日本IT書紀

229 日本語処理

11 嚇躍篇
巻之三十 恢弘

佃 均

第二百二十九

日本語処理

一

ＦＯＳ海外調査団の報告書に呼応するように、新しいオフィス向け情報システムの概念が次々に登場した。

「インテグレーテッド・エレクトリック・オフィス・システム」（ＩＥＯＳ）

「インテグレーテッド・インフォメーション・システム」（ＩＩＳ）

「オフィス・インフォメーション・システム」（ＯＩＳ）

「オフィス・オートメーション・システム」（ＯＡＳ）

といった和製英語が次々に作られた。

六〇年代後半に「ＭＩＳ」が一大センセーションを巻き起こしたように、どういうわけかコンピュータ業界には十年おきに新しいブームが到来した。

旧来のダム端末をインテリジェント・ターミナルに入れ替えてオンライン・システムの高度利用を可能にするという考え方もあれば、オフコンを中核マシンとしてオフィスに小規模なネットワークを構築するという提案もあった。

そのうち――ほぼ一九七九年の後半に入ってだったが――、ＯＣＲとファクシミリ（ＦＡＸ）が脚光を浴びた。

のちにＯＣＲやＦＡＸはソフトウェアとして提供され、パソコンとスキャナーを組み合わせればＯＣＲにもＦＡＸにもなる。こんにちにいたってＦＡＸは、アナログ時代のシンボルのように受け取られるようになった。

しかし七〇年代末の時点では、ＯＣＲ、ＦＡＸの新機種が出たといっては新聞・雑誌が紹介記事を載せ、数台を導入したと聞いては記者が飛び出していく状態だった。

七九年七月十六日付「情報産業新聞」が、

「シー・エス・シーが米スキャンデータ社と提携して大型・高性能ＯＣＲの販売権を取得した」

という記事を載せている。

それを読むと、ページドキュメントＯＣＲ「２２５０―１型」はスタンドアロンでアルファベット、数字、記号を毎秒一千六百文字の速度で読み取ることができ、価格は八千五百万円から一億円とある。毎秒一千六百文字の処理性能には、

「ＯＣＲフォントと呼ばれた変形文字を使えば」

という条件がついた。

八月になると東芝が、

「手書きのカナ文字を読み取ることができる低価格なＯＣＲを発売した」

と発表した。

郵便番号制度に対応し、手書きの数字を読み取って仕分けできる装置を東芝は開発していた。その技術が民生用に生かされたのである。

それは「ＯＣＲ―Ｖ３９５」というマシンで、読み取り速度は毎秒五百字、価格は一千七百十万円だった。

ほぼ同時期のＦＡＸはどうだったかというと、同年七月二日付に

「東芝、高速ＦＡＸ発売／デジタル帯域圧縮型」

という見出しがある。

東京芝浦電気（岩田弐夫社長）はこのほど、原稿読取幅調節機能および縮小電送機能を持つ高速ファクシミリ「ＣＯＰＩＸ８１００」「同９１００」を発売した。ＣＯＰＩＸ８１００はＢ４判標準原稿を約四十秒、同９１００は約二十秒で電送できるデジタル帯域圧縮型の高速ファクシミリで……価格は８１００が三百二十万円、９１００が三百七十万円。

とある。

同年七月二十三日付の紙面を見ると、

「浜名湖競艇企業団がレース情報の配信にＦＡＸを使うことになった」

という記事が見える。（原文ママ）

浜名湖競艇企業団（紫田郁之進理事長）は、七月中旬をめどに、同企業団が開催している浜名湖競艇の全レース情報をファクシミリネットワークにのせる。このため今年四月に東芝の感熱高速ファクシミリＰＢ―４８００およびＫＢ―４８００の採用を決め、準備を進めていた。

浜名湖競艇企業団は浜名湖競艇の主催者で、これまでレース情報をサンケイスポーツ、デイリースポーツ、日刊スポーツ、スポーツニッポン、報知新聞、東京中日新聞、静岡新聞、中日新聞の八紙に送り、レース番組、レース結果を載せているが、このデータを各新聞社に送るのに六人が一時間四十分かかって電話で読み上げていた。

これでは時間がかかるうえ、誤りが出るなどの問題があり、データ送り事務の合理化策に追われていた。（中略）

ファクシミリネットワークの採用によって、同企業団では、従来六人かかっていた連絡事務がわずか一人で、しかも三十分とスピードアップされるほか、電話連絡によるミスがなくなるとしており……。

二

こうした中で課題とされたのは「日本語」というものだった。

高千穂交易にシステム・エンジニアとして勤めていた蓮生重剛が初めてコンピュータで漢字を打ち出すことに成功したのは一九六八年の五月十二日だった。蓮生はインパクト型ラインプリンターの記号「@」を使って、「美」「花」「家」の三文字を打ち出すことができた。（第百三十二「されど漢字」）

それから四年後に本稼動した日本経済新聞社の「ＡＮＮＥＣＳ」、朝日新聞社の「ＮＥＬＳＯＮ」は、最初は行単位の文字列を自動的に鉛の活字にし、次にコンピュータに記憶させた文字フォントをイメージとして出力した。コンピュータを使った写植システムと言い換えてよかった。（第百四十九「ＡＮＮＣＳ」、第百四十八「情報産業」）

たしかにそれはコンピュータによる漢字処理には違いなかったが、自由度がなく、たいへんに高価なものだった。

――もっと気軽に漢字を使いたい。

――日本語の文書を作りたい。

という要望は根強かった。

コンピュータが打ち出す帳票の項目、請求書や納品書など伝票、送り状やダイレクトメール、顧客リストやさまざまな台帳……。カタカナでは読み難いばかりでなく、しばしば間違いが生じた。同姓同名、同音異字ということが、この国では珍しくない。

データエントリ専用機では漢字処理の技術が実用化されていた。メモリー容量の増大と外部記憶装置の小型化によって、漢字フォントと漢字コードをコンピュータに登録することが可能になった。あとはどうやって漢字を出力するかだった。

出力というのには、二つの意味があった。

一つはプリント出力装置だった。インパクト型のラインプリンターでは不可能だった。ゼロックス社のパロアルト研究所で開発されたレーザープリンターが、それを解決した。

七九年七月、電電公社の横須賀通信研究所は毎分一万五千文字を出力できるレーザープリンターの開発に成功していた。八月には日立製作所が「Ｈ―８１９６０―３０」の名で発売した。レンタル価格は月二百三十万円と記録されている。

もう一つの意味は、コンピュータに登録されている漢字フォントを呼び出す方法だった。文字ごとにコードが付い

ていたが、オペレーターがそのコードのすべてを間違いなく記憶するのは不可能に近い。誰もが容易に呼び出すことができる方法が必要だった。

最も簡易なのは和文タイプライター方式だった。漢字そのものを表示した大きな盤を電子ペンでタッチする方法である。

この場合だと当該の漢字を探し出すのに時間がかかった。囲碁の岡目八目ではないが、脇で見ているとどういうわけか目的の文字がどこにあるか分かる。オペレーターが見つけ出すまでイライラして待たなければならない。胃のためにもよくなかったし、それに常用漢字約二千文字が限界だった。

漢字が偏と旁でできていることに注目したのは、インフォレックス社の入力システムを販売していた伊藤忠データシステムだった。同社が扱っていたアメリカ製のミニコン「ＷＡＮＧ」は、実は台湾出身の王という人がアメリカに設立したメーカーで、漢字処理技術を組み込んで台湾や中国に販売していたのである。

「相」という漢字を呼び出すには「木」「目」と入れればいい。同じ音の「愛」なら「ノ」「ツ」「ワ」でいくつか類似の文字が表示されるので、その中から選ぶ。三つの特徴を指定して当該の漢字を呼び出すところから「三角偏号法」と呼ばれた。

一九七一年に川上晃が速記用に開発した「ラインプット」という手法をコンピュータに取り込む動きもあった。基本は速記用タイプライターの鍵盤操作法で、二十個の文字キーを左・右・中の三グループに分け、それぞれに操作する指を固定させるというものだった。

漢字に関連するモノや事がらをカタカナ二文字にして呼び出す方法も考案された。筆者は以前、その具体例をどこかで書いた記憶があるのだが、何せ昔のことなので覚えていない。それで〝喩え〟で書くのだが、つまり「ミラ」と入力すると「鏡」という文字が表示され、「美」という漢字を探すには「ミロ」と入れる。

ミラが鏡なのは、英語で「ミラー」だから、ミロが美なのはミロのビーナス（美ーナス）だからである。駄洒落まがいの入力方式だった。これは「連想入力方式」と呼ばれた。

ただし以上の方式はプロ向きだったし、漢字、ひらがな、カタカナをキャラクターとして扱うに過ぎなかった。オフィス文書を作るのに一文字ずつ呼び出していたのでは、手で書いたほうが早い。また、文字コードはキーオペレーターの頭の中にあって、漢字への変換はつまり人が行っているに過ぎなかった。

三

コンピュータで漢字交じりの文章を生成するということは、コンピュータの技術でいうと自然言語処理、より平明にいえば文法というものをコンピュータに理解させなければならない。

一九六四年ごろから、九州大学でコンピュータによる日本語の文法解析が研究されていた。担当していたのは同大学工学部電子工学科の教授・栗原俊彦の研究室である。同研究室は文節の抽出、単語辞書の使用、構文解析法など、仮名漢字変換の基礎的な手法を開拓した。

ＯＫＩＴＡＣ機を使用した関係から、処理プログラムの開発と単語辞書の作成には沖電気工業が協力していた。少し遅れて京都大学の長尾真も日本語の構文解析をコンピュータで行う研究をスタートさせていた。

一九七三年のこと、日本放送協会（ＮＨＫ）の放送技術研究所に勤務していた相沢輝昭と江原暉将という二人の研究員が「計算機による漢字変換」技術を開発した。九州大学栗原研究室の成果を利用して、海外からテレックスなどで送られてくるニュース文を漢字交じり文に変換するのが目的だった。

カタカナだけでできている文章をそのままアナウンサーに渡しても、すぐに読み上げることができない。それを読解し、漢字交じりの日本語文に清書するだけでもたいへんな労力がかかっていた。

開発したシステムはＩＢＭシステム／３６０で動いた。テレックスで送られてくる電文をコンピュータが解析して文節に分け、品詞と接続詞に分けて漢字交じり文として出力する。カタカナが自動的に漢字に変換できるというのは画期的なことだったが、処理能力の関係から「最長一致法」が採用された。

試しに約七千の文節で構成される新聞記事を試しに入力すると、正しく変換された率は七七・五％という成果を得た。「文節分かち書き変換方式」が具体的な形になった。試作システムとして約八割という変換率は悪くなかった。しかしこの程度では実用に耐えることができない。

一般オフィス業務にコンピュータを適用するには、つまり何が何でも簡易な日本語入力方式と精度の高い変換技術が必要だった。

ここに森謙一という技術者がいる。

一九六二年に東京大学工学部を出て東芝総合研究所に入った。まず取組んだのは磁気コアメモリーの開発、次がＯ

CRの文字認識技術だった。

　OCRフォントと呼ばれる特徴のある文字でなく、手書き文字を正確に読み取るにはどうすればいいか。目の前にあったのは郵便番号制度だった。手書きの数字を読み取ることができれば、OCRの用途は格段に広がる。

　森は文字のパターンをコンピュータが認識する手法を編み出し、ついに「自由手書き郵便番号自動読み取り区分機」を実現した。その基礎技術研究を終えて一息ついていた七一年のこと、新聞社の整理を担当していた人物との雑談のなかで、

　――日本の記者は欧米の記者に比べて記事を書くのが遅い。彼らはタイプライターを使い慣れているが、日本の記者は相変わらず鉛筆を使っている。これを解決する道具を作れないか。

という話が出た。

　――ほう。それは、例えばどんなものですか。

と訊ねると、

　――三つの条件がある。

という。

　一は手で書くより速く、二は一般のオフィスで使え、三は作成した文章を電話回線で送信できること。

　――面白そうだ。

　森は思った。電子技術を生かせば不可能ではあるまい。

　ところが取組んでみると生やさしい話ではなかった。日本語の基礎研究から始めなければならない。

　東芝の研究部門に「アンダー・ザ・テーブル」という制度があった。海のものとも山のものとも分からない技術を模索するために、研究予算の二割を割くものだった。どのテーマにヒト・モノ・カネを割り振るかは技術長・森の権限の内である。

　まず部下で九州大学工学部出身の河田勉という技術者を京都大学に派遣し、長尾研究室で学ばせた。日本語の構文解析から始め、新しい動詞分類を組み上げ、新聞や雑誌に登場する熟語や用語、名詞、動詞などの頻度を丹念に調べていった。

　当時、一般的だった最長一致法では、「ひとは」という入力に対しコンピュータが「人は」を見つけた段階で作業が終了する。最後から一文字ずつ合致する文字列とぶつかるまで落としていくわけだから、「日とは」「火とは」は候補にあがってこない。

　そこで研究班は可能性のあるすべての組み合わせから、文法的にあり得ないものを捨てていく方法をとることにした。

　日本語に多い接頭辞、接尾辞の処理も必要だった。

何よりも力を注いだのは頻度情報を採用することで同音異義語に優先順位をつけることだった。京都大学の長尾研究室との共同研究で、分野や目的によって使用する言葉に偏りがあることが分かっていた。使用した語彙の頻度を機械に記憶させ、使用度の高いものから順に表示させるのである。

「貴社の記者は汽車で帰社する」

という同音異義語の代表例がある。これをクリアできるかどうか。

辞書の整備の次に課題となったのは入力方式と変換方式だった。入力方式では和文タイプライターの文字盤方式、キーボードを用いたプロ用の三角偏号法、連想方式、多段シフト方式などがあった。

それぞれにメリットとデメリットがあり、森はいずれにも決めかねた。コンピュータの素人が簡単に使えなければ役に立たないのだ。

四

――面白い入力方式を開発した会社がある。

という情報がもたらされたのは七七年のことだった。

東京・九段下にある国際プログラムサービス（ＫＰＳ）という会社がローマ字で漢字を表示する仕組みを作った、というのだった。

キーボードは英文タイプライターに準じてアルファベットが配置されている。小学校のときローマ字の表記を学び、キーボードを使い慣れている人にとって、ローマ字であれば容易であるに違いなかった。

それまでもカナ漢字変換方式は存在していたが、元になるカナを入力するために連想方式を用いなければならなかった。ところが新しい方式では、まずアルファベットでローマ字を入力する。

「ＡＩＳＡＴＳＵ」と入れ、変換キーを押す。すると「あいさつ」というカナ文字になる。ひらがな、カタカナの場合はＥＮＴＲＥキーで確定させ、漢字にしたいときはもう一度変換キーを押す。

語彙頻度と熟語辞書によって「挨拶」という熟語が表示される。かくしてここに「ローマ字入力カナ漢字変換方式」が確定した。

製品化を目指してマシンの設計が始まったのは七七年の春だった。

この時点でバトンは森から青梅工場の部長だった溝口哲也に託された。

前後の経緯を溝口は次のように語っている。

「当時、東芝の大型コンピュータ事業は難しい状況にあった。このへんで方向転換を図らなければ、と話し合っていたとき、総研の森さんのグループが面白そうな研究をしている、という話があった。それで早速、自分の目で確かめようと思った」

溝口は川崎駅の売店で新聞を買った。読むためではなかった。その新聞を持って東芝総合研究所の森研究室に乗り込み、自らキーボードを叩いて新聞の論説記事を打ち込んでいった。

「すると漢字まじりの文章がディスプレーに表示される。あのときの感激は忘れられない」

その翌日、興奮冷めやらぬ溝口から報告を聞いたのは、のちに常務となる天羽浩一である。七八年の一月、ＪＩＳに漢字コードが制定された。さっそく森は漢字辞書にＪＩＳコードを組み込んだ。

間もなく試作機はできた。

表示できる文字は六千八百二種だった。だが価格は二千万円以上、大きさはオフコン並みだった。これでは製品として成り立たない。

溝口らは設計をやり直し、削れるものはすべて削った。青梅工場で開発を進めていたミニコンのプロセッサーとオフコンの技術を応用することにして、何とか事務用机と同じ大きさにすることができた。

この間、商品化するかしないかの決定を前にして、突然のように開発中止の命令が出た。痩せても枯れても東芝はコンピュータ・メーカーではないか。文書作成機は事務機である。その事務機にこれ以上の開発費をかけるわけにはいかない、という。

――いちどでいい。実際に見ていただきたい。

溝口は粘った。

エンジニアであれば、自分が味わったのと同じ感激を得るに違いなかった。

一度限り。

という条件で性能テストの許可が出た。

森も必死だった。機械の操作をタイピストではなく、総務の女性事務員に選んだ。大ばくちだった。

以上の物語はＮＨＫが「プロジェクトＸ」という番組で放送した。

七八年十月、東芝はデータショウに初の日本語ワープロを発表した。

「ＴＯＳＷＯＲＤ　ＪＷ―10」である。

価格は六百三十万円だった。

翌七九年の二月に出荷が開始されるや、他メーカーがたちまち参入した。シャープ「書院」、日本電気「文豪」、富

士通「ＯＡＳＹＳ」等々である。

富士通は汎用コンピュータの日本語処理システム「ＪＥＦ」との連携を売り物にし、企業ユーザーに受け入れられ、東芝の「ＪＷシリーズ」は新聞社や中小企業に入っていった。

八二年、マイクロコンピュータの技術がここに生かされた。同年の十月、東芝が発売した「ＪＷ―１」は八ビットマイコン用のＤＯＳ「ＣＰ／Ｍ」を八インチのフロッピーディスクで供給され、漢字フォントは二十四×二十四ドット、プリンターはワイヤードットのインパクト式、入力方式はキーボードと文字盤の二通りから選択できる。価格は五十万円台だった。四年間で性能は大幅に向上し、価格は十分の一以下になった。

日本語ワープロが実現した八〇年代以後も、

――添え状や社内文書は手書きであるべきだ。

という考えかたが根強かった。

しかしビジネスに使う見積書や納品書、仕様書、住所・氏名などは、間違いが発生しないようにするためにコンピュータで打ち出した活字体の文字が好まれた。いちど作った文書をフロッピーディスクやカセットテープに保存しておき、採用できるのもメリットだった。

アメリカ合衆国では表計算ソフトがパソコンの普及につながり、日本では日本語処理ソフトが「日本語ワープロ」という専用パソコンに結びついた。日本には暗算とソロバン、電卓があったからかもしれない。

～～～～　補　注　～～～～

シー・エス・シー　一九六五年四月に設立されコンピュータ関連機器を輸入販売した。主力はＯＣＲ、キーツーディスクのデータ入力装置、カード発行装置などだった。東京・青山のハザマビルに本社があった。

蓮生重剛　はすお・しげつよ：香川県に生まれ高千穂交易を経て日本アウトソースを創業した。第百三十二「されど漢字」参照。

川上　晃　かわかみ・あきら：ローマ字の普及に尽力した田中舘愛橘がフランスから持ち帰ったタイプライター用鍵盤（キーボード）をベースに、一九四二年、裁判や議会の速記用日本語入力方式を開発した。

栗原俊彦　くりはら・としひこ／１９２２～１９７３。文章を分節するには橋本進吉による日本語文法（いわゆる橋本文法）が用いられてきたが、橋本文法では上一段活用の動詞（居る、着る、見る、など）や下一段活用の動詞（得る、蹴る、出る、など）などには語幹がないとされるので、そのままでは単語辞書に登録できない。そこでこれらの動詞は、変化しない部分を語幹とみなして登録することとした。たとえば「着る」では、「き」が語幹で「―、―、る、る、れ、ろ／よ」を語尾とする工夫である。名詞も橋本文法の普通名詞、固有名詞のほか、新たにサ変動詞を接続し得る「サ変名詞」という分類を加えた。これにより、「コウショウスル」に対し「好尚、厚相」などを捨て、「考証、交渉」などに候補をしぼりこむことが可能となった。

最長一致法　入力された文字列全体を内蔵辞書に参照し、フィットする文字列が見つかるまで、最後の文字を一つずつ落としていく。インターネットに掲示されていた例によると、「いわくありげな」というかなが入力されたときシステムはまず「いわくありげな」が自立語かどうかを辞書で調べる。これは自立語ではないので、最後の一字を無視して「いわくありげ」を調べる。以下同様にして「いわくあり」「いわくあ」と調べ、「いわく」まできたときようやく自立語として認識される。ここで「曰くありげな」らしいと判明、今度は文法をチェックし「曰く＋ありげな」で誤りがないことを確認し、最終的に「曰くありげな」を決定する。

長尾　真　ながお・まこと／１９３６～２０２１。三重県に生まれ京都大学工学部を出た。同大学教授、大型計算機センター所長、国立民族学博物館長などを歴任した。コンピュータによる言語処理、機械翻訳システムなどを研究した。

辞書の整備　コンピュータで日本語文を作成するには単語辞書が重要な役割を持つ。当時市販の辞書では事務文書で使用頻度の高い「貴社」「検収」「帳票」「お慶び」といったビジネス用語、姓名、派生語などが収録されていなかった。

国際プログラムサービス　ＫＰＳ：平貞介が六八年十月、東京・九段下に設立した。七一年漢字処理システムの開発を目的に「カンテック」を設立、ここでカナ漢字変換システムをローマ字入力で行える方法を編み出した。七八年に日本システムハウスと共同で日本語ワープロの開発に着手し、専門会社「日本ワードプロセッサー」を設立した。

溝口哲也　みぞぐち・てつや／１９３９～　：福岡県に生まれ六三年東京工業大学理工学部を出て東京芝浦電気に入った。青梅工場で汎用コンピュータ「ＴＯＳＢＡＣ」シリーズの開発に従事し、

七七年日本語ワープロ「ＴＯＳＷＯＲＤ　ＪＷ―10」を作った。また世界初のノートブック型パソコン「ダイナブック」を開発し、東芝のパソコン事業を世界トップクラスに押し上げた。八八年パーソナル・ワークステーション事業部長、九五年パーソナル情報機器事業本部長、九六年取締役、九八年上席常務、二〇〇〇年専務を経て〇三年モバイル放送社長となった。

プロジェクトＸ　二〇〇三年九月三日放送第九十五回「運命の最終テスト～ワープロ・日本語に挑んだ若者たち～」。

ＯＡＳＹＳ　オアシス：富士通独自の「親指シフト」方式が採用されていた。日本語文を入力する速度はプロ向きだったが、専用キーボードが必要だったため、大きなシェアを取れなかった。のち通常のＪＩＳキーボードによるカナ漢字変換方式もサポートし、シャープ、キヤノン、日本電気、東芝などと並ぶ主要な日本語ワープロに数えられた。

▼主要な日本語ワープロ

「書院」（シャープ：七九年九月）
「ＮＷＰ―20」（日本電気：八〇年五月）
「レターメイト80」（沖電気：八〇年五月）
「ＢＷ―20（ワードパル20）」（日立製作所：八一年五月）
「ＷＤ―１００」（シャープ：八二年一月）
「レターメイト８００」（沖電気：八二年二月）
「ＢＷ―10（ワードパル10）」（日立製作所：八二年五月）
「ＭｙＯＡＳＹＳ」（富士通：八二年五月）
「ＶＷＰ―１００」（日本電気：八二年十月）
「ＴＯＳＷＯＲＤ　ＪＷ―１」（東芝：八二年十一月）
「ＷＤ―２４００」（シャープ：八三年一月）
「ＨＷ―１００」（カシオ計算機：八五年五月）
「ＨＷ―30」（ソニー：八五年五月）

以後、音声入力機能、全文一括変換機能、文節変換機能、画像取り込み機能、表作成機能、パソコン通信機能などが装備され、画面表示の大型化、ポータブル型、低価格化が進んで行く。

日本IT書紀　229 日本語処理

著　者：佃 均

発行者：（特非）オープンソースソフトウェア協会
http://www.ossaj.org/
info@ossaj.org

発行日：2023年4月10日

本作品は2004年-2005年ナレイ出版局より刊行された 「日本 IT書紀」全5分冊を底本とし、原著者が一部改定を加えたものを複数の電子書籍に再構成して CC-BY-NC-ND ライセンスにより公開します。