5M－08

Web上の日本語文にふりがなをふるサービス 土竜(もぐら)の試作

大見 嘉弘
東京大学大学院教育学研究科

1．はじめに

日本語で文書を記述する場合、通常は漢字かな混じり文を用いることが大多数である。しかし、漢字かな混じり文が読解できるようになるには、多大な努力を必要とする。日本では、義務教育課程において、漢字を重点的に継続して学習する教育がなされている。また、日本語を学んでいる外国人にとって、漢字の読解は、最も困難な課題の一つである。しかし、最近急速に普及したインターネット上に存在する日本語の情報は、漢字かな混じり文がほとんどを占めており、小学生や日本語を勉強している外国人にとって、大きな障害となっている。そこで、World Wide Web ページ(以下 Web ページとする)中の漢字に振り仮名を機械的に振るサービスである、土竜(もぐら)を試作した。本稿では、その概要について述べる。

2．日本語 Web ページの問題点と振り仮名の付与

現在、インターネット上に存在する日本語の Web ページには、一般の日本人成人が読める程度の漢字が使われていることが多い。しかし、漢字の読みについての知識に乏しい者にとって、それらを読むことは大変な苦労である。国語辞典を引くことと漢和辞典を引くこととの労力や時間の差を考えても分かるように、漢字からその読みを調べるには、普通、非常に手間と時間がかかる。

小学生や外国人向けに仮名だけのページや、やさしい漢字のみを使用したページも存在する。しかし、全ての日本語の Web ページの量と比較すると、それらのページは極めて少ないと言わざるを得ない。そこで Web ページに振り仮名をつければ、小学生や外国人が利用できるページが格段に増えると考えた。もちろん、内容的に難しいページは読みが分かっても理解できないであろうが、読みが分かるだけで読解できるようになるページは、相当多いと推測する。例えば、新聞社の Web ページにあるニュースなどが有用であろう。

3．振り仮名付加サービスの実現

開発は、既存のフリーソフトウェアを活用して、開発時間を低減し、代わりにユーザインターフェースの開発や、変換辞書の整備に重点を置く方針で行っている。

サービス自体は、CGI(Common Gateway Interface)を使用し、CGI プログラムから、漢字かな変換プログラムである KAKASI を呼び出し、振り仮名をつけている。利用者は FORM 形式の Web ページに最初にアクセスし、アクセスしたいページの URL を入力し、「よみにいく」というボタンを押せば、Web サーバ上の CGI が起動される。CGI プログラムは、指定された URL からその Web ページにアクセスし、その内容を KAKASI に渡し、振り仮名を付加した内容をブラウザに返す。また、Web ページ中のリンク先を辿っても、本サービスを使うように URL を変換する。

KAKASI の振り仮名をふる機能は、例えば「**外で遊ぶ**」という文の場合は、「**外[そと]で遊ぶ[あそぶ]**」と変換する。これには、漢字の右側に振り仮名が追加されることで、文書が長くなり体裁の変化が大きく、煩雑で読みづらいという問題がある。また、

MOGURA － a Service which Adds Furigana(Kanji Readings) to Japanese Web Pages
Yoshihiro OHMI
The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo, 113-0033, Japan

形態素の単位で変換するため、送り仮名がある場合、それも振り仮名に含めてしまう。上記の場合、「**外[そと]で遊[あそ]ぶ**」となって欲しい。

そこで、KAKASI のソースコードに手を加え、上記の問題を解決した。まず、振り仮名表示の書式を KAKASI の起動時オプションで自由に変更できるようにし、送り仮名を振り仮名に含めない書式設定が可能となった。この拡張により、Web ページ内の HTML タグが付加できるようになり、次に述べるルビ振りも可能となった。また、この書式指定はデフォルトでは、従来の KAKASI の出力がなされるようにし、完全に上位互換となるよう配慮した。

また、Microsoft Internet Explorer 5 からは、ルビを振る HTML タグが使えるようになったため、このルビ振りに対応した。具体的には、例えば「`<ruby><rb>漢字</rb><rp>[</rp><rt>かんじ</rt><rp>]</rp></ruby>`」と記述すると「漢字(かんじ)」と表示される。また、その他のブラウザでは、「漢字[かんじ]」と表示される。**図１**に、各ブラウザで表示した場合の画面を示す。

また、KAKASI の変換辞書の改良を行い誤変換の低減を進めている。具体的には、各種の Web を変換し、誤変換があると、辞書に単語を登録したり、候補の優先順位調整を行い、できるだけ誤変換が減るようにしている。この結果、ある日の新聞社のページの誤変換率を調べたところ、辞書の整備前で 2.85% だったものが、2.25%に減少した。この誤変換率は、本サービスが十分に実用になることを示していると考える。

４．振り仮名自動付加の根本的問題

上述の通り、誤変換は十分に少なくできるが、いくら努力しても誤変換をゼロにすることは不可能である。例えば、人名には幾通りも読み方があり、漢字のみから読みを断定することはできない。例えば、小山は「おやま」「こやま」のどちらも有り得る。また、本サービスで使用しているような、変換辞書のみを使用し振り仮名を求める方式の場合、前後の文脈を考慮しないための誤変換が多くなる。

このように「誤変換が必ずある」ことを、本サービスの利用者は留意する必要があり、周知させることが重要である。つまり、誤変換された読みを正しいと思い込んで覚えてしまうことのないように配慮する必要がある。

５．今後の課題

本研究には多くの課題があるが、まず、漢字の読みをすべて与えてしまうことが漢字の学習にとって良い効果をもたらすのか？という教育的問題がある。例えば、「漢字にルビが振られることで、逆に、読みを覚えようとしなくなるのではないか？」といったことである。他、技術的課題としては、利用者に有益な機能を追加し、利用者が望む形態にカスタマイズできるようにすることを考えている。

本システムは、ソースコードを公開し、各自が持ち前の Web サーバにインストールして利用していただくことを狙っており、`http://mogura.p.u-tokyo.ac.jp/` にて公開している。

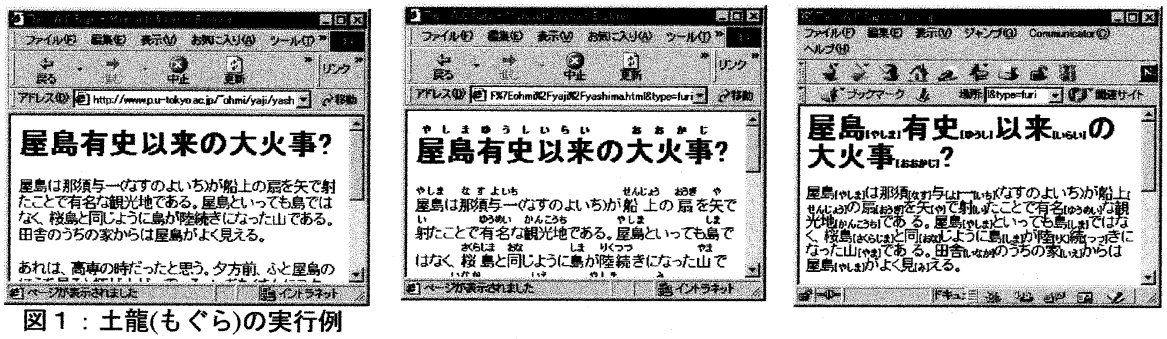

図１：土龍(もぐら)の実行例
(左から、変換前のページ、Internet Exploer 5.0 での表示、Netscape Navigator 4.5 での表示)