。それぞれに五十万円ほどづつ予算が足りなかったのである（予算の見通しを誤ったのも私ですが）。

以上のようにコンピュータに読み取らせ・加工したファイルは，システムが学内ＬＡＮに接続してあるので，研究室や総合情報処理センターの端末から，取り出して，それぞれの目的に応じて利用することができる。

3． 研究に役立てる可能性について

コンピュータに読み取らせたテキスト・文字データをどのように利用するかについては，すでに，多くの方がたが，それぞれに色々なことをしていると思う。コンピュータの利用の仕方が，それぞれの研究の目的・様態，そして研究者のアイディアしだいで変幻自在であるように，データの利用にもやはり枠組などありえない・・・と言って納得していただける方には，以下述べることは蛇足なのだけれど，いくつかの私が見聞きした使い方を簡単に羅列・紹介する。

文献データベース・判例データベース・雑誌記事データベースなどの制作：

それぞれの目的に応じたテキストを読み取り，著者・項目・日時・キーワードなどのインデックスを付けて，リレーショナルデータベース・カード型データベースなどに組み込む。データの量が増えれば増えるほどに，必要な情報を素早く検索し利用することができる。

哲学や文学作品のテキスト・データベースの制作：

文献そのものを読み取り利用する。例えばある哲学者がある用語をどこかで使っていた。それを調べたい。そんな時，コンコーダンスがあると便利であるが，さもないと，記憶が曖昧な場合には（たいてい，記憶というものは曖昧なものである！），こんな作業に莫大な時間がかかってしまう。ところで，もしその哲学者の全集を読み取ったテキスト・データベースがあれば，その個所をコンピュータに探させることができる。また，テキスト・データベースを使って，コンコーダンスを作製するという作業も実際に行われている。または，文の長さ，使用語彙の偏りなどを調べて，文体研究に役立てることもできる。

言語データベースの制作：

ある言語のテキストをコンピュータにできるだけ広範に読み取らせておけば，語彙の使用例を採取したり，または語彙の使用頻度などの言語についての情報を調べることができる。語彙の使用頻度の情報は，語彙を辞書や教材に採否する場合の判断材料になるだろう。

などなど。

私はといえば，ドイツ語学という言語研究が畑なものだから，言語データベースを使って語彙の使用例を集めるという作業に，言語データベースを使っている。「単語の意味」などというのはキッチリとまとまった「固い」ものだ，というのがどうも，広くいきわたっている，素朴な言語観のようだ。だから，学者は，一つ一つの単語（学術用語）の定義にこだわり学生をいじめるし，聖職者や統治者は「ことばの神聖」を汚すありとあらゆる「世俗」に噛みつき，場合によっては弾圧を加えてきた。ところで日常の生活・常識，または自由な知性の目からみれば，「キッチリと固いことばの意味」なんてあるわけがない。実際に言語データベースを使って語彙の使用例を集めて確かめてみると，このことはよくわかる。一つ一つの語彙は，とても気ままに柔軟な使われ方をするものなのだ。

実は，自由なものの捉え方のできる賢い人には，そういうことは確かめなくても自明のことなのだろう。だから学者や宗教者，～主義者の枝葉末節にわたるいさかい（論争）は決して無条件に尊敬されつづけていたわけではない。しかし，いくら賢い人でも，実際に確かめてみなければ，どのようにことばの使用は自由なのかを具体的に示すことはできない。具体的な事例によって誤解を解きほぐしていく方法はいつの世においても科学の正道なのであり，具体的な事例の整理・収集は豊かであればあるほどに意義深いことでありつづける。

長くなるのでこのくらいでやめるが，私自身は具体的な語彙の使用例の分析結果を拠り所にして，人間の言語生活の豊かさを言語システムと知識システムの柔軟なインタラクションに基づいて説明しようとする試みを続けている。「認知科学」の時代といわれる今日では，言語システムと知識システムのインタラクションという考えはすでに新しいものでもなんでもない。しかし，重要なアイディアの中身を具体的な事実の分析に基づいて詰めてゆくという作業は，議論の展開の新しいステップに向けても欠くことができないものだと，私は考えている。

旅費にも，在外研究の機会にもこと欠く地方の（日本の？）国立大学の研究環境は，言語研究，特に日本語以外の言語研究をするものにとって，けっして恵まれたものではない。しかし，コンピュータを利用して言語事実を収集・分析するという具体的な作業の可能性が開かれていることは，少なくとも一つの救いである。今回ＯＣＲによって，言語事実を収集する言語データベースを自前で作製できるようになった。私の研究環境にとっては，これは一つの大きな前進である。

4．　利用の御案内

このＯＣＲのシステムは現在，文学部・法学部・経済学部の三学部の情報処理センター角間実習室に設置してあり，利用するためには，このシステムのワークステーションのユーザになる必要があります。ユーザ登録のメンドウは私がみているので，利用希望の方は文学部・竹内まで連絡ください。

マニュアルなどは，紛失すると困るので，私が保管して（抱え込んで）いますが，「簡単マニュアル」なるものを作製して機械の傍においてあります。利用の仕方そのものはそれほど複雑ではありませんから，「簡単マニュアル」を見ていただければ，使いこなせるようになると思います。といっても，「簡単マニュアル」は，私が暇を見つけてでっち上げたもので，不備が沢山ありますから，ユーザになった方は，「簡単マニュアル」にドンドン書き込みをして改良してください。不明な点は，私の所へきていただければ，正規のマニュアルをお貸しするか，または時間があれば一緒に頭をひねりましょう（，または必要に応じて，利用講習会のようなことを企画してもいいと思います）。

正直に言って，新しいシステムを使いこなすのには，やはりかなりの忍耐がいります。そして，例にたがわず，省力化のために導入したシステムは，新たにその使用に付随した作業を要求します。ですから，非力なシステム管理者の私としては，システムの利用者間に適当なネットワークが形成されて，この「じゃじゃ馬」を乗りこなすための知恵が交換できるようになることを，実は切に願っています。　　（この項だけ「です・ます」で書いた意図をくみ取ってください）

竹内義晴，金沢大学文学部・文学科・ドイツ語学
920－11　金沢市角間町
e-mail take@icews1.ipc.kanazawa-u.ac.jp
TEL(0762)64-5348