

視覚障害者用かな漢字変換結果の音声確認

4T-3

のための電子辞書の研究

鈴木恵美子* 星 恵子* 小川靖彦**

* 東京家政学院筑波短期大学 ** 筑波技術短期大学 視覚部

1 はじめに

ワード・プロセッサの普及に伴い、現在、視覚障害者も一般のワード・プロセッサを用いて手紙やレポートを書いている。その際に用いられている辞書は、ワード・プロセッサに備え付きの辞書で、収録語数は数万語から多くても10万語程度である。視覚障害者は画面で文字を確かめることができないため、カーソル位置の文字については音声で読み上げる方式をとりながら文書を作成できるようなワード・プロセッサを使用しているが、日本語に多い同音異義語を区別できるように、音と訓の読みわけや漢字の部首とつくりをよみわけられるのは限定されたごく一部の語であり、多くは読みを一通りしかもたず、「正規化」を「まさのり・か」と読まれたのを視覚障害者自身が判断して文書作成を行なっているのが現状である。

今回、筆者らは日本電子化辞書研究所で研究・開発された辞書を入手し、調査したところ、この辞書は収録語数約四十数万語、異なり見だし総数約二十数万語で、従来の辞書にない、新しい語を豊富に収容していることがわかったため、これに読みをつけて視覚障害者用ワードプロセッサのユーザインタフェース、特に漢字の読みの改善をめざすこととした。本稿では複数の同音異義語をもつ可能性のある日本語文書作成時において、視覚障害者が音声によってその漢字が確認できるようにする。そのため文書作成の際に視覚障害者が容易に自分の目的とする漢字を変換して音声で確認できるよう、漢字で表記される基本単語(一文字漢字と二文字漢字)を抽出し、それに読みと漢字の詳細な情報を付与した電子辞書を構築する。

2 目的

視覚障害者にとって、ワード・プロセッサでの文書作成においては、ある用語が正しく漢字に変換されたかどうか

かはその漢字の読み方を出力(例えば音声によって)させてみなければならない。一般に漢字の読みは基本単語のみにつけられていて、漢字複合語については基本単語の読みを単純に続けて表示(音声出力)するだけであるため、「今」を「こん」、「取出し」を「しゅだし」のように読まれてしまい、元の漢字を想像することは難しい。また、「かいぎ」と読まれても、それが「会議」なのかそれとも「懷疑」なのか、正しい読みだけでは判定できない。

本稿では、この問題点を解決するため、従来の辞書ではもっていなかった語の情報を、特に基本単語の読みと部首、無線等で使用される「アはアジアのア」といったような語の説明を付与することにより、視覚障害者でも容易に漢字を特定できるようなかな漢字変換のための電子辞書を構築する。このような電子辞書は、将来的には前後の単語との接続関係を情報としてもつことにより、読み情報を表示するだけにとどまらず、墨字文書を分かち書きしたり、分かち書きされた文書の点字文書への変換に応用できると考えられる。本研究に関連すると考えられる研究には「パーソナルコンピュータによるカナ・点字変換システム」の研究や、「かな及び2級英語点字自動翻訳システム」の研究があるがこれらはシステムとして稼働することに主眼がおかれ、実際の使用者である視覚障害者にとっては「システムがどれだけ使い勝手がいいか」ということが現在の問題である。本稿では上記のようなシステムおよび音声出力可能なワード・プロセッサは既に完成しているとして、視覚障害者にとってそのうえでさらにどのような情報を提供することが必要であるかについて考慮して研究を進める。

3 方法

本研究ではまず、EDR辞書インタフェースの内、日英単語辞書中の各項目から、辞書の見出しに相当する語を抽出する。そのためには、

1. 辞書項目をチェックする。
2. 辞書の見出しに相当する語を抽出する。
3. 見出しのうち基本単語を抽出する。

Electronic Dictionary for the Blind with Kanji Information which is used to ensure Kana-to-Kanji Conversion

Emiko SUZUKI*, Keiko HOSHI*, Yasuhiko OGAWA**

*Tokyo Kasei Gakuin Tsukuba Junior College Institute of Information Science

**Tsukuba College of Technology, Division for the visually impaired, Research Center on Educational Media

4. 読み仮名を付ける。

5. 漢字の詳細な情報を付与する。

ここで、基本単語とは漢字一文字の語と漢字二文字の語とする。これは三文字以上の漢字複合語は一部の例外（東西南北や太平洋のような語）を除いて本来一文字あるいは二文字の漢字の組み合わせから構成されていると考えたからである。例えば、「超大型高速電子計算機」といったような語も、分解すると「超」「大型」「高速」「電子」「計算」「機」のように表現できる。実際この方法で漢字複合語を分解することは可能で、既にいくつかの研究発表も行なわれている[4]。このため、(c)で以下の作業をおこなうことにより、基本単位辞書の見出しが抽出できる。

1. EDR 辞書から抽出された見出し語のうち、漢字のみで構成される語を抽出する。
2. (1)で抽出された語の長さを数え、一文字あるいは二文字であればその語は基本単位辞書に格納する。

筆者らが所有している EDR の辞書は読み情報をもっていないので、ここで抽出された基本単位となる一文字漢字も二文字漢字も読みはついていないが、これらについては既に利用されているワード・プロセッサの辞書中に含まれる語が多いと考えられるので、そこから自動的にマッチングをとり、読み情報を付与できるのではないかと考えている。ワード・プロセッサの辞書にない基本単位については、人手で読み情報をつけることとする。

詳細な漢字の情報については一つは EDR 辞書の釈義文を利用することも考えられるが、当面は無線等で利用されている簡単な情報を付けた上で、視覚障害者の意見を聞いてみたい。

4 読み替え情報の登録

現在実験は前節の(1)、(2)、(3)の作業を終了し、読み付けの作業に入っている。(3)では一文字漢字で構成されている見出し、二文字漢字で構成されている見出し、三文字漢字で構成されている見出し、の三種類をファイルに取り出し、それぞれに読み替え情報を付与する。

ここで、先にも述べたように、二文字漢字、三文字漢字は、それぞれ、一文字漢字の組み合わせで構成されていると考え、まず一文字漢字の読み替え情報を付与した(表1参照)。

5 問題点と今後の課題

この研究に用いられた日本電子化辞書研究所の辞書は本研究申請者の利用するオンライン・ネットワーク上のハードディスクに分割して保存されているが、辞書とい

表 1: 1 文字漢字と意味情報

笠	たけかんむりにたつのかさ
檜	しよくぶつのかし
鯀	さかなのかじか
割	わりざんのわり
括	ふごうのかつこのかつ
活	せいかつのかつ
渴	のどがかわくのかわく
滑	えんかつのかつ
葛	かつしかほくさいのかつ
褐	かつしよくのかつ
鯉	さかなのかつお
鞆	にもつをいれるかばん
株	かぶしきがいしゃのかぶ
兜	ほうしのかぶと
釜	こめをたくかま
鎌	かまくらじだいのかま
鴨	とりのかも

う大量のデータを扱う必要と、今後の社会的要請（これからは視覚障害者にもワード・プロセッサでなく、ワークステーション上で音声を伴った快適な作業環境が提供されるようになるであろうこと）を考慮して、今後はワークステーション上で、辞書データを保存して研究を行ないたいと考えている。

なお、この研究は試験研究 1-B 「全盲と晴眼者の文字・図形によるコミュニケーション支援システムの開発」の一環として行われたものである。

参考文献

- [1] 脇田、平山: パーソナルコンピュータによるカナ・点字変換システム 情報処理学会第 30 回全国大会, pp. 1637-1638, 1985.
- [2] 平山: かな及び 2 級英語点字自動翻訳システム 情報処理学会第 32 回全国大会, pp. 1683-1684, 1986.
- [3] 河原: 日本語自動点訳ソフトウェアの開発について 信学技報 HC94-49, 1994.
- [4] 武田、藤崎: 統計的手法を用いた漢字複合語の短単位分割 情報処理学会自然言語処理研究会資料, 48-2, 1985.