論文

用言中心構造モデルによる日本語用言型複合表現の自動抽出

An Automatic Extraction for Verbal Compound Expressions by applying a Structural Model based on Declinable Words

中挾知延子
埼玉大学理工学研究科情報数理科学専攻
島田静雄
ポーランド日本情報工科大学

日本語文章を機械翻訳するために形態素解析した場合，分解されすぎて，意味の対応をつけにくいということが起こる．日本語文章を計算機で処理するとき，述部として機能する文章中の表現で，複数の語の並びとみなしうるものであっても，ひとまとまりとして他の語と切り離して認識させたい．我々はこのひとまとまりの語列を用言型複合表現という．例えば「を通じて」や「であるとは限らない」などは，1つの用言型複合表現として取り扱いたい．我々は日本語文章の用言型複合表現を記述できる用言中心構造モデルを考案し，そのモデルに基づいた抽出手法を開発した．抽出処理は字面からの情報とコンパクトな用言辞書で実現している．本手法を用いて約11万文字の新聞記事を対象に用言型複合表現を抽出した結果，約2600語抽出し，そのうちの9割以上が妥当な用言型複合表現であった．

1. はじめに

本論文では，日本語文章における用言型複合表現を，用言を中心とする構造モデルを用いて抽出する手法について述べる．この手法は我々の開発している日本語処理ツール『文彩』[1]の機能を拡充したものである．我々は，用言型複合表現を抽出することで，機械翻訳の前処理に活用することを目的としている．

日本語には『複合動詞・複合形容詞』と呼ばれる動詞・形容詞が複数個結びついた語がある[2]．英語の場合とちがって，日本語では形容詞だけでも述部を構成することができるので，動詞に形容詞が後接した語も見受けられる．動詞と形容詞が結びついた語と複合動詞・複合形容詞を併せて『複合用言』とすると，『複合用言』は用言の連用形に別の用言が接続した語であり，文法上からみた品詞を指す．例えば，「使いやすい・慎み深い」などは複合用言である．ここで「慎み深い」は，動詞「慎む」の連用形と形容詞「深い」が連接した語である．

『複合用言』に対して，『用言型複合表現』は，文章中の用言を含む複数の語の並びで，述部として機能する表現であり，意味上からみたひとまとまりの語である．ここで「複数の語の並び」と述べているのは，複合用言のように，用言が連用形によって連続して並んでいる語のことではなく，語の並びにおいて，その中の任意の部分文字列が用言ということである．

用言型複合表現には，用言の前後に接続する語と一緒になってひとまとまりの意味を成す表現が多い．例えば，「してもかまわない・に関して」などがある．「に関して」をはじめとする「によって・について」などは，助詞相当語と通常呼ばれるが，それぞれ「関し」

(「関する」の連用形)・「よっ」(「よる」の連用形の音便変化)・「つい」(「つく」の連用形) という用言を含むとして，助詞相当語も用言型複合表現に含める．

日本語の文章解析のためには，分かち書きされない日本語文章を形態素解析によって，意味を持つ最小の単位に分割することが通常行われる．ただ，形態素解析を行った際に，文章が分割され過ぎて，周辺の語句を合わせてひとまとまりの意味を担っている表現を見失ってしまう場合がある．このような表現は文章の区切りや文末に多く見られ，用言である動詞や形容詞を伴っている場合が多い．例えば，「～してしまうかもしれない」という表現において，「し・て・しまう・かも・しれ・ない」をそれぞれ単独で抽出しても意味がなく，用言である「し・しまう・しれ」の周辺にある語句と一緒に抽出して，ひとまとまりの複合表現として意味を把握しておくことが望ましい．

用言型複合表現を文章中で同定することは，機械翻訳の前処理に活用できる．機械翻訳において，形態素解析をする前に，日本語独特の慣用表現や言い回しなどを一つのまとまりとして別枠で処理しておくことで，翻訳の性能を上げることができる．現状では，ユーザが機械翻訳ソフトのマニュアルを参照して，翻訳の精度を上げるために，文章表現を書き換えなければならない場合がある．用言型複合表現には，意味のひとまとまりであることから一語として処理すべきである点と，対応する英語の表現は意訳を要する場合があるという点で，機械翻訳の前処理として前もって抽出し，書き換えを必要とする表現がある．

本論文で抽出する用言型複合表現は，(1) 助詞相当語，(2) 用言を含む述部の表現，の2種類に区別できる．本論文では (1)，(2) のどちらも抽出して，それらの表現を機械翻訳システムで一語として取り扱い，システムの精度を向上させることを目的にしている．とりわけ (2) で言及した用言を含む述部の表現は，機械翻訳の翻訳性能向上のためにはユーザに警告して，冗長であれば簡潔な表現に書き換えてもらう必要がある．

例として，「そういうことも言えなくはない.」と「そういうことが言える.」は，意味的にはほぼ同じであるが，機械翻訳にかけると，訳文は以下のようにかなり違うものになる．

- 「そういうことも言えなくはない。」
  "Without being said even that says so nonexistent."
- 「そういうことが言える。」
  "It is said to say so."

上の例からも分かるように，「言えなくはない」を「言える」にすると言った，少し簡潔な表現にするだけで，妥当な訳文を得ることができる．我々は，用言型複合表現の抽出を行なうことにより，簡潔な表現への書き換え処理の支援環境をユーザに向けて提供できると考える．我々は，用言型複合表現を個別に辞書に登録して必要な時に参照したり，各語に接続する語の内部情報を持たせたりする[3]のではなく，以下のような指針で自動的に文章中から抽出することを試みた．

(1) 用言型複合表現の抽出に必要な，用言の周辺に出現しうる語は，最小限の語群を用意しておく．

　文章中に出現する用言型複合表現は，ひらがなの語の並びが多いとはいえ，出現頻度の多い語は限られている．語群の内容を豊富にしすぎるとかえって無駄が生じるために，ある程度の語群で納めておくことが望ましい．また，対象となる文章の種類や数が多くなればなるほど用言型複合表現の種類は多くなるが，必要に応じて追加・修正をすれば十分である．

(2) 簡易な字面のみの制約で抽出する．

　漢字かな混じり文である日本語文章の特徴を生かした，漢字・ひらがなの区別などの字面上の制約を用いる．ただし用言の同定についてのみ，活用形は生成できるように，字面以外の処理として活用規則は用いる．

文章中における複数の語の並びである複合表現を，ひとまとまりの単位として構造モデルを提案した研究がある[4]．[4] の研究が語と語の連接関係を規定して複合表現を組み立てていくのに対し，本手法では連接規則を使わずに，用言の周辺に用意した語が勝手に出現するということを前提にして抽出を行う．また，大規模コーパスからNグラムを用いて定型表現を抽出する研究[5]や，それを利用した「に関して」などの助詞的定型表現の自動抽出の研究がある[6]．本手法では助詞的定型表現に限定せずに，用言を含む複合表現の抽出を目的としている．また，[6,7] で利用されているNグラムは用いずに，用意した用言と用言の周辺に出現する語群の中から一致する語を取り出してひとまとまりの複合表現を定めている．

さらに本手法を計算機上に実装し，新聞記事を対象にして用言型複合表現の抽出を行ったのでその実験の結果と考察についても述べる．

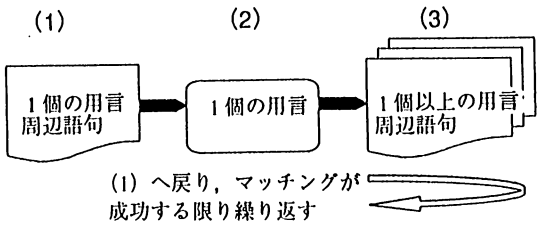

図1　用言中心構造モデル

2. 用言中心構造モデル

用言型複合表現を記述する用言を中心とした構造モデル(用言中心構造モデル) は以下の3つの部分に分けられる．

(1) 用言の前に出現する語群
(2) 用言部
(3) 用言の後ろに出現する語群

図1に沿って説明する．(1) と (3) を合わせて「用言周辺語句」と呼ぶ．用言周辺語句についてであるが，(1) の用言の前に並ぶ語は1語とする．(2) の用言部は，次章で述べる用言辞書を用いて同定される用言である．用言の後ろに並ぶ (3) は，一部の例外を除いて，1語以上の用言周辺語句が位置する．用言周辺語句の一覧を表1に示す．(1) と (3) に用意する語群はまったく同じではなく，表1において，補助用言の語など，図1の (1) には不適切である語は (1) のための語としては用意しない．また，(3) では，1つの一致した用言周辺語句の後ろにも，さらに一致する用言周辺語句が連なっている場合には，一致する限り用言周辺語句を結び付けて抽出する．1つの用言型複合表現を抽出しても，その直後に別の用言型複合表現があれば，(1) からはじまるマッチング処理を繰り返し，用言型複合表現同士をつなげていく．

表1　用言周辺語句

助詞	ながら・ばかり・から・こそ・ さえ・しか・だけ・でし・です・ つつ・とも・ども・など・のみ・ まで・より・ろう・か・が・た・ だ・て・で・と・に・の・は・ ば・へ・も・を
助動詞	みたい・ような・らしい・ けれ・ざる・たい・たく・たら・ たり・べき・べく・べから・まい・ まし・ます・ません・ず・ぬ
抽象名詞	見通し・具合・場合・ 模様・最中・必要・過言・仕方・ 以上・以下・以降・以来・予定・ 結果・無理・方・上・時・得・ いずれ・きらい・とおり・ ところ・うえ・うち・がち・ こと・ごと・ため・とき・ はず・ふう・ほか・ほど・ まま・ゆえ・よう・わけ
補助用言	下さい・ください・ない・ その他・どうか・こう・そう・ どう・ょう・必ず

用言周辺語句の種類は，助詞・助動詞・抽象名詞・補助用言・その他の語とする．補助用言とは，形態上は動詞や形容詞であるが，他の用言に後接して補助的な意味を添え，助動詞のように機能する用言である．その他の語とは，前の4種類に入らない語や，品詞などが明確に決定できない文字列を指す．用言周辺語句の選択にあたっては [8,9,10] を参考にした．

以下に抽出する用言型複合表現の例を挙げる.

(1) (例1)「に関して」
→「に」(助詞) +「関し」(動詞「関する」の連用形) +「て」(助詞)

(2) (例2)「をすることにほかならない」
→「を」(助詞) +「する」(動詞「する」の連体形) +「こと」(抽象名詞) +「に」(助詞) +「ほか」(抽象名詞) +「なら」(動詞「なる」の未然形) +「ない」(補助用言)

(3) (例3)「はやむを得ない」
→「は」(助詞) +「やむ」(動詞「やむ」の終止形) +「を」(助詞) +「得」(動詞「得る」の連用形) +「ない」(補助用言)

3. 用言辞書

抽出のために用いた用言辞書の特徴を以下に示す.

(1) 語幹がひらがなの用言(ひらがな用言)の登録方法
語幹の1字目のひらがなを見出しにしている. ひらがな用言で登録されている語は日常の文章ではひらがなで書かれることが多い語に限っている. 加えて敬語や英語での前置詞句になりやすい表現に含まれるひらがな用言も登録した.

(2) 語幹が漢字の用言(漢字用言)の登録方法
用言の語幹漢字(用言用漢字)を見出しにしている. こうすることで, 語幹漢字を共有する複数の用言を整理している. 例えば「当」であれば「当たる・当てる」という2種類の用言をまとめておく. 見出しの用言用漢字の選択にあたっては, [8,9] を参考にした. 用言型複合表現に含まれる用言としては, 語幹が1字の漢字である語に限っている. 例えば,「に関して」に含まれる用言用漢字は「関」である. なお, 漢字はJIS第1水準漢字の範囲である[11].

(3) 用言活用規則
辞書は, 用言活用規則で操作できるようにしている. 見出しそれぞれについて送り仮名・活用の種類を記し, 用言活用規則を適用して未然形から命令形までの用言の活用形を生成する. それに加えて「られる・れる」が付いた受身形,「させる・せる」が付いた使役形, 音便形, 五段動詞の場合には可能形も生成する.

表2に用言辞書の登録語の内訳を示す. ひらがな用言の登録語は終止形で記す.

表2 用言辞書の登録語の内訳

用言		関対基沿比面照備至当応加際従違
用		伴反除通踏間恐覚於拘過達及初下
漢字		越代限開用巡　(合計36字)
ひら	動詞	あう・あく・あげる・あたる・あてる・ある・いう・いく・いただく・いる・うつ・うる・える・おく・おる・かかわる・かぎる・かける・かなう・・かわる・きく・きまる・きめる・くださる・くだす・くる・くれる・こたえる・さす・さる・したがう・しまう・しる・すぎる・すむ・する・だす・たとえる・たまる・たる・ちがう・ちなむ・つく・つまる・つもる・できる・ともなう・とる・なす・なる・のぞく・はさまる・はさむ・まいる・まとまる・みえる・みせる・みる・めぐる・もつ・もらう・やむ・やる・ゆく・よる・わかる・わたる　(合計69語)
がな	形容詞	いい・ない・にくい・ほしい・やすい・よい　(合計6語)
用言		

4. 用言型複合表現の抽出

4.1 抽出処理

抽出には文章の最初から1文字ずつ読み込み, 用言辞書にある用言用漢字に出会えば, 活用規則にしたがって活用語尾を生成しマッチングを行う. マッチングに成功したら文字列を漢字用言として同定する. また, 文字がひらがなであって, 辞書に登録されているひらがな用言の1字目のひらがなである場合には, 活用規則を適用し活用語尾とのマッチングを行う. マッチングに成功したら, 文字列をひらがな用言として同定する. 次に2章で

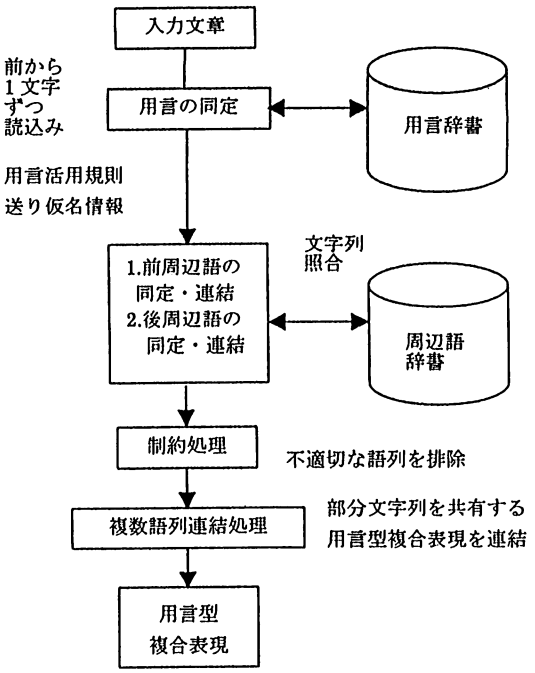

図2　用言型複合表現抽出システム構成図

述べた用言型複合表現の構造にしたがって，(1)，(3) の順に用言周辺語句と前後の文字列とのマッチングを行う．システム構成図は図2のとおりである．ここで「前周辺語・後周辺語」はそれぞれ2章の (1)・(3) に対応している．

抽出には，2つの繰り返し処理がある．1つめは，2章の図1における (3) の部分について，周辺語句とマッチングが成功する限り抽出を繰り返すことである．2つめは，抽出は用言から行われるので，長い文字列中に複数の用言型複合表現が抽出できることがある．前後の用言複合表現の一部が重なっていたり，直後に位置している場合にはそれらを一緒にまとめて抽出する．2つめの繰り返し処理の例を以下にあげる．

- (例1)「というわけでもないのである」「というわけでもないので」,「である」は「で」の部分が重なっているのでまとめて抽出.
- (例2)「にする必要があると」「にする必要が」,「があると」は「が」の部分が重なっているのでまとめて抽出.

表3に（例1）を用いた処理の流れを示す．ここで「抽出候補の語列」，その時点で用言型複合表現の部分列として計算機の内部に蓄えられている語列を指す．また (A)(B)(C) を付けた語列はその時点で得られた用言型複合表現を指す．

表3　用言型複合表現抽出過程の例
(抽出例)「というわけでもないのである」

手順	処理	抽出候補の語列
用言の同定	「いう」の終止形と同定	いう
前周辺語との照合	「と」を同定・連結	という
後周辺語との照合	「わけ」を同定・連結	というわけ
後周辺語との照合	「で」を同定・連結	というわけで
後周辺語との照合	「も」を同定・連結	というわけでも (A)
用言の同定	「ない」の終止形と同定	ない
前周辺語との照合	「も」を同定・連結	もない
後周辺語との照合	「の」を同定・連結	もないの
後周辺語との照合	「で」を同定・連結	もないので (B)
用言の同定	「ある」を終止形と同定	ある
前周辺語との照合	「で」を同定・連結	である
制約処理	制約1を適用	である (C)
複数語列連結処理	(A)(B)(C)を連結	というわけでもないのである

4.2 抽出時の制約

以下の (制約1)〜(制約3) を，抽出処理を行う際に制約として課した．

(1) (制約1) 句読点に関する制約
図1の (3) における説明で，「1語以上の用言周辺語句」とあるが，用言の直後が句読点である場合には，後ろに続く用言周辺語句は0個でもよいことにする．

(2) (制約2)「する名詞」に関する制約
2字以上の漢字熟語で，「する・できる」を伴ってサ変動詞になる語 (する名詞) が

ある．3章で述べた用言辞書には，ひらがな用言として「する・できる」が登録されており，通常ではする動詞の語尾の「する・できる」の活用形も抽出してしまう．これらの語を抽出しないように，「する・できる」の活用形が見つかって，直前の2文字が漢字列であれば抽出しないことにする．

(3) (制約3) 漢字用言の活用・送り仮名に関する制約
複数の用言がある語幹漢字の場合には，複合表現になりやすい用言とそうでない用言がある．例えば，「通」を語幹に持つ用言には「通る」「通す」「通じる」がある．その中で，「通じる」は周辺語句を伴って「を通じて」などの用言型複合表現になりやすい．一方で「通る」は文章中で用いられる場合には「通る」という本来の動詞の意味を表すことが多い．このような，活用の種類によって複合表現のなりやすさを考慮し，1つの漢字に対して複合表現になりやすい活用の種類を限定している．また，漢字用言の送り仮名が用言周辺語句として抽出されないように，用言周辺語句の候補の直前の漢字1字が用言辞書にある用言用漢字以外であれば，その候補の語を送り仮名であるとみなして，用言周辺語句とはしない．例えば「動かせることもある」は，「せる」を用言として「かせることもある」と抽出してしまう恐れがある．この場合「か」が用言周辺語句の候補として考えられるが，1文字前の「動」が用言辞書にないので用言周辺語句とはならず，「もある」を用言型複合表現として抽出する．

5. 抽出実験結果

5.1 抽出実験の内容

実験データは，約11万文字(112,989字)の新聞記事(日本経済新聞1994年の一面記事)である[12]．種類と日付は以下のとおりである．

- 本紙朝刊
  1月1/3/4/5日，7月27〜31日，9月1日
- 本紙夕刊
  1月4/5日，3月1日，7月27〜30日，9月1日
- 大阪夕刊
  1月4日，3月1日，7月28/30日

抽出処理に用いたシステムの大きさは，用言辞書の約3.6KBを含めて約450KBであり，DOS/V機上でMicrosoft Visual Basic^{TM}を用いて開発した．

5.2 抽出結果

5.2.1 制約1〜4を用いての実験結果

表4に抽出実験の結果を示す．本手法を用いて抽出された用言型複合表現の総数は，(A)で示されている．この数字は必ずしも実験データの文章中にある用言型複合表現を網羅しているわけではない．あくまでも本手法を用いて自動的に抽出される用言型複合表現の数である．表中の適合率は(A)の中で，用言型複合表現として妥当であると筆者が判断した語の割合である．結果として92％以上の割合で，抽出された語の中から妥当な用言型複合表現の抽出を行うことができた．図3に抽出出力例を示す．

表4　抽出実験結果

○	2,156語
△	257語
×	26語
▲	170語
抽出語句総数(A)	2,609語
適合率(○+△)/A	92.49％

以下，表中で記されている○，▲などの記号について説明する．

(1) 用言型複合表現である語 … ○
1章で説明した，用言型複合表現とし

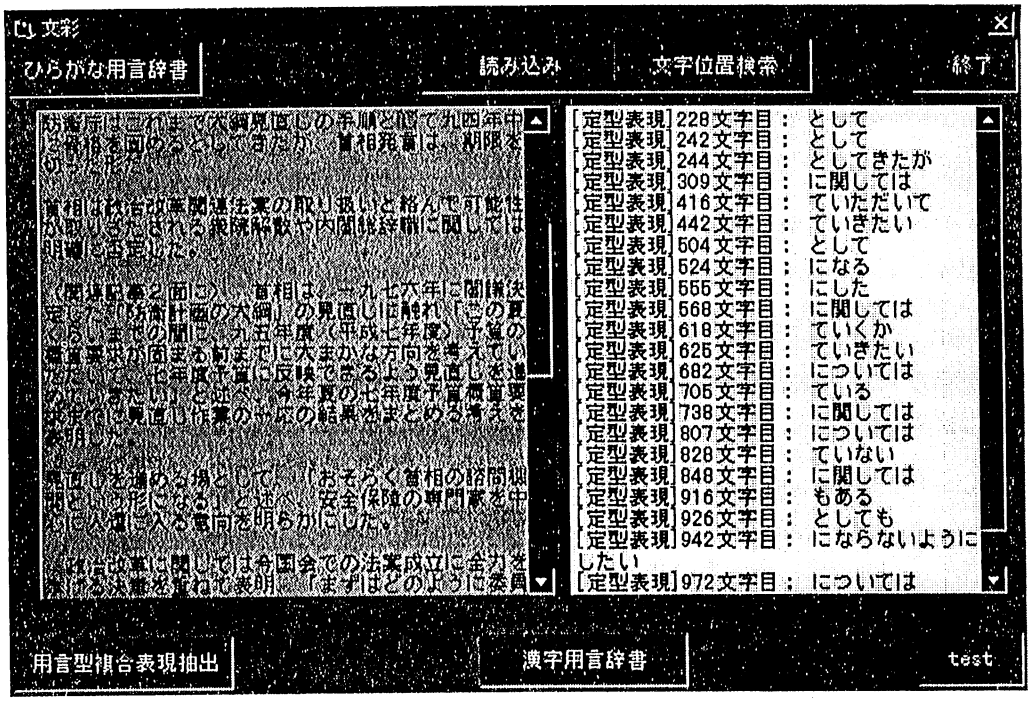

図3　用言型複合表現の抽出出力例

て妥当であると考えられる語を含めた．
○の語は以下のように分類できる．

・日英翻訳において，英語では用言ではなく，接続詞・助動詞・前置詞句にあたる表現

(例)「については・に関しては」
　　→ 'as for'
　　「しなければならない」→ 'must'
　　「に対し」→ 'against'　・他の簡潔な表現で置き換えられる語 (敬語表現も含む)

(例)「があるというわけである」
　　→「がある」
　　「というのである」→「である」
　　「ていただいて」→「て」・日本語独特の言い回しで，英語に翻訳するとき直訳できずに大幅に変更される語

(例)「せざるを得ない」
　　→ 'cannot help ～ing'
　　「ている模様だ」→ 'seem'

(2) 別の動詞に助詞を介して接続し，テンス・アスペクト・モードなどを添える語 … △

(例)「ている」「てきた」「ており」

抽出した語の中で，「ていたにもかかわらず」のような，後ろに用言周辺語句がついている語も多く，本論文では用言型複合表現に含めて差し支えないと考える．

(3) 用言本来の意味を持つ語 … ×

(例)「がめぐる」「を下さい」「も違う」

例えば，「て下さい」は敬語表現なので用言型複合表現に含めるが，「を下さい」の「下さい」は本来の意味を持つ．本論文では後者の場合には用言型複合表現とはみなさない．同様に文章中で頻出しないが，敬語や英語での前置詞句になりやすい表現に含まれるために，用言辞書に登録したひらがな用言は，文脈によっては用言本来の意味を持つ．「めぐる」もその一例であり，「をめぐって」は英語の'through'に対応し，用言型複合表現とみなせるが，前に「が」が付くと，「めぐる」の動詞本来の意味になる．本論文で

は，筆者の判断で用言が本来の意味を持つ語は×とした．　一方，「である・になる・をする」などにおけるひらがな用言「ある・なる・する」の場合，文章の叙述部分での中心的な意味を担うという点では用言本来の意味になっている．しかし，これらは「であるがゆえに・になるかどうかについては」のように用言型複合表現の一部を構成することが多い．日本語文章中で頻繁に出現することを考慮に入れると，このようなひらがな用言を含む語句は，(1) の用言型複合表現とみなす．

(4) 任意の文字列の一部になっている語 … ▲

(例)「かなりの」→ 名詞「かなり」の「なり」を動詞「なる」の連用形として誤抽出

「ことしこそはと」→ 名詞「ことし」の「し」を動詞「する」の連用形として誤抽出．これらは，通常は漢字で書かれることが多い名詞がひらがなで書かれてある場合に多く見られた．

(例)「しかし」→ 接続詞「しかし」の3文字目の「し」を動詞「する」の連用形として誤抽出．

(例)「にこの」→ 代名詞「この」の「こ」を「くる」の未然形として誤抽出

「する」の連用形「し」や，「くる」の未然形「こ」など，活用形によって1字の語になる場合には，用言とはちがう文字列の一部として誤抽出してしまう場合が見られた．

(例)「にこもって」→「こもる」の音便化した「こもっ」という用言を「くる」の未然形「こ」と用言周辺語句「も」＋「る」を結び付けて誤抽出

用言辞書に登録しているひらがなの用言は，表2にもあるように2語ないしは3語の用言が大半を占める．そのため，それらは周辺語句と結びついて，用言辞書に登録されていないひらがな用言の一部として誤抽出されてしまう場合がある．

5.2.2 制約の追加による実験結果

本手法によって，自動的に抽出された文字列の中から，不適切と考えられる語 (×，▲) の抽出を減らすために，4.2で述べた制約1 3に加えて，以下の制約を追加した．4.2での (制約1)〜(制約3) を初期制約，以下の (制約4)〜(制約6) を追加制約とする．

初期制約だけでも表4から92％程度の適合率を得ているが，いくつかの簡単な制約を追加することで，×や▲の語を大幅に除くことができれば，1章で示した本手法の指針には反しないと考える．試みとして，用言周辺語句を中心に，追加制約を設定して実験した．ただ，追加制約を設定していく過程で，簡単であるとはいいにくい部分も出てきたことなど，追加制約の内容や必要性に関しては次の6章で議論する．

- (制約4)　5.2.1の(3)に関連して，用言辞書に登録された漢字用言と，文章中で頻出しないが「敬語や英語での前置詞句になりやすい表現に含まれる」ひらがな用言それぞれに対して，用言本来の意味を持つときに前接する用言周辺語句を除外する．例えば「めぐる」は「が・は・も」などが前接すると用言本来の意味になる．そのため，「めぐる」に前接する用言周辺語句群からこれらの語は除いた．

- (制約5)　5.2.1の(4)に関連して，用言辞書に登録したひらがな用言それぞれにおいて，通常前接しない用言周辺語句を除いた．例えば「これをよりたくさんの …」という文字列の場合，副詞「より」は「よる」の連用形として認識され，「をよりた」を誤って抽出してしまう．動詞の「よる」であれば，「を」は通常前接しないので，「を」を除外しておく．

- (制約6)　5.2.1の(4)に関連して，用言辞書に登録したひらがな用言それぞれにおいて，例えば「くる」の未然形

「こ」のように，活用形が1字になる場合には後接可能な用言周辺語句を制限した．さらに，2文字になる活用形を部分文字列に含む代名詞，接続詞を抽出の際に検査して，該当すれば用言型複合表現として抽出しない．一方，ひらがなの名詞の部分文字列については検査をしていない．理由は，名詞は数も膨大になり，検査の処理が重くなると考えられるためである．また，特定のひらがなの一般名詞が複数の文章にわたって多用されることは少ないので，誤まって抽出しても特殊なケースとして無視できる．それに比べて代名詞や接続詞はひらがなで通常書かれることが多く，名詞に比べて種類が少ないわりには複数の文章にわたって何度も出現することが多いので，検査をすることは誤抽出を減らすために有効であると考える．

表5　制約追加後の抽出実験結果

	制約4追加後	制約5,6追加後
○	2156語(0)	2174語(+18)
△	257語(0)	257語(0)
×	1語(-25)	1語(-25)
▲	166語(-4)	25語(-145)
抽出語句総数(A)	2580語(-29)	2457語(-152)
適合率(○+△)/A	93.53%(+1.04)	98.94%(+6.45)

制約全般についてであるが，初期制約において，制約1は漢字の区別を含む字面のみの判断で行っている．制約2は「する・できる」2語に関する活用規則の適用と漢字かどうかの判別，制約3は複数の用言の語幹になる用言用漢字に対する活用の種類の検査を行っており，活用の種類の情報は得なければならないものの，いずれも簡単なものである．一方，追加制約4，5，6は該当する用言を個々に検査していかなくてはならないので，初期制約に比べると処理も複雑になっている．表5に制約4〜6追加後の抽出実験結果を示す．語数の後ろのかっこは，初期制約のみを用いたときとの増減を示す．また，表6に制約1〜6をすべて適用した実験において，実際に抽出した語句の例を示す．

6. 実験の考察

6.1 総括

本手法では，必ずしもすべての用言型複合表現が抽出されるわけではない．また，用言型複合表現の範囲・定義は難しく，著者にとっては荷が重すぎる．しかし用言型複合表現は日本語表現に独特のものであり，特色でもある．我々は本論文で提案した方法で用言型複合表現を抽出し，分析を試みている．例えば，抽出実験を行う過程で「〜とはいえない」のような表現があった．この場合，前接する用言周辺語句は図1によって1語としているので，用言「いえ」を中心として「は」＋「いえ」＋「ない」のような用言型複合表現が生成される．そのため「は」の前になる「と」が抽出されない．これは「とは・には・では」などの格助詞や副助詞が複数あわさってひとまとまりの意味の語になっているものである．実験データからこのような語を探し，抽出漏れが起きていないかどうかみたところ，「〜であるとはいえない・〜になるには」など用言も一緒に前の文字列に出現しているために，前の用言と一緒に抽出されていた．ただし，「そうとはいえない」のような場合には「はいえない」しか抽出されない．もしも「そうとはいえない」を抽出しようとすれば，図1の用言中心構造モデルにおける(1)の設定を変更しなくてはならない．設定の変更が容易に行えるものであれば，変更していく予定である．

精度については，初期の制約を使用すると適合率は約92.4％であった．一般に抽出の精度が9割以下であれば，抽出漏れを後に人手で検出する方が，人手で作業を行なうよりもコストがかかると言われている．このことを考慮すると，計算機によって本論文で提案した手法により自動抽出することは有効であると考えられる．

本論文では，実験データから提案した手法を使って抽出される用言型複合表現から5.2.1にあるような分類を行う．我々の立場は，できる限り簡易に構成したモデルを用いて，日本語文章中から用言型複合表現の自動的な抽出を試みることである．どの範囲までの文字列が用言型複合表現であるのか，どの範囲までを用言型複合表現に設定すれば文章の解析に最適であるのかは今後の課題である．

6.2 誤って抽出した語(×，▲)について

実験において，×になった「に従って」という語は，文脈によって用言型複合表現とみなせる場合とそうでない場合とがある．実験データにあった「に従って」は用言本来の意味を持ち，英語での'follow'に対応し，○には入らなかった．そのような場合，本抽出手法では文脈を考慮していないために，○か×の判断はできず，候補の表現をあげるというところにとどまっている．これらはユーザとの対話処理を取り入れて，最終的な判断はユーザに任せることでの解決が考えられる．また，×が1語だけというごく少ない結果に終わったが，実験データが新聞記事の1部分という限られた範囲の文章であったせいで，他の種類の文章やデータの数を増やせば×の語はかなり出現すると思われる．

次に▲の語についてであるが，制約5，6を用いると，ある程度の不適切な候補を除外することはできたが，それらすべてを除外することはできない．かえって制約を加えすぎると，用言型複合表現として妥当であると思われる語を除いてしまいかねない．例えば，「かした」は用言周辺語句「か」が前接した形であるが，文章中では「～を強くほのめかした」の一部になっており誤抽出になる．しかし，「する」の連用形「し」に前接する用言周辺語句から「か」を除いてしまうと，「～するとかして」のような用言型複合表現を抽出できなくなる．これについての解決策は今後の課題である．

6.3 制約について

初期制約に制約4，5，6を加えたことで，不適切な文字列をかなり除くことができた．しかし，制約4，5，6は用言に対して個別に用言周辺語句の検査をしなければならない．またこの制約は形態素解析における連接規則の一部とみることができる．できるだけ字面のみで解決しようとしているために，形態素解析で用いるような連接規則を適用することは指針に反する．また，5.1でも述べたように，制約を増やしたからといって抽出精度が必ずしも向上するとはいえず，逆に適切な文字列を除いてしまう恐れもある．制約の内容については今後充分に検討する必要がある．現時点では，初期制約だけでも92％程度の適合率を得ることができており，追加制約のような，個別の語に対する制約を使わなくても，本手法の有効性は示せたと考える．

7. おわりに

日本語文章に頻繁に表われる用言型複合表現を，我々の提案した用言中心構造モデルに基づいた手法を用いて，いくつかの制約は課した上で，9割以上の適合率で抽出できた．用意した用言辞書も，登録する用言の数を少なくし，コンパクトなものにしている．抽出に用いた制約なども字面のみの情報で判断できる場合が多く，簡易な処理で実現できたといえる．今後の課題としては，表4，5で示した適合率を高めるために，制約について改善を重ねることと，新聞記事以外の文章に適用して本手法の有効性を検討することである．

参考文献

[1] 中挾知延子，島田静雄：字面解析による日本語動詞抽出手法，情報処理学会論文誌，Vol.37，No.2，pp.179-187(1996)
[2] 長嶋善郎：複合動詞の構造，日本語の語彙と表現，pp.63-104，大修館書店(1976)
[3] 杉本つとむ，岩淵 匡：日本語学辞典，おうふう(1994)

表6 抽出された用言型複合表現の例

○	がある・にある・のある・もある・である・ていた・てきた・へきて・にした・という・として・ とする・に対し・に伴い・を除く・としての・にしても・によって・によると・について・ てくると・てみよう・てみると・てくれた・に応じて・に沿って・に反して・に比べて・に加えて・ に関して・に関する・に対する・を通じて・が初めて・を問わず・に限って・になって・であります・ としている・がない限り・さえあれば・となる予定の・をみながら・がありそうだ・はあるはずだ・ をめぐって・に代わって・に当たって・を踏まえて・ばならない・にわたって・については・ にとっては・にちなんだ・に違いない・に際しては・に基づいて・に過ぎないが・を通じてしか・ がまとまって・ていただいて・てもらいたい・だけではない・ほどしかない・たことがある・ ている模様だ・いないまま・にしてしまった・をしたところが・たりしている・ができるように・ がない代わりに・が問われている・にあるとはいえ・からではないか・というのである・ べきでないとの・てきたからである・になったといえる・によるものだとも・にしてくれないと・ てきたからである・であるとみている・といわれているが・ができない場合は・ができていることを・ にする必要があると・ていることについて・ているに過ぎない・でいることを踏まえ・ をせざるを得ない・もやむを得ないと・がつかない場合には・ということを除けば・ てしまったことへの・とみることができる・をするかもしれない・のあり方については・ ということになるかも・になるのかにもよるが・をしてきたというのは・があるわけではないが・ ばならないというのが・があるというわけである・をしたことなどによって・必ずしも必要ではないが・ ているというものであり・にするかどうかについては・のありようといえる・のかもしれないが...
△	ていく・ている・ており・でいる ・でみた・でみる・てしまう...
×	に従って
▲	かした・かして・はより・をしか・はやったり・などより・つつきながら・がつきものだ...

[4] 長尾真 編:自然言語処理, 岩波書店(1996)
[5] 首藤公昭, 楢原斗志子, 吉田将:日本語の機械処理のための文節構造モデル, 電子通信学会論文誌, Vol.J62-D, No.12, pp.872-879(1979)
[6] 長尾真, 森信介:大規模日本語テキストのnグラム統計の作り方と語句の自動抽出, 情報処理学会自然言語処理研究会, 96-1, pp.1-8(1993)
[7] 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.32-40(1995)
[8] 劉暁民:日本語・中国語慣用語法辞典, 日本実業出版社(1995)
[9] 森田良行, 松木正恵:日本語表現文型, アルク(1989)
[10] 名柄迪, 広田紀子, 中西家栄子:形式名詞, 荒竹出版(1987)
[11] Japanese Industrial Standards Committee: JIS X 0212-1990 Code of the Japanese Graphic Character Set for Information Interchange(1990)
[12] 日本経済新聞社:日本経済新聞CD-ROM版 1994年版. 1995年.

(1997年8月10日受付)
(1998年2月5日採録)

著者紹介

中挾知延子 (学生会員)
埼玉大学大学院理工学研究科情報数理科学専攻
E-mail: chieko@ke.ics.saitama-u.ac.jp

島田静雄 (正会員)
工学博士
JICAプロジェクトリーダー
ポーランド日本情報工科大学勤務
名古屋大学名誉教授
E-mail: shimada@pjwstk.waw.pl