第 6 章

点字翻訳ボランティアのための対話型分かち書き支援手法

6.1 はじめに

バリアフリーというキーワードの下に各種福祉機器の開発やパソコンソフトの開発が企業や大学で進められている。なかでも視覚障害者向けには点字ピンディスプレイや音声合成装置などを用いて、コンピュータによる積極的な情報処理教育、職業訓練が行われている。このためにはコンピュータのマニュアルや教科書等を点字に翻訳する必要があるが、点字翻訳ボランティアの数は少なく、年間、１人のボランティアが翻訳できる専門書は３、４冊程度である。

日本語を点字に翻訳するシステムは過去にいくつか提案されており、市販されているものもある。日本アイ・ビー・エムの嘉手川らは約７７０００語の基本単語辞書を用いて分かち書きと漢字かな変換を行うシステムを開発した [17]。筑波技術短期大学の河原は市販の点字翻訳プログラムの誤りを解析し、ＩＣＯＴの形態素解析辞書を用いて点字翻訳結果の改良を行うシステムについて報告している [18]。このような状況のなかで、点字翻訳ボランティアにとって最も時間がかかり、難しいとされている分かち書きを自動的に行い、かつ、誤っている可能性のある箇所を指摘して初級点字翻訳ボランティアの分かち書きを支援する方法について考察し、試作システムを構築したのでそれについて報告する [37],[38],[95]。

一方、対話システムとしては、最近、情報機器との自然なコミュニケーションを目指して様々な対話システムやユーザインタフェイスの研究が行われている [56]。ここでは筆者らの提案する対話型システムに最も類似した機能をもつと考えられる、畑田らのＯＣＲの誤り修正支援システムとの比較検討を行う。

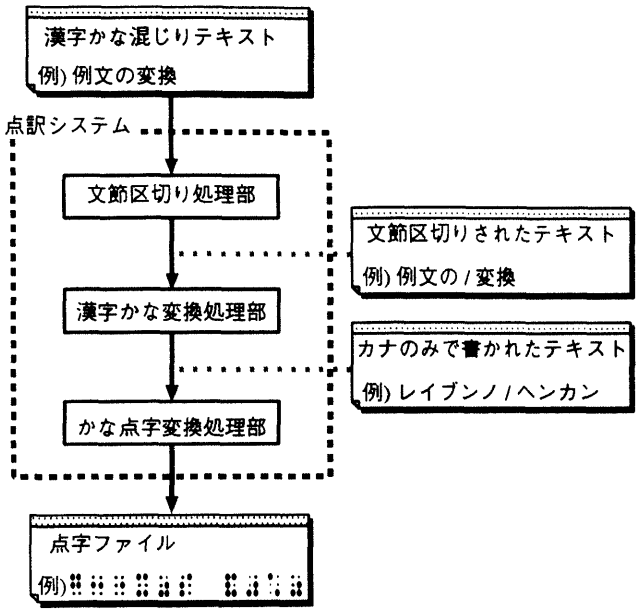

図 6.1: 点訳の手順

6.2 分かち書きの規則と問題点

6.2.1 点字翻訳のための分かち書き

点字翻訳 (以下、本文中では点訳と称する) は一般に図 6.1に示したような手順で行われる。すなわち、漢字かな混じり文を点字の規則に従って分かち書きを行ったのち、漢字に読みをつけ、読みを表音文字体系に変換し、最後に点字出力形態に合わせて点字を出力する。

従来この作業は、点訳ボランティアによりすべて手作業で行われていた。1 字 1 字点筆を用いて打点するため、修正するにはそのページ全体を打ち直す必要があり、非常に時間がかかる。最近、パソコンで動作する点訳プログラムが入手できるようになって点訳の効率は格段に上がった。しかし完全自動化には限界があり、点訳プログラムはあまり利用されていない。その主な原因は、次の 2 点と考えられる。

(1) 通常の自然言語処理における形態素解析や構文解析では点字特有のパターンを全て表現するのは困難であり、文節レベルより細かい区切りが必要とされる。このため、点訳ボランティアは点訳プログラムで区切り処理されたカナあるいは点字の文章全体を再度見直す必要がある。

(2) 日本語の漢字は複数の読みをもつものが多く、読みを一意に決定することは困難である。従って、点訳ボランティアは点訳プログラムが与えた読みについても見直して修正する必要がある。

原因 (1) について補足説明する。第 1 番目の問題として形式名詞の問題がある。形式名詞を要素として含む助動詞「(～する) ことだ」の場合、一般の形態素解析では直前の文節に続けて 1 つの文節とされるが、点字の場合は「こと」の前で区切らなければならない。

第 2 番目に複合語の問題がある。「漢語＋する」の形のサ変名詞は、「研究する」だと点字の分かち書きでは区切らない。しかし、これが複合語化して「共同研究する」となった場合は、「共同」「研究」「する」と 3 つに区切らなければならない。日本語のように漢字を組み合わせて容易に新しい単語を作る言語の場合、従来の形態素解析は複合語を構成する基本となる漢字 2 字熟語、漢字 3 字熟語を辞書に登録しておくことで処理を行う。しかし、点字の分かち書きの場合、その単語が何文字の熟語から構成されているかが認定されるだけでは正しく分かち書きを行うことができない。いくつの漢字 2 字熟語、漢字 3 字熟語が組み合わされているか、という情報が必要になる。

6.2.2 従来方式とその問題点

日本語文の文節区切りは従来、点訳の分かち書き用としてではなく、機械翻訳、日本語によるデータベース検索、文書校正支援等のための形態素解析として、佐藤 [29]、長尾 [52] らによって研究されてきた。これらの手法は大規模な文法情報付きの辞書をそれぞれ独自に用意して形態素解析を行うものであった。

一方、最近では言語資源の共有やモジュラリティの観点から春野 [61]、颯々野 [28]、松本 [73] らがより実用的な形態素解析システムについて提案している。これらのシステムは形態素解析結果の曖昧性については考慮しているが、第 1 候補のみを示す一括処理を基本としている点、文節より細かい分割レベルの解析は行わない点などの理由で、点訳へ直ちに適用するには困難がある。

一般の形態素解析では与えられた入力文に対して単語辞書や文法辞書、それにユーザ辞書などを用いて処理が行われ、文節と認定された単位で区切られる。機械翻訳に用いられる場合と音声合成のために用いられる場合とでは、文節の単位が異なり、一意に決定できるものではない。このような形態素解析方式を点訳に適用するには以下のような問題点があると考えられる。

(1) 点訳に必要な分かち書きの単位としては、一般の形態素解析によって決定される文節よりもさらに細かく分割する必要がある。

(2) 後に続く処理のフィードバックによって前の処理の誤りが発見、修正される一般の自然言語処理と異なり、形態素解析の誤りをその後の点訳処理で回復することが難しい。

(3) 点訳のための分かち書きでは、「連濁」や複合語化によって区切り方が変わるにもかかわらず、従来の形態素解析ではそのような「読み」や複合情報までの詳しい出力が得られない。複合情報とは、点訳のための分かち書きに必要となる連濁や複合語の語数などを示す情報のことである。

6.2.3 提案する手法の基本方針

前述のような問題点をふまえ、筆者らは従来より行われてきた形態素解析に用いられるような大規模な辞書と文法規則を用いず、簡便な表層解析のみを行うことによって分かち書きを行い[37]、対話的に誤りを修正していくことを試みている[38],[95]。対話機能を導入することにより点訳ボランティアの見直しの手間が軽減され、点訳作業にかかる時間の大幅な短縮が期待される。この手法をとり入れた実験的システムを構築することを試みた。システムの基本方針は次の 4 点である。

(1) 分かち書きの処理を自動分割と対話処理の 2 段階で行う。

(2) 文法情報を含まない単語のみからなる 7 種類のテーブルを用いて表層解析を行う。

(3) 自動分割の際に用いる分かち書きの規則を表層情報に基づく知識に書き換え、知識ベース化する (以下、知識ベースAと称する)。

(4) 自動分割による分かち書きが疑わしい箇所および、表層情報では一意に決定できない分かち書き箇所を対話処理でユーザに提示するための規則を知識ベース化する (以下、知識ベースBと称する)。

6.3 実験システムの構成と分かち書き手法

6.3.1 実験システムの構成

今回構築した試作システムの構成を図 6.2に示す。本システムは自動分割部と対話処理部の 2 つから構成される。

自動分割部では以下のような処理を行う。最初に入力文から、各種テーブルと字種情報を用いて表層情報を抽出する。ここで得られた表層情報を基に、知識ベースAの知識を用いて分かち書き箇所を決定する。分かち書きの知識が競合した場合には、優先点数の高い知識を優先する。自動分割部では熟練ボランティアによって作成された正しく分かち書きされたファイル (以後「正解ファイル」と称する) を用いて知識ベースAの知識をチューニングするための機能をもつ。

対話処理部では、自動分割部の処理結果のうち、信頼性が低い箇所を知識ベース B を用いてユーザに示したり、自動分割に用いた知識を表示したりする。ユーザはシステムによって指摘された箇所のみを順にチェックすることにより、容易に分かち書きを行うことができる。

それぞれの処理について順に説明する。

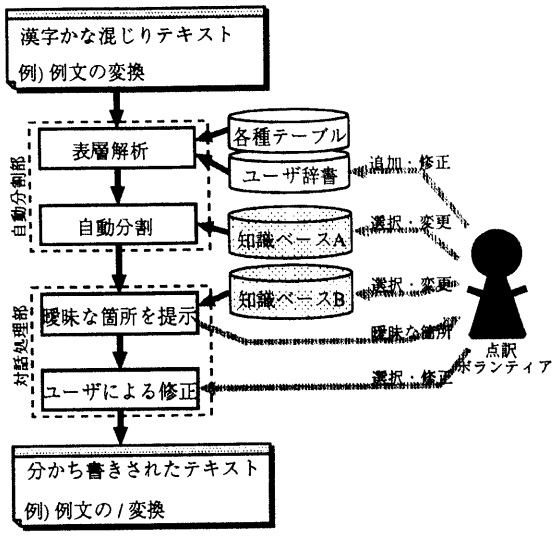

図 6.2: 対話型分かち書き支援システム構成図

表 6.1: 表層解析で用いるテーブル

テーブル名	書式	語数	容量 (kbyte)	例
ひらがな書き自立語	単語:区切り方	250	0.3	しかし：前後, はっきり：前
助詞	単語	25	2.3	と, は
漢字 2 字熟語	単語	62,512	312.6	圧力, 計算
漢字 3 字熟語	単語	24,262	169.3	亜熱帯, 委員長
接頭語	単語	54	0.1	不, 副
接尾語	単語	67	0.2	機, 的
混ぜ書き語	単語	10,826	75.8	引き算, 繰り返

6.3.2 自動分割部

表層解析

提案する分かち書きを行う、表層解析について説明する。ここで表層解析とは、漢字かな混じりテキストの字面から判別できる字種や、以下に示す表層情報を抽出することである。具体的には、表6.1に示す 7 種類のテーブル名が示す情報を表層情報と定義し、文中のどこからどこまでがそれぞれのテーブルの要素と一致したかを調べる。形態素解析では辞書をひいて単語ごとに文法情報を得、文法情報に基づいて単語の連接を詳しく調べていくのに対して、表層解析では、文に出現する文字列の表層情報をテーブルを参照しながらチェックするだけで、前後の語と語の関係については考慮しない。

このように、表層解析で用いるテーブルはここに示した 7 つであり、従来の形態素解析辞書の

ような文法情報は持っていない。ひらがな書きの自立語テーブル以外はすべて単語のみからなる。ひらがな書き自立語テーブルでは、各ひらがな自立語ごとに、その前および後で区切るかどうかの情報を持たせている。

ひらがな書き自立語とは「しかし」「はっきり」のようなすべてひらがなで表記される自立語を指す。これらは表層解析の際、分かち書きの重要な目安となる助詞との区別がつきにくいうえ、ひらがな書き自立語どうしが接続した際、分かち書きを誤る可能性が高い。例えば、「しかし」は独立して出現するので「しかし」の前後は分かち書きを行なう、すなわち「前後で区切る」となる (表6.1の例参照)。「はっきり」の場合、「はっきりと」「はっきりした」「はっきり／示す[1]」というように後続する語によって後ろの分かち書き方が変化する。従って、「はっきり」については「前で区切る」となる (表6.1の例参照)。

これらのテーブルはＥＤＲ日本電子化辞書研究所の「日本語基本単語辞書」から情報処理関連の単語を抽出し、ひらがな書き自立語については区切り方の情報を人手により入力した。また、接頭語、接尾語については文献 [52] を参照し、実際のテキストから抽出して作成した。ひらがな書き自立語以外のテーブルは単語のみからなるため、語の追加・削除が容易である。また、ひらがな書き自立語についてもユーザは容易に修正できる。

[1] '／' は区切りを現す

表 6.2: 知識ベースAの知識一覧

分類	識別番号	知識	優先点数
句読点	知識 1-1	句読点の前で区切らない	9
	知識 1-2	句点の後ろで 2 回区切る	10
	知識 1-3	読点の後ろで区切る	9
括弧	知識 2-1	括弧の後ろは区切らない	8
	知識 2-2	括弧閉じの前は区切らない	11
	知識 2-3	鈎括弧‘「’の前で区切る	8
	知識 2-4	括弧閉じ→{漢字 or カタカナ}間で区切る	1
字種の変わり目	知識 3-1	ひらがな→漢字間で区切る	3
	知識 3-2	ひらがな→カタカナ間で区切る	1
	知識 3-3	カタカナ→漢字間で区切る	1
	知識 3-4	漢字→カタカナ間で区切る	1
	知識 3-5	{漢字 or ひらがな or カタカナ}→アルファベット間で区切る	5
	知識 3-6	アルファベット→{漢字 or ひらがな or カタカナ}間で区切る	5
	知識 3-7	{漢字 or ひらがな or カタカナ}→数字間で区切る	7
	知識 3-8	{漢字 or ひらがな or カタカナ}→(区切る種類の) 記号間で区切る	7
	知識 3-9	(区切る種類の) 記号→{漢字 or ひらがな or カタカナ}間で区切る	5
	知識 3-10	{漢字 or ひらがな or カタカナ}→(区切らない種類の) 記号間で区切る	7
	知識 3-11	(区切らない種類の) 記号→{漢字 or ひらがな or カタカナ}間で区切る	5
助詞	知識 4-1	助詞の前であれば区切らない	4
	知識 4-2	助詞の後ろを区切る	3
自立語	知識 5-1	{前後 or 前}を区切る種類のひらがな書き自立語の前を区切る	7
	知識 5-2	{前後 or 後}を区切る種類のひらがな書き自立語の後ろを区切る	5
	知識 5-3	ひらがな書き自立語の内部は区切らない	4
混ぜ書き語	知識 6-1	混ぜ書き語の前を区切る	5
	知識 6-2	混ぜ書き語の内部は区切らない	6
漢字熟語	知識 7-1	漢字熟語の前を区切る	5
	知識 7-2	漢字熟語の後ろが漢字ならば漢字熟語の後ろで区切る	1
接頭・接尾語	知識 8-1	接頭語の前を区切る	3
	知識 8-2	接尾語の前を区切らない	2
	知識 8-3	接頭語の後ろは区切らない	6
	知識 8-4	接尾語→漢字間で区切る	1
空白	知識 9-1	空白の前を区切らない	9
	知識 9-2	空白の後ろで区切らない	8
その他	知識 10-1	「て、ば」の前に漢字があるなら「て、ば」の後ろで区切らない	4
	知識 10-2	{第 or 約}→数字間では区切らない	8
	知識 10-3	数式の後ろで 2 回区切る	8
	知識 10-4	促音・拗音・撥音の前は区切らない	9
	知識 10-5	「お」→漢字間は区切らない	6
	知識 10-6	「各」の後ろで区切る	5

自動分割のための知識ベース

分かち書きに必要な知識を表6.2に示す。ここで表 6.2の第3列目に表現されている「知識」内のことばについて説明を加えておく。分かち書きには3種類の区切り方がある。1つ目が語と語の間に1文字分のスペースをあけることで、表中、「区切る」とはスペース(空白)をおくことを意味する。とくに「2回区切る」とは、2つのスペースを連続して置くことで、このようなことは句点の後や倒置文において生じる。3つ目がスペースをあけずに続けて書く場合で、これを「区切らない」と書く。

知識の総数は現在のところ３９個である。知識は以下の4種に大別できる。

(1) 句点、かっこ、スペースに関する知識

点訳のための分かち書きを行う上で、必ず守らなければならない知識を導入した。これらの知識には高い優先点数を与えた。

(2) 字種の変わり目に関する知識

字種の変わり目に着目した知識を導入した。様々な字種の組合せについて１１個の知識で対応する。

(3) 漢字熟語、ひらがな書きの自立語、混ぜ書き語、助詞、接頭語・接尾語に関する知識

字種の情報だけでは区切り方の曖昧な箇所について各種テーブルから得られる表層情報を基に区切り方を決定するための知識を導入した。

(4) その他に関する知識

上記 (1) ～ (3) を用いても区切り方が曖昧な箇所に対して導入した。

知識の獲得は、通常のエキスパートシステムにおける知識ベースの構築手順に従って行った。

手順としては、「点訳の手引」等に示されている分かち書きの規則のうち、表層情報を用いて記述できるものを知識として知識ベースに加えた。表6.2のうち、句読点の分類の１－１、１－２、１－３、助詞の４－１、４－２がこれにあたる。また、「点訳の手引」では学校文法風に記述されている規則、例えば、「自立語の前を区切る」という規則を表層情報を用いて記述し、知識ベースに加えた。表6.2における字種の変わり目の分類にある、３－１～３－１１、５－１～５－３、６－１、６－２がこの例である。また、予備実験の結果、表6.2のその他、１０－１～１０－６のような知識を知識ベースに加えた。これらは情報処理関連の文献に出現することが多く、表情情報を用いて表現可能な知識である。このようにして知識を記述していくが、表層情報だけでは記述不可能な規則については対話処理で取り扱う。

知識同士が競合した場合にどの知識を選択するかを決定するために各知識には優先点数を設定した。例えば、「機械的」のような漢字列について考える。「機械」は漢字2字熟語のため、表6.2

の知識7－2により、「械」と「的」の間を区切るという規則が適用される。一方、知識8－2により、接尾語の前は区切らないという規則が適用され、知識が競合する。この際、それぞれの知識に付与されている優先点数をみると、7－2の知識が1点、8－2の知識が2点で、8－2の知識の方が優先度が高いことがわかる。よってこの場合、知識8－2が適用され、「械」と「的」の間は区切らない。これは分かち書きとしては正解となる。

知識ベースAのチューニングのため、筆者らは熟練ボランティアによって正しく分かち書きされた正解ファイルを用意した。正解ファイルを用いることにより、知識ベースAの知識によって正しく区切られる箇所を知ることができ、知識のチューニングを容易に行えるようにした。

例えば表6.2の知識1－3は以下のように簡潔に表現するようにした。

知識1-3　R_3 ＝　(1-3 , 9 ,
if (前の文字 ＝ 読点) ,
then 区切り方 ＝ 注目している文字間で区切る)

ここで、右辺の第1項が知識の識別番号、第2項が優先点数、第3項が知識適用の条件部、第4項がその知識を適用した場合の区切り方である。第4項については前述のとおり「区切らない」「区切る」「2回区切る」の3通りがある。筆者らの提案する分かち書き知識は、文字と文字の間に注目して、その文字間を「区切らない」か、「区切る」か「2回区切る」かを順次判断していく。

自動分割部におけるユーザインタフェイス

自動分割では、ユーザはツールバーに表示されているボタンをクリックすることにより、あるいはメニューバーからプルダウン式に表示されるメニューを選択することにより、「1段落ごと」あるいは「全文一括」のモードで分かち書きを行うことができる。「全文一括」モードでは、文書の全段落の現在何段落目を処理中であるかがダイアログボックスに表示される。

6.3.3 対話処理部

対話処理のための知識ベース

対話処理部における知識の構築の基本方針は次の2つである。

(1) 分かち書きの規則のうち、表層情報だけでは曖昧で区切り方を一意に決定できない規則を知識とした。

(2) 自動分割処理の結果、誤っている可能性が高い区切り箇所の特徴をまとめて知識とした。

(1)については例えば、点訳のための分かち書きの規則として、

- 助動詞の「ない」の前は区切らない
- 形容詞の「ない」の前は区切る

という規則がある。

この場合、「ない」の品詞が決定できないと正しく分かち書きが行われない。筆者らが今回採用している表層解析では「ない」の字種と「ひらがな書き自立語」であるという情報しかもたないため、「ない」の直前の区切りは常に曖昧である。このような分かち書き規則に対し、対話処理で用いる知識ベースBでは「『ない』の直前の区切りは曖昧であるのでユーザに提示する」という結論部によりユーザの指示を待つ。

(2)については、「情報通信」のような漢字熟語の場合、正しくは「情報」と「通信」に分かち書きされなければならない。しかし漢字2字熟語、漢字3字熟語の2つの漢字テーブルを検索すると「情報」、「通信」、「情報通」が登録されている。本来ならばここで、「情報」+「通信」で「情報通信」という組合せが尤もらしいと判断されるべきである。しかし、「流体力学」のような漢字熟語の場合、「流体」「体力」「力学」と分割可能で、熟語が長くなればその組合せはさらに増える。ところで、表層解析では語と語の組合せ(前後の語の接続関係や、オーバーラップしているかどうか、熟語のすべての文字列がテーブルによってカバーされているかどうか、等)についてはチェックしない(6.3.2節参照)。このために「情報通信」の例では、「情報通」が優先されて「情報通」「信」というように誤って分かち書きされてしまう。今回筆者らは簡便な前方最長一致の手法を用いることとしたため、漢字熟語の区切りはこのような誤りが多いことから、知識ベースBに「漢字連続部は分かち書きを誤る可能性が高い」ということでユーザに提示する。この漢字熟語分割手法についてはさらに検討を要する。

対話処理部におけるユーザインタフェース

自動分割における分かち書きが終了すると文書には分かち書きの区切りを表す赤いスラッシュと知識ベースBによって指摘された緑の三角と赤い網かけが表示される。図6.3中、スラッシュ2本が連続している箇所は「2回区切る」すなわち文末であることを表している。ユーザはそれぞれの文字間の区切りを見て、分かち書きに疑問がある場合は区切り箇所にマウスカーソルを合わせ、右ボタンをクリックすることにより、自動分割で使用されている分かち書きの知識を知ることができる。

区切りを削除したい箇所では、削除したい区切りの上にマウスカーソルを合わせてマウスの左ボタンをクリックするだけでよい。反対に区切りを挿入したい箇所についても、区切りたい文字間にマウスカーソルを合わせて左ボタンをクリックすることによって区切りを挿入できる。

図6.3は自動分割処理の後、どのような知識によって区切られているか(いないか)をダイアロ

グボックスを開いて表示している画面である。ユーザはこれを見て、必要に応じて分かち書きの知識の追加、修正、削除を行うこともできる。また実用段階において、もし熟練ボランティアによってあらかじめ正解ファイルが与えられていれば、ユーザが行った区切りと用意された正解の区切りとを比較することによりユーザ自身の見落としやすい箇所や区切り過ぎている箇所を知ることができる。さらに、正解と比較した分かち書きの精度を計算させることも可能である。このような機能は初級点訳ボランティアの教育システムとしての可能性を示しているといえる。

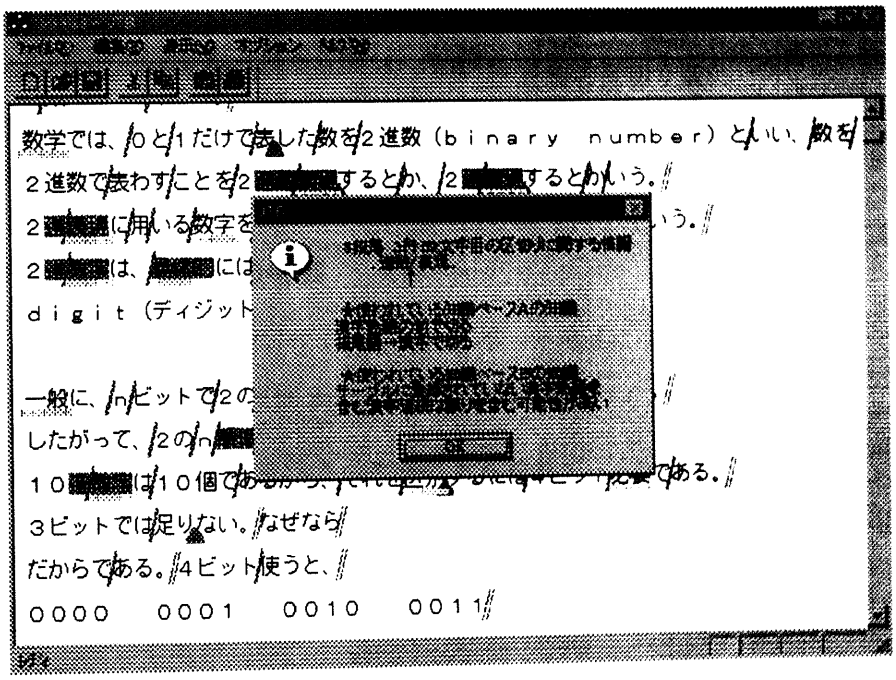

図 6.3: システムによるダイアログボックス表示画面

6.4 評価および考察

6.4.1 実験方法

本システムは一般のボランティアが使用することを考慮して Visual C++を用いて開発を行い、Windows 95 上で動作する対話型システムとした。コンピュータソフトの入門的なテキスト (全4章) を用いて本システムの性能評価実験を行った。このテキストは文章数３４６３文、文字数にして１１３８８４文字であった。このテキスト全4章のうち、前半2章 (文章数１２２６文、６６６７８文字) を用いてそこに出現する語をすべてテーブルに登録し、残りの後半2章については特にテーブルをチューニングせずに、前半・後半それぞれについて正解率を求めた。

分かち書きの正解率は次式によって計算した。なお、'空振り'とは必要のない箇所を区切ってしまった間違い、'見逃し'は区切り忘れの間違いとする。

$$\text{正解率（空振りなし）} = (1 - \frac{\text{空振りの回数}}{\text{正解の区切り数}}) \times 100$$

$$\text{正解率（見逃しなし）} = (1 - \frac{\text{見逃しの回数}}{\text{正解の区切り数}}) \times 100$$

6.4.2 不正解部分に関する考察

本システムで分かち書きが正しく行なわれなかった箇所については、3つのケースに大別される。各ケースの占める割合はほぼ3：1：1である。

(1) 語がテーブルに登録されていないことによる誤り

正 偉大と／いえる。

誤 偉／大と／いえる。

理由 「偉大」が漢字2字熟語に未登録でかつ「大」が接頭語テーブルに登録されているため

(2) 分かち書き手法の不備による誤り

正 …しか／しない。

誤 …しかし／ない。

理由 前方最長一致法を用いているため

(3) 複合情報の不足による誤り

正 コンピュータ会社

誤 コンピュータ／会社

理由 「コンピュータ会社」という熟語は連濁を生じてしまい、「コンピュータガイシャ」と読まれるため

6.4.3 市販の点字翻訳プログラムとの比較結果

市販の点訳プログラム（ＥＸＴＲＡ Ver.3.0) と本システムの分かち書きの精度の比較を示す。ＥＸＴＲＡが用いている分かち書き方式は公表されていないが、形態素解析を行っているものと思われる。

表 6.3より、空振りなし正解率、見逃しなし正解率ともにテーブルにすべて語が登録されている第１・２章においてもテーブルについて特にチューニングを行っていない第３・４章においても、本システムはＥＸＴＲＡと同等程度の精度を得、本システムの有効性を確認できた。テーブルのチューニングでは、ひらがな書き自立語テーブルと接頭語・接尾語テーブルは前半２章に出現する語がすべて含まれるように構築したが、これらのテーブルは語数も少なく、書式も単純なので簡単に更新できる。従って、他の分野のテキストに対しても容易に適用できると考える。

処理時間についてはＥＸＴＲＡは点字表記 (あるいはかな表記) までを一括して処理している。一方、本システムでは分かち書きまでしか行っていないため、単純には比較できない。しかし、ＥＸＴＲＡで翻訳した結果を点訳ボランティアが見直すためには膨大な時間が必要になるのに対し、本システムでは誤り箇所を選択的にユーザに指摘することができる。ユーザは指摘された箇所を選択的にチェックすればよいので、全テキストの分かち書き箇所のうち、Ｔ％が指摘できると仮定し、さらにＣＰＵ時間は無視できる程度に短いとすると、全体のスループットは $\frac{1}{T}$ となる。本システムの分かち書き処理における実験では約Ｔ＝０．１であった。本システムの漢字かな変換処理部、かな点字変換処理部が、現在の分かち書き処理部と同程度の処理能力をもつと仮定すると、点字を出力するまでの全体のスループットは従来の点訳プログラムを利用した場合に比べて約数分の１から十数分の１程度に短縮されることがみこまれる。これはＥＸＴＲＡの一括処理はユーザの介入も必要がなく、コンピュータ処理に要する時間は短いが、その後の見直しに非常に時間がかかり、複数のボランティアが何時間もかけて全文を複数回見直さなければならないのに対し、本システムでは対話処理的に作業が行えるために、見直しの作業に時間がかからないためである。また、実際の点訳ボランティアからは、「漢字かな混じり文のままのほうが分かち書きのチェックがしやすい」との意見もあり、本システムを試用した複数の点訳ボランティアから「誤りの箇所が色別に指摘されるので見やすく、作業がはかどりそう」との感想を得ている。

6.4.4 他の対話システムとの比較

畑田らの開発したＯＣＲ文書の認識誤りを１つずつダイアログボックスで確認して訂正する対話システムとの比較・検討を行う。このシステムでは誤りを含む単語を一斉に表示して正しい単語が第１候補にある場合、ワンタッチで修正を行うもので、第１候補はそれぞれの単語の行の上に表示されている。第１候補にない場合は、マウスカーソルを誤りを含む単語の所へ移動することにより単語の下にプルダウン式に第１候補から第５候補までが表示され、マウスポインタを移

表 6.3: 他のシステムとの精度の比較

	EXTRA		本システム	
	空振りなし	見逃しなし	空振りなし	見逃しなし
第 1・2 章	96.1%	96.0%	98.6%	98.9%
第 3・4 章	96.0%	94.9%	97.7%	98.4%
全体	96.0%	95.4%	98.2%	98.7%

動してダブルクリックすることにより選んだ単語を誤った単語と置き換える。

これに対し、本システムでは現在のところ、曖昧な分かち書き結果をユーザに提示するにとどまり、より正しい分かち書き箇所を提示するには至っていない。これは、スペルチェックのように正しい単語が明らかな場合と異なり、分かち書きの場合、前後にくる単語によって区切り方が異なるためである。

一方、スペルチェックでは同じ単語に対して毎回同じようにスペルミスをするとは限らないが、分かち書きの場合、ある単語の組合せで 1 度修正が行われると次回も同じように修正される可能性が高いことから、ユーザの修正を事例として記録しておいて同じ単語の組合せについては一括、あるいは順次訂正のメッセージを表示し、それ以降の曖昧箇所の指摘から除外する、といった機能を追加することが考えられる。さらに、畑田らも用いている 2 文字、3 文字が隣接して生じる文字の共起関係を利用することは分かち書きにおいても有効である。また、市販の点訳システムでも問題となる専門書に出現する数式やプログラム等、通常の文とは分けて特別扱いしなければならない部分についても本来対話的に処理できることが望ましい。

本システムにおける分かち書きは段落単位で、あるいは全文書を一括に処理することができ、対話の応答時間はユーザの思考を妨げない程度のものである。現在のところ、知識ベース B によってみつけることのできる誤りは誤り全体の約 4 分の 1 程度であり、約 5 %については正確に誤り箇所を指摘することができた。今後、この知識ベース B を充実させることにより、さらにボランティアの見直しの手間を軽減することができると考えられる。

6.5 まとめ

文法情報を含む大規模な辞書の代わりに小規模なテーブルを用いて日本語の分かち書きを行い、曖昧な区切り箇所についてはユーザに問い合わせる日本語文分かち書き手法について検討し，この手法の有効性を確認するため，対話型の分かち書き支援システムを構築した。この実験システムでは表層情報に基づく分かち書きの規則を知識ベース化し、知識に優先度をつけることにより知識の適用順位を変化させることができる。実験システムでは、形態素解析を行わない簡便な

表層解析という方式を利用したが、知識の表現方法を表層情報から形態素情報に変えることにより、従来の形態素解析でも同様に知識を独立させ、知識ベース化することが可能である。

視覚障害者の職域拡大にはコンピュータの利用技術を身につけることが有効であり、そのためには情報処理関連の専門書の点訳が必須である。しかしながら、一般図書と比べて専門書は市販の点訳プログラムで翻訳すると誤りが多く、点訳ボランティアに利用されていない。このような理由から今回は点訳の対象を情報処理関連の専門書としたが、表層解析用のテーブルを変更することで、他の分野のテキストも容易に分かち書き可能である。

現在、対話の有効性を確認するため、本システムの対話処理部で指摘できる誤りの率を上げること、冗長な誤り指摘を減らす手法について検討を進めている。さらに、ユーザの修正に応じてシステムが自動的に誤り箇所を変更・修正したり、以前修正した箇所を覚えておいてユーザに提示するようなシステムの自己学習機能について検討し、対話処理によって分かち書きがどの程度容易になるか、あるいは分かち書き処理の作業時間をどの程度短縮できるか等について調査し、ユーザインタフェイスを向上させていきたい。

今後は、漢字に読みをつけ、点字のフォーマットにあわせて出力できるようにする予定である。また、ユーザの修正箇所と修正結果を記憶しておいて、次に同等あるいは類似の表現が出現したときにユーザの修正を優先するような学習機構の導入が課題である。