日本語自動点訳ソフトの精度について

東京都立北療育医療センター
福井　哲也

1．はじめに

日本語自動点訳ソフトとは一般に、漢字仮名交じりの日本語テキストファイルを自動的に分かち書きされた仮名の点字（普通の日本語点字）に変換するパソコン用ソフトウェアのことをいう。近年、数種類の日本語自動点訳ソフトが開発され、多くの視覚障害者やその関係者に利用されるようになってきた。これらのソフトウェアが、視覚障害者の情報摂取に大いに役立っていることは確かだが、一方でその変換精度が問題となる場合も少なくないように思われる。自動点訳ソフトを利用してみたものの、修正に予想以上に手間がかかってしまったとか、自動点訳された資料が無修正のまま配布されてしまったために非常に読みづらかった、といった話を時々耳にする。確かに自動点訳ソフトでは、漢字の読み下しや分かち書きなどの面である程度の誤りが出てしまうので、その点を踏まえて利用する必要がある。ソフトウェアの機能に対するユーザー側の理解不足が、先のような問題を引き起こすといえよう。ただ、機能を理解するといっても、自動点訳ソフトの変換精度に関する客観的データが少ないこともまた事実である。特に、点字をよく知らない人たちに対して、変換精度がどれぐらいであるかを具体的に示すことは重要ではないかと考えた。また、何種類かあるソフトウェアによって変換精度にどれぐらい差があるのかということにも、多くのユーザーが関心をもっているのではないかと思われた。

そこで筆者は、現在国内で市販ないし頒布されている４種の日本語自動点訳ソフトについて、同じ材料を点訳させ、誤変換の数を誤りの種類別に数える実

験を実施することにした。もとより、このような実験で正確な評価を行うには、多種類の材料について、まとまった量の実験をする必要があろう。その点で、筆者の今回の実験は誠に不十分なものといわざるを得ない。特にソフトウェア間の比較については、「ある限られた視点からの評価に過ぎない」ことをご理解いただきたい。また、実験を行ってから1年以上経過しているので、ソフトウェアによってはその後バージョンアップがなされているものもあることを付記しておく。この「小さな試み」が、ソフトウェアのユーザーや開発者の方々に何かの参考となれば幸いである。

なお本稿は、1993年2月に開かれた第2回視覚障害リハビリテーション研究発表大会で発表した内容に加筆・訂正したものである。

2．実験報告

1）方法

今回実験を行った4種のソフトウェアを、表1に示す。

変換させた文章は下記の2種類で、いずれもMS－DOSのテキストファイルで準備した。

材料1：日本経済新聞のコラム「春秋」1991年12月11日～31日の20回分
(13,026字)

材料2：光村図書出版の小学校4年生用国語教科書、15単元の各冒頭部分
(7,968字)

材料として新聞のコラムを選んだのは文章が比較的平易で様々な分野の話題を含むと思われたからであり、小学校の教科書を使ったのは漢字の少ない文章

表1　実験対象の日本語自動点訳ソフト

名　称	バージョン	メーカー・配布元	価　格
がってんだ	Ver. 2.0 (1991)	㈱言語工学研究所	400,000円
E X T R A	Ver. 1.1 (1992)	㈱アメディア	68,000円
点訳ぷうちゃん	Ver. 1.63 (1992)	松山市視力障害者友の会	20,000円
80　点	Ver. 1.31 (1991)	福祉システム研究会	無　料

の変換精度を見るためである。

［材料１の内容例］

91年12月11日

ゴルバチョフ・ソ連大統領は昨年、ノーベル平和賞を受けた。冷戦の終結、軍縮の進展、ドイツ統一などは、この人がいればこそであった。受賞決定に世界中から歓迎の拍手が送られた。だが、ちょうど一年前の授賞式に大統領は欠席をせざるを得なかった。経済不安をはじめとする重大な国内情勢を抱え、授賞式どころではなかった。

［材料２の内容例］

とびこめ

１そうの船が、世界を航海して、帰りの旅をしていました。おだやかな天気で、船じゅうの人が、みんなかんぱんに出ていました。大きなさるが１ぴき、みんなの中でふざけ回って、人々を楽しませていました。そのさるは、体をよじったり、とびはねたり、顔をこっけいにしかめたりして、みんなをからかいました。みんなが喜んでいるのを知っていたのでしょう。だから、ますますふざけ回っていました。

これらの材料を各ソフトウェアで点字に変換し、誤って変換された所に表２に示す分類に従って印を付けて集計した。点字の表記法に関しては「日本点字表記法1990年版」（日本点字委員会、1990）に定められた規則に従って判断し、分かち書き等の解釈については「最新点字表記辞典」（日本盲人福祉研究会、1991）を参考にした。

この誤変換の分類は、日本語点字の表記規則をベースに筆者が本実験のために考案したものである。16分類のうち、①と②が漢字の読み下しに関わる誤り、③から⑤が仮名遣いに関わる誤り、⑥から⑧が句読点など記号関係の誤り、そして⑨から⑯が分かち書きに関わる誤りである。

一つの単語の中で読みと分かち書きのように異なる種類の誤りが複合している場合には、それが単一の原因により生じたものと思われるものでも、両方にカウントすることとした。

例えば次の例では、「向き出したのか」という文節に３個の誤りが生じてい

表2　誤変換の分類

種　　類	語　例	正	誤
①一般語の読みの誤り	大晦日 水時計	おおみそか みずどけい	だいみそか みずとけい
②固有名詞の読みの誤り	家康 旅順	いえやす りょじゅん	いえこー たびじゅん
③数字とカナの使い分けの誤り	一方的 一番目	いっぽーてき 1ばんめ	1ぽーてき いちばんめ
④長音の処理の誤り	食べよう 吸う	たべよー すう	たべよう すー
⑤助詞「は・へ」の処理の誤り	道は遠い このはがき	みちわ□とおい この□はがき	みちは□とおい この□わがき
⑥記号と文字の間を続ける誤り	来た、と思った	⠗⠁⠞⠁⠂□⠞⠕⠍⠕⠞⠁⠉⠁	⠗⠁⠞⠁⠂□⠞⠕⠍⠕⠞⠁⠉⠁
⑦記号と文字の間に余分な空白	「はい」と言う	⠈⠁⠊⠈⠍⠕⠞⠁□⠊⠞⠁	⠈⠁⠊⠈⠍⠕□⠞⠁□⠊⠞⠁
⑧省略すべき読点・中点を省略しなかった	2、3日 クリスマス・イブ	⠼⠃⠂⠼⠉⠂⠊⠚⠁ ⠅⠗⠊⠎⠍⠁⠎□⠊⠃⠁	⠼⠃⠂□⠼⠉⠂⠊⠚⠁ ⠅⠗⠊⠎⠍⠁⠎□⠊⠃⠁
⑨文節間を続けた	高い山 もう一つ	たかい□やま もー□ひとつ	たかいやま もーひとつ
⑩文節内を切った	カササギ 税金	かささぎ ぜいきん	か□ささぎ ぜい□きん
⑪分かつべき複合語を続けた	資本主義 山田さん	しほん□しゅぎ やまだ□さん	しほんしゅぎ やまださん
⑫続けるべき複合語を切った	走り回る 副社長	はしりまわる ふくしゃちょー	はしり□まわる ふく□しゃちょー
⑬補助用言・形式名詞を続けた	書いたのである 書いたものの	かいたので□ある かいた□ものの	かいたのである かいたものの
⑭助詞・助動詞を切った	花が咲く 試みられる	はなが□さく こころみられる	はな□が□さく こころみ□られる
⑮続けるべき「する」を切った	勉強する お招きする	べんきょーする おまねきする	べんきょー□する おまねき□する
⑯分かつべき「する」を続けた	きらきらする 自費出版する	きらきら□する じひ□しゅっぱん□する	きらきらする じひ□しゅっぱんする

ることになる。（ここでは便宜的に、点字を平仮名に置き換えて表す。□はマスあけを示す。）

バブルがはじけて人心が「文化」に向き出したのか、

→ばぶるが□はじけて□じんしんが□「ぶんか」に□むきで□したの□か

①一般語の読みの誤り

⑩文節内を切る誤り

⑭助詞を切る誤り

2）結果

各ソフトウェアによる変換結果の一例を次に示す。この例は、どちらかというと誤変換の多い部分である。○の中の数字は、誤変換の分類番号である。（以下、“点訳ぶうちゃん”は“ぶうちゃん”と略記する。）

［原　　文］　若い客が目立つ。広いテーブルを取り巻いて男女半々ぐらいが、楽しそうに話し合っている。一杯のみ屋での若い女性姿も、いつの間にか板に着いて何の違和感もない。聞いていると、話題も明るい。その場にいない何某君の失敗談をマンガ風に再現してみせ、みんなで屈託なく笑ったりしている。

［がってんだ］　□□わかい□きゃくが□めだつ。□□ひろい□てーぶるを
とりまいて□だんじょ□はんはんぐらい□⑭が、□たのしそーに
はなしあって□いる。□□⑨いっぱい⑩のみ□⑭⑪おくでの□わかい
じょせい□すがたも、□いつのまにか□いたに□①きいて□なんの
いわかんも□ない。□□きいて□いると、□わだいも□あかるい。
その□ばに□いない□①なんぼー□②くんの□しっぱいだんを
まんが□⑭⑪かぜに□さいげんし□⑭てみ⑨せ、□みんなで
くったく⑨なく□わらったり□して□いる。

［EXTRA］　□□わかい□きゃくが□めだつ。□□ひろい□てーぶるを
とりまいて□だんじょ□はんはんぐらいが、□たのしそーに
はなしあって□いる。□□1ばい⑩のみ□⑨やでの□わかい
じょせい⑪すがたも、□いつの□⑭⑪あいだにか□いたに□ついて
なんの□いわかんも□ない。□□きいて□いると、□わだいも

あかるい。□□その[⑬]ばに□いない□[①]なんぼーくんの
しっぱいだんを□まんがふーに□さいげんして□みせ、□みん
[⑭]なで□くったく□なく□わらったり□して□いる。

［ぶうちゃん］　□□わかい□きゃくが□めだつ。□□ひろい□てーぶるを□とり
[㉒①]かんいて□だんじょ□はんはん□[⑫]ぐらいが、□たのしそーに
はなし□[⑫]あって□いる。□□1ぱい[⑪]のみ□[⑦]やでの□わかい
じょせい□すがたも、□いつの□[㉒①]あいだにか□いたに□ついて
[①]なにの□いわかんも□ない。□□きいて□いると、□わだいも
あかるい。□□その□[①]じょーに□いない□[①]なにぼー□[㉒①]きみの
しっぱいだんを□まんが□[㉒①]かぜに□さいげん□[⑬]して□みせ、
みんなで□くったく[⑭]なく□わらったり□して□いる。

［80　点］　□□わかい□きゃくが□めだつ。□□ひろい□てーぶるを□とり
[⑫]まいて□だんじょ□はんはんぐらいが、□たのしそーに
はなしあって□いる。□□[⑤]いっぱい[⑪]のみやでの□わかい
じょせい□すがたも、□いつの□[㉒①]あいだにか□いたに□ついて
なんの□いわかんも□ない。□□きいて□いると、□わだいも
あかるい。□□その□ばに□いない□[①]なにぼー□[⑫]くんの
しっぱいだんを□まんが[①]かぜに□さいげんして□みせ、
みんなで□くったく[⑭]なく□わらったり□して□いる。

2種の材料を4種のソフトウェアで変換した結果について、誤りの種類別に誤変換の数を合計し、それを原文1万字あたりに換算したのが表3である。なお、合計欄の数字が若干合わないように見えるのは、除算の結果を四捨五入しているためである。

ここで注意しておきたいのは、誤変換の量を個数で表しているのであって、パーセンテージではないということである。例えば、原文1万字あたり誤変換400個という数値は、誤変換率4％と読みかえるべきではない。自動点訳ソフトのうたい文句にはよく「精度99％」などという表現が見られるが、これは分子と分母にそれぞれ何をとっているのかがきわめて不明確である。そこで、筆者はあえてパーセンテージによる表現を避けることにしたのである。

表3　誤変換の種類別個数（原文1万字あたり）

誤変換の種類	材料1（新聞コラム）				材料2（小4教科書）			
	がってんだ	EXTRA	ぶうちゃん	80点	がってんだ	EXTRA	ぶうちゃん	80点
①一般語の読みの誤り	93.7	99.0	166.6	99.8	61.5	61.5	113.0	62.8
②固有名詞の読みの誤り	16.9	13.8	14.6	14.6	1.3	7.5	7.5	2.5
③数字とカナの使い分けの誤り	2.3	4.6	9.2	26.9	3.8	0	3.8	3.8
④長音の処理の誤り	1.5	2.3	9.2	0	46.4	56.5	11.3	37.7
⑤助詞「は・へ」の処理の誤り	3.1	1.5	10.0	3.1	13.8	12.6	18.8	5.0
⑥記号と文字の間を続ける誤り	6.9	6.1	0.8	4.6	5.0	1.3	2.5	0
⑦記号と文字の間に余分な空白	18.4	3.8	29.2	1.5	15.0	2.5	43.9	0
⑧読点・中点を省略しなかった	5.4	8.4	3.1	8.4	1.3	1.3	1.3	1.3
⑨文節間を続けた	19.2	45.3	62.2	83.7	52.7	135.5	139.3	177.0
⑩文節内を切った	64.5	33.0	69.1	23.8	160.6	90.4	126.8	40.2
⑪分かつべき複合語を続けた	24.6	25.3	41.5	35.3	35.1	26.4	11.3	8.8
⑫続けるべき複合語を切った	36.8	43.0	147.4	61.4	20.1	51.5	72.8	26.4
⑬補助用言・形式名詞を続けた	47.6	6.1	60.6	10.0	60.2	7.5	81.6	22.6
⑭助詞・助動詞を切った	54.5	16.1	30.7	11.5	84.1	36.4	33.9	8.8
⑮続けるべき「する」を切った	1.5	0.8	32.2	0.8	2.5	3.8	13.8	2.5
⑯分かつべき「する」を続けた	1.5	1.5	1.5	1.5	0	0	0	1.3
合　　計	398.4	310.9	687.9	386.9	563.5	494.5	681.5	409.1

3）考察

(1)誤変換の合計数で各ソフトウェアの変換精度を比較すると、材料1では、最も精度が高かったのが“EXTRA”、2位が“80点”で、僅差で“がってんだ”が続いている。材料2では、最も精度が高かったのが“80点”、以下“EXTRA”“がってんだ”の順となっている。4種のソフトウェア間にはかなりの価格差があるが、この実験で見るかぎり、高価なソフトウェアほど精度が高いとはいえないようである。

(2)①の一般語の読みの誤りは、“ぶうちゃん”を除く3つのソフトウェアはだいたい同じ値となっている。また、②の固有名詞の読みの誤りは、4つのソフトウェアとも際立った違いは見られなかった。漢字の読み下しに関しては、どのソフトウェアも似たような水準にあるようである。

(3)⑨の文節間を続ける誤りと⑩の文節内を切る誤りを比較してみると、ソフ

トウェアごとの特色が出ている。⑨と⑩の比率は、“がってんだ”がおおむね1：3であるのに対して、“80点”はおおむね4：1である。すなわち、“がってんだ”は「切りすぎ」、“80点”は「つなげすぎ」ということができよう。他の2つのソフトウェアについては、このような大きな偏りは見られない。もっとも、「切りすぎ」と「つなげすぎ」はどちらが読みにくいとも一概にいえないと思うので、この点についてソフトウェア間の優劣はつけがたい。

(4)⑬の補助用言・形式名詞を続ける誤りは、“がってんだ”と“ぶうちゃん”が他のソフトウェアに比べて際立って多い。また、“がってんだ”は⑭の助詞・助動詞を切る誤りも他に比べて高い数字となっている。⑮のサ変動詞の語尾「する」を切る誤りは、“ぶうちゃん”だけが多い。補助用言・形式名詞・助詞・助動詞は、種類が限られていて、接続の仕方もほぼ決まっている。語尾の「する」についても同様で、これらはソフトウェアの改善により比較的容易に誤変換を減らせるのではないかと推測される。逆にいうと、“がってんだ”と“ぶうちゃん”は他のソフトウェアに比べて、分かち書きに必要な基本的文法処理の詰めが甘いのではないかと思われる。

(5)材料1と材料2を比較してみると、①の一般語の読みの誤りは、どのソフトウェアでも材料1より材料2の方が少なくなっている。逆に材料2の方が顕著に多くなっているのは、⑨の文節間を続ける誤りと⑩の文節内を切る誤りである。これは文章中の漢字の数と関連があるものと思われる。字数全体に対する漢字の割合は、材料1が13,026字中4,245字で32.6%、材料2が7,968字中1,219字で15.3%であった。原文に漢字が少なければ読みの誤りは減るが、文節の区切りを解析する手がかりを漢字に頼っているために、文節の切り誤りはかえって増加するのであろう。文節を切り誤った点字は判読に苦労することも少なくない。次に一例を示す。この例ではどのソフトウェアでも、「はい登りました」の「は」が助詞の「は」と解釈されたため、読みが「わ」となったばかりか、分かち書きも大幅にくるってしまっている。

［原　　文］　あっという間に、ロープを伝ってはい登りました。

［がってんだ］　あっと□いう□まに、□ろーぷを□つたって⑭⑮わ□
　　⑭いのぼりました。

［ＥＸＴＲＡ］　あっと□いう□あいだに、□ろーぷを□でんってわ□
いのぼりました。

［ぶうちゃん］　あっと□いう□あいだに、□ろーぷを□つたってわい□
のぼりました。

［80　　点］　あっと□いう□あいだに、□ろーぷを□でんってわ□い□
のぼりました。

３．現状と課題

１）精度に対する評価

日本語自動点訳ソフトの精度は、残念ながらまだまだ低いといわざるを得ない。原文１万字あたり300から600個の誤変換があるということは、１行32マスの点字にして、２行に１個、ないし１行に１個の割合ということになる。もちろん、原文の内容によってはもっと誤りが少ない場合もあるが、逆に用語や文体のちょっとした違いで変換精度が大幅に落ちてしまうこともある。

変換精度が上がらない根源的な理由はいうまでもなく漢字の存在である。漢字の読み方は一通りではないので、正しく仮名に置き換えるのがまず難しい。加えて、文節間の分かち書きの問題も大きい。日本語の墨字は、漢字と仮名の組み合わせで書かれるので、文節間にスペースがなくても十分に読める。これに対して点字では文節間にスペースが必要なので、墨字を点字に変換する際は文節の区切りを見出さなければならない。これも、自動点訳の大きな障害となっている。

日本語に比べれば、英語自動点訳ソフトの完成度は非常に高いといえよう。英語では墨字でも単語間にスペースを置くので、日本語であるような分かち書きの誤りは起こらない。各単語は、フルスペルの墨字から略字を用いた第２種英語点字に変換しなければならないが、これもきわめて稀な例外を除けば、単語ごとに一対一に対応している。筆者が使用している英語自動点訳ソフト“VIEW２”（点字ディスプレイ装置ナビゲータに付属，テレセンサリー社製）では、特殊な固有名詞を別として、略字の用法に関する誤りはほとんど起こらない。あとは記号関係とレイアウトを整えるだけで修正は容易に完了するのである。

ここで筆者は、日本語自動点訳ソフトの開発者の技術レベルが低いなどというつもりはさらさらない。技術が低いのではなくて、取り組むべき課題の方があまりに困難なのだ。これはまさに漢字使用国の悲劇といえよう。

いずれにしても、今の日本語自動点訳ソフトは、まだ「出しっぱなしではきちんとした点字資料にはならない」ということをしっかり認識すべきである。点字を常用する視覚障害者はこのことを容易に実感できるが、点字を知らない晴眼者に正しく理解してもらうにはときに苦労する。「点訳おまかせソフト」（“がってんだ”取扱説明書，表紙）とか「詩や歌のような特殊なものでなければ、99パーセント以上の正確さです」（“EXTRA”使用説明書，P.8）のような表現は、できれば避けてほしいというのが筆者の本音である。この点“80点”のドキュメントファイルには、「自動点訳とは言っても、半自動。手元の日本語文書ファイルを80％くらい正しく（あくまで目安）自動で点訳し……」と記述されており、誠に謙虚な説明といえよう。

もう一つ、自動点訳ソフトではレイアウトのくずれという問題も大きい。例えば、タイトルがセンタリングされ、１文字ごとに空白が挿入されているようなもの、段階ごとに頭を下げて箇条書きされたもの、表形式のものなどをそのまま自動点訳にかけると非常に読みづらい点字になってしまう。

2）実用性と使い方

日本語自動点訳ソフトは、精度の面でまだまだ不満が残るが、かといって実用性がないわけではない。むしろ、上手に活用すれば視覚障害者の情報環境改善の有力な武器となりうるのである。すなわち、墨字の原文がテキストファイルで入手可能な条件があれば、とりあえず今すぐ内容を知りたいという要求にはかなり答えてくれる。また、修正を前提とすれば、点字資料の作成作業を効率化することもできる。筆者自身も、このような目的で日本語自動点訳ソフトを頻繁に利用している一人である。

ただ、点字を知らない晴眼者が自動点訳ソフトを操作することは、できるだけ避けるべきだというのが筆者の主張である。点字の知識がなく、出力結果の良否を判断できない場合には、墨字のテキストファイルをそのまま視覚障害者に渡す方がずっと親切なのではなかろうか。受け取った墨字ファイルを自動点

訳にかけるのか、音声出力で読むのかといったことは、視覚障害者側の判断に委ねてほしいと思う。

特に最近、盲学校や視覚障害者リハビリテーション施設に勤務する晴眼者の一部に、点字の勉強をおざなりにして、安易に自動点訳ソフトに頼ろうという傾向が見られることは誠に残念でならない。また、自動点訳ソフトだけでなく、点字が画面に平仮名で表示されるような点字エディタ（“BASE”“ブレイルスター”など）を点字の知識なしに使用することも危険である。視覚障害者（児）に関わる職業についた以上、パソコンを使うより前に、まず点字器や点字タイプライターで点字が書け、紙に書かれた点字が読めるようになることを最優先にするべきである。

3）ソフトウェアの改良に向けて

最後に、今後のソフトウェアの改良について一言付け加えたい。素人の筆者が想像するに、現在の自動点訳ソフトの性能を向上させるためには、漢字辞書の充実だけでなく、平仮名の並びの解析や、さらに文意に踏み込んだ解析の研究が必要なように思われる。日本語の世界でも「出しっぱなしでも安心して使える自動点訳ソフト」を是非とも実現させてほしい。そのためには、コンピュータの専門家、日本語の専門家と並んで点字の専門家の参画が不可欠ではなかろうか。点字を大切にする一人の視覚障害者として、関係者の一層の努力に期待する次第である。

《インフォメーション2　研究雑誌－2：1993年3月～1993年9月》

視覚障害乳幼児スペシャリストとして（渡辺万里子）障害者の福祉　通巻144号　Pp.17－19　1993年7月

ネパールの盲人と障害者福祉（福山博）障害者の福祉　通巻144号　Pp.25－28　1993年7月

視覚障害者へ新しいツアー企画「手で見る世界の芸術」（手で見る世界の芸術実行委員会）障害者の福祉　通巻145号　Pp.13－16　1993年8月

大成功だったマッサージ・セミナー――十か国・地域から百五十人参加――（牧田克輔）障害者の福祉　通巻146号　Pp.20－22　1993年9月

「小さな凸」の提案　誰でも一緒に遊べるおもちゃをめざして（星川安之）障害者の福祉　通巻146号　Pp.42－45　1993年9月