ヒューマンインタフェース 56－8
（1994．9．8）

日本語自動点訳ソフトウェアの開発について

河原正治†

† 筑波技術短期大学 教育方法開発センター

現在ある4種類の日本語自動点訳ソフトウェアにおいては，原文1万字当たり300～600個の誤変換があり，まだまだ改善すべき点が多くあると報告されている。本論文では，まず，既存点訳プログラムの誤変換の内容を分析して改善のための指針を示す。つぎに，この指針に基づいて開発した点訳プログラムの精度について報告する。この種のプログラムは，幅広いユーザの使用からのフィードバックが不可欠であり，近くフリーソフトウェアとして公開することを予定している。

Yet Another Japanese Braille Translation Program

Masaji KAWAHARA†

† Research Center on Educational Media
Tsukuba College of Technology
4-12 Kasuga, Tsukuba-shi, Ibaraki, 305, JAPAN

There are four existing programs that translate Japanese sentences into braille. They are reported to have 300-600 errors per 10,000 characters. First, we analyze problems associated with these programs. Secondly we describe the methods and the results of a new braille translation program. We would distribute the program as free software to get bug reports and improve the program.

1. はじめに

近年，パーソナルコンピュータで動作する日本語自動点訳ソフトウェアが，いくつか開発され広く利用されるようになった。ここで「日本語自動点訳」とは，かな漢字混じりのべた書き文を6ビット記号体系の触知文字(点字)に変換することを意味している。この処理の最も困難な部分は，かな漢字混じり文を点字のルール[1, 2]に基づき，かなの分かち書きにすることにある。このような機能を持つプログラムの精度を比較した報告によると，以下のような改善すべき点が指摘されている[3]。

- 原文1万字あたり300〜600個の誤変換がある。
- 特に，ひらがなの並びの解析に不備がみられる。
- より精度を向上させるためには文意に踏み込んだ解析が必要である。

2.で示すように，筆者らの調査においても，新聞記事データベースに対して市販の点訳プログラムを実行した結果，原文1万字あたり300個程度の誤りが検出されている。また，誤変換の内容についても上記指摘と同様な結果を得た。しかし，このような点訳ソフトウェアは，人手による前後処理を適切に行えば点訳作業の効率化に大きく寄与していることも事実である。

本論文では，新たに開発した点訳ソフトウェアについて述べる。開発にあたって，以下のような目標を定めた。

(1) 市販ソフトウェアよりも精度の高い点訳プログラムを提供すること。
(2) さまざまな既存ツールが統合可能であること。
(3) 辞書およびその他のツールを含めて，フリーソフトウェアとして公開できるものとなること。

最先端の自然言語処理技術を容易に取り込めるように，開発にあたっては高性能ワークステーションの使用を前提とした。言うまでもなく，より広範囲のユーザを得るためにはパーソナルコンピュータへの移植が必須である。プログラムは，すべてC言語で書かれており比較的移植しやすい構造になっている。パーソナルコンピュータへの移植が完了するまでの便宜を目的として，また，広く視覚障害者の福祉に貢献するために「点訳メールサーバ」の運用を計画している。

以下では，点訳システムの中で大きな役割を果たす分かち書きプログラムを中心に解説する。

2. 既存点訳ソフトウェアに関する検討

2.1 日本語処理技術としての点訳

自然言語処理は学会での主要なテーマの一つであり，数多くの論文が発表されている。しかし，かな漢字変換や機械翻訳技術の動向から明らかなように，実用レベルではまだまだ人手に依存する部分が大きく，これから改善されるべき部分が多く残されている分野である。また，研究段階においても製品開発においても費やすコストが大きいにも関わらず，その成果が広く公開・共有され，さまざまな応用をみているものが少ないことも特徴の一つである。

たとえば，市販の日英機械翻訳ソフトウェアのマニュアルには，よい英文を得るために人間がしなければならないこととしてつぎのような記述がある[4]。

- 長い文章は，50文字以内の短い文章にする。
- 主語や目的語を明確にする。
- 係り受け関係を[ ]を使って指示する。

また，パーソナルコンピュータ用の安価な翻訳ソフトウェアが販売されるようになってきたが，ほとんどのものは英語から日本語の翻訳である。このような機械翻訳ソフトウェアを一度使用してみれば，日本語自然言語処理技術の現状と山積している課題が容易に理解できよう。

以上のことは，かな漢字混じりの日本語の解析が，いかに困難であるかを明示している。英文の点訳は文字列の置換とほぼ同じ程度の複雑さしかないのに対して，日本語の点訳は多くの困難を含んでいる。同時に，日本語音声認識・合成技術，機械翻訳技術に関する研究が活発に進められており，そのような最新の研究成果を点訳技術に反映していく必要がある。

2.2 既存ソフトウェアの精度について

現在，広く利用されている日本語自動点訳ソフトウェアには以下の4種類がある。

- がってんだ ((株) 管理工学研究所)
- EXTRA((株) アメディア)
- 点訳ぶうちゃん (松山市視力障害者友の会)
- 80点 (福祉システム研究会)

この4種類のソフトウェアの精度を比較した結果の報告[3]によると，4種類のソフトウェアの精度に大差はない。ここでは，上記ソフトウェアのうち，EXTRA(Ver.1.1)[5]について，その変換精度を調べた結果を報告する。

誤変換の個数について表1にまとめた。変換対象のデータとして，朝日新聞 (朝刊) データベースの1993年1月1日より80記事分，64,351文字を利用した。誤変換の分類については，福井の分類[3]に従った。

	誤変換の類型	個数
1	一般語の読みの誤り	465
2	固有名詞に関する誤り	226
3	数字とかなの使い分けの誤り	32
4	長音の処理の誤り	20
5	助詞の処理の誤り	12
6	記号と文字の間を続ける誤り	33
7	記号と文字の間に余分な空白が入る	7
8	省略すべき読点・中点を省略しない	79
9	文節間を続けた	341
10	文節内を切った	88
11	分かつべき複合語を続けた	197
12	続けるべき複合語を切った	246
13	補助用言を続けた	7
14	助詞，助動詞を切った	25
15	続けるべき「する」を切った	16
16	分かつべき「する」を続けた	8
17	数字の表し方の誤り	72
18	その他	40
	合計	1914

表1: 市販点訳ソフトウェアの誤変換の個数
(分類方法は福井[3]による)

表2に典型的な誤りの内容を示す。

このような誤変換の内容から，既存の自動点訳ソフトウェアは，品詞情報は考慮せず，後述する右方向最長一致法を中心にして，読み切れなかったものは単漢字として強引に読みを与えていると思われる。

上述のような誤変換の内容を検討した結果，つぎのような改良の指針を立てた。

(1) 辞書の語数を増やすことで一般語および固有名詞の読み誤りを減少させる (表1の1,2)。
(2) ひらがなの解析精度を向上させる (表1の9,10,14)。
(3) ごく簡単な係り受け解析を行うことで誤変換を減少させる (表1の2,3,17)。
(4) 品詞の接続情報を解析することで誤変換を減少させる (表1の11,12)。

分かち書きを実現するために，まず原文を形態素の列に分割し，読みと品詞の候補を特定して，点字の分かち書きルールに基づいて形態素を連結する方法をとることにした。以下では，形態素解析に用いる辞書の開発について述べた後，形態素解析プログラムについて解説する。

語　例	読み・分かちの誤り
大晦日	だいみそか
前触れ	ぜん　ふれ
酉年	ゆうねん
丸餅を	がんへいを
隣町	りんちょう
本田技研	ほんでん　ぎけん
竹下登	たけしたと
久米豊	くめほう
若杉弘	じゃくさんこう
響灘	きょーなだ
きのうまでとは	きの　うまでとは
建設にからんで	けんせつにからんで
批判しておきながら	ひはんしておきながら
開業にこぎつけたい	かいぎょーにこぎ　つけたい
剥奪（はくだつ）が	はくだつ（わ　くだつ）が
社会正義	しゃかいまさよし
七三％で	しちぞう％で
５０００万人	５０００ばんにん
第十一次	だい１０１じ
第十五回	だい１０５かい
弱みが	じゃくみが
度合いに	どあいいに
年明けと	としあきけと
揺さぶりが	よーさぶりが

表2: 市販点訳ソフトウェアの典型的誤りの例

3. 分かち書き用形態素解析辞書の開発

3.1 システム辞書の開発

1.で述べたような目標を満たすと同時に，2.2の指針に従うには，以下のような条件を満足する形態素解析辞書が必要である。

- 必要充分な語数を持つ。
- 追加・改造が容易である。
- 将来もライセンスなどの制約を受ける恐れがない。
- 先端の形態素解析アルゴリズムに対応しているかまたは対応可能である。

ICOT(新世代コンピュータ技術開発機構) では，その研究成果として，変更・再配付自由な形態素解析辞書 (15 万語) を公開している[1]。ICOT 版形態素解析辞書は以下のような特徴を持つ[6]。

- 冊子体の国語辞典よりはるかに多い 15 万語を収録。
- 通常の国語辞典では見出し語にならない付属語，接頭・接尾語を収録。
- 形態素の分類には接続関係を考慮した独自の分類体系を採用。
- テキスト形式の辞書から TRIE 型へ変換するユーティリティが付属。

この辞書を入手し，つぎのような拡張を加えた。

(1) 単語の補充 (6 万語程度)
補充した単語は，後述する分かち書きプログラム実行時に抽出した未定義語を独自に登録したものである。これによって現在約 21 万語の形態素解析辞書が完成している。これは，分かち書きプログラムとあわせて再配付することができるようになっている。たとえば，前出の自動点訳ソフトウェアの EXTRA の辞書は約 9 万語，かな漢字変換プログラムの ATOK8 の辞書は約 15 万語，岩波広辞苑の見出し語は約 22 万語であることなどを考慮すると，現在完成している形態素解析辞書は一般的な使用に対しては十分な規模と言える。辞書はテキスト形式で提供されるので，各ユーザが追加・修正することが容易である。したがって，各ユーザの要求に応じた専門用語などの追加が容易に行える。

(2) 品詞コードの拡張
品詞コードについては ICOT のものを踏襲しているが一部拡張した部分がある。たとえば，人名については，ICOT の辞書では，姓と名を区別する手段がないが，これを区別するために品詞コードを拡張している。これに伴い，品詞コードを 2 バイトから 4 バイトに変更し，将来の拡張に対応可能なようにした。この拡張した部分には意味的係り受けを実現するための情報を組み入れていく予定である。

(3) 複合語の読みを分かち書きに変更
点字表記には独特の分かち書き規則がある。このため，ICOT 版辞書に登録されている複合語などに対して，分かち書き規則に基づいて読みを分割して登録し直す作業を行った。

テキスト辞書の記述例は以下の通りである。

単語	読み	品詞・係り受けコード
来	く	00000083
杭	くい	00000010
空	くう	00000010
狂	くる	00000040
胡桃	くるみ	00000010

これを TRIE 型に変換した後，形態素解析処理での検索に用いる。TRIE 型は，図 1のような構造を持っており，形態素解析実行時に頻繁に行われる最左部分列の切り出しが効率的に行われるようになっている[7, 8]。

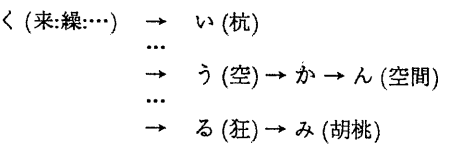

図 1: TRIE 型辞書の論理的構造

3.2 ユーザ辞書

ユーザ辞書の形式はつぎの通りである。

[1] これは，ftp://icot.or.jp/ifs/natural-lang/unix/morphdic.tar.Z で入手可能である。

	単語	読み	品詞コード	係り受けコード
-	大学	ガイガク	0010	
+	大学	ダイガク	0010	
-	正治	マサハル	0014	0001
+	正治	マサジ	0014	0001

ここで，＋は追加語を，－は削除語を，数字は品詞コードと係り受けコードである。係り受けコードは省略可能である。

ユーザ辞書は，現在の仕様では，追加語，削除語をそれぞれ1万語まで登録できるようになっている。ユーザ固有の単語を登録したい場合の他に，システム辞書に誤りがあるときの応急処置として利用することもできる。

4. 点訳用分かち書きプログラムの開発

4.1 形態素解析アルゴリズムについて

形態素解析アルゴリズムとしては，現在さまざまな方式が提案され，より高い精度を目指した新しい方式が検討されているが，実用レベルで利用できるようになっているアルゴリズムは，右方向最長一致法，字種区切り法，接続表による解析法，文節数最小法などがある[7]。また，これらのアルゴリズムに意味的係り受けを考慮した方式もある。

右方向最長一致法は実現が容易であるが，たとえば，「新高（にいたか）」という固有名詞が辞書に登録されている場合は，「新高速道」という単語は，「にいたか　そくどう」という読みになる。このように，右方向最長一致法では，誤分割が避けられず，経験的に20%程度の誤りがあるといわれている。

字種区切り法は，ひらがなから他の字種への変わり目などを分割点とするものである。この方法においても，たとえば「変わり目」は「かわり／め」と余計なところで分割してしまうという欠点がある。

接続表に基づく解析法は，得られた分割の候補について，その前後の形態素の情報からそれが妥当かどうかを判断する方法である。

以上のように，決定的なアルゴリズムはなく，各アルゴリズムを組み合わせた上に，経験則を加味して，実現されているのが実状である。

4.2 分かち書きプログラムの実現と精度

今回開発した分かち書きプログラムは，接続表による解析を行い，文節数最小法に経験則を加味した優先順位付けを行うものである。

すなわち，最初に，可能なすべての形態素の分割を行い，接続情報と合致した分割を選別する。その形態素の列について文節数最小法に経験則を加味して優先順位をつける。

また，ごく簡単な係り受け解析を行なっている。たとえば，「関東平野」は「関東」と「平野」という2つの形態素の列に分割され，「平野社長」は「平野」と「社長」に分割される。形態素「平野」には「へいや」と「ひらの」という読みの可能性がある。後者の分割においては，「人名＋人名につく接尾語」の並びを優先することによって「ひらの」の読みを第一候補として出力する。さらに，数詞と数助詞の連続，氏名の姓と名の連続などについても考慮し，解析精度を向上させた。

2.2で述べた既存点訳ソフトウェアの誤変換について，新たに開発した形態素解析プログラムの出力例を表4，表5，表6に示す。また，表1で示した市販ソフトウェアの誤変換のうち40記事分の383種類の誤りについて，今回開発したプログラムでの解析結果を表3にまとめた。383種類の誤変換のうち330(86%)については完全に分かち書きできた。

完全に分かち書きできるもの	330
失敗するもの	7
最適解として特定できないが候補にはあがるもの	46
合　計	383

表3: 分かち書きを実行した結果

今回開発した分かち書きプログラムは，既存のものと比べて，以下のような特徴がある。

- 辞書を整備した結果，一般語や固有名詞を未定義語として読み間違えることが少なくなった。
- 読みが一つの候補に絞れない場合でもほとんどの場合に正解が選択肢にあがる。

	原文	上段に市販ソフトウェアの誤変換 下段に開発したソフトウェアの解析結果
1	新高速道	ニイタカソクドー 新 [シン, (接頭語，名詞につくもの)] ◇ 高速道 [コーソクドー, (名詞，一般)]
2	その場合	ソノバ◇ゴー その [ソノ, (名詞，一般)] ◇ 場合 [バアイ, (名詞，一般)]
3	立候補届	リッコーホカイ 立候補 [リッコーホ, (名詞，サ変動詞性)] ◇ 届 [トドケ, (名詞，一般)]
4	５０００万人	５０００バンニン ５０００万 [５０００マン, (数詞)] 人 [ニン, (接尾語，数詞につくもの)]
5	第十一次	ダイ１０１ジ 第 [ダイ, (接頭語，数詞につくもの)] 十一 [１１, (数詞)] 次 [ジ, (接尾語，数詞につくもの)]

表 4: 市販ソフトウェアの誤変換と開発したプログラムによる解析結果 (1)

	原文	上段に市販ソフトウェアの誤変換 下段に開発したソフトウェアの解析結果
6	きのうまでとは	キノ◇ウマデトワ きのう [キノー, (名詞，一般)] まで [マデ, (副助詞，用言の連体形につくもの)] とは [トワ, (助詞相当語，名詞，係助詞，用言の終止形につく)]
7	年頭のねぎごとが	ネントーノネギゴトガ 年頭 [ネントー, (名詞，一般)] の [ノ, (格助詞，用言の連体形につくもの)] ◇ ねぎごと [ネギゴト, (名詞，一般)] が [ガ, (格助詞，名詞につき，格，副，係助詞がつかないもの)]
8	建設にからんで	ケンセツニカランデ 建設 [ケンセツ, (名詞，サ変動詞性)] に [ニ, (格助詞，用言の連用形につくもの)] ◇ から [カラ, (動詞の語幹，五段活用，マ行)] ん [ン, (動詞の活用語尾，五段活用：連用２，撥音便)] で [デ, (接続助詞，用言の連用形につくもの，撥音便になる)]
9	批判しておきながら	ヒハンシテオキナガラ 批判 [ヒハン, (名詞，サ変動詞性)] し [シ, (動詞の活用語尾,「する」を接続するもの：連用１)] て [テ, (接続助詞，用言の連用形につく，撥音便にならない)] ◇ お [オ, (動詞の語幹，五段活用，カ行,「行く」を除く)] き [キ, (動詞の活用語尾，五段活用，カ行：連用１)] ながら [ナガラ, (副助詞，動詞の連用形につくもの)]

表 5: 市販ソフトウェアの誤変換と開発したプログラムによる解析結果 (2)

	原文	上段に市販ソフトウェアの誤変換 下段に開発したソフトウェアの解析結果
10	中江利忠	ナカエリチュー
		中江 [ナカエ, (名詞，固有，人名)] ◇ 利忠 [トシタダ, (名詞，固有，人名)]
11	槙文彦	テンフミヒコ
		槙 [マキ, (名詞，固有，人名)] ◇ 文彦 [フミヒコ, (名詞，固有，人名)]
12	田辺誠	タナベセイ
		田辺 [タナベ, (名詞，固有，人名)] ◇ 誠 [マコト, (名詞，固有，人名)]
13	久保亘	クボコー
		久保 [クボ, (名詞，固有，人名)] ◇ 亘 [ワタル, (名詞，固有，人名)]
14	響灘	キョーナダ
		響灘 [ヒビキナダ, (名詞，固有，人名以外)]
15	男鳥	ダントー
		男鳥 [オシマ, (名詞，固有，人名以外)]

表 6: 市販ソフトウェアの誤変換と開発したプログラムによる解析結果 (3)

本論文では，既存点訳プログラムの誤変換に対する結果について述べた。現在，実用化に向けて，本論文のプログラムによって，多様な文書を大量に処理した場合の精度について調査しており，つぎの論文で報告する予定である。

5. 点訳プログラムの実行例

現在，点訳プログラムはUNIX ワークステーションで動作している。今後，MS-DOS パーソナルコンピュータへ移植していくことも検討している。まず，以下のようなコマンドを実行することによって，点訳用分かち書きプログラムを起動して，かな漢字混じりの日本語を分かち書きすることができる。

```
$ wakati -m2 test.txt
```

続いて，tcode コマンドを用いて分かち書きデータを点字データに変換することができる。ここで，$ はUNIX のプロンプトである。これを実行している例を図2に示す。図2は，以下の原文を分かち書きしたものである。

> コンピュータで日本語文書を点字に変換するソフトウェアの改良の指針を示すために、人手による点訳作業の過程を解析し、ソフトウェア技術の可能性を考える。当面は、点訳作業を人手による部分と計算機による部分に分割し、それに合わせたソフトウェアの開発と前後処理をする専門的点訳エディタの養成が必要である。

現在までに，分かち書きプログラムのオプションとして以下のようなものを実現した。

-m1	右方向最長一致法で解析
-m2	本論文の方法 (4.2参照) で解析
-v	音声合成装置による結果読み上げ

UNIX における標準的なエディタの一つであり，辞書検索などの多様な機能を容易に組み込めるという特徴を持つ emacs エディタによるインタフェースも利用できる。これは，emacs のバッファおよびリージョン単位，またはセンテンス単位で分かち書きし，結果を別のバッファに表示することができるものである。emacs を点訳のために使用可能とすることによって，文書作成から点訳までを一貫して行なうことができるようになると考えている。

6. むすび

かな漢字混じり文を点字に変換するソフトウェアについて，分かち書きの機能を中心に解説し，筆者の得た成果について報告した。筆者の開発した分かち書きプログラムを利用することにより，

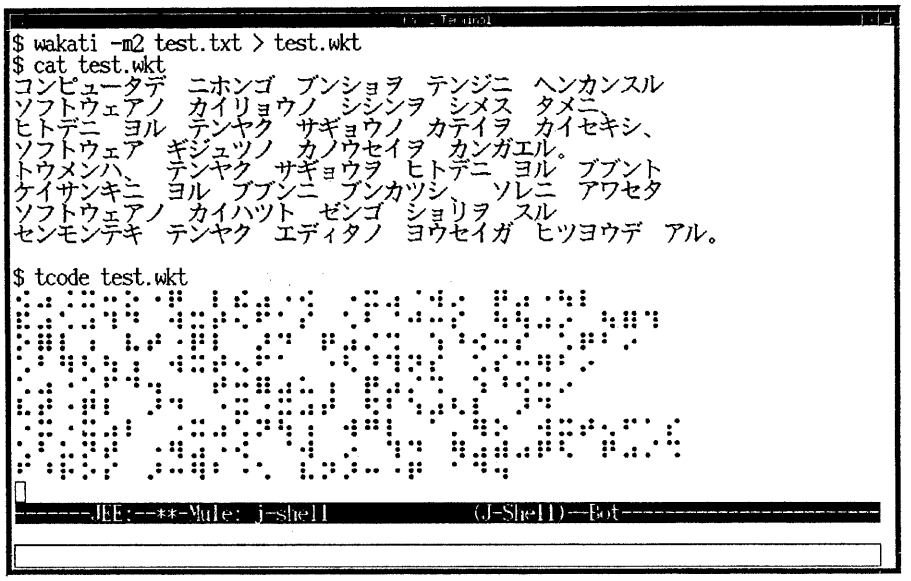

図 2: UNIX による点訳の実行

新聞記事データに対する既存点訳プログラムの誤変換のうち 86%が解決できることを示した。

この種のシステムを実用化するためには，多くのユーザによって使用され，その結果がフィードバックされ，改善されるというサイクルが必須である。このような観点から，点訳システムをフリーソフトウェアとして公開し，幅広いユーザの支援をうけ，改善していきたいと考えている。

また，電子メールでデータを受け取り，分かち書きデータまたは点字データを返送する点訳メールサーバの運用も計画している[1]。

本システムを核にして，さまざまな視覚障害支援ツールが構築・統合されることによって，視覚障害者の助けとなることを願っている。

謝辞　本研究を進める上で，ご指導頂いた森亮一筑波大学名誉教授に深謝いたします。

参考文献

[1] 日本点字委員会：“日本点字表記法” (1990)

[2] 日本盲人福祉研究会：“最新 点字表記辞典”，博文館新社 (1991)

[3] 福井哲也：“日本語自動点訳ソフト 4 種の精度の比較”，第 2 回視覚障害リハビリテーション研究発表大会論文集, pp.114-117 (1993)

[4] 富士通株式会社：“FUJITSU S ファミリー ATLAS 翻訳の手引”, pp.12-19 (1991)

[5] 株式会社アメディア：“EXTRA 使用説明書”, p.133 (1991)

[6] (財) 新世代コンピュータ技術開発機構：“TRIE 構造日本語辞書とその形態素分類体系の概要” (1992)

[7] 田中穂積：“自然言語解析の基礎”, pp.134-153, 産業図書 (1989)

[8] 青江順一：“トライとその応用”, 情報処理, Vol.34, No.2, pp.244-251 (1993)

[1] 点訳メールサーバに関する最新情報は，http://www.k.tsukuba-tech.ac.jp/tools/braille-serve.html を参照のこと。