日本語文章推敲支援ツール『推敲』におけるとりたて詞「は」の抽出法とその評価†

菅沼 明†† 牛島和夫††

日本語文章推敲支援ツール『推敲』は日本語文章を字面だけで解析し，推敲に役立つ情報を書き手に提供することを目的としてわれわれが開発したツールである．『推敲』には現在，受身，接続助詞「が」，指示詞「これ，それ，…」，とりたて詞（副助詞，係助詞の一部），否定表現などの候補を指摘する機能がある．文章中でそれらを使用していれば，『推敲』がそれを指摘し，書き手に推敲する手がかりを提供する．本論文では，とりたて詞「は」について，それを指摘する意義と，それを抽出する字面解析手法の構築およびその評価に関して述べている．とりたて詞「は」とは副助詞，係助詞「は」のことである．このとりたて詞「は」が文章中に現れる際に文が読み難くなる場合がある．そのために，文章を推敲する際にとりたて詞「は」に注意を払うことは有用である．『推敲』でとりたて詞「は」を指摘するために，字面解析でそれを抽出する方法を構築した．構築するに当たっては，日本語文章約 70 万字を実際に調査し，その結果を参考とした．さらに，別の文章に適用して評価を行った．構築した抽出法は文字についての簡単な条件をいくつか適用するだけの形になっている．これは，「指摘した候補を書き手が必ず吟味する」を『推敲』の開発方針としているために，「とりたて詞でない「は」も候補の中に含まれてしまう」という誤りをある程度許しているからである．実際に構築した抽出法でとりたて詞「は」の候補を抽出すると，候補の中にいくつかのとりたて詞でない表現も含まれる．しかし，抽出精度（実際のとりたて詞の件数／総指摘件数）は，98% 以上である．『推敲』で字面解析を採用したのは「実用規模の文章を待ち遠しくない時間で処理して欲しい」ためである．パソコン（PC-9801）上に実現した『推敲』で処理時間を測定すると，実用規模（1万字）の文章からすべてのとりたて詞「は」の候補を1秒以内で抽出できる．さらにこの抽出法は，解析対象の文章を一度しか走査しないので，検索時間は文章の長さに比例する．

1. はじめに

最近，日本語ワードプロセッサの普及にともなって，計算機可読の日本語文章が増えている．これらの文章は計算機可読でありながら，その利点が必ずしも有効に活用されていない．日本語ワードプロセッサは，本質的には，文書の入力，書式の設定，清書出力，文書の保存を行うものであって，文章の推敲作業を支援するような高度のテキスト処理は行わない．一方，英語の文章に対しては，IBM の EPISTLE-CRITIQUE[1),2)] や，UNIX 上の Writer's Workbench[3),4)] のような作文支援システムが実用に供されている．

人間が行う推敲作業は，文章を読み返して問題となる箇所を探し，その部分を検討して，必要であれば書き直すという形で行う．この作業のうち，問題となる箇所を探す部分を計算機で支援できれば，人間は計算機が指摘したものを吟味して推敲を行うことができるので，人間の作業を軽減できる．このことからわれわれは，

①文章中に問題となりそうな箇所があればそれを指摘できればよい．（実際に推敲するのは書き手である）

②実用規模（1万字程度：図や表を含めると論文誌刷り上がり7～8ページの文字数）の文章を待ち遠しくない時間で処理して欲しい．

という方針を定め，それに従って日本語文章推敲支援ツール『推敲』[5)] の開発を行っている．このツールは日本語文章（漢字かな混じりの科学技術文章を主な対象とする）を字面だけで解析し，推敲に役立つ情報を書き手に提供することを目的としたものである．『推敲』は，いくつかのプロトタイプの段階を経て，現在パーソナルコンピュータ上に実現している[6)]．

『推敲』には指示詞，受身形，接続助詞「が」，否定表現など，推敲に役立つ情報を抽出する機能がある．本論文では，推敲に役立つ情報として，とりたて詞「は」に着目する．本論文の構成は，まず2章でとりたて詞「は」を取り上げる意義について述べる．3章では『推敲』で採用している字面解析法について，それを採用した理由などを説明する．4章では，機械可読の日本語文章を調査した結果を基にして，字面解析

† Construction and Evaluation of a Textual Analysis Method to Extract Particle "は" (wa) in Writing Tools for Japanese Documents by AKIRA SUGANUMA and KAZUO USHIJIMA (Department of Computer Science and Communication Engineering, Faculty of Engineering, Kyushu University).

†† 九州大学工学部情報工学科

によるとりたて詞「は」の抽出法を構築する．さらに5章では，4章で使用したものとは別の文章（約280万字と約2,000万字の文章）に適用して，抽出法を評価する．

2. とりたて詞「は」

とりたて詞「は」は，文中のある構成要素を取り出して提示するはたらきをする副助詞または係助詞である[7),8)]．助詞「は」の機能をさらに詳しく分けると題目（提題）と対照との二つになる．

題目の機能を果たす「は」は文のいろいろな成分に付いて，その成分が題目（テーマ）であることを示すものである．たとえば，「この飛行機は新型機です」という文では，「この飛行機について述べると」というように「この飛行機」をテーマとした文であることを表している．しかし，「この飛行機」をテーマとして取り上げたくない場合には，「この飛行機が新型機です」としたほうがよくなることもある．また，埋め込み文にするときには，「彼女に，この飛行機が新型機であることを教えた」のように「は」が「が」となる．

対照の「は」はある脈絡のなかにあるものを指定して，それと関連するものを対比することで強調するものである．この対比されるものは文脈に明示されている場合もあれば，文脈に含まれている場合もある．

日本語文では，1文中にとりたて詞「は」が複数含まれることが多くある．題目の「は」はその性質上1文に一つしか存在できないために，複数のとりたて詞「は」がある場合は，二つめ以降は対照の「は」と考えることができる．対照の「は」は言葉の上ではいくらでも重ねることができる．しかし，数が多くなると対比するものが膨大な量となるために，わかりにくい文になりやすい．

『推敲』を一部大学関係者に試用してもらい評価をしてもらった．その中に「主格を表す助詞「が」と「は」の誤用を指摘できないだろうか」という中国人留学生からの要求があった．「は」と「が」の使用法には微妙な違いがあり，その使い分けがこの外国人留学生にはとても難しいとのことである．また，日本人も助詞「が」や「は」を繰り返し使用し，分かりにくい文を書く場合もしばしばある．

このように，文章中のとりたて詞「は」を指摘することは文章を推敲する上で有効な情報となりうる．

3. 字面解析

日本語の文章は分かち書きされていない．そのため，日本語の文章を解析するには，まず辞書に基づいて形態素解析を行い，構文解析を行うのが普通である．しかし，それらを行っても必ずしも一意に解析できるとは限らない．さらに，辞書を使った文法処理を行うと，実用規模の文章では解析時間がかかりすぎて『推敲』の開発方針②を満たさない．そのため，われわれは日本語文章を字面だけで解析する方法を採用した．

文章を字面だけで解析する方法を採ったため，解析精度（指摘するものの数に対する指摘すべきものの割合）は文法解析を行う場合よりも低いことが予想できる．しかし，『推敲』の開発方針①から，『推敲』が指摘したものは書き手が一度は目を通すことになるので，文章の解析精度はこの開発方針を満たす程度でよい．

文章から問題となる箇所を抽出する場合に犯す可能性がある誤りには2種類ある．そのうち，第一種の誤り「指摘に漏れがある」は犯してはならないけれど，第二種の誤り「指摘すべきでないものまで指摘してしまう」はある程度許容できるとしている．字面だけの解析で第一種の誤りを犯さずに開発方針を満たす程度の精度が得られれば，『推敲』で採用する解析手法としては十分である．

すでに，われわれは指示詞（「これ，それ，あれ」など），受身などの助動詞「れる，られる」，接続助詞「が」，否定表現の抽出を上記の条件を満たす形で実現することに成功した[9)~11)]．指示詞の抽出は単に文字列照合だけであるけれども，受身の抽出，接続助詞「が」の抽出，否定表現の抽出では機械可読な日本語文章約70万字を実際に調査して，抽出法を構築した．これらの字面解析手法による抽出では，第一種の誤りを犯さずに約95%以上の精度で目的のものを抽出する．

4. 助詞「は」の抽出

4.1 文字「は」の調査

この章では字面だけの情報で助詞「は」を抽出する方法について述べる．まず，文章中に現れる文字「は」は助詞「は」の候補である．この候補には，たとえば「はじめに」「はっきり」という単語（第二種の誤り）が含まれてしまうので，それらをふるい落と

す条件を設けて助詞「は」の候補を絞り込む．

これまでに構築してきた字面解析手法では，学校文法から設けた判定条件（抽出すべきものがどのような語に接続するか）が有効であった．たとえば，受身形の抽出では，学校文法から設けた判定条件だけで候補を約半分に絞ることができ[9]，接続助詞「が」の場合では，候補を約10分の1以下に絞ることができた[10]．そこで，助詞「は」の場合にも同様に，学校文法から文字「は」の1文字前の条件を設けることを期待した．

学校文法によると，助詞「は」の接続は，

（1） いろいろな語に接続する．

（2） 用言（動詞，形容詞，形容動詞）には，連用形に接続する

とある．接続(2)の性質からは「は」の1文字前の条件を設けることができる．しかし，接続(1)の性質があるために，学校文法から「は」の1文字前の条件を設けることはできない．

研究室に蓄積している機械可読の日本語文章（総文字数 683,867 字，論文，レポートなどから成る漢字かな混じりの文章）中から文字「は」を検索し，その「は」が実際に助詞として使われているか否か，「は」がどんな文字に接続しているかを調査した．その結果を表1に示す．

表1に示したように，約 68 万字の文章中に文字「は」が 10,521 個出現し，そのうちの 10,105 個 (96.0%) が助詞「は」であるという結果を得た．これは受身や接続助詞「が」の場合と比べて高い割合を示している*．抽出の精度（指摘した候補の数に対する指摘すべきものの割合）は高いにこしたことはない．開発方針①からすると抽出の精度は 95% 以上くらいあれば十分だとしているので，文字列照合だけによる抽出でもよさそうである．しかし，抽出精度としては満足いく値が得られても指摘される候補の絶対数が多いので，第二種の誤りが目だつ．そのために，第二種の誤りを除去するいくつかの判定条件を検討する．

表 1 68 万字の文献中の「は」の数
Table 1 A number of occurrences of the character "は."

「は」の1文字前	総 数	助 詞	割 合
記号，英数字など	1,351	1,311	97.2%
カタカナ	1,522	1,522	100.0%
漢 字	3,112	3,053	98.1%
ひらがな	4,536	4,219	93.0%
合 計	10,521	10,105	96.0%

* 割合は総数に対する助詞の数の割合

* たとえば，接続助詞「が」の場合には，文中に含まれる「が」の数が 6,978 で，そのうち本当の接続助詞は 398 個 (5.7%) であった．

{記号，英数字など}＋「は」の形で出現した「は」のうち助詞でないものは，「はじめに」が半数以上 (72.5%) を占めていた．これは，対象とした日本語文章のデータが論文，レポートなどを多く含むためだと思われる．「はじめに」という語の性質から，文頭や，数字＋ピリオドの後（たとえば，「1. はじめに」）に出てくる場合が多いので表1のような結果になったものと思われる．

助詞「は」は付属語であるので「は」の前には自立語または付属語があるはずである．そのことから，「は」の1文字前が空白や改行，句読点（ピリオド，コンマも含む）などの場合には，その「は」は助詞ではないと判断できるように思える．しかし，下に示すように文字「は」の1文字前が読点であるものや改行記号であるものもこの調査でみられた．

「但し，P，Q，R…，はn変数関数である」

「たとえば，

$X \leftarrow (A * B) + (C * D)$；［改行記号］

は，次のようにコンパイルされる」

これらの「は」は，明らかに助詞である．したがって，「は」の1文字前が空白や改行，ピリオドといった条件で除去すると第一種の誤りを犯す可能性があるので，別の方法を考えなければならない．

日本語文章中に現れるカタカナはそれだけで一つの単語となることが多い．今回の調査で出現した {カタカナ}＋「は」のパターンでは，カタカナの部分がすべて名詞またはその一部であり，カタカナに続く「は」はすべて助詞であった．

{漢字}＋「は」のパターンは，{カタカナ}＋「は」と同様に，名詞もしくは名詞相当語句の一部に助詞の「は」が付いたものがほとんど（約 98%）であった．助詞でなかった「は」は，すべて「又は」（接続詞）であり，他のものは現れなかった．

{ひらがな}＋「は」については，前述のケースとは異なりこれといった傾向はみられず，「あるいは」，「または」などの接続詞や，名詞，動詞の一部になっているものなど様々であった．

今回の調査で出現した「は」のうちで助詞でないものの内訳を表2に示す．

表 2 助詞でない「は」の内訳
Table 2 A list of "は" which occurs as a non-particle.

項　目	数	割 合
あるいは，或いは	158	38.0%
または，又は	120	28.8%
はじめに，はじめる はじめて，はじめの はじまる，はじまり	37	8.9%
はっきり	21	5.0%
「を」+「は～」	13*	3.1%
そ の 他	74	17.8%
合　計	416	—

* 「を」+「は～」は，13個あったがそのうちの7個は「はじめる」，「はっきり」と重複している．

4.2 「は」で終わる接続詞

文章中に現れる文字「は」には，"ha" と発音する場合と "wa" と発音する場合との二通りがある．先に述べたとりたて詞「は」は，"wa" と発音する「は」である．"wa" と発音する「は」には，そのほかに「あるいは」「または」など「は」で終わる接続詞もある．2章で述べた1文中に含まれる複数の「は」の問題は，"wa" と発音する「は」が複数含まれる文でも起こりうる．特に，文節の終わりに "wa" と発音する「は」が数多く出現する次のような文，

「あるいは，この学校では私は，～」

は耳障りである．

このように "wa" と発音する「は」を重ねた場合も，とりたて詞「は」を重ねた場合と同様に文が読みにくくなることがある．したがって本論文で述べる抽出法では，助詞の「は」だけに限定しないで，「は」で終わる接続詞も含めて，"wa" と発音する「は」を抽出の対象とする．

4.3 助詞でない「は」の除去

◎「は」の1文字前の条件

今回の調査で出現した「を」+(「は」で始まる語)についてはそれを取り除く判定条件を設けることができる．文中に現れる助詞「を」は目的格を表す．この目的格の単語を強調するには「を」の後ろに「は」を付けるのではなく，「を」を「は」に変えるだけでよい．そのために格助詞の「を」に副助詞の「は」が続くことはない．また「を」で終わる単語もない．このため，以下の条件を判定条件に加える．

判定条件1：「は」の1文字前が「を」である場合，その「は」は助詞でない．

◎「は」の1文字後の条件

「は」を助詞と仮定すると「は」の後に続く文字は自立語の最初の文字である．ところが，促音(「っ」)や撥音(「ん」)で始まる自立語はないことから，「は」の1文字後が促音かまたは撥音である場合，この「は」は助詞でないといえる．そのため以下の条件を判定条件とする．

判定条件2：「は」の1文字後が促音，撥音である場合，その「は」は助詞でない．

以上，二つの判定条件を設けることができた．しかし，それらの判定条件では今回の調査に出現した「はじめに」など出現頻度が多い第二種の誤りを取り除くことができない．そのため，以下の条件も判定条件に付け加える．

◎はじめ*，はじま*

検索した「は」が助詞の「は」であれば，「は」に続く単語は自立語である．したがって，「じめ」，「じま」で始まる自立語を公用データベース日本語単語辞書[12)]* で調べた．その結果を**表3**に示す．このことから，以下の二つの判定条件を設けることができる．

判定条件3：「は」の後に「じめ」と続いた場合，「め」の1文字後が「い，じ，つ，ん」のいずれかの場合だけ，その「は」を助詞の候補とする．

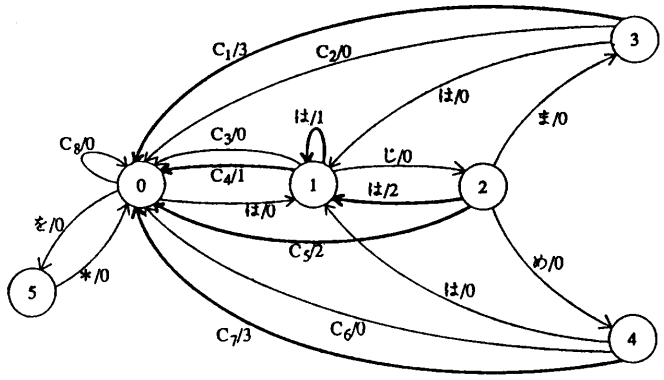
判定条件4：「は」の後に「じま」と続いた場合，「ま」の1文字後が「い，え，く，ま，わ，ん」のいずれかの場合だけ，その「は」を助詞の候補とする．

以上，四つの判定条件を使用して助詞「は」の候補を抽出する．これをプログラムとして動かす場合には，**図1**に示すようなパターンマッチングマシン(PMM)を利用する．このPMMは，状態0を初期状態として，文字を1文字ずつ受け取り，内部状態を遷移し，outputに示してある値を返す．0以外の値が返ってきたときに "wa" と発音する「は」の候補が検出されたことを表す．この0以外の戻り値は，最

表 3 「じめ」「じま」で始まる語
Table 3 A list of words beginning string "じめ" or "じま."

じめい（自明），じめじめ，じめつ（自滅）
じめん（字面），じめん（地面）
じまい（地米），じまえ（自前），じまく（字幕）
じまま，じまわり（地回り），じまん（自慢）

* 見出し語総数約18万7千語からなる自立語だけの辞書を使用した．



input / output
input : 入力文字
＊は任意の文字
$C_1 \in$ {い,え,く,ま,わ,ん}　$C_2 \in$ {い,え,く,は,ま,わ,ん}
$C_3 \in$ {ん,っ}　$C_4 \in$ {じ,は,ん,っ}　$C_5 \in$ {は,ま,め}
$C_6 \in$ {い,じ,つ,は,ん}　$C_7 \in$ {い,じ,つ,ん}　$C_8 \in$ {は,を}
output : 出力値
0以外のとき候補を検出
(0以外の値は候補が何文字前にあるかを表す)

図 1 "wa" と発音する「は」を抽出するパターンマッチングマシン
Fig. 1 The pattern matching machine to extract particles "は."

後に PMM に渡した文字の何文字前に候補があるかを表している．この PMM を使用することで，文章中にある文字を1度しか走査しないので，"wa" と発音する「は」を抽出する時間計算量は $O(n)$ になる．

4.4 抽出アルゴリズムの精度

前節で述べた四つの判定条件を用いると，"wa" と発音する「は」の候補として抽出する数が 10,455 個である．そのうち正しい候補("wa" と発音する「は」)が 10,386 個で，その精度は約 99.3% となる．これは，実用規模（約1万字）の文章で考えると，平均 152.9 個の指摘があり，そのうち誤って指摘してしまうものが，約 1.01 個しか含まれないことになる．

推敲作業を支援する情報としてわれわれが抽出したいものは，とりたて詞「は」を重複して含んでいる文である．そこで，とりたて詞「は」の候補を1文中に複数個含む文の数を調査した．その結果を表4に示す．前記の 68 万字の文章では，文が 18,381 あり，そのうちとりたて詞「は」の候補を複数含む文は 1,825 (9.9%) であった．この指摘には，

- 「実現の詳細の部分は，代数的仕様記述は記述できる必要はない」
- 「主プログラムは仮想的な主タスクから呼び出されると知っておくのは大切である」

のような聞き苦しい文もあれば，

- 「この方法は検証の網羅性がなく，検証された妥当性は部分的なものにすぎない」
- 「論理リンク層のサービスデータ単位はフレームであり，データ長は固定である」

のような自然な文も含まれている．しかし，とりたて詞「は」の候補を複数含む文を探すだけで推敲の対象となる文の数を 10 分の1に減らすことができることに注意されたい．

表 4 とりたて詞「は」の候補を複数含む文
Table 4 A number of sentences which include some candidates of the particle "は."

「は」の数	文の数	割合
0	10,113	55.0%
1	6,443	35.1%
2	1,511	8.2%
3	277	1.5%
4	26	0.1%
5	11	0.1%
合計	18,381	100.0%

5. 抽出アルゴリズムの評価

5.1 抽出精度

4.3 節で述べた抽出アルゴリズムは約 68 万字の日本語文章を調査して構築したものである．このアルゴリズムが他の日本語文章にも有効であることを確認するために，別の文章で評価を行った．評価に使用した日本語文章は論文の抄録（総文字数約 280 万字）と新聞記事（総文字数約 2,000 万字）である．

評価の方法は構築した判定条件に従って「は」を抽出し，"wa" と発音する「は」であるか否かを調査した．さらに，各判定条件によって除去される「は」について，第一種の誤りを犯していないかどうかを調べた．

5.1.1 抄録による評価

推敲支援ツール『推敲』では科学技術文章を主な解析対象と考えている．4.3 節で述べた抽出アルゴリズムは，『推敲』に組み込むので，一般の科学技術文章を高い精度で解析してほしい．また，第一種の誤りを犯さないことも確認したい．そのために，論文の抄録を使用して各抽出アルゴリズムの評価を行った．使用したものは，JICST 科学技術文献ファイルの管理システム編の表題と抄録の部分で，文献数は 14,380 文

表 5 文字「は」の調査結果（論文抄録）
Table 5 A result of extracting the candidates of the particle "は" (abstract).

総文字数	2,842,062
「は」の数	33,783
判定条件を満たす「は」	33,011
"wa" と発音する「は」	32,489
助詞「は」	31,550
「は」で終わる接続詞	939

表 6 第二種の誤り（論文抄録）
Table 6 Errors of the second kind appearing in the evaluation of the textual analysis method (abstract).

項目	数	割合
「はあく，は握」	243	46.6%
名詞の一部	59	11.3%
「あてはまる」	31	5.9%
「はるかに」	25	4.8%
その他	164	31.4%
合計	522	—

献，総文字数は 2,842,062 文字である．

調査の結果を**表 5** に，判定条件を満たす第二種の誤りを**表 6** に示す．評価の結果，約 280 万字の文章中に「は」が 33,783 個含まれており，そのうち 33,011 個の「は」が四つの判定条件を満たす．目視によって調べた結果，助詞の「は」は 31,550 個であり，「は」で終わる接続詞は 939 個出現した．このことから，抽出の目的である "wa" と発音する「は」の数は 32,489 個となり，抽出の精度は 98.4% となる．実用規模（約 1 万字）の文章で考えると，平均 116.2 個の指摘があり，その中に約 1.8 個の誤りを含むことになる．抄録を使用した評価で得た抽出精度は，抽出アルゴリズムを構築する際の調査結果と比べると 1% 程度悪い値となっている．この原因は，表 6 にあるように，「はあく，は握」が平仮名書きされていたために，その「は」を "wa" と発音する「は」の候補として抽出してしまうからである．漢字「把」「握」はともに常用漢字であるために，「はあく，は握」のすべてが「把握」と書かれていたと仮定すると，抽出精度は 99.1%となる．

文章中から文字「は」を検出する場合に，"wa" と発音する「は」の取りこぼしは起こらない．さらに，判定条件で取り除かれた「は」772 個（33,783−33,011）を調べた結果，その中に "wa" と発音する「は」が含まれていないことを確認した．すなわち，上記判定条件を使用しても第一種の誤りを犯していない．

表 7 文字「は」の調査結果（朝日新聞記事）
Table 7 A result of extracting the candidates of the particle "は" (articles).

総文字数	20,969,926
「は」の数	353,786
判定条件を満たす「は」	348,964
"wa" と発音する」は」	341,950
助詞「は」	340,234
「は」で終わる接続詞	1,716

5.1.2 新聞記事による評価

抽出アルゴリズムを構築する際に調査した文章も，前項で述べた評価に使用した文章も科学技術文章がほとんどであった．ここでは，科学技術文章とは異なる新聞記事を使用した．これは，新聞記事の文章が科学技術文章に比べて日常書かれる文章に近いと考えられるからである．さらに，大量の文章を機械可読な形で入手することができたためでもある．使用したのは朝日新聞の半年分（1988 年前半）の記事データで，総文字数は約 2,000 万字である．

調査した結果を**表 7** に示す．約 2,000 万字の文章中に文字「は」が 353,786 個含まれており，そのうち "wa" と発音する「は」は 341,950 個（96.7%）であった．この値は，抽出アルゴリズムを構築する際の調査や抄録での評価の場合とあまり変わらない．文字「は」に対して 4.3 節で構築した判定条件を適用して，条件を満たすものを取り出した結果，"wa" と発音する「は」の候補は 348,964 個であった．これより，抽出精度は 98.0% となり，前述の二つの調査と比べるとやや下がる．

また，第一種の誤りを犯してないことを 5.1.1 項と同様にして確かめた．

5.2 "wa" と発音する「は」を複数含む文

4.2 節で述べたように，"wa" と発音する「は」を複数含む文は読み手にとって耳障りになる場合がある．今回構築した抽出アルゴリズムを使用して，"wa" と発音する「は」を複数含む文の候補を取り出して書き手に提供したい．その場合，問題でない文を数多く指摘してしまうことが考えられる．このやり方でどの

表 8 "wa" と発音する「は」の候補を複数含む文
Table 8 A result of counting the sentences which include some candidates of the particle "は."

「は」の数	JICST 論文抄録		新聞記事データ	
	文の数	割合	文の数	割合
0	31,955	58.0%	220,514	46.0%
1	16,122	29.3%	187,399	39.2%
2	5,058	9.2%	54,497	11.4%
3	1,401	2.5%	12,612	2.6%
4	390	0.7%	2,671	0.6%
5以上	177	0.3%	754	0.2%
合計	55,103	100.0%	478,447	100.0%

くらい推敲の範囲を絞り込むことができるかを知るために，前項の評価で使用した論文抄録と新聞記事データとを使用して，"wa" と発音する「は」の候補を複数含む文の調査を行った．

調査結果を表 8 に示す．論文抄録を使用した場合，文の総数が 55,103 であり，"wa" と発音する「は」の候補を複数含む文が 7,026 (12.8%) という結果を得た．また，新聞記事データを使用した場合には，文の総数が 478,447 で，"wa" と発音する「は」を複数含む文が 70,534 (14.7%) であった．この指摘には，

- 「今日では，この種の方法は，一時的あるいは特定結果を目指した場合には有効だが，その他の場合には必ずしも有効ではないと判明した」
- 「同次官は今後の動向については，11月実績は10月実績があまりにも悪すぎたための反動ともいえ，このあとは10月ほどは悪くないが，11月ほどは良くないという数字が続くであろうと語った」

のような聞き苦しい文もあれば，

- 「設備室には昼間は警備員が1人常駐し，夜はパトロールをしているが，午後7時35分ごろに警備員が見回ったときは，設備室付近には異状はなかった」

のように対比するものが多くて読み手にとって繁雑に感じる文もある．また，

- 「得られた式は $ST=a+tT+dD+fDT$ で ST は検索時間，T はプロファイルタームの数，D は文献の数，a は定数，t，d，f は回帰係数」

のような自然な文も含まれている．しかし，とりたて詞「は」の候補を複数含む文を探すだけで推敲の対象となる文を全体の約 15% に減らすことができることに注意されたい．このことから，今回構築した抽出法を利用して，"wa" と発音する「は」を複数含む文の推敲範囲を，十分絞り込むことができると考えている．

5.3 応答時間

『推敲』を開発する際の方針の一つとして『実用規模の文章を待ち遠しくない時間で処理してほしい』をあげている．ここでは，上で述べた抽出法がこの方針を満たしているかどうかを評価する．

『推敲』は現在パソコン上に実現している．ユーザがキーボードからコマンドを発すると，『推敲』はすべての候補を検索し，結果の先頭部分を画面に表示する．画面からあふれた結果に対して，ユーザは画面をスクロールさせて候補を一つ一つ吟味していく．このように，『推敲』を使用する際には，コマンドを発してから先頭部分の画面が表示されるまでの時間（応答

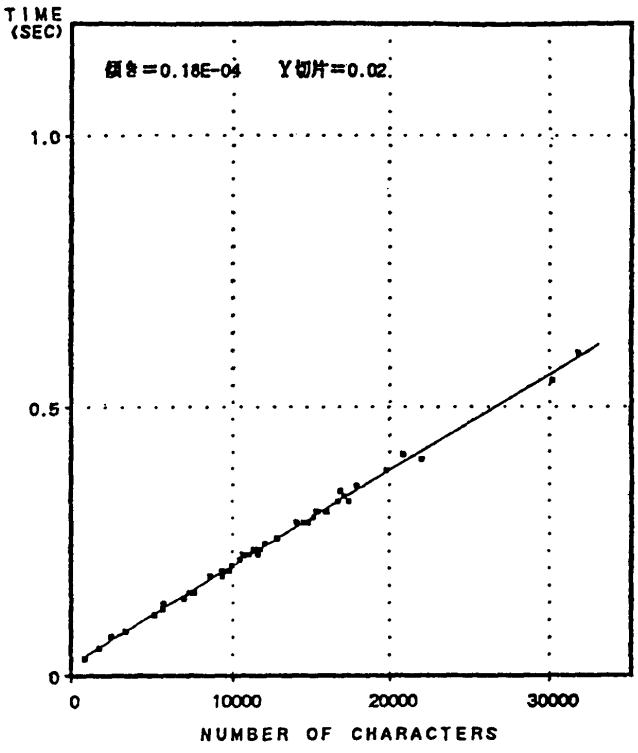


図 2 応答時間の測定結果
Fig. 2 The response time of the textual analysis method.

時間）が問題となる．

4.3 節で述べた判定条件を『推敲』に組み込んで，とりたて詞「は」の候補を抽出するときの応答時間を測定した[13]．測定には，PC-9801 VX (CPU i 80286, CLOCK 10 MHz) を使用した．その結果を図2に示す．図からわかるように応答時間は文章の文字数にほぼ比例する．これは，解析対象の文章を一度しか走査しないためである．

図から，入力文章が3万字程度であっても応答時間が1秒以内であることがわかる．すなわち，『推敲』に課した「実用規模の文章を待ち遠しくない時間で処理して欲しい」を十分満たしているといえる．また，"wa" と発音する「は」を複数含む文を抽出して指摘するのに要する時間（同様に応答時間を測定）も測定した．その結果，とりたて詞「は」の候補を抽出するときの応答時間とほとんど同じ値を得た．

6. おわりに

文章中に存在する "wa" と発音する「は」を字面解析だけで抽出する方法を構築した．抽出の精度は約99%である，さらに．字面解析手法を構築する際に使用した文章とは異なる文章を使って評価した結果でも抽出精度は 98% 程度の値が得られた．このことから，上で述べた "wa" と発音する「は」の候補の抽出法は，『推敲』の開発方針①からすると，十分実用的であるとみなすことができよう．しかし問題となるのは，指摘の数そのものが多いということである．そのため，「は」の抽出を『推敲』に組み込む場合には，それを基本コマンド[6]として実現し，通常のユーザに対しては他のコマンドと組み合わせた使いやすいマクロコマンド（たとえば "wa" と発音する「は」を複数含む文を抽出せよなど）を提供しなければならない．

『推敲』に組み込む字面解析手法を構築することを考えて，文章中に現れる文字「は」の調査を行った．この調査で得られた結果から，漢字かな混じり文章中で文字「は」を見たら，それは助詞の「は」であるというヒューリスティックスを使用しても，それによる誤りは 4% 程度であることが確かめられた．

謝辞 本研究を進めるにあたり，JICST 廃棄テープの使用について姫路短大の田中康仁教授（現在，愛知淑徳大学）に便宜をはかっていただいた．朝日新聞ニューメディア本部には，貴重な新聞記事データの使用を許していただいた．また応答時間の測定には製品科学研究所の森川浩氏から提供していただいた打鍵データ収集システムを使用した．ここに記して謝意を表する．

この研究は一部，文部省科学研究費補助金試験研究（課題番号 01880008）の援助を受けた．

参考文献

1) Heidorn, G. E., Jensen, L. T., Miller, L. A., Byrd, R. J. and Chodorow, M. S.: The EPISTLE Text-critiquing System, *IBM Syst. J.*, Vol. 21, No. 3, pp. 305-326 (1982).
2) Richardson, S. D.: Enhanced Text Critiquing Using a Natural Language Parser, IBM Research Report, #51041 (1985).
3) Cherry, L.: Writing Tools, *IEEE Trans. Communications*, Vol. COM-30, No. 1, pp. 100-104 (1982).
4) Macdonald, H. H.: The UNIX Writer's Workbench Software—Rational and Design, *Bell Syst. Tech. J.*, Vol. 62, No. 6, pp. 1891-1908 (1983).
5) 牛島和夫，日並順二，尹 志熙，高木利久：日本語文章推敲支援ツール『推敲』のプロトタイピング，コンピュータソフトウェア，Vol. 3, No. 1, pp. 35-46 (1986).
6) 倉田昌典，菅沼 明，牛島和夫：日本語文章推敲支援ツール『推敲』のパソコン上での実用化，コンピュータソフトウェア，Vol. 6, No. 4, pp. 55-67 (1989).
7) 本多勝一：日本語の作文技術，朝日新聞社 (1976).
8) 中島文雄：日本語の構造—英語との対比—，岩波新書 (1987).
9) 牛島和夫，石田真美，尹 志熙，高木利久：日本語文章推敲支援ツールにおける受身形の抽出法，情報処理学会論文誌，Vol. 28, No. 8, pp. 894-897 (1987).
10) 菅沼 明，牛島和夫：日本語文章推敲支援ツール『推敲』における字面解析手法とその評価，自然言語処理研究会報告，No. 68, 68-8 (1988).
11) 菅沼 明，倉田昌典，牛島和夫：日本語文章推敲支援ツール『推敲』における否定表現の抽出法，情報処理学会論文誌，Vol. 31, No. 6, pp. 792-800 (1990).
12) 吉田 将，日高 達，稲永紘之，田中武美，吉村賢治：公用データベース日本語単語辞書の使用について，九州大学大型計算機センター広報，Vol. 16, No. 4, pp. 335-361 (1983).
13) 森川 浩：キーボードエミュレーションを行うとき望まれる BIOS の機能について，情報処理学会文書処理とヒューマンインタフェース研究会，16-2 (1988).

(平成3年5月27日受付)
(平成3年9月12日採録)

菅沼　明（正会員）

1961年生．1986年九州大学工学部情報工学科卒業．1988年同大学大学院工学研究科情報工学専攻修士課程修了．1991年同博士課程修了．同年九州大学工学部助手，現在に至る．工学博士．日本語処理，ユーザインタフェースに興味を持つ．日本ソフトウェア科学会会員．

牛島　和夫（正会員）

1937年生．1961年東京大学工学部応用物理学科（数理工学）卒業．1963年同大学院修士課程修了．同年九州大学中央計数施設勤務．1977年九州大学工学部情報工学科教授（計算機ソフトウェア講座担当），現在に至る．1990年4月から九州大学大型計算機センター長を兼務．工学博士．日本ソフトウェア科学会，電子情報通信学会，ACM 各会員．