字面解析による日本語動詞抽出手法

中挾 知延子† 島田 静雄†

我々は，字面解析による文章中からの日本語動詞の抽出のための手法を開発した．そのための動詞抽出ツール『文彩（ぶんさい）』による抽出の結果を示し，これらを考察する．日本語の動詞を，漢字・ひらがなで書かれている動詞（以下，それぞれ漢字動詞・ひらがな動詞とする）に分類した．日本語動詞については，広辞苑[1]を参考にし，標準的な日本語動詞を含めた動詞辞書を作成した．また，この手法において，動詞として抽出される語句の中で，文中での役割としては，補助的なものや英語での前置詞句・助動詞・副詞句に対応するものも見つかる．そのため，動詞を単独で抽出するのではなく，文中での役割に応じて意味的なまとまりを持った形で抽出するようにした．『文彩』を用いて論文文章を対象に抽出実験を行った結果，99.7%の適合率で抽出できた．一方，動詞辞書に登録していない複合動詞が出現した場合には，ユーザに対して判定を求めるが，これを含めると，100%の再現率になる．なお，『文彩』は，システム全体でも1MB弱で，パソコン上で実現している．

An Extraction Method of Japanese Verbs Using Textual Analysis

CHIEKO NAKABASAMI† and SHIZUO SHIMADA†

We developed a new extraction method of Japanese verbs in machine-readable Japanese documents only using textual analyses, and implemented it as an extraction tool, called "BUNSAI". This paper describes the results of the extraction tests. First, we classified Japanese verbs into two categories, the one which is written with a Chinese character KANJI, and the other, is a KANA, Japanese phonetic character. Under this categories, we compiled a dictionary of standard Japanese verbs which are basically included in "KOJIEN [1]". With the extraction method we can detect and extract each verb's role and function with verb phrases at every possible occurrence. When an unknown compound verb appears in a document, "BUNSAI" can ask a user whether it is a verb or other articles.

1. はじめに

我々は，日本語文章中から動詞を抽出して，動詞辞書を作り，それを利用して字面解析のみで日本語の文章解析をすすめている[2),3)]．このための動詞抽出のツールとして『文彩』を開発した．日本語処理のための電子化辞書は，形態素解析用の辞書をはじめとして多数開発されているが，動詞に限定した辞書は今までに開発されていない．また，簡易な字面処理によって日本語の助詞的定型表現の抽出の研究がある[4)]が，文章中からの動詞の抽出については行われていなかった．本手法は動詞を抽出することが目的であり，形態素解析で行われるような，文章を形態素に分解し分かち書きをするという解析ではない．形態素解析が文章を言語として解析するのに対して，本手法は文章を文字という記号の集合としてとらえており，その見方からの文章解析を「字面解析」としている．よって，動詞を抽出する際に隣接する語句の意味的関係を無視し，動詞で用いる文字記号列として文字を一致させることによって，小規模な辞書で抽出処理を実現している．

日本語処理に用いる辞書には，キーワード抽出を目的として作成される名詞辞書等がある[5)]．コンピュータが，名詞をどのくらい知っていれば満足のいく処理ができるかについては，人間の大人の平均的な語彙数を参考にすると，4万語くらいになる[6)]．それに対して，動詞をコンピュータに教えておく場合，名詞のように数多く知っていればそれだけ有利である必要はなく，動詞辞書は名詞辞書とは別枠で用意するべきである．日本語の動詞は名詞に比べて，活用に関して明確なルールがある．また動詞に用いられる漢字も漢字全体から見れば数が少ない．我々はこれらの事実から，文章中からの語句の抽出を動詞に限った場合，日本語文章で通常用いられる動詞の語幹を辞書として用意し，それに活用変化のルール等を持たせておけば十分であると考えた．そのためには，日本語の動詞の分類と，

† 埼玉大学工学部情報システム工学科
Department of Information and Computer Sciences, Saitama University

動詞辞書の構成・利用方法が重要になる．

そこで我々は，日本語の動詞を，漢字・ひらがなで書かれている動詞（以下，それぞれ漢字動詞・ひらがな動詞とする）に分類した．漢字動詞については，JIS第1水準の漢字動詞のすべてと，第2水準の漢字については動詞として用いられる漢字で文章中での出現頻度の多いものを採用し[7]，広辞苑[1]で確認をした．漢字動詞は，漢字1字を用いるもの・2字熟語で「～する」と使うもの・複合動詞に分類した．また，ひらがな動詞については，文章中で通常はひらがなで書かれるものとした．ひらがな動詞は，2字の基本動詞・習慣的にひらがなで書かれるものに分類される．習慣的に書かれるものについては，出現頻度の高いものにしぼり，88個採用した（**表1**）．我々が採用したこれらの漢字・ひらがな動詞を合わせて，日本語文章中の動詞は網羅できると考えている．作者の勝手に造った動詞や，流行として使われている，元来動詞ではない俗語は除いた．また，採用していない複合動詞が出現した場合は，ユーザに対して動詞であるかどうかを質問して，決定をすることで補足している．漢字動詞に関しては，例外を除いてほぼ100％文章中から抽出できる結果を得ている[3]．

表1 ひらがな動詞一覧表
Table 1 List of verbs written in phonetic characters.

登録したひらがな動詞（終止形）
あう・ある・あえる・あがる・あげる・あたる・あてる・あいまる
いう・いく・いる・いえる・いける・いだす・いただく
うる
える
おく・おる・おける・おれる・おきる
かかる・かまう・かかれる・かまえる・かかれる
きる
くる・くれる・くださる
さす・させる・される
しる・しまう・しれる・したがう・したがえる・しまえる
すむ・する
たまる
だす・だせる
ちなむ
つる・つく・つくる・つける・つめる・つれる・つもる・つくれる・つもれる
づく
でる・できる
とる・とれる・とどまる・ともなう
なす・なる・なせる・なれる
はさむ・はさまる・はさめる
みる・みえる・みせる
めく・めぐる
もつ・もてる・もらう・もたせる・もらえる
やむ・やる・やめる・やれる
よる・よれる
わかる・わたる・わたれる

漢字動詞は，後ろに助詞等の付属語をともなって動詞句となり，述語としての役割を持つほかに，「～に関して」のように英語の前置詞句にあたるものの2つの主たる用法がある．ひらがな動詞は，それに加えてさらにさまざまな出現形態がある．我々は，ひらがな動詞が文章中に出現するときの特徴を以下のようにとらえた．

(1) 漢字動詞の後ろに付いて補助動詞になる．
(2) 名詞の後ろに付いて接尾辞の働きをする．
(3) 文章中に出現するひらがな動詞は，文字数が少なく，多くの意味を持つ基本動詞が多い[8]．
(4) ひらがな動詞を複数つなぎ合わせて，文末を長くできる．
(5) 周辺の語句とともに用いられる句を英語に対応させると，動詞に対応するもののほかに，前置詞句・助動詞・副詞句等に対応するものがある．

これらの特徴を，動詞を抽出する場合に適用すると，単独で抽出するよりも，動詞句として抽出すれば，以下の利点が得られる．

- 機械翻訳をする場合に日本語動詞を，翻訳される言語の動詞に対応するもの・助動詞・前置詞句・慣用句に対応するものにあらかじめ区別しておける．
- 日本語文章は文の終わりまで読まないと作者の意図が判断しにくく，文末表現は表現のしかたで意味が微妙に異なる．文末には動詞句が出現することが多く，動詞句は日本語文章の意味を決定する重要な部分といえる．機械翻訳を，学術論文・新聞の解説記事・マニュアル等の技術文書に用いる場合，複数の解釈がなされる箇所があってはならない．そのため機械翻訳の前処理として，冗長な表現や意味の曖昧な表現を明確にしたり，後処理として用語の統一や翻訳調の表現を日本語らしい表現に校正する必要がある．そのような校正すべき箇所は文末表現にも多く見られる．現状では人手によってそれらの作業が行われており，かなりの手間と時間を要する．本手法を用いて，文章中の動詞句を抽出して，校正すべき候補を人間に提示できれば，それらの作業の支援になる．

我々は，これらの利点を生かすために，動詞句として抽出することにした．このとき動詞句の動詞以外の語句は周辺語辞書に登録し，動詞句の生成に利用した．動詞句を抽出するメカニズムを提供することは，他の用言である形容詞句・形容動詞句の場合にも応用できる．なぜならば，用言は複雑な文末表現を作り出すが，語構成の規則は動詞が一番複雑であり，動詞句

抽出のメカニズムが確立できれば，他の用言にも容易に移行できるからである．また，我々は動詞の抽出と並行して名詞の抽出に関する研究も行っており，日本語文章で重要な骨格をなす名詞・動詞の抽出ができれば，文章の解析において役に立つと考える．抽出は簡易な字面処理で行い，パソコン上で実現している．抽出に使った辞書・ツールをすべて合わせても，1 MB弱の大きさである．

2章では動詞の役割と我々が設定した動詞句の構成を述べる．3章では動詞辞書と，動詞の前後に出現しやすい助詞・接尾辞等の語句を登録した周辺語辞書の内容を述べる．そして4章では，動詞句を文章中から抽出するツール『文彩（ぶんさい）』について述べ，5章にその抽出結果を示す．6章には5章の結果の考察を述べ，7章にまとめと今後の課題を述べる．

2. 動詞の役割と動詞句の構成

2.1 文章中での動詞の役割

動詞句は文章中で述部の役割をするほかに，以下の役割がある．

(1) 補助動詞の役割
ひらがな動詞は，漢字動詞の後ろに助詞をともなって出現し，補助動詞として漢字動詞に補助的な意味を添える．たとえば，「考えている」は，漢字動詞「考え」に助詞「て」とひらがな動詞「いる」で構成されている．この場合，「考え」の部分が文章中の動詞として主たる意味を持ち，「ている」は進行中であるという意味を「考え」に付加している．また，漢字動詞にも，「～かも知れない」・「～をなし得る」のように補助的に意味を添える使い方がある．

(2) 文末の表現を形成
動詞は文末に多く出現し，最後まで文章を読まないと文全体の意味がつかめない．また，婉曲な言い回しのためにも，文末の表現が冗長になる．特に，ひらがな動詞は，2～3 文字程度の単純な基本動詞が多く，文末表現に多く見られる．たとえば，
- ～が考えられることになるかもしれない
- ～においてもされることがある

等である．複数個が同時に出現し，その前後に付属語や「こと」等の抽象名詞をともなう．また，付属語も複数個が同時に出現してひらがな動詞に付く．上の例は，
- 「こと」＋「に」＋「なる」＋「かも」
（ひらがな動詞「なる」の連体形の周辺の構造）
- 「かも」＋「しれ」＋「ない」
（ひらがな動詞「しる」の仮定形の周辺の構造）
- 「に」＋「おい」＋「ても」
（ひらがな動詞「おく」の連用形が音便変化したものの周辺の構造）
- 「ても」＋「される」＋「こと」＋「が」
（ひらがな動詞「する」が使役の助動詞をともなった形の周辺の構造）
- 「が」＋「ある」
（ひらがな動詞「ある」の終止形の周辺の構造）

のようになっている．

(3) 前置詞句・助動詞・副詞句の働き
動詞は単独で出現して動詞の役割をする以外に，周辺語句と結び付いて英語の前置詞句・助動詞・副詞句の役割をする．たとえば，
- 「に関して」→「about」（前置詞句）
- 「として」→「as」（前置詞句）
- 「～かもしれない」→「may」（助動詞）
- 「～にもかかわらず」→「in spite of」（副詞句）

等がある．

2.2 抽出のための語句の構成

2.1をもとに，動詞を抽出するための構成要素を以下のように設定した．

(1) 動詞の前に出現する語句群
終助詞を除く助詞と「もの」・「こと」・「よう」等の抽象名詞．

(2) 動詞部
動詞辞書に登録されている，活用形も含めた漢字動詞・ひらがな動詞．

(3) 動詞の後に出現する語句群
動詞に後続可能な助動詞・助詞・抽象名詞．(1)と異なる点は，終助詞・助動詞・形容詞の「よい」・「ない」とそれらの活用形も含む．

構成の内容を**図1**に示す．また，(1)と(3)に含まれる語句については，9)，10)に載っている日本語表現の例を参考にした．

3. 動詞辞書と周辺語辞書

動詞辞書と，前節で述べた構成に沿って，必要となる語句を登録した周辺語辞書を作成した．周辺語辞書には該当する周辺語句を見出しとして登録した．動詞辞書に関しては，見出しといくつかの項目を設定した．

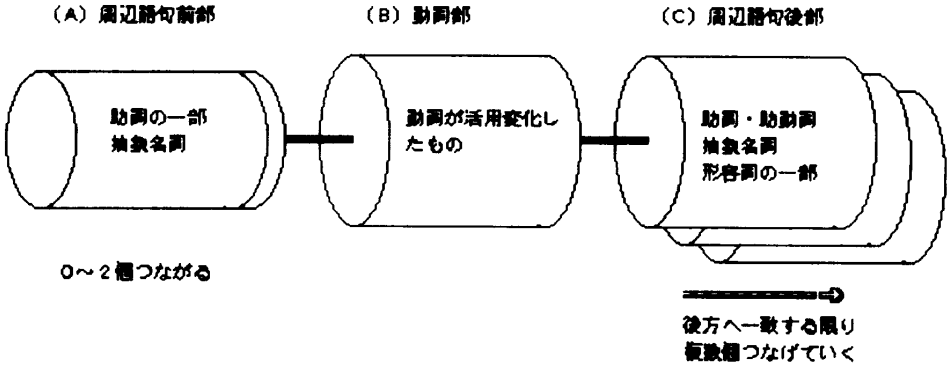

図1 動詞句の構造
Fig. 1 The structure of phrasal verbs.

以下に項目の内容を述べる.

(1) 動詞辞書

- 見出し語
  漢字動詞は，動詞に用いられる漢字であるJIS第1水準1450字・第2水準81字の合計1531字である．ひらがな動詞は，88語の見出し語が登録してある．
- 読み
  見出し語が漢字の場合，動詞で用いるときの読みを登録した.「する」動詞になるときの読みも含む.
- 送りがな
  動詞の語幹の送りがなの部分を登録した.
- 活用の種類
  動詞の活用の種類を登録した．未然形から命令形までに加えて，受身形・使役形・「～て」をともなうときの音便形を含む．また，5段活用の動詞に関しては，下1段活用になったとき可能の意味になるので，それらも含めた．たとえば,「書く」と同時に「書ける」・「ある」と同時に「あえる」も抽出できる．また，自動詞・他動詞の区別も登録した.
- 後続可能語句
  「つめる」というひらがな動詞ならば「込む」という漢字動詞を後ろに付けて複合動詞「つめ込む」を作る．このように，後続して複合動詞になりうる漢字動詞の漢字部分を登録した．また，漢字2字で「する」動詞になるとき，2字目の漢字を登録した.

図2，**図3** に動詞辞書の実際の画面を示す．図2では「集」で始まる漢字動詞が，図3では「も」で始まるひらがな動詞が登録されている.

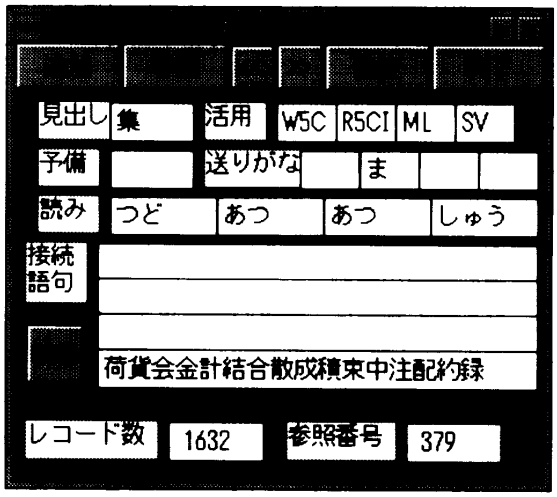

図2 動詞辞書画面（漢字動詞）
Fig. 2 Dictionary of verbs on screen (KANJI).

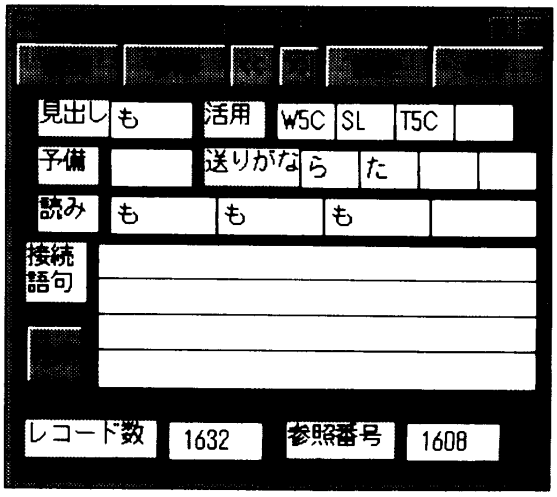

図3 動詞辞書画面（ひらがな動詞）
Fig. 3 Dictionary of verbs on screen (KANA).

(2) 周辺語辞書

動詞の前後に接続可能な助詞・助動詞・抽象名

表2 周辺語句一覧表
Table 2 List of words attached to a verb.

助詞	が・の・に・を・へ・と・で・や・ば・て・は・も・か・つつ・より・まで・から・こそ・さえ・とも・ども・だけ・のみ・しか・なの・ながら・ばかり
助動詞	う・ず・た・だ・る・たい・たら・たり・まい・まう・よう・られ・ろう・べし・べく・べき・べから・ような
抽象名詞	こと・うち・わけ・とき・ため・よう・最中・必要・以上・以下・過言・ところ
接尾辞	すぎ・ふう・です・ます・でし・まし・まま
形容詞	ない・なか・なく・いい・よい・よか・よく・いない・なけれ・よけれ・なかっ・よかっ・やすい・やすく

詞・形容詞を登録した.

表1に登録したひらがな動詞の一覧を，表2に周辺語句の一覧を示す.

4. 動詞抽出ツール「文彩」

動詞を抽出するツール『文彩』を開発した.『文彩』は図1で示した動詞句を文章中から字面のみで一致させて抽出する．図4に抽出のしくみを示す.

また，図5に『文彩』の動詞抽出画面を示す．図4のように，動詞抽出処理は大きく分けて，

- 動詞知識ベース
- 動詞抽出エンジン
- 動詞辞書・周辺語辞書

の3つの部分から成る．3番目の辞書については3章で述べてあるので，それ以外の項目について説明する.

4.1 動詞知識ベース

動詞知識ベースは「動詞活用ルール」が主体となっており，我々は「動詞活用ルール」を従来の形態素解析で用いられているルールに比べ，より簡素化した．ここで簡素化したというのは，活用ルールそのもののアルゴリズムを高速にしたのではなく，活用ルールを適用する語句を絞り込んだ点と，複合動詞の抽出の際に活用ルールを適用するに至るまでを自動化したという意味での簡素化であり，以下に詳細を述べる.

(1) 形態素解析では，読み込んだ文字に対して最初から品詞を絞り込まずに辞書と照らし合わせて可能な場合をすべて列挙していく．それに対して本手法は動詞の抽出が目的であるため，辞書には見出し語として，動詞に用いられる語だけを登録しているので，動詞でない語句であれば見出し語の検索の段階で排除され，動詞でない語に活用ルールを用いてむだな一致を試みる時間が省かれる.

(2) 複合動詞は形態素解析用の辞書には，出現頻度の高い動詞が登録されていて，読み込んだ文字が複合動詞の先頭の動詞と一致した場合，その動詞から始まる複合動詞の見出し全部に対して検索をしなければならない．それに対して本手法では，先頭の動詞が連用形であると判断すると，再帰的に動詞の抽出処理を行い，一致する限り3語以上の複合動詞の抽出もできる．そのため辞書に複合動詞を明記しておかなくても，自動的に複合動詞を生成して，活用ルールを最後尾の動詞に対して用いて文中の語句との一致を試みる.

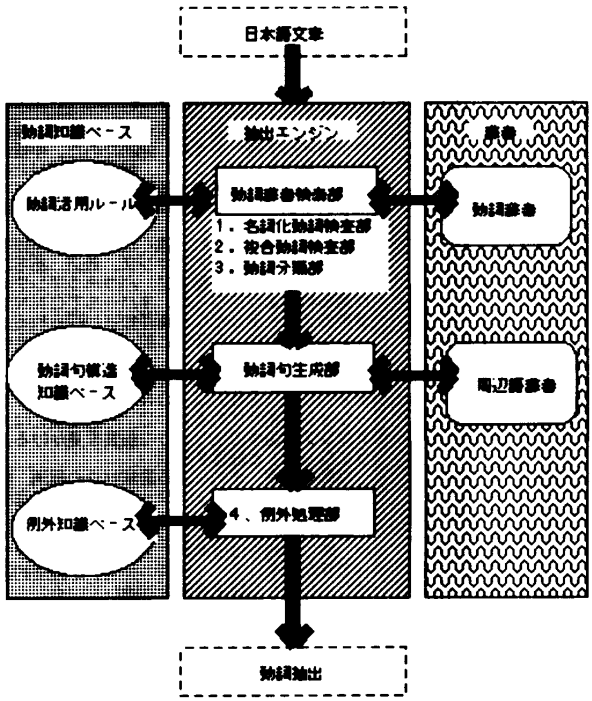

図4 『文彩』における動詞抽出のしくみ
Fig. 4 Scheme of "BUNSAI" system.

4.2 動詞抽出エンジン

『文彩』の抽出エンジンは，日本語文章の入力を先頭から1文字ずつ読み込み，動詞辞書と動詞活用ルールを用いて文章中の動詞を単独で抽出する．そのときに図4で示した1から3の検査をする．以下にその説明を述べる.

4.2.1 名詞化動詞検査部

抽出した動詞で連用形の中に，名詞として使われる場合がある．たとえば,「試み」は「試みる」の連用形であるが，名詞として使われることがある.『文彩』はこれらの名詞を，動詞として抽出しないために，動詞の連用形の直後に出現する語句によって検査をする．直後の語句として名詞に続く助詞・名詞に続く接尾辞が出現していると，動詞としてそれらを抽出しない.

4.2.2 複合動詞検査部

日本語の動詞は連用形に他の動詞をともなって，容

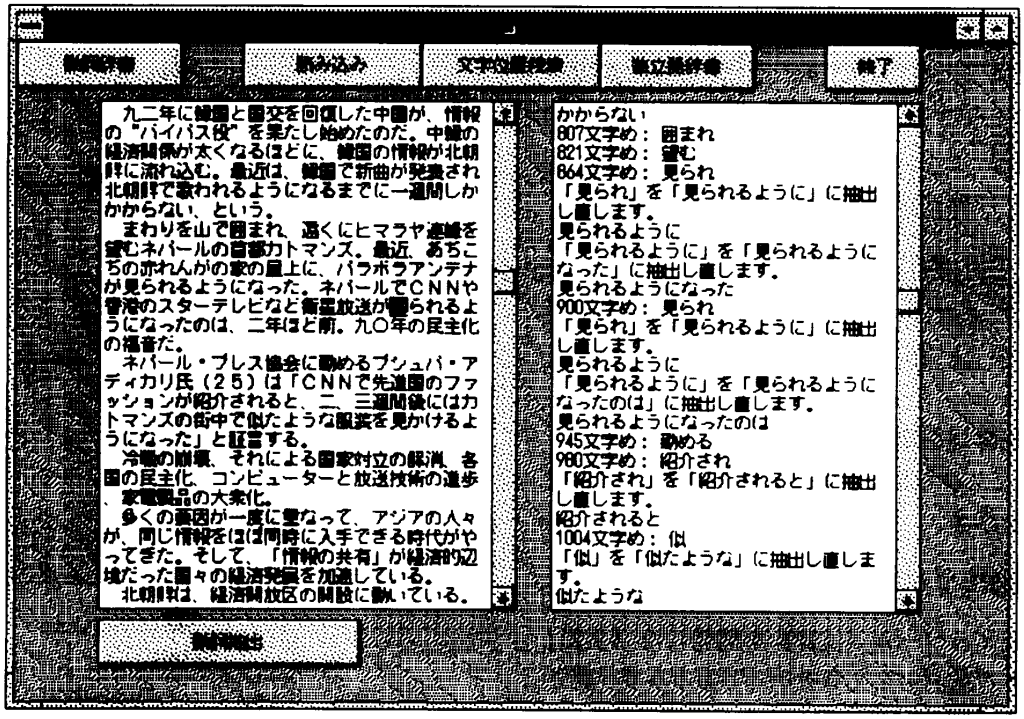

図5 「文彩」における動詞抽出画面
Fig. 5 Interactive window of "BUNSAI".

易に複合動詞になりうる．文章中には作者自身が勝手に造り出した複合動詞も見られる．動詞辞書の「後続可能語句」項目を利用することで，一般によく使われる複合動詞は抽出できるようにしてあるが，登録語句の数も限界があり，すべての複合動詞を生成することはできない．そこで，登録されていない複合動詞は，ユーザとの対話により決定する．もしも，動詞の連用形の直後に別の動詞があれば，いったん複合動詞として生成し，ユーザに対して生成した複合動詞が動詞であるのかを問い正すようにした．抽出できる組み合わせには以下のものがある．

- 「ひらがな」＋「ひらがな」
  （例）「できあがる」·「とりもつ」
- 「ひらがな」＋「漢字」
  （例）「し直す」·「めぐり来る」
- 「漢字」＋「ひらがな」
  （例）「見つける」·「考えだす」
- 「漢字」＋「漢字」
  （例）「見破る」·「拡大し続ける」

4.2.3 動詞分類部

抽出された動詞を以下のように分類する．

(1) 補助動詞でなく主動詞として使われているもの
(2) 補助動詞として使われているもの
(3) 動詞としてよりも，英語の前置詞句・助動詞・副詞句として使われているもの

これら3つに分けた後，次処理の「動詞句生成部」に進むわけである．そのとき，(1)であれば，図1の(B)＋(C)を抽出し，(2)と(3)は(A)＋(B)＋(C)を抽出する．

動詞句を生成した後，それらに対して例外処理を行う．4の例外処理部は動詞が連体詞・接続詞・副詞の一部と一致する場合を省いた．

『文彩』での抽出にあたって設定した制約と留意点を以下に述べる．これらは「動詞句生成部」において用いられている．

図1で示した（A）の部分は，動詞の前に配置できる語句を組み合わせることで抽出が行われるが，特定の動詞に対して組み合わせが許されるための制約を設けた．

制約1 ひらがな動詞が主動詞となる場合

ひらがな動詞が主動詞となるときに配置される周辺語句は除いた．「いる」を例にとると，「太郎がいる」·「お金がいる」·「豆をいる」のように，前に「が」·「を」をともなうときには，補助動詞ではなく主動詞であり，それぞれ「居る」·「要る」·「煎る」という漢字で書ける．このような場合は周辺語句を含めて「いる」を抽出せず，単独に抽出する．

制約2 前に配置できない周辺語句の排除

それぞれの動詞に対して，前に出現しない周辺語

句は除いた．表1で示した動詞に対して，「ちなむ」を例にとると，「～にちなんで」のように，前に「に」をともなって出現はするが，「～でちなんで」の形では出現しない．そのため，「ちなむ」が出現するときには，前に配置する周辺語句として「で」は排除した．

制約3 動詞活用形制約

「つく」を例にとると，前に位置するの語句として「に」を設定したとき，「～について」・「～につき」があり，それぞれ英語の前置詞句になる．一方，「～につく」のように，か行5段動詞「つく」の終止形「つく」は，「に」を前にともなうときには主動詞になる．このように，活用形によって前に周辺語句をともなって意味的なまとまりを持つときと，そうならずに単独で主動詞のはたらきをするときがある．そのため，それぞれの周辺語句に対してひらがな動詞が後続して意味的にまとまるときの活用形を限定した．

留意点 文末表現の抽出

動詞が文末等に補助動詞として複数同時に出現している場合には，別々に抽出せずに，すべてを結合して抽出する．たとえば，「それらを～が考えられることになるかもしれない」が文章中に出現した場合には，「考えられることになるかも」と「かもしれない」のようには抽出せず，「考えられることになるかもしれない」のように抽出する．

5. 抽出結果

『文彩』を用いて，実験データとして科学技術論文・新聞の解説記事の5万文字文章中から動詞の抽出実験を行った．実験に用いたマシンはDX4/100 MHzのDOS/V機で，主記憶装置の大きさは12 MB，抽出に用いた辞書の大きさは約850 KBで辞書はハードディスク上において実験した．抽出にかかった時間は，ユーザとの対話を含めなければ原稿用紙1枚（400字）あたり約3.11秒，対話を含めれば約3.68秒であった．**表3**に実験での詳しい条件と再現率と適合率を，**表4**～**表6**に文章中から抽出した文末表現・動詞句として抽出したときの付属語の部分・英語の前置詞句・助動詞・副詞句にあたるものの一覧を示す．

6. 実験の考察

表3の再現率について，すべての動詞を抽出してはいるが，その中の2.24％は処理を一時中断してユーザに判定してもらっている．『文彩』の我々の字面処理による抽出方法では，作者が勝手に造った複合動詞を見分けられない．動詞の抽出をすべて自動でするために，対象となる分野特有の複合動詞辞書を作成する必要がある．またその方法に加えて，例外として，『文彩』の側でも動詞の前後に接続して複合動詞を造りやすい動詞（例：「出る」・「続ける」）を知識として持ち，それらが複合動詞の一部として含まれている場合には，自動的に抽出することも考えている．

また，今回は論文文章を対象に抽出を行ったため，述語は漢字で書かれている場合が多く，ひらがな動詞

表3 抽出実験結果
Table 3 The results of extraction tests.

全文字数（文字）	58,626
(A) 抽出すべき動詞句の語数（単語）	3,388
(B) 抽出した動詞句の語数（単語）	3,398
(C) 抽出できた動詞句の語数（単語）	3,388
(D) (A)の中でユーザとの対話で排除した語数（単語）	17
(E) (B)の中でユーザとの対話で補足した語数（単語）	76
(F) 再現率 ((A)/(C)) (%)	100.00
(G) 適合率 ((A)/(B)) (%)	99.70
(F) においてユーザとの対話で補足した語数の割合 ((E)/(C)) (%)	2.24
(G) においてユーザとの対話で排除した語数の割合 ((D)/(B)) (%)	0.50

表4 抽出した文末表現
Table 4 List of verbal expressions.

意味	文末表現
断定	～ことがある・～ことである・～ものがある・～ものがあった・～ものである・～ものであった・～ことである・～ことによる・～である・～であった・～のであった・～のである・～となる・～がある・～である・～がいる・～とされる・～とする・～としている・～にしている・～ことになる・～と言うことになる・～ているともい（言）える・～であったりする・～ということになる・～ということである・～ものに過ぎない・～てくる・～たものであるということである・～があることである・～のであった・～と言われている
推量	～とい（言）えるのかもし（知）れない・～とい（言）えるであろう・～とい（言）ってよいであろう・～ことになる・～であるとい（言）うこともできる・～てよいだろう・～ないであろう・～できよう・～があったと思われる・～があると思われる・～ように思える・～であろう・～なのであろう・～おくべきであろう・～てよい・～になっていく・～になっているということでもある
義務	～必要である・～必要がある・～ことが必要ではないか・～べきである・～必要があると言ってよい・～てはならない・～なければならない・～ねばならない
継続	～ている
疑問	～てよいのだろうか・～のではないか・～であろうか・～のであろうか・～ということになるのではないだろうか
仮定	～ことにしよう・～てみよう・～とする
否定	～ているわけではない・～というわけのものではない・～ことはない・～ことはなかった・～と言えるわけではない
理由	～たものではないからである
可能	～ことができる

表5 抽出した文末以外の動詞句の中の付属語の部分
Table 5 List of words attached to a verb.

意味	文末以外の表現
仮定	～ば・～ためには
強調	～が・～ことは・～ことを・～ものが・～ものは・～のが・～のは・～ことが・～とは・～ことに・～ことの・～だけでは・～のも
継続	～と・～ながら・～つつ・～たり
否定	～ない・～ないが・～ず・～わけでなく・～わけではなく・～のではないことで・～ないときに・～ないのが
目的	～べく・～ための・～ために・～ためにも・～よう
例示	～ような・～ように・～ことで・～とき・～ときの・～ときに
並列	～とともに
譲歩	～ものの・～ても
理由	～ので・～から・～ことで
過去	～た
意志	～よう
疑問	～ではないか

は動詞辞書に登録されているもので十分であった．しかし，対象を物語や随筆にした場合，述語となるひらがな動詞が多く出現し，抽出漏れが予想される．また，辞書には複合動詞抽出のための後続可能語句は漢字に限っているので，ひらがな動詞を含む複合動詞の場合には，本手法において自動的に生成した後，それが動詞であるかどうかをすべての場合にユーザと対話しなければならない．今後ユーザとの対話を必要最小限にとどめるように，ひらがな動詞に関する辞書の充実が必要である．なお，今回の抽出実験でユーザと対話して抽出したひらがな動詞の例を以下に示す．

- 漢字動詞と組み合わさって複合動詞を作る場合
  結びつく（「結ぶ」＋「つく」）
  見いだす（「見る」＋「いだす」）
  見つける（「見る」＋「つける」）
  聞きなれない（「聞く」＋「なれる」）
- ひらがな動詞どうしで複合動詞を作る場合
  とりいれる（「とる」＋「いれる」）
  つきあたる（「つく」＋「あたる」）
- 「する」動詞になることのできる漢字2字の熟語に後続する場合
  関係づける（「関係」＋「づく」）
  位置づける（「位置」＋「づく」）

次に適合率であるが，余分に抽出したのはすべてひらがな動詞であった．以下に抽出例を分類し，例をあげて示す．

- 助詞＋副詞の一部をひらがな動詞として抽出する場合
  これらはさまざまな … → 「はさま」を「はさむ」の未然形として抽出

表6 抽出した補助動詞と英語での前置詞句・助動詞句・副詞句にあたる表現
Table 6 List of phrasal expressions.

漢字動詞を含むもの	から言って・て言えば・で言う・と言えば・と言える・とは言えず・とは言えない・とは言うまでもなく・と言われるが・と言うよりも・にすることから言って・であると言える・に対して・に対しても・に対する・に向けて・に関する・に関わる・を通じた・という極めて・と言っても・に基づいて・に基づく・と思われるが・ように思われる・を除けば・ことはあり得るが・に限っていえば・を始めて
ひらがな動詞を含むもの	おけば・があり・があるが・があることは・があると・があれば・かという・がなくなり・からといって・きたものではあるが・ことがある・ことができないという・ことであった・ことである・こととして・ことについて・ことによって・ことにより・ことによるだけならば・ことになる・されるのが・すれば・するのは・たのだとすれば・たものであって・であったものが・であったり・であって・であり・である・であるが・のである・のであるが・であることが・であることは・であることに・であることには・であることを・であることの・であると・であるという・であるというのが・であるとすれば・であるのに・であれば・ではあるが・ていき・ていく・ていけば・ていった・ている・ているが・ていると・ていることは・ていることが・ていることを・ているとすれば・ているのだとすれば・ているように・ているにもかかわらず・ており・てきた・できることを・てくる・てしまう・でなければならず・ではないかという・という・というのが・というのがあって・というのは・というべきもので・というものを・というよりは・といって・といっても・とされる・として・としての・としながら・とした・としたら・としても・とする・とするとき・とするときは・とすれば・となった・となって・とはいえ・なければならないと・ならば・となり・となる・にある・においては・における・にかかわる・にした・にしたが・にしている・にしているはず・にする・にするのは・にするのが・にせよ・について・についての・については・についても・につき・にとって・にとっての・にとっても・になり・になることが・になるという・によって・によらず・による・によるという・により・によれば・にわたる・のであって・はあるが・はできず・べきものであり・みれば・ものであって・ものである・ものであるが・ものであることが・ものがあることを・ものとして・ものとしての・ようとするのが・をするものと・をもって・をもった

これにしたがって … → 「にした」を周辺語句「に」＋「する」の連用形＋周辺語句「た」の組み合わせとして抽出

- 副詞・接続詞の一部をひらがな動詞として抽出する場合
  むしろ … → 「しろ」を「する」の命令形として抽出
  もっとも … → 「もっ」を「もつ」の連用形が音便変化した形として抽出
  いかに … → 「いか」を「いく」の未然形と

して抽出

- ひらがな動詞を組み合わせて生成した複合動詞が文中の動詞でない語と一致して抽出する場合
  そうでないことを … → 「いこ」を「いる」の連用形＋「くる」の未然形として抽出
  これが大きい … → 「きい」を「くる」の連用形＋「いる」の連用形として抽出
  とりあえず … → 「とりあえ」を「とる」の連用形＋「あえる」の連用形として抽出

これらについては例外知識の充実と副詞・接続詞等を登録した「独立語辞書」の整備が必要である.

7. ま と め

字面処理によって，文章中から日本語動詞をユーザとの対話を含めれば，100％の割合で抽出できた．今後の課題は，抽出の対象となる文章の数と種類を増やして，抽出の性能を評価するとともに,「独立語辞書」と例外知識の充実を進めていくことである．また,『文彩』に動詞抽出機能のみにとどまらず，字面処理によってできうる他の種類の語句の抽出機能も持たせて，日本語文章の統合解析ツールにする予定である．

参 考 文 献

1) 新村 出編：広辞苑，岩波書店 (1955).
2) 中挾知延子，島田静雄：外国人のための日本語文章校正システム，第7回テクニカルコミュニケーションシンポジウム研究発表論文集，pp.15–19 (1995).
3) 中挾知延子，島田静雄：動詞辞書の提案とその利用についての一考察，第51回情報処理学会全国大会講演論文集 (3)，3-41 (1995).
4) 新納浩幸，井佐原均：疑似Nグラムを用いた助詞的定型表現の自動抽出，情報処理学会論文誌，Vol.36, No.1, pp.31–41 (1995).
5) 水野 聡，島田静雄：日本語キーワードの自動抽出手法，情報処理学会研究報告 92-NL-91，Vol.92, No.74, pp.41–46 (1992).
6) 宮島達夫他：岩波講座 — 日本語 9，語彙と意味，岩波書店 (1977).
7) Japanese Industrial Standards Committee: JIS X 0212-1990 Code of the Japanese Graphic Character Set for Information Interchange (1990).
8) 樺島忠夫他：ことば読本 — やまとことば，河出書房新社 (1989)
9) 松山羊一：中学国文法，昇龍堂 (1992)
10) 劉 暁民：日本語・中国語慣用語法辞典，日本実業出版社 (1995)

(平成7年 8 月 28 日受付)
(平成7年 12 月 8 日採録)

中挾知延子（正会員）

昭和40年生．昭和63年大阪大学工学部応用物理学科卒業．現在，東京国際大学で助手として勤務するとともに，埼玉大学大学院理工学研究科情報数理科学専攻博士後期課程在学中．日本語情報処理，知識情報処理の研究に従事．

島田 静雄（正会員）

1931年生．1954年東京大学工学部卒業．工学博士．1959年東京大学工学部土木工学科助手，講師を経て，1963年名古屋大学工学部助教授，教授．1983年から工学部共通講座情報検索学に配置替．1990年から埼玉大学工学部情報工学科教授．自動作図とグラフィクス，情報管理，データベースの研究に従事．土木学会などの会員．