数学文書点訳における
日本語分かち書き変換に関して

橘美紗　　鈴木昌和（九州大学）

2007年2月22日

1　数学文書の点訳現状

1.1　日本語の点訳

日本語文書を点訳する際，日本語の点字は仮名文字のみで表現するため，墨字の際によく使われる漢字が全く使えない．漢字情報がないため，文書をそのまま点字に変換しただけでは単語の区切り等がわかり難く，内容を理解しにくい．そこで，日本語文書の点訳では，まず初めに文書を「分かち書き」という形に変換する．

「分かち書き」とは，ある規則に従って文節毎に文書を区切る形である．例えば，

右辺にこれを代入する
↓
うへんに これを だいにゅー する

といった形の変換になる．この例文がもし区切られずに，

うへんにこれをだいにゅーする

と言った形で読まれたら意味が理解しにくくなるのがわかるだろう．この例文は比較的短い文であるが，これがもしもっと長い文になれば，意味の理解を深める為の「分かち書き」の必要性が強くなるのがわかるであろう．

今回は，数学文書におけるこの「分かち書き」変換に関しての研究成果を発表する．また，今回の分かち書き規則は,「点訳のてびき」という全国視覚障害者情報提供施設協会が出版している本に準拠している．

1.2　現存するソフトウェアによる点訳現状

現存する自動点訳ソフトウェアには，Extra等の安定したソフトウェアがある．これらのソフトウェアは，一般文書の点訳においては大変性能が良く，またユー

ザーがユーザー辞書登録をすることによって，専門用語等にも対応できる仕組みになっている．

しかし，数学専門用語を全て辞書登録した状態で自動点訳をさせてみても，やはり誤りが目立つ状態となる．数学文書には，数式といった特殊なものがあることも原因の一つではあるが，日本語部分においても，分かち書きの誤りが見られる．そこで Extra の数学文書日本語分かち書き変換に関して詳しく見てみた．

1.3 Extra による数学文書日本語分かち書き

Extra とは，現存する点訳ソフトウェアで，大変優秀なものである．一般文書に対する点訳処理は，その点訳スピード，正確性共に納得のいくものである．しかし，数学文書など専門書の点訳となると，正解率は落ちる．その Extra の数学文書点訳における日本語分かち書き変換についての現状を以下に述べる．

Extra には，ユーザーによる辞書登録の機能がある．その機能によって，専門書の点訳の際に問題となる専門用語の不足が補えるだろうと思い，数学専門用語を１万語ほど用意した．そして，ユーザー辞書登録してみた場合と辞書登録していない場合の正解率を調べてみた．以下は，その実験に使用したデータの詳細である．

表 1: 実験データの詳細

データ	ページ数	文節数 (数式除く)	数式数	文字数
書籍 1	32P	2456	848	12814
書籍 2	13P	1538	228	6753
書籍 3	14P	1120	462	6042

これらの実験データは，異なる分野の数学書籍から 1 章分ずつ抜き出したデータである．表内の文節数とは，分かち書きされた際に分けられる最小のものを 1 文節とし，数式や数字などの日本語を全く含まない文節の総数を表している．数式数とは，実験データに含まれている数式の総数を表し，文節数には含まれていない．文字数とは，実験データに含まれている数式を除く部分の文字総数で，分かち書き変換する前の文字総数である．

次にこの実験データの正解率を以下に記す．

表 2: Extra による数学文書分かち書き正解率 (2006 年 12 月の実験値)

Extra	正解率
ユーザー辞書登録なし	**87.34%** 高-90.05%　低-85.23%
ユーザー辞書登録あり (数学専門用語)	**86.33%** 高-88.61%　低-84.92%

ユーザー辞書登録機能により数学専門用語の登録を行わない場合の方が正解率が高いという結果がでている．ユーザー辞書登録を行うことにより正解率が高くなるであろうと思っていたのでこれは驚きである．そこで，その詳細を調べてみた．

1.4 Extra のユーザー辞書登録機能による効果

Extra のユーザー辞書登録機能により，数学専門用語を１万語ほど登録したにも関わらず，数学専門文書の分かち書き変換正解率は低くなってしまった．その原因を見てみると，次のようなことが起こっていた．

数学専門用語を登録する前の Extra による分かち書き変換での誤りは，数学専門用語部分に関するものがほとんどであった．そこで数学専門用語を辞書登録したのであるが，そうすると今度はそれまで正しく変換できていた部分で誤りがでてきた．具体的な例を以下に記す．

表 3: 数学用語登録後に誤りが出る単語例

単語	登録前	登録後	単語	登録前	登録後
結果	けっか	けつか	場合	ばあい	ばごー
とおく	と おく	とおく	解く	とく	かいく
行なう	おこなう	ぎょーなう	組み合わせ	くみあわせ	くみみ あわせ
数	かず	すー	結ぶ	むすぶ	けつぶ
解決	かいけつ	かい けっ	結局	けっきょく	けつきょく
完全形	かんぜんけい	かんぜん かたち	積み上げる	つみあげる	せきみ あげる
真ん中	まんなか	しんん なか	曲面	きょくめん	くめん
強さ	つよさ	きょーさ	数えて	かぞえて	すー えて
図示せよ	ずしせよ	ず しめせよ	解いて	といて	かい いて

この表によると，ユーザー辞書登録前にはうまく読めていたものが，登録後には漢字の読み部分が音読みが強くなるなど，誤りが出ている．これらの変化によって，先の正解率が期待通りにいかなかったことがわかる．

もちろん，数学専門用語に関しては，ユーザー辞書登録により，正しく変換されたものもある．その具体的な例は以下の通りである．

一般文書では,「根」などは,「ね」と読むことばかりであるが，数学専門文書においては,「ね」と読むよりも「こん」と読む場合がほとんどである．このようなものに対しては，このユーザー辞書登録の効果は大きいが，読みが誤るものが圧倒的に多い為，結果として登録後の方が正解率が低くなったと思われる．

表 4: 数学用語登録後正しくなる単語例

単語	登録前	登録後
複素	ふくもと	ふくそ
根	ね	こん
複素平面	ふくそへい めん	ふくそ へいめん
共焦点	とも しょーてん	きょーしょーてん
高階	たかしな	こーかい

1.5 数学専門文書用点訳ソフトの必要性

これらのデータを見てみると，現存する点訳ソフトウェアでは数学専門文書の点訳の正解率を上げるのは困難だと思われる．その為，自動点訳ソフトウェアをせっかく用いても，数学文書の分かち書き変換には人的修正作業に頼る部分がかなりあることになる．そこで，今回我々は数学文書に適した日本語分かち書き変換プログラムの作成にあたった．その処理内容に関して述べていく．

2 仮名分かち書き変換処理

2.1 分かち書き処理の概要

今回我々が数学文書日本語分かち書き変換に利用した手法は，辞書による 2 文節最長一致法である．数学文書などの専門文書は，一般文書には出てこないような言葉が多いが，出てくる言葉自体はあまり変わらない．全ての用語を取ったとしても，一般文書ほど様々な言葉は出てこないのである．そこで，出てくる可能性のある単語を辞書に全て登録し，その辞書を用いて全ての変換を行うこととした．

辞書には，墨字とそれに対応する分かち書き文字を登録し，最小単位は分かち書きした際の 1 文節とした．そして，辞書に登録された単語と一致するものを 2 文節最長一致法により求めていく．

変換文字は，墨字データと分かち書きデータと，0 か 1 を値として取るフラグデータを持った形で残した．このフラグデータは人的修正作業の効率化の為の値で，0 であればその単語をマーク付け（エディターソフトにて赤く表示）して表示することにした．そして，人的修正作業の際には，マーク付けされた部分のみ見ることで作業が効率化されることを目指した．

また，辞書の容量を減らす為に，辞書は用途に合わせて，五種類に分けて登録した．

2.2　辞書の種類

日本語の単語は品詞によって分類分けされるが，今回の点訳処理においては，品詞までは考えずに行い，以下の六種類の辞書を用いた．

第一の辞書は，中心となる辞書で，分かち書きした際の 1 文節を 1 単位とした単語を登録した辞書である．また，動詞や形容詞などの活用形がある単語に関しては活用形も全て登録するような辞書を作った．使用したデータは，独立行政法人 情報処理推進機構（IPA）が配布している辞書データ，数種類の数学専門書籍（200 ページ程の数学書籍 10 冊分）と数学専門用語集から引用した．登録用語総数は，563,277 語である．

第二の辞書として，上の辞書と組み合わせることによって 1 文節となる語尾を登録した辞書を作った．この語尾は，第一の辞書に登録されているもの全てと組み合わせるように設定した．しかし実際は，動詞や形容詞の活用形につくことはあるはずもないのだが，今回のプログラムの手法上，問題ないと思われたので，何に続く可能性があるかといった細かな点は見ないことにした．登録用語総数は，109 語である．

第三，第四の辞書として，一文字の漢字辞書を作った．それは，音読み，訓読みを出てくる場面によって使い分け可能なものを登録したもので，音読み辞書，訓読み辞書とした．違いとしては，同じ漢字でも，数字や漢字に続く場合に音読みになるもの，語尾が続く場合に訓読みになるものを登録した．具体例では,「場」に対して，音読み辞書には「じょー」，訓読み辞書には「ば」を対応させるといった形で登録した．訓読み辞書には，語尾が続く場合の一文字漢字を登録したので，音読みはないが訓読みはあるものがある.「綾 (あや)」がその例である．登録用語総数は，音読み辞書が 106 語，訓読み辞書が 225 語となった．

第五の辞書として，文脈に依存して読みが変わる一文字の漢字辞書を作った．漢字の読みがその前後の文章内容を見ることでしか判断できないようなものを登録した．具体例としては,「根」に対して「ね」と「こん」,「底」に対して「そこ」と「てい」といったもので，登録用語総数は 22 語（一文字漢字につき読みが 2 語ずつ対応するので一文字漢字は 11 個）となった．

最後に，漢字に続く語尾漢字 (一文字) を登録した辞書を作った．これは具体例で言うと,「的」,「力」などである．登録用語総数は 32 語となった．

2.3　2 文節最長一致法

2 文節最長一致法とは，辞書にある単語と一致したもので最長のものを 2 文節分まで見て，1 文節を決めていく手法である．

まず，この一致するものを見る際に，変換対象である文の初めからではなく，文末から見ていくことにした．それはこの方法では，辞書にない単語が出てきた際に，その後がずっとずれていく，といったことが起こる．その為，もし最初に辞書

にない単語が出てきた際には，その一文が丸ごとずれてしまう可能性がある．しかし，文末に出てくる単語は限られているので，文頭からするよりは，初めに辞書にない単語が出てくる可能性は低いだろうと判断し，この方法を適用した．

初め，1 文節での最長一致法を用いて処理してみたが，それだと次のような場合にうまく変換できなかった.「もしかけるならば」という部分を 1 文節最長一致法で変換しようとすると,「も しかけるならば」という風に変換される．確かに後ろの文節の文字数は最長ではあるが，前の文節は明らかに辞書登録されておらず，一文字で変換されたものとなる．これを，2 文節最長一致法で行うと,「もし かけたならば」という方が，どちらも辞書に入っているため，長い方として採用される．よって，後ろの文節は,「かけたならば」に決まる．

辞書にある単語と一致したもの…と言っても，上で述べた五種類の辞書の中に優先順位がある．まずは，第一の辞書である．第一の辞書と，その内容に語尾（第二の辞書）がついたもので，一致するものを探す．それがなければ，漢字であれば，第五の辞書の文脈に依存する漢字であるかどうかを見る．それになければ，音読み・訓読み辞書に登録されていないかを見る．登録されている場合は，前後の文字種類により，どちらかを採用する．

これらの辞書にない場合は，辞書になかった文字として，1 文字ずつ変換する．1 文字変換したらまた辞書検索をし，またなければ 1 文字変換…といった風にしていく．

2.4 記号による特別処理

上の方法で，変換しスペースを入れるのだが，分かち書き文書には，記号によってスペースの入れ方に規則がある．例えば,「.」であれば後ろを 2 つあける，括弧類は開き括弧なら続ける，など．

故に，記号が出てきた際には，変換後，その記号種類によって，前をあける・あけない，後ろをあける・あけない，後ろをいくつあける…といった処理を行う．

2.5 人的修正作業の効率化

上で述べた分かち書き変換を用いて，数学文書の日本語分かち書き変換を行い，数式部分に関しては，そのまま点訳するようにした．(筑波技術大学 金堀氏がこの点訳処理部分を作成した．)

まずは，分かち書き変換されたデータにより，辞書に登録されておらず 1 文字で変換された部分と，前後の文脈により変換が変わる文字 (第五の辞書に登録されている語) 部分は，赤くマーク付けして表示し，一目で確認すべき箇所がわかるようにした．

自動点訳ソフトウェアにおいては，正しく変換されたかの人的確認がやはり一

番時間がかかる．また，分かち書き文書は仮名文字のみを使用するので見づらい為，見落としも出る可能性が高い．そこで，今回の変換処理では，誤り可能性のあるものはマーク付けすることで，注目ポイントを絞り，効率化をはかった．

3　分かち書き変換の効果

3.1　分かち書き変換正解率

今回作成したプログラムによる分かち書き変換を，先に述べた Extra における実験の際に使ったデータと同じデータに対して行なった実験結果を以下に記す．これらのデータは，今回の辞書作成の際に用いたものではない．実験データの詳細と変換にかかった時間もそれぞれ下に記している．但し，今回は数学文書における日本語部分の分かち書き変換に関してのみ評価している．

表 5: 変換所要時間と正解率 (2006 年 12 月の実験値)

実験データ	変換所要時間	正解率
書籍 1	23.0 秒	95.68%
書籍 2	12.5 秒	96.63%
書籍 3	10.6 秒	95.46%

Extra においてはこれらの同じデータに対する分かち書き変換正解率が 80%後半であったが，今回のプログラムにおいては，95%近くの正解率となっているのである．これは数学文書に有効な分かち書き変換プログラムと言えるであろう．

3.2　辞書追加登録後との比較

今回の分かち書き変換処理は，辞書を中心に行うプログラムであったので，先の実験データに対して，足りない単語を辞書に追加登録することによる正解率の変化を見てみる．

表 6: 辞書追加登録前と後の正解率 (2006 年 12 月の実験値)

実験データ	前の正解率	後の正解率
書籍 1	95.68%	99.18%
書籍 2	96.63%	99.28%
書籍 3	95.84%	99.55%

このデータによると，辞書登録がされていれば，正解率は 100 %に限りなく近づくことがわかる．では，辞書登録も万全にされている状態で，どのような分か

ち書き変換誤りがあるのだろうか．そのことに関して詳しく見てみると，次のような誤りが見られた．

3.3　実験結果-誤りの種類

分かち書き変換における誤りは，次の２種類がある．

まず一つ目は，マーク付けされた部分の誤りである．つまり，辞書登録されていなかったか，一文字漢字に対する読みの誤りである．分かち書き変換処理内容について述べた際に辞書の種類が５種類ほどあったが，その中の第五の辞書として述べた，文章の前後によって読みが変わる漢字 文脈依存文字 がこれである．これは，先にも述べたが，人的修正作業を効率化させる為のフラグにより，マーク付けされ，発見が容易である．具体的な例としては，先に述べた「根」数学文書においては「こん」と読むことが多いが,「ね」の可能性も捨てられない．

もう一つの誤りは，マーク付けされていない部分の誤りである．今回の実験による，このような誤りの詳細を見てみると，それらは平仮名が続く場合の誤りであった．具体例としては,「～となり,」は，正しくは「～と　なり」という分かち書きとなってほしいのに,「～となり,」という風になってしまう．それは，文末からの辞書内２文節最長一致単語を探し，その最長一致単語に変換するプログラムであるので，後ろから見た際に,「～と」と「なり」の２文節の長さと,「～」と「となり」の２文節の長さが一緒であるので，判定がつかずに誤ったのである．

今回の実験では，１文字漢字に対する読みの誤りはなかったため，マーク付け部分における誤りは辞書登録されていない語のみであった．故に，辞書追加登録後の誤りは後者の方の誤りしかなかったため、人的修正作業のためのフラグ部分を除く正解率は，先の辞書追加登録後の正解率と等しくなった．つまり、マーク付けされた部分のみの修正で，先の正解率が得られたのである．

3.4　辞書追加登録の注意点

今回の実験結果より，辞書登録の際には，平仮名単語の登録に注意する必要があることがわかった．漢字であれば誤るはずのないものでも，平仮名であると，助詞の一部である平仮名との区別がつきにくくなる．しかし，かといって平仮名単語の登録を避け，漢字で書けるものは漢字での登録を優先したのでは，書籍内での単語が，できるだけ漢字で書かれている保障はないので，全ての文書に対応できなくなる．つまり，平仮名での単語登録の際には，その必要性を考える必要がある．

今回の辞書による分かち書き変換プログラムの作成にあたっては，まず辞書データを収集し，なるべく多く登録した．特に，今回品詞等の区別はせずに，辞書を作成したので，活用形のあるものは全て登録するといった方法に出た．また，漢

字の読みを他に構えるといったこともしていないので，漢字に対してその読みに対応するものもなるべく登録するようにした．すると，先に述べたような誤りがたくさんでてきた．

例えば,「〜するとつぎのような」という文は，初め,「〜する とつぎのよーな」といった形に変換された．正しくは,「〜すると　つぎのよーな」と変換したかった．これは,「とつぐ（嫁ぐ）」と言う動詞の変形で,「とつぎ（嫁ぎ）」を登録していた為に起きたのである．漢字であれば誤るはずはないのだが，平仮名であるとこういった誤りがでてしまう．これに関しては，数学文書においては特に，この平仮名での「とつぐ」という動詞は出てこないだろうと判断し，辞書から外した．

4　今後の課題

4.1　プログラムの現状

実験数値からすると，今回の数学文書分かち書き換処理はかなり満足のあるものになっているだろう．しかし，初めに述べた人的修正作業の効率化に関しては，まだ完全には実現されていない．というのも，誤り部分としてでてきた，平仮名が続く場合の誤りは，マーク付けがされていない．もう一つの文脈依存一文字漢字に関してはマーク付けしてあるので，注意して作業すれば発見できるが，平仮名が続く場合は，変換されたものが辞書に登録されているので，マーク付けされていないのである．

また，辞書未登録部分のマーク付けに関しても，数学文書に対する単語辞書登録実験がまだ足りない為，マーク付けされていない箇所でも辞書登録の必要のある単語がある可能性がある．例えば,「図示せよ」という文だと,「ずし せよ」そのものが入っていなければ,「ず しめせよ」といった分かち書きになってしまう．この場合,「図示」と「せよ」が辞書に登録されていたとしても,「図」と「示せよ」が入っていれば，後ろからの 2 文節最長一致法による変換だと，どちらの場合も長さが同じであるので，判定しきれず,「ず しめせよ」と変換される

また，人名，地名などの固有名詞等は，数学専門用語以外は出てきたものしか辞書に登録されていない為，うまく変換できない．但し，その場合は，辞書にないということでマーク付けされているので，人的修正作業により容易に補えるであるだろう．

4.2　今後の課題

今回のプログラムをより信頼性の高いものにするにはやはり，辞書登録データを更に増やす必要があるだろう．但し，その際には先に述べた平仮名単語の登録には十分に注意する必要がある．

実用性のあるプログラムとしては，誤り可能性部分をマーク付けすることで全ての分かち書き変換誤りが自動判定できるのが望ましい．その為には，助詞と平仮名の誤りを解決するために，助詞の解析を強化しようと思う．また,「図示せよ」の例を考えると，一文字で一単語となっている部分の解析も強化しようと思う．

現在，このマーク付け強化のために，Extra(数学専門用語の登録は行なわない状態) による結果との比較機能を追加した．この比較による異なる部分には，誤り可能性部分としてのマーク付けをした．このマーク付けによる効果を検証しているところである．

自動点訳処理には，人的修正作業は欠かせないと思われるので完全に自動点訳処理での正解率を 100 %にすることはできないだろう．それならせめて，人的修正作業効率化のためのフラグ部分のみの修正で正解率が 100 %となるように目指して研究していこうと思う．

参考文献

[1] 特定非営利活動法人　全国視覚障害者情報提供施設協会，(株) 大活字，東京，1981

[2] Extra のホームページ，http://www.extra.co.jp/