自動点訳システムIBUKI-TENと点訳用辞書

兵藤安昭*　横平貫志*　早川哲史*　生川武史*
太田秀昭**　池田尚志*

*岐阜大学工学部
{hyodo,kanji,satoshi,narukawa,ikeda}@ikd.info.gifu-u.ac.jp
**(財) ソフトピアジャパン
ota@softopia.pref.gifu.jp

1　はじめに

本論文では，日本語テキストからの点字翻訳編集システムIBUKI-TENについて述べる．自動点訳システムは既に市販され広く使われているものもあるが，さらに高精度の点訳が望まれ，また特に自動点訳後の後編集を支援する機能の充実が望まれている．我々は，現在，IBUKI-TENの開発を行っており[1][2]，評価版を2000年9月にWWW上で公開し[3]，各地の視覚障害者，点訳ボランティアの間で利用されている．

日本語テキストに対する点字翻訳は，通常，漢字を仮名に変換し，分かち書きを行う必要がある．漢字を仮名に変換する場合は，点字の表記法に従って書き表す必要があり，分かち書きについても，適当に分かち書きすればよいというものではなく，読みやすさ，理解のしやすさという観点から，点字用の分かち書き規則が細かく決められている．IBUKI-TENでは，我々が現在，開発中の日本語文節解析システムIBUKIの文節解析結果をベースとして点訳を行う．点字翻訳は，基本的には，IBUKIが参照する辞書中に点訳規則を記述することで実現した．以下，まず始めに，点字翻訳における課題を述べ，次にIBUKI-TENシステムの概要を示す．次に，IBUKI-TENの点訳用辞書の詳細について述べ，最後に今後の課題について述べる．

2　点字翻訳での課題

現在，日本で広く用いられている6点点字では，1マスあたり63通りの文字しか表現できないため，1マスで日本語文字すべてを表すことはできない．そのため，漢字仮名混じり表記を仮名表記に変換する必要がある．しかし，仮名表記のみによるテキストは，読みにくいばかりでなく，意味が正確に伝わらない場合も多い．そこで，点字翻訳では，漢字を仮名に変換するだけではなく，分かち書きを行う必要もある．また，分かち書きは適当に分かち書きすればよいというものではなく，読みやすさ，理解のしやすさという観点から，点字用の分かち書き規則が細かく決められている．

2.1　点字分かち書き

点字分かち書きでは，「読みやすい，理解しやすい，誤読しない」の観点から，語の区切り目に関する2つの規則がある．第1規則は，文節ごとの基本的な区切りに関する規則である．第2規則は，複合語や固有名詞などの内部の区切りに関する規則である[4][5][6]．複合語に関する規則を以下に示すが，個別的・例外的な規則も多く複雑である．

1. 3拍以上の単語が2つ以上あれば，その境目で区切りを入れる．
   - 「建設業界」 … 建設 / 業界
   - 「仮名文字（2拍+2拍）」 … 仮名文字
2. ただし，2拍以下でも，自立性が強く意味の理解を助ける場合には区切って書き表す．
   - 「都市国家」 … 都市 / 国家
   - 「歯科医師」 … 歯科 / 医師
3. 連濁が生じる場合は区切らない．
   - 「株式会社（かぶしき がいしゃ）」… 株式会社
   - 「柱時計（はしら どけい）」… 柱時計
4. 接辞は続けて書き表す．
   - 「新 校長」… 新校長
5. ただし，意味を助け，発音上の切れ目が考慮できるものは，意味を明らかにするために区切って書き表す．
   - 「元 議員」… 元 / 議員

2.2 漢字仮名・点字表記変換

漢字仮名変換では,「行った」が「いった，おこなった」,「通った」が「とおった，かよった」のように文脈に応じて仮名表記が異なる場合がある．この場合は，前後の文法情報や文脈を考慮して意味的に正しい仮名表記を選択する必要がある．点字表記変換では，助詞「ハ，ヘ」を発音通りに「ワ，エ」と,「宇宙」などのウ列の長音は「ウチュー」と表すなど，点字規則で決められた表記への変換を行う．また，漢数字を含む単語の場合，数量の意味が強い単語は数字を用いて表すため,「四輪駆動」は「4りん/くどー」となる．しかし「再三再四」の場合は「さいさん/さいし」と表す必要がある．

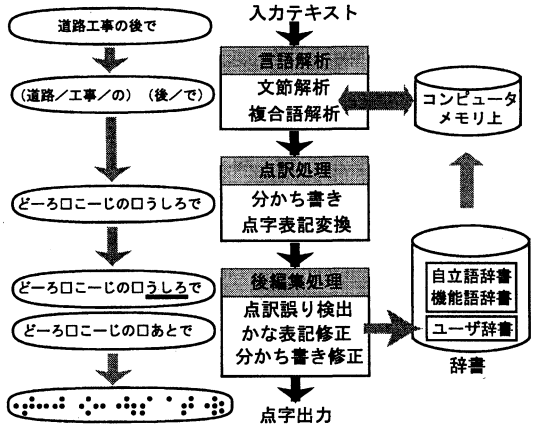

図 1: システム構成

3 点字翻訳システム IBUKI-TEN

3.1 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日本語解析システム IBUKI[7] による文節解析によって，文節単位の切り出しを行い，漢字連続文字は，複合語解析によって名詞，接辞等に分割する．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プログラム処理と，辞書に記述されている分かち書き規則により判別される．後編集処理では，自動点訳誤りの修正等を IBUKI-TEN のインタフェース上で行う．その際，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には，インターフェース上に色を変えて表示される．ユーザは，これらの情報を参考にしながら，点訳結果の後編集を行う．本システムは Windows 上で動作し，VisualC++を用いて開発を行っている．点訳実行時間 (点字分かち書きと漢字仮名変換) は，1,000 文 (47 文字/文) あたり 13 秒であった (PentiumIII650MHz, Memory:128MB，OS:Windows2000)．

3.2 点訳誤り箇所の検出

点字分かち書き誤りは，解析システム IBUKI の文節区切りに対する誤りと，複合語内の分かち書き誤りの 2 つに分類される．文節区切り誤りについては,「機能語を伴わない自立語のみの文節」「未登録語を含む文節」等を誤りの可能性があるとして抽出する．

また，2.1 節で述べたように，複合語は 3 拍以上の単語が 2 つ以上あれば，その境目で区切りを入れるが，2 拍以下でも自立性が強く意味の理解を助ける場合には区切って表すという規則がある．そこで，例えば，複合語に 2 拍以下の単語を含む場合は，区切ることとし，切りすぎの可能性があるとして指摘するようにした．

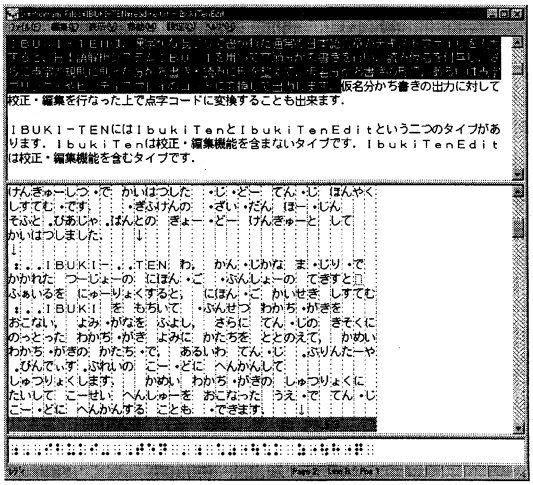

図 2: システムインターフェース

3.3 システムインタフェース

本システムのインタフェースを図2に示す．インターフェースは，上から順に，原文表示ウインドウ，点訳結果表示ウインドウ，点字表示ウインドウの3つから構成されている．点訳結果表示ウインドウには，自動点訳されたデータを，かな分かち書き形式で表示する．また6点による表示も可能である．点訳結果は，あらかじめ設定した1行の文字数，1ページの行数に従って表示されるので，ユーザはページレイアウトを考えながら編集が可能である．点字表示ウインドウでは，点訳結果表示ウインドウ上のカーソル行の点字を確認できる．

4 点訳用辞書

本システムの辞書構造を図3に示す．自立語辞書は，EDR 日本語単語辞書[1]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．

表記	点訳規則	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
再三再四	さいさん／さいし	副詞		0
にもかかわらず	にも／かかわらず	機能語		0
にも関わらず	にも／かかわらず	機能語		0
である	で／ある	機能語		0
たろう	だろー	機能語		0
でさえ	でさえ	機能語		0
・・・	・・・	・・・		

'／'：分かち書きの区切り

図 3: 辞書上の点訳規則

4.1 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字表記として登録してある．仮名表記は2.2節で述べたように点字規則に従った表記法で登録した．また，漢数字を含む単語の場合，数量の意味が強いかどうかにより表記を区別しなければならない．そこで，EDR 日本語単語辞書の各単語に付与されている「英語概念説明」中で，「one,first,second...」等の記述が現れる単語については，数字による表記とした．具体的には，現在，使用している単語中で，漢数字が含まれる単語は4,529語あったが，該当する単語は1,732語であった．

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．

4.2 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[9]．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．

[1]EDR 日本語単語辞書 [8] の公開を許可下さいました (株) 日本電子化辞書研究所の皆様に深く感謝いたします．

4.3 ユーザ辞書

IBUKI-TEN の点訳精度は最終的には辞書に依存する．IBUKI-TEN の自立語辞書は EDR 日本語単語辞書をベースに作成しているが，これで点訳が要求されるあらゆる分野に対応できるわけではない．そこで IBUKI-TEN では，ユーザが独自に，新たな単語の追加したり，システム辞書上の単語に対して，単語の削除，仮名表記の追加/変更，区切り情報の変更が容易にできるようにした．

現在，IBUKI-TEN の WWW 上での掲示板や E-Mail 等でシステム辞書内の点訳規則の誤り情報等を収集しており，システムのバージョンアップとともにシステム辞書の更新も進めている．

5 おわりに

高精度の日本語自動点訳を目指し，また自動点訳後の後編集支援機機能の一つとして誤り個所指摘機能を持たせ，視覚障害者も利用可能な点字翻訳編集システム IBUKI-TEN について述べた．本システムの評価版を 2000 年 9 月に WWW 上で公開し，各地の視覚障害者，点訳ボランティアの間で利用されている．

今後の課題は，インターネットを通じてユーザが登録した辞書を収集するなどして，固有名詞辞書や分野毎の辞書を整備し，点訳精度の向上を図りたいと考えている．

参考文献

[1] 兵藤，横平，早川，池田，辞書データ主導型の自動点字翻訳システム IBUKI-TEN，電子情報通信学会技術研究報告，WIT99-22，pp131–136，1999.

[2] 横平，兵藤，早川，生川，村上，太田，池田，自動点字翻訳システム IBUKI-TEN の校正支援機能，電子情報通信学会技術研究報告，WIT00-18，pp37–42，2000.

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

[4] 全国視覚障害者情報提供施設協議会，点訳のてびき第 2 版，1991.

[5] 日本点字委員会，日本点字表記法 1990 年版，1990.

[6] 阿佐博監修，遠藤謙一著，初心者のための点字・点訳完全マスター，1999.

[7] 兵藤，池田，文節単位のコストに基づく日本語文節解析システム，言語処理学会第 5 回年次大会，pp.502–504，1999.

[8] 日本電子化辞書研究所，EDR 電子化辞書仕様説明書，1995.

[9] 兵藤，池田，文節解析のための長単位機能語辞書，言語処理学会第 6 回年次大会，pp.407–410，2000.