2A10

日本語ワードプロセッサの開発

（株）東芝　総合研究所

1．はじめに

１９８０年代は計算機の日本語処理時代であるともいわれている。身近にある日本語ワードプロセッサはもとより、大型計算機からオフィス・コンピュータやパーソナル・コンピュータまでが日本語処理機能をもっている。

１９７８年の春と秋に、東芝から漢字オフィス・コンピュータと日本語ワードプロセッサが市販されたのがきっかけになって、この潜在的な市場に一気に火がつき、今日のような爆発的な普及が始まった。自国の言語で計算機を利用できるようにするのは、技術者として当然の目標であるが、長い間これが実用化に至らなかったのは、日本語処理の技術的な難しさと、日本語の入出力装置のコストの壁が厚く、この壁を突破するのに２０年近くの時間を要したからである。

2．かな漢字変換の研究のスタート

当社では、何か新しい研究をはじめるときにアンダー・ザ・テーブルの研究としてスタートすることが多い。これに対するのがオン・ザ・テーブルの研究で、これは研究企画書が正式に認められ、リソースが投入されるとともに、計画通りに研究が進行しているかの報告、フォローが行われるテーマである。あるテーマの研究を行なうべきだという強い信念と、それがもし達成されたとすると社会的に大きな貢献できるという使命感があるが、まだ基本的なアイディアが発見されず固まっていない段階がある。この段階では大勢の研究者や研究費を投入したからといって、アイディアが生まれるのを加速できるわけではない。少数であっても、強い信念と使命感に導かれて、夢の中でも考えるほどに執念深くアイディアを追及してゆく段階である。これをアンダー・ザ・テーブルの研究と呼んでいる。日本語の処理の研究もアンダー・ザ・テーブルの研究としてはじまった。

昭和４６年頃、新聞社の方々と雑談をしていたとき、欧米の新聞記者に比較して、日本の記者は記事を書くのが遅いことが話題になった。どうすれば速く記事を書けるのか、どうすれば速く事件現場を取材した記者のニュースを新聞紙面にのせることができるのか。これらの要求を技術の言葉に翻訳すると次の３点になる。

（１）手で書くより速く記事をタイプできること。
（２）タイプした内容を電話を通じて遠隔伝送できること。
（３）装置はポータブルにして、どこへでも持運べること。

後の２項目はハードウェアの問題であるが、最初の項目は日本語入力の問題そのものである。新聞記者のように専門家でない人が使って、手書きするより速く文章が作成でき、

かつ将来ポータブル型にすることができるようにしなければならないとすると、日本語入力方式もかなり厳しい条件がつくことになる。そのため１０本指で操作できるローマ字鍵盤か、かな鍵盤を用いてかな文字を漢字かな混じり文へ変換する技術がどうしても解決されねばならない目標ということになった。

3．かな漢字変換の研究

かな漢字変換の研究は１８６０年頃から行なわれており、大学や民間研究所で努力が続けられていた。しかしながら日本語に多い同音異義語の処理が不十分で、変換率は８０～８５％程度だった。この変換率を向上させることが最大の問題点であった。国語辞典が必ずしも参考にならないこともわかってきた。日本語の辞典は“文書を読むため”に使うものであって、“文章を書く”ためのものではない。基礎的な単語とむずかしい単語はのっているが、日本人なら誰でも知っていて辞書も見る必要のない単語や、容易に類推できる単語は国語辞典にはのっていない。このため、事務文書や手紙などを作成するときに必要な次のようなタイプの単語を新たに見つけ出さねばならなかった。

（イ）他の単語を知っていれば、その意味が容易に理解できる単語
（ロ）事務文書ではよく使われるが、他の分野ではあまり使われない単語
（ハ）人名や地名などの固有名詞
（ニ）接辞を含んだ派生語

一般に日本語ワードプロセッサの単語辞書には、３～１０万語が登録されている。しかしその中の全部を一様に使うということは決してない。日本語ワードプロセッサを使う個人によって、よく使う単語や、辞書に入っていても全く使わない単語の差が出てくる。つまり単語の使用頻度に偏りが生ずる。平均的には人間は一生の間に自分で作る文章のなかで３万語程度の異なる単語（固有名詞を除く）を使うといわれている。もし１０万語の単語辞書を持っている日本語ワードプロセッサがあるとすると、その７万語はまったく使わないばかりか、同音異義語の発生率は３万語の場合の３倍以上になってしまう。

そこで日本語ワードプロセッサ自身に、利用者が使用した単語の頻度を自動的に計数させ、使用頻度の高い単語から表示する方法を考案した。こうすれば単語辞書に何万語入っていても、利用者のよく使う単語を同音異義語の中から選び出し、最初に表示することができる。したがって変換率はその利用者にとって高くなる。最初に表示された単語以外の語を入力したいときは“次候補キー”を押せば、次の頻度順位の単語が表示される。単語辞書に登録してある単語が無駄にならないばかりか、入力速度、変換率が大幅に向上する。

さらに、ある文書を入力中に同音異義語の中の１つを選択すると、その単語は臨時的に表示順序が変更されて、次に同じ同音異義語が出現したときには、優先的にその単語を表示する機能も加えた。この２種の頻度情報のダイナミックな利用方式が確立され、変換率は従来の８０～８５％から９５％以上に飛躍的に向上し、日本語ワードプロセッサのための“かな漢字変換”が実用に供せるほどにタフなものになったのである。