自動点訳/編集システムIBUKI-TENと辞書開発支援システムIBUKI-TOOL

岸井 謙一* 辻子 純央* 渥美 誠* 丹羽 智彦* 山田 祥広* 太田 秀昭** 池田 尚志*

*岐阜大学工学部 {kishii,tsujiko,atsumi,niwa,yamada,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，テキスト文から点字への自動点訳システムIBUKI-TEN[3]と，テキストを解析して辞書開発を支援するシステムIBUKI-TOOL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．我々は自動点訳システムIBUKI-TENの開発を行い，さらに，評価版をWWW上で公開してユーザから得られたフィードバックをもとにシステムの改良を進めてきた．

IBUKI-TENは，我々が現在開発中の，日本語解析システムIBUKIの文節解析結果をベースとして点訳を行うシステムである．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．IBUKI-TENシステムのアプリケーションとしてIbukiTen・IbukiTenEditを開発した．IbukiTenは，自動点訳のみを行うもので，視覚障害者も簡単な操作で自動点訳が可能である．また，IbukiTenEditは，主に点訳ボランティア向けに開発したもので，IbukiTenに点訳結果の後編集機能を加えたものである．これら2つについては現在も改良を進めていて，英語2級点字への点訳・IbukiTenEditの後編集機能・PrintServer機能，音声インターフェース化など改良・機能拡充を進めている．また，IBUKI-TENの点訳エンジンを利用した自動点訳サーバibukiTenServerや，HTML文書の自動点訳システムも開発した．

IBUKI-TENの実用化においては，解析誤り箇所や未登録語を見い出し，辞書の修正や登録をしていく等の作業が欠かせない．そこで我々は，任意のテキストデータを入力として，IBUKIで解析された結果をさまざまな角度から眺めることができ，また辞書登録，再解析が簡単に行えるなどの機能を備えたテキスト解析ツール（ibukiTool）を開発した．

以下，IBUKI-TENの概要と拡充内容を述べ，次にIBUKI-TOOLについて紹介する．

2 IBUKI-TENの概要

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[1]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対し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．

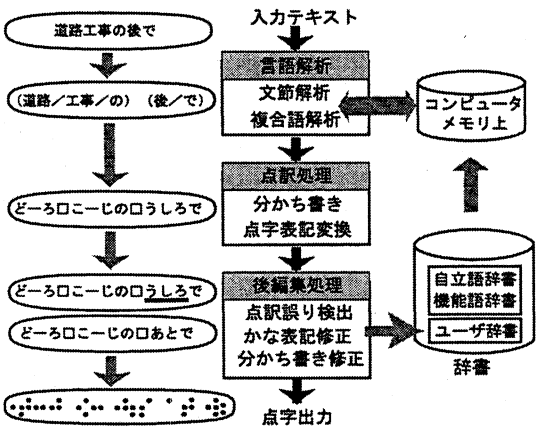

図 1: システム構成

3 IBUKI-TENの機能拡充

3.1 英語2級点字への点訳

3.1.1 IBUKI-TENにおける英語点訳処理

IBUKI-TENシステムの点訳処理（図1）部分に英語点訳モジュールを追加することで，2級英語への自動点訳を実現した．

2級点字とは，アルファベットや記号類1文字に対

し，ひとつの点字記号を割り当てて点字表記を行う１級点字に，文章量削減や読解速度の向上を目的として特定の単語や文字列を省略する規則約２００種類を追加した英語点字の表記体系である．

IBUKI-TEN での点訳処理中にアルファベット表記文が出現した場合は，まず単語単位への切り出しを行う．次に切り出された単語文字列に対して適応可能な省略規則を，規則を記述したファイルを参照して全て検出する．それぞれの規則には適用に際しての制限や規則間の優先順位があるため，見つかった規則の中から最適なものを採用して，それを点訳結果とする．最適な規則の決定はプログラム処理で行っているが，誤った点訳となってしまう単語もあるため，それらは別登録することで対応している．

3.1.2 評価と考察

市販の自動点訳ソフトとの比較実験を行った．使用したデータは英語版「不思議の国のアリス」(総単語数 26560 語) である．表 1 はその結果である．

	語数		単語異なり数	
総数	26560		2673	
異なる点訳結果	1880	7.10%	56	2.10%
市販ソフトの誤訳	917	3.45%	14	0.50%
IBUKI-TEN の誤訳	971	3.66%	44	1.65%
(ear) に関する誤訳	97	0.37%	17	0.64%

表 1: 点訳精度比較結果

この結果から，異なり検出から推測できる今回使用したデータの点訳精度は，市販ソフトで 96.55%，IBUKI-TEN で 96.34%であり，ほぼ同程度の精度であることがわかった．誤訳の原因を調べたところ，文字列 [ear] の点訳ミス (0.37%) の占める割合が多く，これを修正すれば正解率が 96.71%まで向上することもわかった．

3.2 HTML 文書の点字での閲覧システム

視覚障害者のホームページ閲覧を支援するシステムを開発した．このシステムはインターネット上の HTML 文書を取得し，点字コードに変換するものである．まずインターネット上から取得するが，HTML 文書は Shift-JIS や EUC など異なる文字コードで記述されているため，この時 Shift-JIS コードへの変換も行う．次に HTML 文書を，点字表記したときに読みやすい形にするために，HTML 文書を構成している各種要素についてタグの取り除きや代替処理を行う．

ex) ＜IMG＞… ［画像 (alt 属性値)］

こうして整形されたものを IBUKI-TEN を利用して点訳することで，HTML 文書のピンディスプレイによる閲覧や，音声読み上げ・点字印刷などが可能となった．

3.3 IbukiTenIbukiTenEdit などの機能拡充

IbukiTenEdit における後編集を容易にするため，機能の拡充を行った．

点訳結果文に対する検索・置換機能を追加した．一般の文字検索と異なる点として，点訳結果が点字表記になっている場合には，６点入力による検索を行う機能がある．また点訳結果が分かち書き表記になっている場合の配慮として，「が」と入力されたとき，自動的に「か゛」を検索できるようにする機能もある．

ユーザからの要望を受け，後編集時に使用するショートカット機能の整備を行い，ユーザ環境に応じたカスタマイズも可能にした．

テキスト画面に対して読み上げを行う 95/2000Reader や PC-Talker のような画面読み上げソフトがある．これらを利用して，点訳結果を様々な範囲や方法で読み上げる機能や，読み上げを途中で中止する機能を追加した．これらを利用することで視覚障害者が点訳結果を確認することや，後編集を行うことが容易になると考えられる．

IBUKI-TEN は盲学校などでも利用されているが，要望として校内ＬＡＮを利用した点字印刷機能の追加というものがあり，プリントサーバとクライアントソフトの開発を行った．

4 ibukiTool

我々は，任意のテキストデータを入力として，IBUKI で解析された結果をさまざまな角度から眺めることができ，また辞書登録，再解析が簡単に行えるなどの機能を備えたテキスト解析ツールを開発した．解析結果は市販の RDB に格納しており，RDB の諸機能を利用してツールを構成している．ibukiTool は，未登録語の処理など辞書作成ツールとして活用できるばかりでなく，小説などの著作物の語彙統計を取るなど文書解析ツールとして便利に活用できる．

4.1 IBUKI での解析

IBUKI はまず形態素解析を行い，ついで構文解析を行う [1]．文節解析では，文節として可能性のあるものをすべて求めた上で，文節単位のコスト最小法で解を求めている．文節には文節カテゴリ（構文的な観点から文節を分類したもの）を付与する．係り受け解析では文節カテゴリを元に，係り受け可能な文節を求め文節間の関係を解析している．ibukiTool は現在

のところ IBUKI の形態素解析を利用している段階である．

4.2 ibukiTool の構成

本研究で開発した ibukiTool は IBUKI の形態素解析を利用したテキスト解析ツールである市販のデータベースと組み合わせることで単語や複合語，未知語をはじめとした語彙統計や条件による抽出等を容易に行える．また各種規則をデータベースの形式に変換し，後述する辞書や規則の編集を行うことができる．操作等は DB 上で行い，IBUKI に関わる処理は DLL 化しており，必要に応じて DB 上よりコールすることになる．

ibukiTool の構成を図 2 に示す．

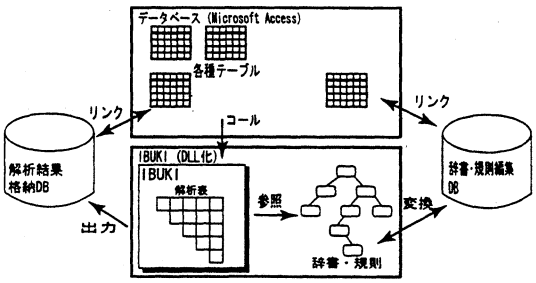

図 2: ibukiTool システム構成

ibukiTool の操作は市販のデータベース上のフォームから行う．これには Microsoft 社の Access を用いている．入力テキストの解析等，IBUKI を用いた処理については DLL を作成し DB（VBA）からコールする形で行っている．IBUKI による解析結果もまた DB に格納され，これにより統計や抽出等の作業が容易に行える．これは ibukiTool 本体とは別のファイルとして作成され本体からのリンクによりそのデータを表示・操作することができ，データの保存も容易である．

4.3 ibukiTool の諸機能

ibukiTool の諸機能としては以下が挙げられる．

1. 入力テキストファイルを指定して形態素解析
2. 任意のキー入力されたテキストを形態素解析
3. 解析された語の一覧を表示（延べと異なり）
4. 解析された複合語の一覧を表示（延べと異なり）
5. 未登録語として解析された語とその出現文脈を表示（延べと異なり）
6. 語の一覧を，出現順，読みの順，品詞，出現頻度順などで整列して表示
7. 適当にフィルタリングして，条件に合うものだけを表示
8. 解析された語が出現した文を表示
9. 解析された語の辞書情報を表示
10. 解析誤りの可能性ありと判定した箇所を表示
11. メモリ上の辞書に語を登録
12. メモリ上の辞書から語を削除
13. 新しい辞書でテキストを再解析
14. メモリ上の辞書をファイルに保存
15. メモリ上の辞書をデータベースに変換
16. データベースに変換した辞書を編集
17. 編集した DB の内容を辞書ファイルに変換
18. メモリ上の接続規則をデータベースに変換
19. データベースに変換した接続規則を編集
20. 編集した DB の内容を接続規則ファイルに変換

解析誤りの可能性に関しては，独立文節（機能語を伴わない文節）や，終端文字がひらがな小文字の文節等，統計的に誤りの多い箇所を抽出しており，適合率は 14 %，再現率は 70 %程度である [2]．なお，IBUKI の解析精度は，EDR 日本語コーパスのデータに対して 98 %程度である [1]．また辞書としては，自立語については EDR の辞書をベースにしており，機能語については長単位の表現を採用するという方針で我々の研究室で作成したものを用いている．

4.4 辞書編集

特徴的な機能として辞書編集を説明する．IBUKI は辞書ファイルを読み込むと，メモリ上にパトリシアという木構造の辞書を構築する．ibukiTool では辞書編集として 2 種類の方法を実装した．一方はメモリ上のパトリシアに情報を追加していく動的な方法であり，もう一方はパトリシアを RDB に出力して RDB 上で編集しその後テキストファイルに変換するという静的な方法である．前者は変更が即時に行われるが辞書の中身が把握しにくく，後者は直観的な編集作業が可能だが変換処理等に時間がかかるという特徴がある．

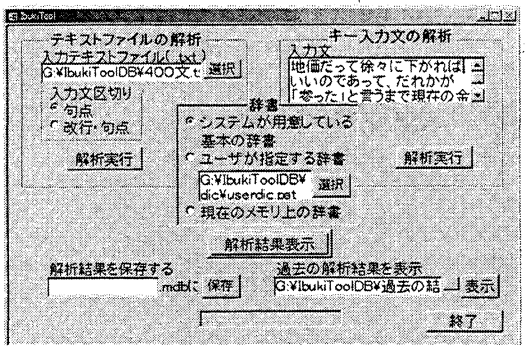

図 3: ibukiTool

図 4: 解析結果

5 おわりに

自動点訳システム IBUKI-TEN に 2 級点字への英語点訳機能を追加した．HTML 文書の点字による閲覧システムを開発した．自動点訳ソフト IbukiTen・IbukiTenEdit の機能拡充を行った

辞書編集機能をもつテキスト解析ツールを開発した．文書データを解析し，単語や複合語の一覧・語彙統計，未登録語，誤り可能性のある文節等の情報を抽出・表示でき，辞書や解析規則の整備機能を実装した日本語テキスト解析ツールを開発した．

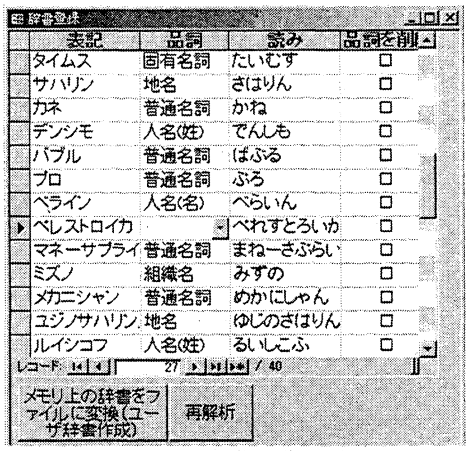

図 5: 辞書編集作業 (動的)

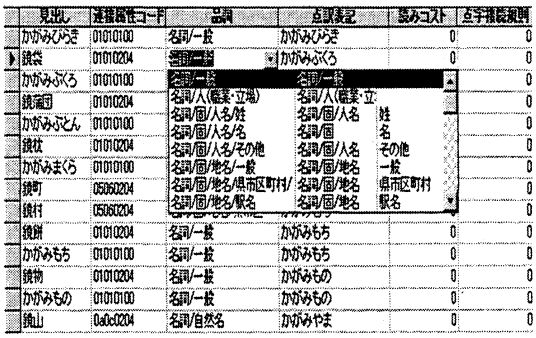

図 6: 辞書編集作業 (静的)

参考文献

[1] 兵藤安昭, 池田尚志：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 (1999)

[2] 村上裕, 神光太郎, 兵藤安昭, 池田尚志：複合語・文節解析誤り個所の検出, 言語処理学会 第 7 回年次大会 (2001)

[3] 兵藤安昭, 横平貫志, 早川哲史, 村上裕, 池田尚志, 誤り検出機能を備え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.J84D1, No.7, pp1102-1111, 2001

[4] 福井哲也, 初歩から学ぶ英語点訳, 日本点字図書館, 1991

[5] 辻子純央, 池田尚志, テキスト解析ツール ibukiTool について, 情報処理学会第 63 回 (平成 13 年後期) 全国大会, 2-21/22, 2001