日本語文章推敲支援ツール『推敲』における助詞「が」の抽出について

下園幸一† 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開発したツールである．本論文では，助詞「が」について，それを指摘する意義と抽出法の構築および評価に関して述べている．助詞「が」には接続助詞「が」と格助詞「が」とがある．文章を推敲する場合に，あいまいな接続助詞「が」や，格助詞「が」ととりたて詞「は」の使い分けを熟考することは重要である．まず，実際の文章中の文字「が」を調査し，格助詞，接続助詞，その他に分類した．その結果を基に，文章中の「が」からその他の「が」と接続助詞とをふるい落とすことで，格助詞「が」を抽出する抽出法を構築した．その際に，既存の接続助詞「が」の抽出法を使用することを検討した．しかし，この抽出法では，接続助詞でない「が」も候補に含んでしまうため，それをそのまま格助詞「が」抽出に使用すると，候補に含まれない格助詞「が」が存在してしまう．今回，用言，助動詞だけを要素として持つ辞書を構築し，それと字面解析手法を使用して，接続助詞「が」の抽出法の再構築を行った．その結果，新聞記事データから95.7％の精度で接続助詞「が」を抽出できることを確認した．この抽出法を利用して格助詞「が」の抽出を行った．また，格助詞「が」の抽出法と以前構築したとりたて詞「は」の抽出法とを利用して，「が」と「は」を複数含む文を高速に抽出して書き手に提示することができる．

Extraction Methods for Nominative/Conjunctive Particle "*ga*" in the Writing Tool "SUIKOU" for Japanese Documents

Koichi Shimozono,† Akira Suganuma†† and Kazuo Ushijima††

A system designed as a writing tool, called "SUIKOU", analyzes a machine-readable Japanese document only textually and provides writers with useful information for polishing it. In Japanese, it is difficult to choose the suitable particle from the two possible options: "*ga*" (nominative/conjunctive particle) and "*wa*" (topic/contrastive particle). Thus, to point out the occurrences of these particles in a document is helpful for polishing Japanese documents. This paper describes extraction methods for particle "*ga*". First, the authors investigate Japanese document samples and categorize the occurrences of character "*ga*" into three classes: nominative particle; conjunctive particle; and non-particle. Using this result, the authors formulate a simple method to detect the occurrences of nominative particle "*ga*". It filters out the occurrences of "*ga*" which are possibly conjunctive or non-particle. However, the method also filters out some of the occurrences of nominative "*ga*". Therefore, the authors compile a small dictionary consisting of verbs, adjectives and auxiliary verbs, and link it with the textual analysis method proposed first. Combining the extraction method for nominative particle "*ga*" and that for topic/contrastive particle "*wa*" the authors previously constructed, "SUIKOU" can detect every possible occurrence of those particle and display sentences containing them very quickly.

1. はじめに

日本語ワードプロセッサの普及は著しい．この日本語ワードプロセッサの基本機能は，文章の入力，書式の設定，印刷，文章の保存である．文章が機械可読な日本語テキストになっているのであるから，もっと高度なテキスト処理を施したい．われわれは，文章の推敲作業に着目し，日本語文章推敲支援ツール『推敲』を開発してきた[1),2)]．

文章を推敲する際，書き手は，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ったことを行う．そこで，問題とな

† 九州大学情報処理教育センター
Educational Center for Information Processing, Kyushu University
†† 九州大学工学部情報工学科
Department of Computer Science and Communication Engineering, Faculty of Engineering, Kyushu University

る箇所を探す」という部分を計算機で可能なかぎり支援し，書き手は計算機が指摘したものを吟味して必要があれば書き直すという方法をとれば，書き手の推敲作業をより質の高いものにできると考えた．

われわれは日本語文章推敲支援ツール『推敲』を開発するにあたり，以下の二つの方針を立てた．

（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（2〜3秒程度）で処理して欲しい．

この方針に従い，現在『推敲』はパーソナルコンピュータ上に実現している[2]．

『推敲』には指示詞，受身形，接続助詞「が」，否定表現などの候補を抽出する機能がある．本稿では，まず用言のみを要素とする辞書と，助動詞のみを要素とする辞書を使って，以前構築した接続助詞「が」の抽出法[3]の精度を上げる方法について述べる．次にこの接続助詞「が」の抽出法を基に格助詞「が」を抽出する方法について述べる．最後に，この方法を適用して，1文中に含まれる格助詞「が」ととりたて詞「は」の数の合計が複数となる文を高速に抽出する機能の実現について述べる．

2. 助詞「が」を抽出する意義

助詞「が」には接続助詞と格助詞がある．接続助詞「が」は，順接，逆接，ただ二つの句をつなぐだけ，という三つの用法を持っている．以下にその例を示す．

順接：今日は暖かかった**が**，明日も暖かいであろう．

逆接：来いと言った**が**，来ていない．

句をつなぐだけ：パッケージの仕様中にある関数 IS_NULL である**が**，これは抽象データ型 lists で使うものである．

つまり，接続助詞「が」は，どのような関係にある二つの句でも接続することができる．この性質により，接続助詞「が」を文章中に用いると書き手の意図が読み手に正しく伝わらないことが起こりうる[4]．

格助詞「が」は体言に付いて，その体言が用言に対して主格の関係にあることを示す．この格助詞が複数出現する文は読みにくくなる可能性がある．また，助詞「は」と「が」の使い分けは微妙であり，1文中に同じ助詞が繰り返されると耳障りであったり文がわかりにくくなったりするので推敲の対象として重要である．

3. 字面解析手法とその精度

従来『推敲』で採用している日本語文章解析法は，字面解析のみによる手法である．この手法は，文中のある特定の文字列（助詞「が」の抽出の場合では文字「が」）に注目して，その文字列の前後の文字に条件を付けることによって，抽出したいものであるかどうかを決定する．この条件は学校文法にある単語の接続条件やわれわれの研究室で書かれた機械可読な科学技術論文（総文字数669,842文字：以下**67万文字文章**と称す）での調査結果を利用して設けてきた[3),5),6)]．

文章を字面だけで解析する方法を採ったため，抽出精度は文法解析を行う場合よりも低いことが予想される．この抽出精度の指標として，情報検索の分野で使用されている再現率と適合率とを使用する．再現率，適合率は以下の式で定義される．

$$\text{再現率}=\frac{\text{候補中に含まれる抽出すべき対象の数}}{\text{文章中の抽出すべき対象の数}}$$

$$\text{適合率}=\frac{\text{候補中に含まれる抽出すべき対象の数}}{\text{抽出された候補の数}}$$

文章から問題となる箇所を抽出する場合に犯す誤りには2種類ある．第一種の誤り「指摘に洩れがある」と第二種の誤り「指摘すべきでないものまで指摘してしまう」である．第一種の誤りを犯せば，再現率が下がる．第二種の誤りを犯せば，適合率が下がる．現在までに構築してきた字面解析手法は，第一種の誤りを犯さないこととしているので，再現率は100%である．一方，『推敲』の開発方針（1）から，『推敲』が指摘したものは書き手が目を通すことになるので，第二種の誤りを犯すことはある程度許容する．しかし，誤りは少なければ少ないほどよい．すなわち，再現率100% のもとで適合率をできるだけ高くできるような字面解析手法を構築してきた．

4. 接続助詞「が」の抽出法

『推敲』[2]で採用している字面のみによる接続助詞「が」の抽出法を**表1**に示す[3]．以下これを字面抽出法と呼ぶ．この表にある各判定条件を設けた根拠は以下のとおりである．

判定条件 1：接続助詞「が」は用言および助動詞の終止形に接続する．用言の活用をみると，形容詞の終止形は「い」，形容動詞の終止形は「だ」でそれ

ぞれ終わる．また，五段活用動詞はカ行，サ行，タ行，ナ行，マ行，ラ行，ワ行，ガ行，バ行にしか存在しないために，五段活用動詞の終止形は「う，く，す，つ，ぬ，む，る，ぐ，ぶ」のいずれかの文字で終わる．さらに，上一段活用動詞，下一段活用動詞，カ行変格活用動詞，サ行変格活用動詞の終止形はいずれも「る」で終わる．

助動詞はその活用のしかたで動詞型，形容詞型，形容動詞型，特殊活用型，語形変化なしの5種類に分けることができる．特殊活用型の助動詞のうちの「ます，です，ぬ」と，語形変化のない助動詞「う，よう，まい」は，その終止形が動詞，形容詞の終止形と同じ文字で終わる．過去の助動詞「た」と否定の助動詞「ん」の二つだけが動詞，形容詞，形容動詞の終止形の文字とは異なる文字で言い切る．以上のことから，文章中で「が」を見つけたとき，その「が」が接続助詞であるためには少なくとも判定条件1を満たさなければならない．

判定条件 2：「が」を接続助詞であると仮定すると，「が」の後に続く文字は，自立語の最初の文字である．ところが，促音や撥音で始まる自立語はないことから，「が」の1文字後が促音か撥音の場合，明らかにその「が」は接続助詞でない．したがって，この「が」は接続助詞の候補から外すことができる．

判定条件 3：文頭の「だが」に関しては明らかに接続助詞でない．

判定条件 4, 5：抽出法を構築する際に調査した67万文字文章で判定条件1を満たす誤りの中から，比較的出現頻度が高く，調査対象の文章に特有でない「一つが」「〜ほうが」を取り除くために設けた．

接続助詞「が」の1文字前が「つ」である場合，その「つ」はタ行五段活用動詞の活用語尾でなければならない．しかし，数字，漢数字は動詞の語幹には現れない．そのため，判定条件4を設けることができる．

接続助詞「が」の1文字前が「う」である場合，その「う」はワ行五段活用動詞の活用語尾または，助動詞「う，よう」の活用語尾のいずれかでなければならない．しかし，ワ行五段活用動詞には「ほ」+「う」で終わるものはない．さらに，助動詞「う」の接続は五段活用動詞，形容詞，形容動詞の未然形であり，それらの未然形は「ほ」で終わらない．したがって，「ほ」+助動詞「う」となることはない．また，助動詞「よう」は「う」の1文字前が明らかに「ほ」でない．以上のことから判定条件5を設けることができる．

表 1 字面のみによる接続助詞「が」の抽出法（字面抽出法）
Table 1 The textual analysis method to extract the conjunctive particle "*ga*".

判定条件1	「が」が接続助詞であるためには「が」の1文字前が「う，く，す，つ，ぬ，む，る，ぐ，ぶ，い，だ，た，ん」のいずれかでなければならない．
判定条件2	「が」の1文字後が促音，撥音である場合，その「が」は接続助詞でない．
判定条件3	文頭が「だが」であるとき，その「が」は接続助詞でない．
判定条件4	「が」の1文字前が「う」であるとき，その「う」の1文字前が「ほ」であれば，その「が」は接続助詞でない．
判定条件5	「が」の1文字前が「つ」であるとき，その「つ」の1文字前が数字または漢数字であれば，その「が」は接続助詞でない．

字面抽出法を構築する際に使用した文章とは別の文章を使用してこの字面抽出法の評価を行う．使用した文章は朝日新聞記事データ約1か月分（総文字数3,494,993文字：以下 **350万文字文章** と称す）である．その結果，上記判定条件をすべて満たす6,794個の候補中に接続助詞「が」が5,716個，格助詞「が」が869個，その他の「が」が209個含まれていた．この抽出法では接続助詞「が」の指摘漏れを起こさないため，再現率は100%である．また，適合率は84.1%となった．この結果を1万字の文章に当てはめると19.4個の接続助詞「が」の候補を抽出し，その中に3.1個の第二種の誤りを含む．1万字の文章に対して第二種の誤りが3個程度であるならば，一つ一つ吟味しながら推敲するのに邪魔にならないという理由で，『推敲』では表1に挙げた字面抽出法を採用している．

5. 格助詞「が」の抽出法の構築

5.1 文字「が」の分類と調査

文章中に現れる文字「が」は，格助詞，接続助詞，助詞以外に分類することができる．格助詞「が」は体言につく．つまり，格助詞「が」の前に来る文字は体言の末尾となりうる文字ならば何であってもよい．そのため文字「が」の前の文字情報を利用して，その「が」が格助詞であるかどうかを判定することは難しい．また文字「が」の後の文字に注目してみると，「が」が格助詞であれば，次の文字は単語の先頭の文字，また

は読点である．そのため文字「が」の後の文字でその「が」が格助詞であるかどうかを判定することも難しい．

文章中に現れる文字「が」で，格助詞以外の「が」を考えてみる．一つには，文字「が」を含む自立語がある．この自立語を公用データベース日本語辞書[7]で調べてみると約6,600個ある．さらに前述の接続助詞「が」がある．

まず，日本語文章中に出現する文字「が」を調査した．調査に使用した文章は前出の67万文字文章と350万文字文章である．この二つの文章を文字「が」をキーにKWIC表示させ，目視でその「が」を**格助詞**，**接続助詞**，**その他**に分類した．調査結果を**表2**，**3**に示す．これより，67万文字文章中に現れる文字「が」の9割が，また350万文字文章では8割が格助詞「が」であることがわかった．

この結果から，文章中から文字「が」を探し，その中から**その他**と**接続助詞**を除くことによって**格助詞「が」**を抽出することを考えた．

5.2 その他の「が」の除去

350万文字文章中にはその他の「が」が，4,908個ある．その内訳を出現頻度順に**表4**に示す．この結果から非助詞「が」除去法を設ける．

「ながら」の除去：350万文字文章では，文字列「ながら」として出現する文字「が」のうち助詞であるものは一つもなかった．そのため，この文字列が出現した場合それを助詞の候補から外すことにする．これにより「～しながら」のような表現を候補から外すことができる．「ながら」という文字列を候補から落とすことで，格助詞「が」の抽出の場合，第一種の誤りを犯す可能性がある．

促音，撥音の除去：字面抽出法（表1）の判定条件2と同じ理由により「が」の1文字後が促音か撥音の場合，この「が」は助詞の候補から外すことができる．

文頭の「が」，「だが」の除去：文頭の「が」は，接続詞「が」または，単語の先頭の「が」と考えられるので候補から外す．文頭にある「だが」は，すべてが接続詞「だが」であった．これも助詞の候補から外すことにする．

5.3 接続助詞「が」の除去

4章で述べたように字面抽出法（表1）では，第二種の誤りを犯すため，接続助詞「が」の候補の中に格助詞「が」も含まれている．この字面抽出法を用いて，文字「が」の集合から接続助詞「が」を除去しようとすると，格助詞「が」まで除去してしまうので，格助詞「が」の抽出法としては第一種の誤りを犯すことになる．このため，字面抽出法を再検討しなければならない．

6. 用言辞書，助動詞辞書の構築と使用

6.1 方　　針

現在の字面解析手法ではある特定の文字列を判定するためにその前後の1～3文字を見て判定を行っている．前後の判定すべき文字列を長くすれば解析精度は上がると考えられる．しかし，長くすればそれだけ文字に対する条件は多くなる．この文字列と文字に対する条件の集まりを辞書の形式で用意することを考える．

一般に，形態素解析や仮名漢字変換に用いられる機

表2 文字「が」の分類（67万文字文章）
Table 2 A number of occurrences of the character "*ga*" (about 670K characters).

品　詞	個　数	割合(%)
格助詞	6,249	89.6
接続助詞	401	5.7
その他	328	4.7
合　計	6,978	100.0

表3 文字「が」の分類（350万文字文章）
Table 3 A number of occurrences of the character "*ga*" (about 3.5M characters).

品　詞	個　数	割合(%)
格助詞	42,575	80.0
接続助詞	5,716	10.7
その他	4,908	9.2
合　計	53,199	100.0

表4 その他の「が」の内訳（350万文字文章）
Table 4 A list of "*ga*" which occurs as a non-particle.

分　類	個　数
「～ながら」	913
「が」の後が促音，撥音	741
「わが」	437
文頭の「だが」	365
「ところが」	212
「上がる」の活用形	195
文頭の「が」	150
その他	1,895
合　計	4,908

械可読辞書は，その内容のほとんどが，名詞（サ変名詞も含む）である．形態素解析は文章全体を解析するためにすべての品詞に対して辞書を持っていなければならない[8]．しかし，名詞は，実世界でどんどん増えていく．そのため，辞書に含まれない単語が多く存在する．一方，用言はあまり増えないと考えられる．これより，ある抽出法を構築する際に，“名詞に接続する”といった条件ではなく，“用言に接続する”という条件を作ることができれば，用言のみを要素として持つ辞書で十分解析は可能である．用言のみを辞書として持つことにすれば，形態素解析用辞書と比べて辞書の大きさをかなり小さくできると考えた．辞書の大きさを小さくすることができれば，主記憶上に辞書全体を載せることができ，高速な辞書アクセスが可能となる．

実際に文章の解析を行う場合には，助動詞の処理も必要となることがある．助動詞は用言ではないが，用言と同様に活用し，その活用型も非常に似ている．したがって，助動詞の処理も必要である．

6.2 辞書の構築

まず，公用データベース日本語単語辞書[7]から用言のみを選び出した．ここで，形容動詞は名詞に断定の助動詞「だ」が付いたものと活用が同じである．そのため形容動詞は用言辞書に含めないことにした．また，「勉強する」などはサ変名詞にサ変動詞「する」が付いたものと考えることができるので，これも用言辞書に含めないことにした．サ変動詞「する」とカ変動詞「くる」とは語幹がない．また，活用が不規則であるので，特例として後述の助動詞辞書に含めた．さらに，公用データベース日本語単語辞書の見出し語には複合語も含まれている．複合語とは，単語のうちでさらに造語成分，接辞などに分割できるものである．複合語の場合は，その基本となる部分（「食い違う」ならば「違う」）だけを要素とし，基本となる部分が同じ複合語は一つにまとめた．用言辞書の見出し語数は，4,120 語である．品詞ごとの見出し語数を**表 5** に示す．

助動詞同士が接続する場合には，ある決まった順にしか接続しない．この処理を文章解析時に行うと処理が繁雑になり，『推敲』の開発方針（2）を満たせないと考えた．そこで，いくつかの助動詞同士が接続した辞書をあらかじめ用意しておく．助動詞辞書の構築には新世代コンピュータ開発機構（ICOT）のフリーソフトウエア集に収められている形態素解析用辞書を用いた．この辞書と辞書付属の品詞接続テーブルを用いて助動詞辞書を構築した．助動詞辞書の見出し語数は 933 個になった．

プロトタイプは UNIX 上で構築した．辞書は各語の語幹を可変長のレコードとして持ち，リスト状につなげたものとした．また，用言辞書に関しては，それぞれの品詞の最初のエントリの位置を別途格納し，特定の品詞の語幹のみを検索できるようにした．用言辞書の大きさは，約 29 Kbytes，助動詞辞書の大きさは約 8 Kbytes である．辞書の大きさは十分小さいので，パソコン版『推敲』が実現されている MS-DOS 上でも主記憶に載せることができる．

6.3 辞書の使用

辞書を用いた抽出法の処理の流れを**図 1** に示す．文章中に特定の文字列（接続助詞「が」の抽出ならば文字「が」）が現われた場合，まず字面解析部で字面解析を行う．候補でないと判定した場合，それが結果とな

表 5 構築した用言辞書の見出し語数
Table 5 Classification of words which has the constructed dictionary.

品詞	見出し語数
ワ行五段動詞	223
カ行五段動詞	324
ガ行五段動詞	88
サ行五段動詞	468
タ行五段動詞	84
ナ行五段動詞	2
バ行五段動詞	44
マ行五段動詞	256
ラ行五段動詞	778
上一段動詞	177
下一段動詞	945
形容詞	731
総計	4,120

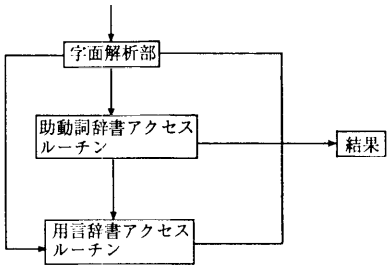


図 1 辞書を用いた抽出法の処理の流れ
Fig. 1 The flow of the method using the dictionary.

る．字面解析手法を用いた場合，どのような文字列が現われると抽出精度が悪くなるかを調べておく．抽出精度を悪くする文字列が現われた場合にのみ助動詞辞書，用言辞書アクセスルーチンに処理をわたす．助動詞辞書アクセスルーチンでは2文字以上の見出し語と照合した場合は候補となる．1文字の見出し語にしか照合しなかった場合と見出し語に照合しなかった場合は用言辞書アクセスルーチンに処理がわたる．用言辞書アクセスルーチンでは，用言辞書を品詞ごとに先頭から検索して，見出し語と照合した場合候補となる．

6.4 接続助詞「が」抽出法への適用

4章で調べた350万文字文章から字面抽出法によって抽出される接続助詞「が」の候補（6,794個）とその中に含まれる格助詞「が」（869個）との内訳を調査する．接続助詞「が」の候補をその1文字前の文字で分類した結果を**表6**に示す．表から，格助詞「が」を抽出してしまう誤り（第二種の誤り）は「が」の1文字前が「ん」の場合と，「い」の場合に集中していることがわかる．また，「く，つ」も出現数は少ないが誤りの割合が大きい．この結果から字面抽出法を適用する際に文字「が」の1文字前が「い」「ん」「く」「つ」の場合に辞書を援用する．

接続助詞「が」は用言，助動詞の終止形に接続する．そのため，「が」の1文字前が「い」である場合，その「が」が接続助詞であるためには，「い」は形容詞の終止形，および助動詞「まい」「たい」の終止形の末尾の「い」でなければならない．同様に「が」の1文字前が「ん」である場合，その「が」が接続助詞であるためには「ん」は助動詞「ん」でなければならない．また「く」，「つ」の場合は，それぞれカ行五段動詞，タ行五段動詞の終止形の最後の文字でなければならない．

表6 接続助詞「が」の候補の1文字前の文字
Table 6 Classification of characters before the candidates for conjunctive particle "*ga*".

文字	候補の数（個）	候補中の格助詞の数（個）	候補中の格助詞の割合（%）
た	2,075	26	1.3
る	1,494	5	0.3
だ	1,069	0	0
い	921	349	37.9
ん	436	367	84.2
す	412	9	2.2
う	238	26	10.9
く	76	55	72.4
つ	54	31	57.4
む	10	1	10.0
ぶ	5	0	0
ぬ	3	0	0
ぐ	1	0	0
合計	6,794	869	

まず，字面抽出法（表1）を適用する．その際，取り出された候補に対してさらに以下の条件を適用する．

条件：「が」の1文字前が「い」である場合，用言辞書，助動詞辞書を引いてそれが終止形の「い」であるかどうかを判定する．形容詞，形容詞活用変化をする助動詞の終止形と判定できる場合だけ，その「が」を接続助詞の候補に残す．

条件：「が」の1文字前が「ん」である場合，以前構築した否定表現抽出法[5),9)]を適用する．しかし否定表現抽出法も「ん」に関して精度が悪いため[9)]に，「ん」の1文字前が「か，が，さ，た」である場合，用言辞書を引きにいく．その結果「ん」が否定の助動詞であると判定できる場合だけ，その「が」を接続助詞の候補に残す．

条件：「が」の1文字前が「く，つ」である場合，用言辞書を引いてそれらがそれぞれカ行五段動詞，タ行五段動詞の終止形の活用語尾「く，つ」であるかどうかを判定する．終止形と判定できる場合だけ，その「が」を接続助詞の候補に残す．

この結果，接続助詞「が」の抽出法によって誤って抽出する格助詞「が」の数を869個から97個にまで減らすことができた．また，その他の「が」も209個から162個になった．したがって，接続助詞「が」の抽出法の適合率は95.7%となる．また，用言辞書，助動詞辞書を用いても再現率は100%である．

7. 格助詞「が」の抽出精度と抽出速度

5.1節で述べたように文中の文字「が」の中から5.2節の非助詞「が」除去法と6.4節の接続助詞「が」の抽出法とを利用して格助詞「が」を抽出した．その結果を**表7**に示す．ここで，第一種の誤りとは，抽出してこなかった格助詞「が」の数であり，第二種の誤りとは，候補に含まれてしまった格助詞以外の「が」の数である．この表から再現率を計算すると99.8%になる．また適合率は94.3%である．第一種の誤りはすべて接続助詞「が」の抽出で接続助詞の候補となってしまった格助詞である．その内訳を**表8**に示す．また，第二種の誤りはその他の「が」である．

別の朝日新聞記事データ（総文字数 1,981,950 文字，以下**200万文字文章**と称す）で格助詞「が」の抽出法の評価を行った．その結果を**表9**に示す．再現率は99.8%であり，適合率は94.4%であった．抽出法構築に使用した350万文字文章の場合と同様の結果を得た．この結果を1万字の文章に当てはめてみると，格助詞「が」が123.9個含まれる，そのうち0.3個を見落とす，7.3個を誤って拾ってくる，ということになる．

格助詞「が」の抽出法（第一種の誤りを含む）をPC-9801版『推敲』に試験的に組み込んで，検索速度を計測した．メニューで「格助詞「が」を複数含む文」を選択し候補が表示されるまでの時間は，PC-286 VF (CPU：80286/12 MHz) で，1万字の文章に対して，0.48秒であり，1章で述べた開発方針(2)を十分満たしている．

この抽出法を実際に『推敲』の機能として実現する場合，第一種の誤りを犯すので『推敲』の開発方針(1)を犯してしまう．そこで，以下の方法を考えてみる．

(a) 格助詞「が」抽出法では抽出洩れを犯すことをユーザに伝えるために，メッセージを表示する．
(b) 必要ならばユーザの支援を受けて抽出洩れを復元する．

(b)は，格助詞「が」抽出の第一種の誤りが接続助詞「が」抽出における第二種の誤りであるということから

step 1：接続助詞「が」の候補をユーザに示す．
step 2：ユーザは接続助詞「が」の候補から**目視**で格助詞「が」を指摘する．

という手順を踏む．step 2で指摘されたものが格助詞「が」の抽出における第一種の誤りである．1万字あたりの接続助詞「が」の候補数は17個程度，そのうち格助詞は0.3個程度なのでユーザの負担は小さいと考えられる．しかし，字数の多い文章の場合は問題が残る．

8. 格助詞「が」ととりたて詞「は」について

今回構築した格助詞「が」の抽出法（第一種の誤りを犯す）と，とりたて詞「は」の抽出法[6]とを使って200万文字文章から格助詞「が」ととりたて詞「は」の候補を複数含む文を抽出した．結果を**表10**に示す．このうち，「が」または「は」，あるいは「が」および「は」が1文中に複数含まれる文は，16,194文あり，全体の文の約35.8%である．『推敲』が実用規模としている文章（1万字程度）に当てはめると，総文数は約230文に対して，「が」または「は」，あるいは「が」および「は」を複数含む文は，約82文となる．1文中の「が」の数と「は」の数とを足したものが4個以上の文を抽出すれば，全体の文の約6.0%（1万字文章中約14文）となる．

以下に，200万文字文章から抽出したものの例を示す．下線つき文字は抽出法によって得られた候補である．各例の最後に実際の格助詞「が」，とりたて詞「は」の個数を示す．

例文 1：米国側には，しかし，その事実を承知のうえで，日本や欧州の輸出が伸びたのは，米国の強い

表7 格助詞「が」の抽出結果（350万文字文章）
Table 7 A result of extracting the nominative particle "*ga*" (about 3.5M characters).

分　類	個　数
格助詞「が」の候補	45,055
候補中の格助詞「が」	42,478
第一種の誤り	97
第二種の誤り	2,577
文章中の格助詞「が」	42,575

表8 格助詞「が」の抽出における第一種の誤り（350万文字文章）
Table 8 A list of the nominative particles "*ga*" which cannot be extracted (about 3.5M characters).

分　類	個　数
あなたが	21
～さんが	13
～ようがない	9
いすが	5
しかたが	3
けいれんが	3
向こうが	3
その他	40

表9 格助詞「が」の抽出結果（200万文字文章）
Table 9 A result of extracting the nominative particle "*ga*" (about 2M characters).

分　類	個　数
格助詞「が」の候補	25,952
候補中の格助詞「が」	24,505
第一種の誤り	55
第二種の誤り	1,447
文章中の格助詞「が」	24,564

表 10 格助詞「が」ととりたて詞「は」の候補を複数含む文
Table 10 A number of sentences which include some candidates of the nominative particle "*ga*" and/or the particle "*wa*".

は＼が	0	1	2	3	4	5以上	合 計	割 合
0	13,675	5,249	1,490	376	79	20	20,889	46.1
1	10,157	5,267	1,625	419	115	46	17,629	38.9
2	2,468	1,702	724	221	56	27	5,198	11.5
3	527	412	179	62	24	11	1,215	2.7
4	93	87	61	16	8	4	269	0.6
5以上	20	33	11	10	0	1	75	0.2
合 計	26,940	12,750	4,090	1,104	282	109	45,275	100.0
割 合	59.5	28.2	9.0	2.4	0.6	0.2	100.0	

個人消費と開放された市場があったためではないか，そちらこそ内需拡大，輸入増，市場開放をもっと進めるべきではないか，という考え方が根強い．(「は」4個，「が」3個)

例文 2：米国では以前から不当表示規制に関する連邦規則によって注射剤，非経口剤については添加剤の内容表示が義務づけられていたが，全成分の表示を要求するさまざまな働きかけが行われた結果，1985年末からは製薬団体自身の手で全成分表示のガイドラインが作られ，今日では原則としてすべての医療用医薬品の添加剤が表示されるようになった．(「は」4個，「が」4個)

例文 3：民間企業の規模別では，従業員 1000 人以上の大企業の組織率が前年より 2.4 ポイント上がって 68% になったのに対し，100 人以上 1000 人未満は 27.4%，30 人以上 100 人未満は 6.2%，30 人未満は 0.4% で，企業規模が大きいほど労組組織率が高い傾向がさらに進んだ．(「は」4個，「が」4個)

9. ま と め

実際の文章に現れる文字「が」を調査し，その「が」の 8～9 割が格助詞であることがわかった．このことから，格助詞以外の文字「が」を候補から外すことにより格助詞「が」の抽出法を構築した．『推敲』で用いている他の抽出法と比較すると幾分抽出精度が悪い．これは，その他の「が」を有効に候補から落とすことができなかったためである．また，今回の抽出法では第一種の誤りをわずかながら犯してしまう．この第一種の誤りをユーザの支援を受けて除去する方法も考察した．

また，今回構築した用言辞書，助動詞辞書を他の文字列抽出にも利用する予定である．

謝辞 朝日新聞ニューメディア本部には，新聞記事データの使用を許していただいた．ここに記して謝意を表する．

なお，本研究の一部は文部省科学研究費補助金試験研究(B)(2)課題番号 01880008 "日本語文章推敲支援ツールの機能拡張と移植および利用分野の拡大に関する研究" および奨励研究(A)課題番号 04780037 "字面解析を応用した日本語文章の推敲支援法の研究" の補助を受けた．

参 考 文 献

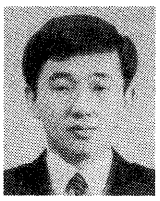
1) 倉田昌典，菅沼 明，牛島和夫：日本語文章推敲支援ツール『推敲』のパソコン上での実用化，コンピュータソフトウェア，Vol. 6, No. 4, pp. 55-67 (1989).
2) 牛島和夫，菅沼 明：日本語文章推敲支援ツール『推敲』(Version 1.9) 使用手引書，九州大学工学部情報工学科計算機ソフトウェア研究室 (1992).
3) 菅沼 明，石田朗子，倉田昌典，牛島和夫：日本語文章推敲支援ツール『推敲』における字面解析手法とその評価，情報処理学会自然言語処理研究会報告，68-8 (1988).
4) 清水幾太郎：論文の書き方，岩波新書 (1959).
5) 菅沼 明，倉田昌典，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).
6) 菅沼 明，牛島和夫：日本語文章推敲支援ツール『推敲』におけるとりたて詞「は」の抽出法とその評価，情報処理学会論文誌，Vol. 32, No. 11, pp. 1392-1400 (1991).
7) 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).

8) 稲永紘之，平井誠人，吉田　将：仮名文字文処理のための機械辞書システム，信学技報，NLC 92-11 (1992).
9) 下園幸一，菅沼　明，牛島和夫：字面解析手法を用いた否定表現抽出法の評価—朝日新聞記事データへの適用—，第 44 回情報処理学会全国大会論文集，5C-4 (1992).

(平成5年6月18日受付)
(平成6年4月21日採録)

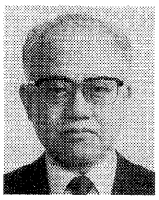
下園　幸一（正会員）

1968 年生．1991 年九州大学工学部情報工学科卒業．1993 年同大学院情報工学専攻修士課程修了．同年より九州大学情報処理教育センターに助手として勤務．現在に至る．日本語テキスト処理，計算機ネットワークに興味を持つ．



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 年同大学院工学研究科情報工学専攻修士課程修了．1991 年同博士後期課程修了．同年九州大学工学部情報工学科助手勤務．1993 年同大学工学部情報工学科講師，現在に至る．工学博士．日本語処理，ユーザインタフェース，ニューラルネットワークの応用などに興味を持つ．1994 年情報処理学会奨励賞受賞．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（数理工学コース）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 1994 年3月九州大学大型計算機センター長を兼務．1991 年度情報処理学会九州支部長．工学博士．ソフトウェア科学会，電子情報通信学会，ACM 各会員．