自動点訳システムIBUKI-TENと新点字規則への対応

髙松大地　岸井謙一　伊佐治和哉　松本忠博　池田尚志

岐阜大学工学部

1　はじめに

本報告では，我々が開発してきた自動点訳システムIBUKI-TENの概要，現在行っている改良の状況や課題，WWW上での公開の状況などについて述べる．

2　IBUKI-TENの概要

IBUKI-TENは，我々が開発している文節解析システムIBUKIを自動点訳システムに応用したものである[5]．本システムは，入力された日本語テキスト文書に対して，点字規則に沿って分かち書きすること，漢字かな混じり表記を点字規則に沿ってかな表記に変換すること，また数符など必要な点字符号を挿入することなどの点字用処理を施して，点字文書を得るソフトウェアである．さらに得られた点字文書を編集する機能も持っており校正作業の支援が出来るようになっている．また点字プリンタや点字ピンディスプレイに出力する機能も備えている．

IBUKI-TENはWWW上に公開し，全国で多くのユーザに使われている．ユーザからは多くの要望や質問が寄せられており，改良・改版の契機になっている．

2.1　システム概要

システム構成を図1に示す．入力されたテキストは，1文毎に，我々が現在開発している文節解析システムIBUKIによって，文節単位の切り出し，複合語解析による漢字連続文字の名詞／接辞等への分割などを行う．この文節解析により抽出された文節単位に点訳処理を実行する．点訳処理では，文節内の単語の前を切るか続けるか，単語内を切るかといった分かち書き処理と，点字の表記法に従ったひらがな表記への変換を行う．分かち書きは，例えば「する」の前を切るか続けるかといった一般規則を記述した規則表に基づく処理と，辞書に記述されている個別の分かち書き規則とで行う．

また自動点訳誤りの修正，レイアウト用の処理などを行うためのエディタがある．エディタ上では分かち書きや漢字仮名変換について誤りの可能性がある箇所が，区別された色・記号で表示されるようになっており，ユーザの後編集作業を支援する仕組みを備えている．

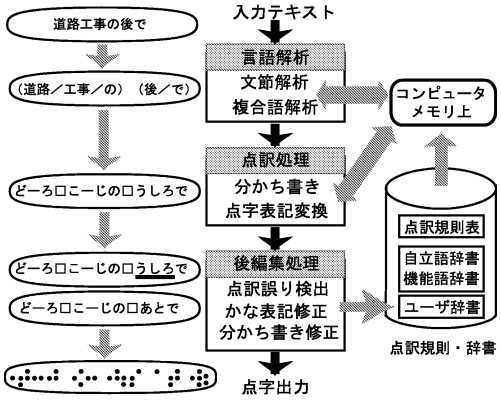

図 1: システム構成

2.2　点訳用辞書

辞書上の点訳のための情報の一例を図2.2に示す．点訳規則の”/”は分かち書きの区切りを示す．自立語辞書は，EDR日本語単語辞書をベースに作成し，機能語辞書については我々が独自に作成した．

表 1: 辞書上の点訳規則

表記	点訳表記	品詞	連濁	読みコスト
急に	きゅーに	副詞		0
会社	かいしゃ	普通名詞	○	0
今日	きょー	時詞		1
今日	こんにち	時詞		2
都市国家	とし/こっか	普通名詞		0
四輪駆動	４りん/くどー	普通名詞		0
にもかかわらず	にも/かかわらず	機能語		0
である	で/ある	機能語		0
だろう	だろー	機能語		0

2.2.1　自立語辞書

点訳規則フィールドには，各単語毎に単語の前を切るか続けるか，単語内を切るかどうかを示す区切り情報を含む単語の仮名表記が点訳規則として登録してある．

連濁フィールドは，複合名詞を構成する2番目以降の単語が連濁する可能性があるかどうかを示すものである．現在のところ，EDR日本語単語辞書の4文字以上の名詞，サ変名詞を複合語解析し，元の単語の仮名表記と分割された単語の仮名表記を比較し連濁している単語を収集した（「会社，時計，黒子，茶碗，菓子」など約100語）．

読みコストフィールドには，仮名表記の優先順位を表すコストを登録してある．複数の読みを持つ単語に対して，使われる読みの程度に応じて次の 3 段階のコストをつけることで候補の絞込みを行っている．

コスト 1 複数ある場合に通常選択する仮名表記
……「今日」の場合「きょう」

コスト 2 その他に使う可能性がある仮名表記候補
……「今日」の場合「こんにち」

コスト 3 通常では使わない仮名表記
……「今日」の場合「こんじつ，こんち」

EDR 単語辞書中で IBUKI-TEN が使用している漢字表記は異なり語数で 193,604 語あり，同じ品詞，活用型で 1 つの仮名表記しか持たない単語は 185,167 語であった．複数の仮名表記を持つ単語，8,437 語（延べ仮名表記数 18,322 語）に対して，手作業で上記のコスト付けを行った．

2.2.2 機能語辞書

IBUKI では文節を意味的なまとまりに従って切り出すために，できるだけ長い単位で機能語を登録している．例えば「〜ておかねばならぬ」を 1 つの文節として切り出す．これを点字規則に従った短い単位に分解することは，辞書にその分解規則を書いておくだけで処理できるので，短く区切られたものを，より長い単位にまとめ上げていくことより容易である．すなわち，辞書を参照することで,「〜て/おかねば/ならぬ」と点字規則に基づいた分かち書きを直ちに得ることができる．

2.3 点訳規則

図 2 に点訳規則適用の様子を示す．

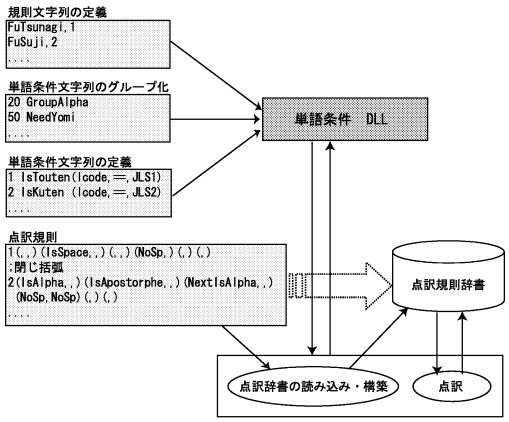

図 2: 点訳規則の概要

2.3.1 単語のグループ化

文節解析によって得られた単語の情報を用いて単語のグループを判定する規則を作った．現在 43 の規則がある．その一部を以下に示す．

```
;GroupAlpha (1-19)
1 IsTouten (lcode,==,JLS1)
2 IsKuten (lcode,==,JLS2)
（中略）
21 IsSpace (lcode,==,SPACE)
22 IsOpenBracket OR(word,==,（)(word,==,〔)
23 IsOpenKagi AND(lcode,==,JLS4)(rcode,==,JRS4)
24 IsCloseBracket OR(word,==,）)(word,==,〕)
25 IsCloseKagi (lcode,==,JLS3)
26 IsKigouAt (word,==,@)
27 IsKigouTen OR(word,==,…)(word,==,‥‥)
（中略）
52 IsSuru OR(lcode,==,JSV2)(lcode,==,JSV3)
(以下略)
```

”IsTouten””IsSpace”がそれぞれのグループ名で，これが次節にある点訳規則の「単語条件文字列」となる．その後に書かれている部分が，単語を分類するための規則を表現している．

2.3.2 点訳規則

点訳規則の規則数は現在 85 ある．その一部を以下に示す．

```
1 (,,)(IsSpace,,)(,,)(NoSp,)(,)(,)
（中略）
; 記号
20 (,,)(IsKigouAlpha,,)(GroupAlpha,,)(,NoSp)(,)(,)
21 (,,)(IsKigouAlpha,,)(,,)(,Sp)(,)(,)
22 (,,)(IsKigouTen,,)(,,)(Sp,Sp)(,)(,)
23 (GroupAlpha,,)(IsKigouAt,,)(GroupAlpha,,)
   (NoSp,NoSp)(,)(,)
(中略)
28 (,,)(IsKigouFuseji,,)(,,)(Sp,Sp)(,)(,)
29 (,,)(IsKigouBatu,,)(,,)(Sp,Sp)(,)(,)
30 (,,)(IsKigouSymbol,,)(,,)(,Sp)(,)(,)
(中略)
; アルファベット
50 (IsNumber,,)(IsAlpha,,)(IsSuffix,,)(NoSp,NoSp)
   (FuEiji,FuTsunagi)(,)
（中略）
; 数字
70 (IsPrefix,,)(IsNumber,,)(,,)(NoSp,)(,)(,)
（中略）
; カタカナ＋カタカナ
137 (IsKata,,)(IsKata,,)(,,)(ASpKataKata,)(,)(,)
（以下略）
```

点訳規則は IBUKI の文節解析によって切り出された文節内の単語に対して記述している．この規則は現在の単語とその前後の単語によって決まり，現在の単語に対して区切る／区切らない，後ろを区切る／区切ら

ない等を記述している．点訳規則の書式を以下に示す．

点訳規則優先順位ＩＤ
（前の単語条件文字列，不等号，単語サイズ）
（現在の単語条件文字列，不等号，単語サイズ）
（次の単語条件文字列，不等号，単語サイズ）
（単語の前の区切り，後の区切り（誤り可能性含む））
（単語の前の符号，後の符号）
（読みの曖昧性，内部の分かち書きの曖昧性）

点訳規則は，点訳規則順位 ID と 6 つの括弧で構成されている．それぞれの役割を以下の表 2 に示す．

表 2: 点訳規則

書式	役　割
点訳規則順位 ID	規則が複数選択された場合の優先順位．若い番号の規則を優先．
括弧１，２，３	単語の種類を記述している．「単語の条件文字列」には，単語グループなどを記述する．「不等号」と「単語サイズ」については点訳規則に単語の拍数または見出し語の文字数が影響する場合，単語サイズにはその文字数を書き，不等号には単語サイズに対する不等号を記述する．
括弧４	単語の前後の区切りに関する規則．（Sp：強制的に区切る　NoSp：強制的に区切らない など）
括弧５	単語の前後につく数符，外字符，つなぎ符等の記号に関する規則．
括弧６	単語の読みの曖昧性，単語内部の分かち書きに関して曖昧がある場合の規則．

3　現在の修正・改良の状況

3.1　日本点字表記法 2001 年版への対応など

IBUKI-TEN は日本点字委員会の「日本点字表記法」に基づいて点訳を行っている．「日本点字表記法」の最新版は 2001 年版になっている．IBUKI-TEN を作成したときは「日本点字表記法」1990 年版を元にしたが，最新の表記に対応していく必要があり，その対応をおこなっている．また 1990 年版では対応できていなかった部分に関しての修正も行っている．その主な点を以下に示す．

- **「～する」の表記**
  「～する」の点訳については点字表記法 2001 年版で改正があり，IBUKI-TEN では基本的に点訳規則によって対応することができた．以下にその点訳規則を示す．

```
;「する」について
90 (IsSahenN,,)(IsSuru,,)(,,)(1,)(,)(,)
91 (IsNounSuffix,,)(IsSuru,,)(,,)(0,)(,)(,)
92 (IsNumSuffix,,)(IsSuru,,)(,,)(0,)(,)(,)
93 (IsVerb,,)(IsSuru,,)(,,)(0,)(,)(,)
94 (IsKata,,)(IsSuru,,)(,,)(1,)(,)(,)
95 (,,)(IsSuru,,)(,,)(1,)(,)(,)
```

- **伏字，シンボルマークの表記**
  伏字（☆，〇等）については後に 1 マス空ける，また伏字に囲まれた場合（☆☆☆　連絡先　☆☆☆）は囲まれた文字列の前後 1 マスを空けるというルールがある．また，シンボルマーク（〒，℡等）の場合もマークの後に 1 マス空けるルールがある．そこで，伏字として利用されることがあるだろう記号（☆〇◇□△▽等）をまとめて，単語グループとして登録し．また，”×”は単独で単語グループとして登録した．”×”は算用記号としても扱われており，現在は伏字と同じ処理を行うようになっているが，将来的に規則を変更するときに分けてあるほうがよいと考えたためである．シンボルマークも同様に単語グループとして登録した．

```
33 IsKigouFuseji (lcode,==JLN8)
   (OR(wordcode,>=, ☆)(wordcode,<=, ▼)
34 IsKigouBatu (lcode,==,JLN8)(wordcode,==, ×)
35 IsKigouSymbol OR(word,==, 〒)
   (word,==, ℡)(AND(lcode,==,JLN8)
   (wordcode,>=, ㍾)(wordcode,<=, ㍼))
```

これらの単語グループを用いて点訳規則の修正を行い，伏字については前後 1 マスずつ空ける，またシンボルマークについては後の 1 マスあけるように修正した．(前頁の点訳規則 28-30)

- **URL，E メールアドレスの表記**
  ホームページの URL や E メールアドレス等を書き表す場合は，情報処理用点字の囲み符号を用いる必要がある．以下の図 3 にその表記を示す．

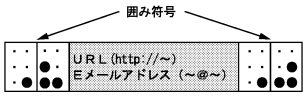

図 3: URL，E メールアドレスの表記

この表記については，IBUKI の修正も行い，URL，E-Mail の場合は URL という単語として IBUKI-TEN で受け取るようにして，点訳規則，単語グループ等を用いて対応した．その結果、図 4 のように点訳を行うようにした。

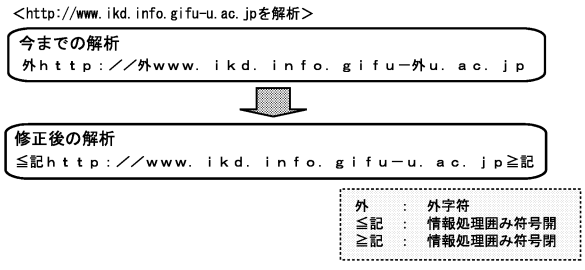

図 4: 点訳結果

3.2　ユーザからの要望への対応など

ダウンロード時に頂くメッセージの他に，IBUKI-TEN 掲示板や E-Mail で，ユーザからの質問や要望，意見を頂いている．これまでにおよそ 230 件，現在月におよそ 15 件のご指摘，要望，質問等を頂いている．以下にそのいくつかを紹介し，また今後の課題等について述べる．

- **点字プリンタをネットワークで共有したいので，プリンタ出力内容をファイル化して欲しい**．

  点字プリンタへは，
  １）点字コード（NABCC コード）
  ２）プリンタを制御するための共通コード（印字命令，改行，改ページなど）
  ３）プリンタの機種毎の制御コード（文字数，行数，紙サイズなど）
  を送信している．そのうち，１と２をファイル化する機能を追加した．(対応済)

- **時間を表示するとき，コロンのあとに数符が余分に入ってしまう**．

  時間表示「13:00」の場合，IBUKI-TEN では，「数符」「１」「３」「コロン」「数符」「０」「０」のようになっていたが，本来はコロンの後に数符は必要ない．このことについてはコロンの後に数符がつかないように修正し，またコロンの前に数字（数符）がない場合でコロンの後に数字がある場合は数符が通常通り付加されるように修正した．(対応済)

- **名前 : や住所 : のようにコロンを使用した場合，読みやすくするために，コロンの表記を変更して欲しい**．

  名前：や住所：のようにコロンを使う場合があるが，読みやすくするためにコロンの表示を変更して，表記させる必要がある．(検討中)

- **情報処理点訳の追加**

- **音声読み上げ機能の改修**

  IBUKI-TEN では，市販のソフトを利用して音声で点訳結果を読み上げることが可能であるが，読み上げるときに，文書の中の単語を飛ばして読み上げてしまうなどの問題が残っている．

- **インターフェースの使用性の向上（コピー＆ペースト等）**

4　IBUKI-TEN の公開状況

4.1　ダウンロード状況

IBUKI-TEN は 2000 年 9 月に WWW 上に公開し (バージョン 0.1)，ユーザからの指摘・要望・意見を取り入れながら数多くの改変・機能追加を行ってきた．途中で一時公開を中断したが (2001.12〜2002.11)，現在バージョン 0.53 になっている．これまでに我々が把握している 2004 年 1 月末現在のダウンロードの延べ数は 4,320，把握できる異なり数はおよそ 2,500 である．(途中でログを取り損ねた時期あり (2002.12〜2003.4).

各地の視覚障害者，盲学校，点訳ボランティアの間で使われており，また最近では中・高校で総合学習の授業教材として用いられているケースもある．

4.2　専門用語辞書の公開

IBUKI-TEN には初期辞書があるが，その他にユーザが単語追加等を行って作成するユーザ辞書を用いることができる．

先日，ユーザのお一人から鍼灸の専門用語辞書の提供，公開の申し出をいただいた．この辞書を IBUKI-TEN ホームページで公開し，公開してから 4ヶ月ほどたっているが，およそ 250 件ダウンロードされている．

課題の一つとして，ユーザ辞書に簡単に取り込めるようなツールを用意することも考えている．

5　おわりに

自然言語処理に基づく自動点訳システム IBUKI-TEN の概要について述べ，また公開の状況，現状の問題点，今後の課題などについて述べた．

今後も，ユーザからの協力を頂きながら辞書の充実（専門用語辞書，一般辞書）に努め，また IBUKI-TEN の不具合の改修，機能の向上，点訳精度の向上について研究開発を進めていきたい．

参考文献

[1] 自動点字翻訳編集システム IBUKI-TEN,
http://www.ikd.info.gifu-u.ac.jp/IBUKI-TEN/

[2] 日本点字委員会，日本点字表記法 1990 年版，1990.

[3] 日本点字委員会，日本点字表記法 2001 年版，2001.

[4] 日本電子化辞書研究所，EDR 電子化辞書仕様説明書，1995.

[5] 兵藤，横平，早川，村上，池田，誤り箇所指摘機能をもたせた点字翻訳編集システム IBUKI-TEN，電子情報通信学会論文誌 VOL.J84-D-I No.7，pp.1102–1111，2001.