

5 Q-1

OS/omicron 仮名漢字変換システム第2版の設計思想

下村秀樹 本宮志江 酒井貴子 並木美太郎 高橋延匡
 (東京農工大学 工学部 電子情報工学科)

1. はじめに

我々の研究室では、日本語情報処理を対象としたオペレーティングシステムOS/omicronを開発している [1]。仮名漢字変換は、この日本語情報処理環境の中核を成すものである。OS/omicronでは、1987年度に仮名漢字変換システム第1版を開発し、現在第2版を開発中である [2] [3]。本稿では、このOS/omicron 仮名漢字変換システム第2版の設計思想を述べる。

2. 文書作成環境と仮名漢字変換

仮名漢字変換は、文章エディタへの日本語の入力手段である。現在のパーソナルコンピュータなどでは、仮名漢字変換をフロントエンドプロセッサとしてシステムに組み込む形式が主流となっている。この方式は、文章エディタ側が仮名漢字変換を全く意識しなくてもよいという利点がある反面、キーの割当てや変換作業中の文字列の表示位置などに制限が加わり、文章エディタの自由度が失われるという問題がある。

また近年では、文章エディタを単なる入力・編集の道具にとらえるだけでなく、文章の訂正箇所検出機能などを結合して、インテリジェント化するための研究も盛んである [4]。我々も、文章を形態素解析・構文解析して、文章の訂正箇所を検出する研究を進めている [5]。しかし、これらの解析は多くの計算機パワーを必要とするので、実用的な機能を提供するためには、なんらかの方法でこの解析を高速化することが必要であるとの認識を持っている。我々は、このための一つの方法として、仮名漢字変換での解析の情報を利用できないかと考えた。仮名漢字変換は、入力文字列が仮名であるか漢字仮名混じりであるか、という点を除けば、形態素解析とほぼ同じ処理を行っている。したがって、仮名漢字変換での解析情報を取り出すことができれば、その後での文章解析の負担を軽減することができるはずである。

3. 日本語プログラミング環境と仮名漢字変換

一口に『日本語プログラミング』と言っても、さまざまなレベルがある。C言語を例にとっても、コメントだけに日本語が許される、プログラムの識別子に日本語が許される、などが考えられる。我々の研究室では、計算機上で英字も仮名も漢字も同等に扱いたい、との考えか

ら、OS/omicronの基本コードをJIS X 0208のフル2バイトコードとした。そして、コメントだけでなく識別子にも日本語(もちろん英字も含む)が使える言語CコンパイラCATを開発した [1]。

このような環境において仮名漢字変換は、文書作成時のそれとは違った機能を要求される。すなわち、英字と平仮名、漢字が混在したものを、頻繁に入力する必要があるということである。現存の仮名漢字変換では、英字と仮名の入力の際に、モード切り替えを必要とするものがほとんどであった。一般の文書のように英字の含有率の低い場合と違って、プログラム作成時におけるこの手間は、プログラム入力者にとって非常に大きな負担となる。

また、仮名漢字変換をフロントエンドプロセッサとしてシステムに組み込んでしまうと、2章で述べた文章エディタでのキー割当て、表示の自由度の問題と同じことが、プログラミング環境でのプログラムエディタでも当然起こる。

4. OS/omicron 仮名漢字変換システムの設計思想

OS/omicronの仮名漢字変換システム第2版は、前記のような日本語情報処理アプリケーションからの要求を考慮して、次のような基本方針をもって設計した。

(1) 仮名漢字変換(文字列変換)機能をライブラリ関数として用意する

これは、仮名漢字変換を利用するアプリケーションに、自由度を与えることを目的としている。仮名漢字変換システムは、文字列変換機能(連文節変換やローマ字仮名変換など)をC言語の関数レベルで提供する。アプリケーションは、この文字列変換機能を利用して変換を行い、変換に関するキーの割当てや変換操作中の文字列の表示は自由に行える。ただし、このような自由度よりも変換機能を容易に利用できることを望むアプリケーションもあるだろう。そのようなアプリケーションのために、システムにはこの仮名漢字変換を利用したフロントエンドプロセッサも組み込む(図1)。

(2) 仮名漢字変換を使う作業に応じて、変換の環境を設定できるようにする

仮名漢字変換は、使う人や使う場面によって、要求されるものが違う。例えば、文書を作成する場合に使う辞

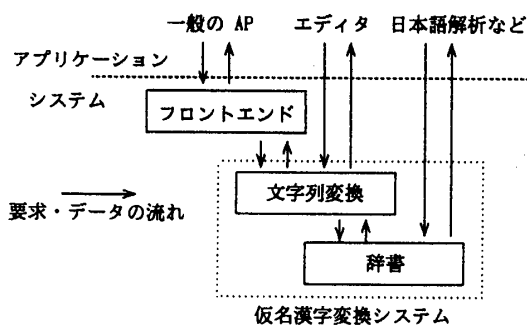


図1 仮名漢字変換システムの位置づけ

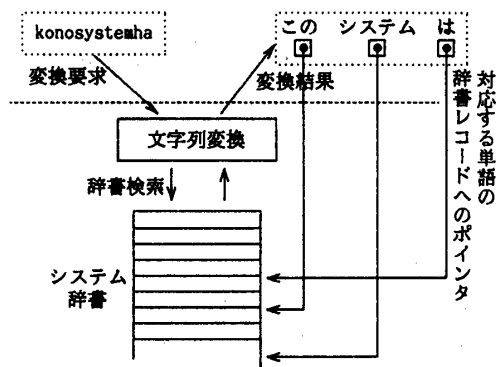


図2 仮名漢字変換での解析情報の出力

書と、プログラムを入力する場合の辞書は、その内容が異なっていて当然である。この仮名漢字変換システムでは、辞書などの変換に関する情報を、作業ごとにそれぞれ設定することができるようにする。つまり、計算機を使う人間やその作業内容に応じて、仮名漢字変換環境を自由に切り替えることができる。

(3) ローマ字表記や英単語表記からの変換も許す

これは、次のようにローマ字（あるいは、辞書に登録してある英単語）から直接漢字仮名混じり文字列に変換するという機能である。

sisutemuno (変換) → システムの
sinraisite (変換) → 信頼して
systemno (変換) → システムの, systemの

この方式では、仮名漢字変換システム内部でローマ字仮名変換を行うので、英字を入力する場合にも入力モードを変更する必要がなくなる。英字と仮名・漢字の比率に近いテキスト（プログラムや理工系の文書）を入力する場合には、有効な機能であろう。例えば、プログラムでよく使う一時変数の *i* や *j* などをそのまま辞書に登録しておけば、

iha (変換) → *i* は
jkaraimade (変換) → *j* から *i* まで

のような変換もできる。ただしこのとき、

iha (変換) → 胃は、意は、...

という候補も変換結果として出てくる。

(4) 仮名漢字変換で使用する辞書を日本語解析などのアプリケーションにも開放する

仮名漢字変換や日本語解析のアプリケーションには、日本語辞書が必要である。もちろん、それぞれのアプリケーションが目的に応じて別々の辞書を持ってよいが、辞書はデータ量が多いので、共有できる情報は共有すべきである。さらに、辞書を検索するライブラリなどが用意されていれば、辞書を利用するアプリケーションの構築が容易になる。

OS/omicon では、このような目的のために、システム側で日本語辞書 (OS/omicon 日本語システム辞書) を用意し、管理している。辞書には、単語についての品詞、意味、使用頻度など、さまざまな情報が記載されており、また辞書の検索関数もシステムが用意している。仮名漢

字変換をはじめとする辞書を利用するアプリケーションは、その目的に応じて必要な情報を取り出して利用することができる (図1参照)。

このように辞書を共有するというシステム構成をとると、仮名漢字変換での解析情報を、比較的簡単に日本語解析 (形態素解析など) で利用することができる。例えば、仮名漢字変換システムが、仮名漢字変換結果の各単語の属性の一つとして、その単語に対応する辞書レコードへのポインタ (論理的なポインタ) を返すようにしたとする。こうすれば、変換結果の各単語についての詳細な情報は、辞書検索のライブラリ関数を利用して簡単に得ることができる。形態素解析や構文解析などは、必要に応じて論理的なポインタによって辞書レコードを読み出して、解析処理に利用すればよい (図2)。

5. おわりに

本稿では、OS/omicon における仮名漢字変換システム第2版の設計思想を述べた。仮名漢字変換システムを実現した後は、それを利用したアプリケーションを開発したいと考えている。

参考文献

- [1] 高橋: "研究プロジェクト総説 OS/omicon の開発", 情報処理学会オペレーティングシステム研究会報告 39-5, 1988
- [2] 下村, 並木, 高橋: "OS/omicon 仮名漢字変換システムとユーザ辞書", 情報処理学会第38回全国大会講演論文集 4N-7, 1989
- [3] 本宮, 下村, 他: "OS/omicon 仮名漢字変換システム第2版の設計と実現", 情報処理学会第42回全国大会講演論文集, 1991 (予定)
- [4] 武田, 鈴木, 他: "日本語文書校正支援システム CRITAC", 情報処理学会第32回全国大会講演論文集 4T-12, 1986
- [5] 下村, 高橋: "形態素解析を利用した日本語スペルチェッカ", 情報処理学会第31回プログラミングシンポジウム報告集, pp. 117-125, 1990