日本語文章における主格省略の自動検出

北村達也[a], 中村慶太[a], 川村よし子[b]

[a] 甲南大学 知能情報学部 知能情報学科
兵庫県神戸市東灘区岡本８－９－１, 658-8501
[b] 東京国際大学 言語コミュニケーション学部 英語コミュニケーション学科
埼玉県川越市的場北１－１３－１, 350-1197

(受理日 2012 年 10 月 12 日)

概 要

本研究では，一般の日本語母語話者が日本語非母語話者向けの平易な文書を作る作業を支援するため，日本語非母語話者にとって文の難易度を高める要因の 1 つである主格省略を検出する技術を提案する．提案法では，構文解析技術と Web 検索エンジンを利用して主格省略判定を行う．まず，入力文中の述語とそれに係っている名詞を組み合わせてガ格とヲ格の検索文「(名詞) が (述語)」と「(名詞) を (述語)」を作り，両者について Web 検索エンジンを用いて検索する．その結果，各々が検索された件数を比較することによって，当該の名詞が当該の述語の主格となるか否かを判定する．文中に主格となりうる名詞が存在しない場合には主格省略と判定し，主格が省略されていない場合には，主格となる文節がどれかも判定する．評価実験の結果，人が主格省略と判定した述語のうち 84.6 %を検出でき，人が主格ありと判定した述語の 72.3 %を検出できた．

キーワード： 日本語教育, 主格省略, 構文解析, Web 検索エンジン, 用例検索

1 はじめに

日本語教師が教材を作成する際には，対象となる学習者のレベルや知識に応じて文を書き換えて難易度を調節したり解説を加えたりしている．また，公共機関等で日本語非母語話者を対象として日本語で広報を行う際にも，内容理解に支障をきたさない平易な文への書き換えを行う必要がある．

このような広報活動は，行政，職場，学校，町内会など日常生活のあらゆる場面で必要となる．とりわけ災害時にはわかりやすい日本語に関する配慮が不可欠であり [1], [2]，情報伝達のスピードが生死を分けることさえある．

ところが，一般の日本語母語話者は，文の難易度に寄与する要因に関する知識や書き換えのノウハウを十分に持たない．これは，日本語を分析的に見ることができる日本語教師や日本語研究者と異なる点である．そのため，一般の日本語母語話者は，日本語非母語話者にとってどのような文が難しく，それをどのように書き換えれば平易になるのかという基準がはっきりしない．

このような問題に対して，ワープロソフトのスペルチェッカーや文章校正ツールのように，文の難易度を高めてしまう要因を示すツールがあれば有用となるはずである．そこで，我々は，一般の日本語母語話者が日本語非母語話者にとって理解し易い文章を書くため，もしくは既存の文章を理解し易い文章にするための支援ツールに関する研究を行っている．本研究では，文の難易度に寄与する要因の 1 つである主格省略を検出するシステムを開発する．

川村ら [3] は，ヨーロッパおよびアジアの非漢字圏出身の学習者 382 名を対象にして文の難易度判定実験を行った．この実験では，単語の難易度や構文の複雑さが異なる文を対象にして初級，中級，上級の学習者に各文の難易度判定と母語への翻訳を行わせた．因子分析の結果，文の難易度には単語の難易度と構文の複雑さが寄与することが明らかになった．その他，初級，中級学習者にとってはゼロ格 (特に主格省略)，モダリティやアスペクトに関係する補助動詞，視点の移動，慣用表現等が文の難易度に寄与することが示された．従って，文を平易に書き直すための支援ツールとしては，これらを自動検出する技術が求められる．

学習者にとって難しい構文を抽出するシステムに関しては，以下のような先行研究がある．Yamura-Takei ら [4] および竹井ら [5] はゼロ格を自動検出するシステム Zero Detector を開発した．Zero Detector はゼロ格の有無の判定を『日本語語彙大系』[6] に収められた単語の意味属性や文型パターンに基づいて行う．

また，水嶋ら [7] は，中止法および名詞修飾を自動検出するシステムを開発した．中止法は文を途中で一旦区切って直後の文につなぐ用法である．一方，名詞修飾は被修飾名詞の前に修飾語が置かれる用法である．彼らは形態素解析と構文解析を利用してこれらの自動検出を実現している．

内田ほか [8] は主格省略の自動検出法を検討した．主格省略を検出するには，係り受け関係にある名詞と動詞において，当該の名詞が当該の動詞の主格となるか否かを何らかの方法で判定する必要がある．例えば,「パンが食べたい」という文が与えられた場合，ガ格の「パンが」は「食べたい」の主格ではない．こうした文で主格が省略されていることを正しく判定できなければならない．

彼らはこの問題に対処するため，旧日本語能力試験 4 級の出題基準の動詞を対象に各動詞がどのカテゴリの名詞を主格として取り得るかを手作業でデータベース化した．名詞のカテゴリは形態素解析システムによって得られるので，このデータベースがあれば主格省略が検出できる．しかし，すべての動詞に対してデータベースを作成するには膨大な人力を要するばかりか，形態素解析システムによって得られる名詞のカテゴリの範囲が広すぎるため，主格省略検出の精度が低いという問題が生じた．

そこで，本研究では，異なるアプローチによって主格省略の自動検出法について検討した．本研究では，当該の名詞が当該の動詞の主格になり得るか否かの判定をインターネット上の Web 検索エンジンを用いた用例検索に基づいて行う．これにより，内田ら [8] のようなデータベースを作成せずに主格省略の自動検出を実現できる．本稿では，まず提案法について述べ，次に評価実験によりその有効性を明らかにする．なお，本稿は，中村ら [9] の報告に対して，主格省略の自動検出が可能な文のタイプを増やし，さらに評価実験の対象を拡大させたものであることを付記する．

2　主格省略の自動検出

2.1　処理の流れ

本研究で提案する方法は，日本語として正しい文が入力され，構文解析によって係り受け構造が正しく得られることを前提とする．

最初に，入力文を構文解析し係り受け構造を得る．次に，文中の述語に係る名詞とその後続の助詞を列挙する．そして，その助詞が主格を取り得るものだった場合に，当該の名詞が当該の述語の主格となり得るか否かを判定する．主格となり得る名詞が存在すれば「主格有り」，存在しなければ「主格省略」と判定する．以上の処理を文中のすべての述語に対して実行し，それぞれに対して主格省略の有無を出力する．ただし，命令文は主格を必要としないので，主格省略の判定を行わない．

2.2　係り受け構造の分析

入力文の各形態素の係り受け構造は構文解析システム KNP ver. 3.01 [10] を用いて分析する．KNP は形態素解析システム JUMAN による形態素解析結果に基づいて係り受け構造を解析する．例えば，「彼に借りた本を読んだ.」という文の場合，図 1 のような構文木の形で分析結果が得られる．

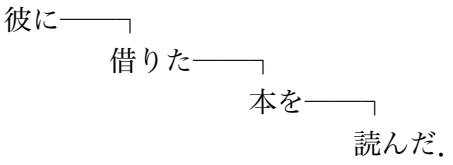

図 1:「彼に借りた本を読んだ.」を KNP ver. 3.01 で分析した結果

ここで，文節を結ぶ線は係り受け関係を示しており,「彼に」が「借りた」に,「借りた」が「本を」に,「本を」が「読んだ」に係っていることを表している．なお，KNP による分析により，各形態素の品詞の情報も得られる．

上の文は正しく分析されているが，この文で「彼に」を「彼は」に変えると不具合が生じる.「彼は借りた本を読んだ.」という文の分析結果を図 2 に示す.

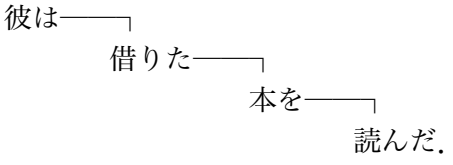

図 2:「彼は借りた本を読んだ.」を KNP ver. 3.01 で分析した結果

この結果では「彼は」が「借りた」に係っていることになっているが，正しくは「彼は」は「読んだ」に係っていると分析する必要がある．KNP ではこの文に限らず係助詞の「は」の後に名詞修飾

節が現れると上の例のように誤って分析される場合がある．この問題に対する解決方法として，我々は係助詞の「は」の後に読点を付与することによって文意を変えずに正しく分析できることを見いだした．そこで，形態素解析によって助詞であると判定された形態素「は」がある場合，その後に読点がなければ読点を挿入した上で構文解析を実行するようにした．上の例の場合,「彼は，借りた本を読んだ.」と変形して KNP で分析する．このような前処理を施した文の分析結果を図 3 に示す．

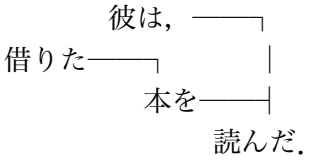

図 3: 「彼は，借りた本を読んだ.」を KNP ver. 3.01 で分析した結果

この結果では,「彼は,」が「読んだ」に係っていると正しく分析されている．また，この前処理が「借りる」や「借りた」以外の動詞を含む名詞修飾節のある文の構文解析にも悪影響を与えないことを確認した．

一方，図 1 などに示した構文木は，人間が視覚的に文の構造を把握するには好都合であるが，コンピュータによる処理には不向きである．そこで，実際の処理においては，係り受け構造のほか各形態素の読みや品詞などの各種情報をリスト形式で表現した図 4 のような「tab 出力」を利用する．この出力形式は KNP に “-tab” というオプションを付与して実行すると得られる．

2.3 主格の有無の判定

構文解析の後，入力文のすべての述語に関して主格の有無を判定する．まず，当該の述語に係るすべての文節を列挙する．そして，いずれかの文節に主格を取り得る助詞が含まれれば，それが主格か否かを判定する処理を行う．この際，述語が動態述語の場合と状態述語の場合では異なる方法を適用する．一方，すべての文節に主格を取り得る助詞が含まれなければ，主格省略と判定する．本研究では，主格を取り得る主な助詞として表 1 に示す 12 種を選択した．これは必要に応じて増やすことができる．

表 1: 主格を取り得る主な助詞のリスト

が	しか	ばかり
ぐらい	だけ	ほど
こそ	でも	まで
さえ	は	も

```
# S-ID:1 KNP:3.01-CF1.0 DATE:2011/11/30 SCORE:-4.11180
* 1D <文頭><ヲ><助詞><体言><一文字漢字><係:ヲ格><区切:0-0><RID:1121><格要素><連用要素><正規化代表表記:本/ほん><主辞代表表記:本/ほん>
+ 1D <文頭><ヲ><助詞><体言><一文字漢字><係:ヲ格><区切:0-0><RID:1121><格要素><連用要素><名詞項候補><先行詞候補><正規化代表表記:本/ほん><解析格:ヲ>
本 ほん 本 名詞 6 普通名詞 1 * 0 * 0 "代表表記:本/ほん 漢字読み:音 カテゴリ:人工物-その他; 抽象物" <代表表記:本/ほん><漢字読み:音><カテゴリ:人工物-その他; 抽象物><正規化代表表記:本/ほん><文頭><漢字><かな漢字><名詞相当語><自立><内容語><タグ単位始><文節始><文節主辞>
を を を 助詞 9 格助詞 1 * 0 * 0 NIL <かな漢字><ひらがな><付属>
* -1D <文末><時制-過去><句点><用言:動><レベル:C><区切:5-5><ID:(文末)><RID:1498><係:文末><提題受:30><主節><格要素><連用要素><正規化代表表記:読む/よむ><主辞代表表記:読む/よむ>
+ -1D <文末><時制-過去><句点><用言:動><レベル:C><区切:5-5><ID:(文末)><RID:1498><係:文末><提題受:30><主節><格要素><連用要素><正規化代表表記:読む/よむ><用言代表表記:読む/よむ><格要素-ガ:NIL><格要素-ヲ:本><格要素-ニ:NIL><格要素-ト:NIL><格要素-デ:NIL><格要素-カラ:NIL><格要素-ヨリ:NIL><格要素-マデ:NIL><格要素-ヘ:NIL><格要素-時間:NIL><格要素-ノ:NIL><格要素-修飾:NIL><格要素-トイウ:NIL><格要素-外の関係:NIL><動態述語><主題格:一人称優位><格関係 0:ヲ:本><格解析結果:読む/よむ:動 2:ガ/U/-/-/-/-; ヲ/C/本/0/0/1; ニ/U/-/-/-/-; ト/U/-/-/-/-; デ/U/-/-/-/-; カラ/U/-/-/-/-; ヨリ/U/-/-/-/-; マデ/U/-/-/-/-; ヘ/U/-/-/-/-; 時間/U/-/-/-/-; ノ/U/-/-/-/-; 修飾/U/-/-/-/-; トイウ/U/-/-/-/-; 外の関係/U/-/-/-/->
読んだ よんだ 読む 動詞 2 * 0 子音動詞マ行 9 タ形 10 "代表表記:読む/よむ" <代表表記:読む/よむ><正規化代表表記:読む/よむ><表現文末><かな漢字><活用語><自立><内容語><タグ単位始><文節始><文節主辞>
。 。 。 特殊 1 句点 1 * 0 * 0 NIL <文末><英記号><記号><付属>
EOS
```

図 4: 「本を読んだ.」を KNP ver. 3.01 で分析した結果 (tab 出力)

2.3.1 一般の動態述語の場合

動態述語は一般のものと可能を表すものとに分けて処理を行う．述語が一般の動態述語の場合，Web 検索エンジンを利用した用例検索に基づいてそれに係っている文節 (名詞節) が主格か否かを判定する．

Google とヤフーは Web 検索エンジンをユーザーのプログラムから利用できる仕組み (application programming interface, API) を提供しており，これを用いることによって検索語や検索文の検索数を知ることができる．本研究ではこの検索数に基づいて判定を行う．ただし，Google，ヤフー両社とも API からの検索の利用回数に制限を設けているため，提案するシステムは無制限に利用できるわけではない．

提案法では，当該の名詞と述語をガ格でつないだ文とヲ格でつないだ文の検索数を比較することによって，当該の述語がその名詞を主格として取り得るか否かを判定する．なお，述語は基本型で検索する．ガ格でつないだ文の検索数は,「“(名詞) が (述語)”」が検索された数，ヲ格でつないだ文の検索数は「“(名詞) を (述語)”」が検索された数とした．検索文をダブルクォーテーションではさんでいるのはフレーズ検索と呼ばれる検索法で，ダブルクォーテーションの間の文字列を形態素に分解せずまとまりとして検索することができる．例えば,「彼が読んだ.」の場合，主格を取り得る助詞とつながっている名詞は「彼」なので,「“彼が読む”」の検索数と「“彼を読む”」の検索数を比較する．ただし,「彼は本を読んだ.」のように述語 (「読んだ」) に表 1 の主格を取りうる格の文節以外にヲ格の文節 (「本を」) が係っている場合，上のような検索数の比較をせずとも「彼は」が主格であると判定することができる．

ガ格でつないだ検索文とヲ格でつないだ検索文の検索数 (それぞれ N_{ga}，N_{wo} とする) が得られた後，以下の規則で当該の名詞が当該の述語の主格になり得るか否かを判定する．

規則 (1) $N_{ga} < \alpha$ かつ $N_{wo} < \alpha$ ならば，判定不能である

規則 (2) $(1+\beta)N_{ga} < N_{wo}$ ならば，主格にならない

規則 (3) $(1-\beta)N_{ga} \geq N_{wo}$ かつ $(1+\beta)N_{ga} \leq N_{wo}$ ならば，主格になる可能性がある

規則 (4) $(1-\beta)N_{ga} > N_{wo}$ ならば，主格になる

ここで，α，β は変数であり，これらを調節することによって判定結果が変化する．なお，β は 1 以下の値をとる．

規則 (1) は検索数が少なすぎる場合に対応するためのものである．インターネット上には誤用も含まれるので，N_{ga} と N_{wo} がともに小さければ検索結果は信頼できない．そこで，そのような場合には判定不能とする．N_{ga} と N_{wo} のいずれかもしくは両方が α を越える場合には規則 (2) 以降の規則を適用して主格になるか否かを判定する．

規則 (2)，(3)，(4) に関しては，図 5 に基づいて説明する．この図の横軸は検索数を表し，これらの規則が適用される N_{ga} と N_{wo} の関係を表している．なお，この図は $N_{ga} \leq \alpha$ を想定している．まず，規則 (2) は，N_{wo} が N_{ga} の $1+\beta$ 倍より大きい場合，すなわち N_{wo} が図 5 における規則 (2) の範囲に含まれる場合には主格にならないと判定することを意味している．ここで単に「$N_{ga} < N_{wo}$ ならば」としていないのは，N_{wo} と N_{ga} が拮抗している場合には信頼性のある判定ができないためである．β の値を大きく設定すればより確実な判定ができるが，その一方で，β の値を大きくしすぎると本来「主格にならない」と判定されるべき場合にこの規則が適用できないという問題が生じる．

規則 (3) では,「主格になる可能性がある」という曖昧な判定をしている．上述のように N_{wo} と N_{ga} が拮抗していれば，信頼性のある判定ができないためである．このような場合はあえて判定を曖昧に

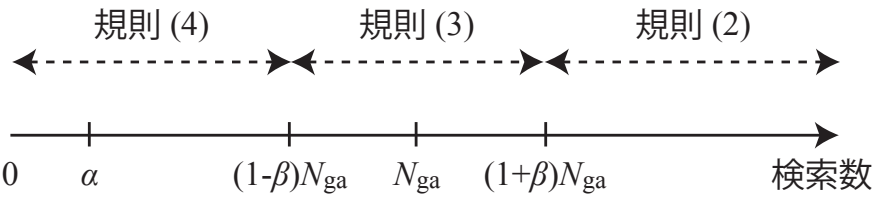

図 5: 判定基準

して最終的な判断は利用者にゆだねるようにしている．ただし，インターネット上では日々テキストデータが増加しているので，時間の経過とともに判定精度の向上が期待できる．

規則 (4) は，N_{wo} が図 5 における規則 (4) の範囲に含まれる場合には，当該の名詞が当該の述語の主格であると判定することを示したものである．

以上の処理で判定不能 (規則 (1)) あるいは主格になる可能性がある (規則 (3)) と判定された場合には，ガ格と述語およびヲ格と述語の間に単語が入る可能性を考慮して改めて検索を行う．このときのガ格の検索文は「“(名詞) が (述語) ”」と「“(名詞) が * (述語)”」，ヲ格の検索文は「“(名詞) を (述語)”」と「“(名詞) を * (述語)”」である．“*” は正規表現のワイルドカードで，Web 検索エンジンを用いた場合には 1 つの単語 (形態素) を意味する．このような検索により，ガ格やヲ格の後に副詞等の語が挿入されているパターンを検索対象に含めることができる．そして，これらの検索文の検索数に対して上記の規則を適用して判定する．それでも規則 (1) や規則 (3) が適用される場合には，助詞を伴う文節等が挿入されている場合を想定して「“(名詞) が ** (述語)”」と「“(名詞) を ** (述語)”」も検索文に追加して上記の処理を繰り返す．検索文には任意の数のワイルドカードを挿入できるが，本研究では挿入するワイルドカードの上限は 2 個とした．

以上の処理を入力文中の各述語に対して，それに係るすべて「主格を取り得る助詞」が含まれるすべての文節に対して実行する．その結果，いずれかの文節が主格になると判定されれば，当該の述語の「主格あり」と判定する．逆に，すべての文節が主格にならないと判定されれば，当該の述語の「主格が省略されている」と判定する．また，すべての文節に関して判定不能の場合には，当該の述語の「主格省略は判定不能」とする．これら以外の場合，すなわち，主格になると判定された文節がなく，かつ主格になる可能性があると判定された文節がある場合には，当該の述語の「主格が文中に存在する可能性がある」と判定する．

2.3.2 可能を表す動態述語の場合

「できる」,「できない」など可能を表す動態述語の場合，2.3.1 節に示した方法では主格の有無を判定できない．なぜなら，これらの動態述語は以下のように表 1 の助詞を含む文節が主格にならない文を作り得るからである．

(1) 「彼は 英語が できる.」

(2) 「英語も 話すことができない.」

ここで，下線部は表 1 の助詞を含むにもかかわらず主格にならない文節である．本研究では，このような場合に，以下のような形で Web 検索エンジンを用いて当該の文節の名詞が主格となるか否かを判定する．検索文は「“(名詞) は * が (述語)”」である．上記の文 (1) を例にとると，それぞれ「“彼は * ができる”」と「“英語は * ができる”」を検索する．そして，その検索数が閾値 γ を越えた名詞はその動態述語の主格であると判定する．閾値 γ の値を小さくするとインターネット上の誤用による判定誤りが生じる可能性が増えるため，10,000 などの比較的大きい値に設定する必要がある．

2.3.3 状態述語の場合

状態述語の場合も，一般の状態述語と後述する例外とを分けて処理する．一般の状態述語の場合，述語に表 1 の主格を取り得る助詞が含まれる文節が係っている場合には「主格あり (主格省略なし)」，係っていない場合には「主格が省略されている」と判定する．例えば,「空が青い.」という文であれば，状態述語「青い」に係る文節「空が」には主格を取り得る動詞「が」が含まれるので「主格あり」と判定される．

一方,「好き」,「嫌い」,「上手」,「下手」などを状態述語に含む文においては上記の規則では不都合が生じる．これらは，以下のように表 1 の助詞を含む文節が主格にならない文を作り得るからである．

(1) 「彼は 英語が 好きだ.」

(2) 「彼は 英語は 嫌いだ.」

(3) 「彼は 英語だけ 上手い.」

(4) 「彼は 英語も 下手だ.」

ここで，下線部は表 1 の助詞を含むにもかかわらず主格にならない文節である．このような場合には 2.3.2 節と同様に Web 検索エンジンを用いて当該の文節の名詞が主格となるか否かを判定する．検索文は「“(名詞) は * が (状態述語)”」である．上記の文 (1) を例にとると,「“彼は * が好き”」と「“英語は * が好き”」を検索する．そして，その検索数が閾値 γ を越えた名詞はその状態述語の主格であると判定する．本研究でこのような処理の対象とした状態述語を表 2 に示す．

表 2: 主格を取り得る主な助詞のリスト

好き	上手	苦手
嫌い	下手	専門
欲しい	得意	

「上手」は「じょうず」と「うま (い)」の双方を対象とした．

上記の状態述語のうち「好き」,「嫌い」,「欲しい」は以下のような文も作り得る.

(1) 「犬が 好きだ.」

(2) 「犬は 嫌いだ.」

(3) 「犬も 欲しい.」

これらの文の下線部が主格か否かを判定するのは非常に難しい．そこで，本研究では「好き」「嫌い」「欲しい」の状態述語に係っている文節が 1 つの場合には,「主格または目的格が省略されている」と判定することにした．このような文に対する主格省略判定は将来の課題とする.

2.3.4 名詞修飾節への対応

名詞修飾節には特別な対応が必要である.「壊れた車」における「壊れた」ものは「車」であるが,「壊した車」における「壊した」ものは「車」ではないからである．前者では主格の有無を考える必要がないが，後者では考える必要がある．本研究ではこの区別にも Web 検索エンジンを利用した.

まず，当該の述語が名詞修飾節に含まれているか否かを判定する．これは述語の係り先が名詞か否かを調べることで容易に実現できる．当該の述語が名詞修飾節に含まれている場合，当該の述語と係り先の名詞で 2.3.1 節と同様にガ格とヲ格の検索文を作り検索数を比較する．そして，2.3.1 節の規則 (1)〜(4) で当該の名詞が主格にならないと判定されれば主格の有無を判定する必要がなく，それ以外に判定されれば主格の有無を判定するようにした.

例えば,「壊れた車」の場合，検索文「車が壊れた」と「車を壊れた」を検索すると前者が多く検索されるので，主格の有無を判定する必要がない名詞修飾節と判定される．一方,「壊した車」の場合,検索文「車が壊した」と「車を壊した」を検索すると後者が多く検索されるので,「壊した」の主格の有無を判定する処理を実行する.

3 処理例

以下にいくつかの文に対する処理結果を示す．α，β，γ の値は，それぞれ 100，0.2，10,000 に設定した.

(1) 「彼は東京に行った.」
→ 主格あり：[彼は] は [行った] の主格になる

(2) 「彼と美しい彼女は東京に行った.」
→ 主格あり：[彼と彼女は] は [行った] の主格になる

(3) 「英語が話せる.」
→ [話せる] の主格は省略されている

(4) 「英語が学びたい.」
→［学びたい］の主格は省略されている

(5) 「英語ができる.」
→［できる］の主格は省略されている

(6) 「彼は英語が好きだ.」
→主格あり：［彼は］は［好きだ］の主格になる

(7) 「本が欲しい.」
→［欲しい］の主格または目的格が省略されている

(8) 「彼は話しながら歩いた.」
→［話しながら］の主格は［歩いた］の主格と同じ
→ 主格あり：［彼は］は［歩いた］の主格になる

文 (1) には係助詞の「は」が含まれるので，その後に読点が付与された形 (「彼は，東京に行った.」) で，構文解析される.「行った」には表 1 に含まれる助詞「は」を含む文節「彼は」が係っているので，その文節が「行った」の主格になるかを判定する必要がある．そこで，ガ格の検索文「彼が行く」の検索数 (N_{ga}) とヲ格の検索文「彼を行く」の検索数 (N_{wo}) を比較する．この処理を行った 2011 年 12 月 19 日においては前者が 2,410 件，後者が 121 件であった．従って，2.3.1 節の規則 (4) の条件に合致し,「［彼は］は［行った］の主格になる」という結果が得られる．

文 (2) は「彼」と「彼女」が並列構造になっている文である．KNP により文中の並列構造を知ることができるので，その情報を利用して「彼と彼女は」が主格であるという結果を得ることができる．

文 (3) は可能を表す文でガ格が主格とならない文である．もし，文 (3) の動態述語をそのままの形で用いてガ格の検索文を作ると「英語が話せる」となる．これらは頻繁に使われる表現なので，多数検索され，これらの文節が主格であると判定されてしまう．しかし，本研究では述語の基本型を用いて検索するようにしているため，ガ格の検索文は「英語が話す」となる．これらはほとんど用いられない表現なので，結果的にヲ格 (「英語を話す」) の検索数がガ格の検索数を大幅に上回り，主格省略と正しく判定される．希望を表す文 (4) についても同様である．

文 (5)，(6)，(7) は 2.3.3 節に述べた手続きにより主格の有無が判定されている．

文 (8) は述語 (「話しながら」) に接続助詞 (「ながら」) が含まれる例である．この文は図 6 のように構文解析されるため,「話しながら」に係る文節がない．そのため，上記の手法ではこの動態述語の主格が省略されていると判定されてしまう．しかし,「話しながら」と「歩いた」の主格はともに「彼は」であると考えられる．そこで，本研究では述語が述語に係る場合，その述語が係っている述語の主格と同じと判定することにした．

4 評価実験

提案法によって得られる結果は，主にパラメータ α，β の値に依存する．そこで，適切な α，β の値を検討するとともに，提案法の性能を評価する．

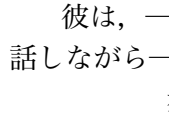

図 6: 「彼は，話しながら歩いた．」を KNP ver. 3.01 で分析した結果

4.1 方法

以下の 10 文書を対象にして人手と提案法による判定結果を比較した．これらの文書はできるだけ多様なジャンル，多様な文体の文書が集まるよう配慮して収集した．今回の実験対象は，以下の文書に含まれていた 195 文である．

1. “新しい公共の世紀へ：市民の力で社会を変える,” 朝日新聞 2011 年 9 月 26 日社説
2. サエキけんぞう, “パソコンが好きだ,” 週刊アスキー 11 月 8 日号，p. 118 (2011)
3. 小田順子, “あとがき,” 公務員の文章・メール術，pp. 212–213 (2011)
4. 著者不明, “飽きのこない食べ物,” 冷え性改善 冷え取り LABO, http://www.kenkolabo.net/hietori/
5. 阿部珠樹, “追悼・スーパークリーク 武豊を G1 ジョッキーへと導いた名馬,” Love Sports web Sportiva, http://sportiva.shueisha.co.jp/ (2010)
6. 著者不明, “異人館からのインフォメーション,” 異人館ネット, http://www.ijinkan.net/
7. 水谷信子, “レッスン 2 地下鉄,” 総合日本語中級前期，p. 26 (1989)
8. のり, “人生が深まるクラシック音楽入門,” Woman.excite, http://woman.excite.co.jp/ (2011)
9. 著者不明, “何のために生きるのか,” 心理コラムの xSUNx, http://www.xsunx.org/ (2007)
10. 芥川龍之介, “蜘蛛の糸,” 青空文庫, http://www.aozora.gr.jp/ (1892)

人による判定では，日本語を母語とする大学生 2 名に対して，述語に印を付けた文書を与え，各述語に関して主格が省略されているか，省略されていない場合には主格はどれかを回答させた．2 名の判定が異なる部分は著者らが議論して決定した．

Web 検索エンジンは 2012 年 1 月 10 日現在のヤフーを利用した．形態素解析の段階での誤りによる悪影響を避けるため，これらの文書に現れる固有名詞については，あらかじめ JUMAN の辞書に追加した．α と β の値は以下の 5 組を設定し，γ は 10,000 に固定した．

1. $\alpha = 10,\ \beta = 0.25$
2. $\alpha = 100,\ \beta = 0.25$
3. $\alpha = 500,\ \beta = 0.25$
4. $\alpha = 100,\ \beta = 0.0$
5. $\alpha = 100,\ \beta = 0.5$

4.2 結果

人が「主格が省略されている」と判定した述語に関する提案法の判定結果を表 3 に示す．また,「主格あり」と判定された述語に関する結果を表 4 に示す.

表 3 に示した結果において α，β の値による違いは見られなかった．これは，この評価実験においては N_{ga} と N_{wo} の差が大きく，β の値が影響しなかったことによる．また，判定不能となった場合の検索数は 0 や 1 が多く，α が 10 以上では結果に影響しなかった.

主格省略の判定が人と一致したのは 84.6 %であり，判定不能であったのは 10.3 %であった．判定不能となったのは,「スーパークリーク」(馬の名前) などの固有名詞が現れる場合が多い．固有名詞と述語と組み合わせた検索文は検索数が少なくなる傾向があり，提案法では判定不能となってしまう．また，人が主格省略と判定したにもかかわらず提案法により主格ありと判定された述語は 5.1 %であった.「人間が見る」,「人間を見る」のようにガ格文とヲ格文がともに頻繁に使用される組み合わせでは誤判定が生じることがあった.

表 4 に示した結果において α，β の値による違いは見られなかった．これは，この評価実験においては N_{ga} と N_{wo} の差が大きく，β の値が影響しなかったことによる．また，判定不能となった場合の検索数は 0 や 1 が多く，α が 10 以上では結果に影響しなかった.

提案法により主格ありと判定された結果のうち主格が人と一致しなかったのはわずか 1 件 (0.7 %) であり，それ以外では主格の有無ばかりでなく，主格がどれかまで正確に判定できた.

以上の結果から提案法はおおむね有効に機能すると結論できる．また，この評価実験からは α は 10 と設定するのが適当で β は結果に影響しないという結論を得たが，β を 0 と設定すると N_{ga} と N_{wo} が僅差のときに不都合が生じるため 0.25 程度に設定しておくのがよいと考えられる.

表 3: 人により主格が省略されていると判定された述語 (39 個) に関する実験結果

	提案法による結果			
	判定不能	主格省略	可能性あり	主格あり
$\alpha = 10, \beta = 0.25$	10.3%	84.6%	0%	5.1%
$\alpha = 100, \beta = 0.25$	10.3%	84.6%	0%	5.1%
$\alpha = 500, \beta = 0.25$	10.3%	84.6%	0%	5.1%
$\alpha = 100, \beta = 0.0$	10.3%	84.6%	0%	5.1%
$\alpha = 100, \beta = 0.5$	10.3%	84.6%	0%	5.1%

表 4: 人により主格ありと判定された述語 (137 個) に関する実験結果

	提案法による結果			
	判定不能	主格省略	可能性あり	主格あり
$\alpha = 10, \beta = 0.25$	17.8%	4.4%	6.6%	72.3%
$\alpha = 100, \beta = 0.25$	18.2%	4.4%	6.6%	70.8%
$\alpha = 500, \beta = 0.25$	20.4%	4.4%	6.6%	68.6%
$\alpha = 100, \beta = 0.0$	18.2%	4.4%	6.6%	70.8%
$\alpha = 100, \beta = 0.5$	18.2%	4.4%	6.6%	70.8%

5 おわりに

本研究では，構文解析技術と Web 検索エンジンを利用して主格省略判定を行う技術を提案した．提案法は，入力文中の述語とそれに係る名詞を組み合わせて検索文を作り，その文が Web 検索エンジンにて検索される数を利用して主格の有無を判定する．

提案法の評価においては，様々なジャンルから選択した 10 文書を対象にして人による判定との比較を行った．その結果，提案法は人が主格省略と判定した述語のうち 84.6 %を検出でき，人が主格ありと判定した述語の 72.3 %を検出できた．さらに，人が主格ありと判定した述語の 71.6 %に関しては，主格の文節も正確に判定することができた．本研究は，日本語教育分野が培ってきた日本語に関する知識を情報技術と融合することによって，有用な支援システムが構築できることを示したものといえる．

インターネット上のテキストデータは日々増加していく．したがって，提案法の精度も日々向上していくものと期待できる．しかし，インターネットスラングや誤用の影響を受けてしまう可能性があることは否定できない．この問題に関しては，提案法の開発当初に，検索対象を新聞社の Web ページに限定することによって対処を試みた．しかし，新聞社の Web ページに現れる文は用いられる単語に制限があったり，言い回しに偏りがあったりするため，新聞記事以外の文書が入力された場合主格の有無を判定できないというケースが続出した．そのため，現在は検索対象の制限は行っていない．

また，提案法を支援システムとしてより有用なものにするためには，日本語非母語話者にとって理解が難しい主格省略のパターンを明らかにする必要がある．主格省略のすべてが難しいわけではなく，あらゆる述語に対して主格を補う必要はないはずである．この課題に関しては日本語教育分野からの提案を期待する．

謝辞

本研究の一部は平成 23 年度科学研究費補助金 (21320095) 及び私立大学等経常費補助金の支援を得て行われた．

参考文献

[1] 佐藤和之, “市民が支える外国人への情報伝達と「やさしい日本語」化支援,” 地方自治職員研修, vol. 44, no. 9, pp. 49–257, 2011.

[2] 田中英輝, 美野秀弥, “「やさしい日本語」ニュースの理解度テスト：ニュースのための「やさしい日本語」の設計に向けて,” 電子情報通信学会技術報告, vol. 111, no. 228, NLC2011-22, pp. 1–6, 2011.

[3] 川村よし子, 前田ジョイス, 保原麗, 川村ヒサオ, “文章の難易度判定システム構築のための基礎調査,” ヨーロッパ日本語教育, vol. 15, pp. 171–178, 2011.

[4] M. Yamura-Takei, M. Fujisawa, M. Yoshie, and T. Aizawa, “Automatic linguistic analysis for language teachers. The case of zeros,” in *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 1114–1120, 2002.

[5] 竹井光子, 相沢輝昭, 藤沢美保, “読解支援システムへの認知的・第二言語習得理論的アプローチ：ゼロ代名詞による結束性,” 教育システム情報学会誌, vol. 21, no. 3, pp. 205–213, 2004.

[6] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦, 日本語語彙大系. 岩波書店, 東京, 1997.

[7] 水嶋博志, 内田聖也, 北村達也, 川村よし子, “学習者にとって難解な構文の自動検出,” 日本語教育方法研究会誌, vol. 18, no. 1, pp. 64–65, 2011.

[8] 内田聖也, 北村達也, 川村よし子, “文章難易度に寄与する構文の自動検出システムの開発,” 2011 年度日本語教育学会春季大会予稿集, pp. 289–290, 2011.

[9] 中村慶太, 北村達也, 川村よし子, “検索エンジンを用いた主格省略文の自動判定,” 日本語教育方法研究会誌, vol. 19, no. 1, pp. 4–5, 2012.

[10] 黒橋禎夫, “結構やるな，KNP,” 情報処理, vol. 41, no. 11, pp. 1215–1220, 2000.