# 構文解析器を利用した主語・述語抽出の検討

川村, 晃市

(Citation)
国際文化学,31:51-64

(Issue Date)
2018-03-20

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81010137

(URL)
https://hdl.handle.net/20.500.14094/81010137

# 構文解析器を利用した主語・述語抽出の検討

# Examining a Subject and Predicate Extraction Using Parsers

川村　晃市
KAWAMURA Koichi

**概要**

初級学習者の英語作文には多くの誤用が含まれる。誤用の中でも、特に global error は文の全体構造に影響して内容の理解を阻害するため、添削の際に見落とすことはできない。しかし、教師が global error を検出するには誤用文の文構造および学習者の意図を理解する必要があり、相当な労力を要する。そのため教師の負担を軽減する添削支援ツールの必要性は高いと考えられる。従来の添削支援ツールの多くは英語作文のみを対象としており、学習者の意図を考慮しておらず、global error に十分対処できていない。それゆえに日本語についても対象とすべきであると考えられる。そこで本研究では、global error も検出可能なシステムの構築を目指し、その誤用判定基準に使用する主語および述語（動詞）を構文解析器による依存構造情報に独自の抽出条件を適用してどの程度正確に抽出できるのか検証を行った。その結果、抽出の正確性は高いこと、また選定方法に条件を加えることで抽出の正確性がより高くなる可能性があることが判明した。これらの結果を踏まえると、今後、依存構造情報を利用して日本語と英語の主語と述語（動詞）を自動抽出し、比較することで英語作文の正誤判定を容易に行うことが可能になると考えられる。

**キーワード**

構文解析器、主語、述語、英語、日本語

## I はじめに

近年、コンピュータ技術の発達に伴って、膨大な言語資料を容易に分析することができるようになり、これまでの手作業による研究では気付くことのできなかった情報を抽出することが可能となった。外国語教育の分野では、学習者コーパスのような学習者の産出言語データの分析により、単語頻度情報、共起情報、過剰使用・過少使用といった特徴語情報などが明らかとなり、学習者の言語習得過程が解明されつつある。また、情報科学の分野では、

対訳コーパスのような対応性を持った異なる二言語のデータを形態素解析、構文解析、意味解析、文脈解析といった自然言語処理技術を利用して解析し、スペル訂正、用例検索、誤用検出、文章校正、機械翻訳などを取り込んだ学習支援システムが開発されている。

このようにコンピュータ技術の発達により、学習者の言語習得に関わる問題の解明や言語学習支援システムの開発が行われてきたにもかかわらず、外国語教育の現場において学習支援システムが十分に活用されているとは言い難い状況がある。その原因として、従来の添削支援ツールの多くは英語作文のみを対象としており、学習者の意図を考慮できず、日本語から英語に文構造を正しく変換できない学習者の誤用を見落としてしまう可能性が高いことなどが考えられる。それゆえに英語のみならず日本語についても対象とする必要性があると言える。そこで本研究では、学習者の global error も検出可能なシステムの構築を目指し、その誤用判定基準に使用する主語および述語（動詞）を構文解析器による解析結果に独自の抽出条件を適用してどの程度正確に抽出できるのか検証を行った。具体的には、初級英語学習者の産出した原言語である日本語と目標言語である英語の各作文から人手により判定した主語および述語（動詞）と、構文解析器（日本語：Cabocha, 英語：Stanford Parser）による依存構造情報から独自の抽出条件により選定した主語候補および述語（動詞）候補を比較し、どの程度正確に主語および述語が抽出できるのか検証を行った。

## II　先行研究

言語を習得する上で、その言語の文型、語順、形態的特徴といった文法的特徴を理解することは重要である。しかしながら、原言語と目標言語の言語間距離が大きい場合、言語の文法的特徴の不理解から誤用が生じることが多い。本研究では日本人英語学習者の英語基本文構造である主語や述語動詞に着目した誤用研究について、応用言語学と自然言語処理の二つの視点から概観する。

応用言語学の立場で分析した研究として、坂内・佐々木（2005）、水品・麻柄（2007）、野地（2008）、工藤（2009）、山内・内田（2011）、Trent（2012）が挙げられる。坂内・佐々木（2005）は、日本人高校生のもつ主語名詞句と述語動詞間の一致の知識について、主語が三人称の単数形、かつ動詞が現在形である場合に動詞に付与される形態素“-s”（いわゆる三単現の“-s”）の誤用の観点から調査した。その結果、主語と動詞との間に入れられる副詞や関係節の位置は形態素“-s”の使用または脱落に影響することを明らかにした。水品・麻柄（2007）は、日本人中高生の英語の主語把握の誤りについて、母語である日本語の知識が干渉しているのか調査したところ、多くの中高生が母語干渉によって日本語の「～は」をそのまま英語の主語として誤用する実態が明らかとなった。野地（2008）は、日本人英語学習者がおかす英語の目的語格標示の誤用の原因は母語である日本語主格目的語構文の知識にあるのか検証した。その結果、日本語の主格目的語構文の知識が英語の格標示に影響しているという明示的証拠はなく、Schwartz（1988）や Schwartz・Sprouse（1994, 1996）の唱える完全転移説とは異なる結論を得た。工藤（2009）は、英語ライティング能力の異なる日本人高校生の global error の特徴について調査し、ライティング能力が高い学習者の誤用は低い学習者と比べて内容的エラーが多く出現すること、またライティング能力が低い学習者の

誤用は高い学習者と比べて主述関係を含めた文全体の構造が不明確な言語的エラーが多く出現することを明らかにした。山内・内田（2011）は、日本人大学生が産出する英語作文の誤用を分析し、母語である日本語の統語的特徴の影響について調査した。その結果、日本語の主題／主格については「X は Y だ」構造(X，Y は名詞）が借用可能になる条件が厳しいこと、また、品詞／述語項構造については「X は（が）＋感情・感覚を表す述語」という構造が品詞や述語項構造で一致していないために借用可能性が低いこと、そして、ゼロ照応については「省略している」という意識をせず、日本語と同様に使用している可能性があることなどを明らかにした。Trent（2012）は、初級レベル大学生の主語のエラーについて分析し、初級学習者は日本語と英語のどちらの言語においても正しい文主語の概念を持っていないと結論づけた。

自然言語処理の観点で行われた研究としては、永田（2013）が挙げられる。永田（2013）は、英語文における主語の人称・数と動詞の人称・数とが一致していない誤りを検出する手法について検討した。構文解析を利用する代わりに品詞解析と句解析を利用し、英語の性質に基づいた規則により主語と動詞の関係を抽出した方が正確さは高くなること、そして、文長が短い学習者の英語文に対して有効であることを確認した。

上述のように、外国語教育の分野では基本的な文構造である主語や述語の理解は目標言語（英語）の正確性に大きく影響を及ぼしていることが明らかとなっており、その知見を教育の現場に生かすために様々な研究がなされている。しかしながら、自然言語処理技術を用いた学習者の主語および述語動詞に関する誤用分析や誤用検出の研究はまだ少なく、原言語情報を取り入れた研究は管見の限りない。そこで、本研究では学習者の基本文構造に関連する誤用を容易に検出するためのシステムを作成することを目指し、その誤用判定基準に使用する主語および述語（動詞）を構文解析器による解析結果に独自の抽出条件を適用してどの程度正確に抽出できるのか調査を行う。

## III　研究の枠組み

### 3.1 本研究の目的

本研究では初級英語学習者の日本語作文データを原言語、それに対応する英語作文データを目標言語として利用する。そして、それぞれの作文データに対して既存の日本語・英語構文解析器（日本語：Cabocha，英語：Stanford Parser）を利用して構文の解析結果を取得し、そこから独自の条件に従って選定した主語候補および述語（動詞）候補が、人手によって抽出した主語および述語（動詞）と比較してどの程度一致するのか、そしてその正確性は文構造部分の誤用検出を行ううえで実用レベルにあるのか検証する。これにより、将来的に構文解析器を利用した新たな英語作文の誤用検出手法の確立を目指す。具体的なリサーチクエスチョンは以下の二点である。

RQ1　日本語作文に対する構文解析器の解析結果から選定した主語候補および述語候補は、人手によって抽出した主語および述語と比較すると、どの程度一致しているのか。

RQ2　英語作文に対する構文解析器の解析結果から選定した主語候補および述語動詞候補は、人手によって抽出した主語および述語と比較すると、どの程度一致しているのか。

簡潔な記述のため、以降、人による判定を人手判定、構文解析器による抽出を自動抽出とする。また、本来、文構造の誤りを検出するには主語・述語に加えて目的語などの成分についても調査する必要があるが、今回は主語・述語（動詞）に限定して調査を行う。

### 3.2 使用する構文解析器と主語候補および述語（動詞）候補の抽出条件

日本語および英語、それぞれの主語候補および述語（動詞）候補については、構文解析器を利用して取得した解析結果から、下記にある独自の条件に従って選定する。

まず、日本語作文に対して構文解析を行い、その解析結果から得た最終文節を述語候補とし、その述語候補と依存関係にある格助詞の「が」、または係助詞の「は」、「も」を含む文節を主語候補としている。手順としては、構文解析器の結果から述語候補とリンクするすべての文節を対象として、これらから「は」、「が」、「も」を含むもののみを抽出する。

つぎに、英語作文に対して構文解析を行い、その解析結果から得た依存構造情報より“nsubj”（nominal subject）で関連付けした二つの単語を抽出すると同時に、これらの単語に関連する“cop”（copula）などの部分も併せて抽出する。ここでは、“nsubj”と“cop”は依存構造を示すタグであり、“nsubj”は、述語と述語にかかる主格の名詞句の依存構造情報を示したものである。これにより、be 動詞や will などの助動詞、否定の表現も併せて抽出される。また、“nsubj”によって関連付けした二つの単語の組み合わせが複数抽出された場合、最初の組み合わせを主語候補と述語動詞候補として選択する。

例文）今日は良い天気です。

```
<sentence>
<chunk id="0" link="2" rel="D" score="-1.137013" head="0" func="1">
<tok id="0" feature="名詞,副詞可能,*,*,*,*,今日,キョウ,キョー">今日</tok>
<tok id="1" feature="助詞,係助詞,*,*,*,*,は,ハ,ワ">は</tok>
</chunk>
<chunk id="1" link="2" rel="D" score="-1.137013" head="2" func="2">
<tok id="2" feature="形容詞,自立,*,*,形容詞・アウオ段,基本形,良い,ヨイ,ヨイ">良い</tok>
</chunk>
<chunk id="2" link="-1" rel="D" score="0.000000" head="3" func="4">
<tok id="3" feature="名詞,一般,*,*,*,*,天気,テンキ,テンキ">天気</tok>
<tok id="4" feature="助動詞,*,*,*,特殊・デス,基本形,です,デス,デス">です</tok>
</chunk>
</sentence>
```

図 1　Cabocha による構文解析結果例

| 例文）It is fine today. |
| --- |
| ((u'fine', u'JJ'), u'nsubj', (u'It', u'PRP'))<br>((u'fine', u'JJ'), u'cop', (u'is', u'VBZ'))<br>((u'fine', u'JJ'), u'nmod:tmod', (u'today', u'NN')) |

図 2　Stanford Parser による構文解析結果例

本研究では、Cabocha 0.69（Kudo and Matsumoto, 2003）と Stanford Parser 3.6.0（Klein and Manning, 2003）をそれぞれ日本語と英語の構文解析器として使用する。また、一文は一句点までの文とし、抽出は一文毎に行う。

ここでは、日本語構文解析器 Cabocha および英語構文解析器 Stanford Parser を利用した主語候補と述語（動詞）候補の抽出について、日本語作文「今日は良い天気です。」と英語作文“It is fine today.”を例として示す。

まず、日本語作文からの主語候補および述語候補の抽出では、日本語作文「今日は良い天気です。」を構文解析器 Cabocha によって解析し、その依存構造情報を得る（図 1）。そして、最終文節、ここでは最も大きい文節番号（chunk id = 2）を持つ「天気です」を述語候補として抽出し、その文節番号（chunk id = 2）を係り先の文節番号（link = 2）としている文節「今日は」と「良い」のうち、係助詞の「は」を含む文節「今日は」を主語候補として抽出する。

つぎに、英語作文からの主語候補および述語動詞候補の抽出では、英語作文“It is fine today.”を構文解析器 Stanford Parser によって解析し、その依存構造情報を得る（図 2）。まず、“nsubj”によって関連付けされた“fine”と“It”を抽出し、指示代名詞（PRP）の“It”を主語候補、“fine”を述語動詞候補の一部として抽出する。そして“cop”として表示された“is”と併せた“is fine”を述語動詞候補として抽出する。

### 3.3 人手判定による主語および述語（動詞）の判定方法

#### 3.3.1 主語および述語（動詞）の判定基準

日本語作文中の主語および述語と英語作文中の主語および述語動詞を人手判定する際には、それぞれの定義を以下のように定める。

日本語の主語に関しては、主語とは何か、ということについて様々な学説が存在し、主語の定義には多くの問題が存在する。そこで本研究では、学校文法に従い格助詞の「が」、係助詞「は」（複合助詞「には」を含む）と並立助詞「も」を伴った文節を主語と定義する。ただし、一文節では意味が不明瞭である場合、意味をなす最小限の複数の文節の集合を主語とする。例えば「周りの人の方が」などが該当する。また、「～することは」や「～するのが」などの動詞を名詞化する表現についても、動詞を含む複数の文節の集合を主語とする「優先すべきことは」などが該当する。本研究において助詞を三種類に限定した理由は、「が」、「は」、「も」は主格の役割をする助詞のうち高頻度で使用されている助詞である（花岡，2012）からである。留意点として、格助詞「が」の用法には主体と対象があるが、

対象として使用される場合（ガ格名詞句）には主語とは認めないこととする。また、主語が省略されている場合、無主語が存在すると考える。よって本研究で扱う日本語の主語は「が」、「は」、「も」を伴った一文節または複数の文節、および無主語の四種類となる。また構造的な観点から主語をみると、節が一組ある文、および節が二組以上ありそれらの節が主従の関係で構成された文においては、主語は一つまたは存在しない。節が二組以上あり、それらが対等の資格で並列されて構成された文においては、主語が複数存在する場合もある。

日本語の述語に関しては、文末に置かれる動作や状態を表す用言（動詞，形容詞）＋補助用言＋コピュラ、および名詞＋コピュラを述語とする。形容動詞＋補助用言＋コピュラも述語と考える説もあるが、本研究では時枝（1950）の文法に従って形容動詞を品詞として扱わず、体言＋助動詞（「である」、「だ」、「なり」）とする。このことによりコピュラは全て助動詞となる。また、主語と同様に述語も意味をなす最小限の複数の文節の集合体とする。例えば、「かじっているわけにはいきません」、「悪いものです」などが該当する。

英語の主語および述語動詞に関しては、状態や動作を表す語および心的態度を表す語を述語動詞とし、その述語動詞に対応する動作主（名詞、代名詞、名詞句および名詞節）を主語とする。

### 3.3.2 主語および述語（動詞）の判定方法

日本語の主語および述語の判定と英語の主語および述語動詞の判定は、前述の判定基準に沿って行う。

人手判定による主語および述語（動詞）抽出にあたっては、学習者の誤用による影響や文構造上の理由により主語や述語（動詞）が省略されている場合や、複数存在する場合があるため、以下の抽出条件を定めておく。

1. 主語と述語は一組だけ抽出する。
   本来、二つ以上の節が対等の資格で並列されて構成された文、いわゆる重文である場合、節の数だけ主語と述語が存在する可能性があるが、本研究では他の文の種類と同様に主語と述語を一組だけ抽出する。
2. 節が一組あるタイプ I、節が二組以上あるタイプ II のいずれの場合においても、
   ・主語が省略されている場合、述語のみ選択する。（日本語）
   ・誤用により主語および述語動詞の候補が複数ある場合、文意を考慮して最も適切と考えられる組み合わせを選択する。（英語）

なお、文の種類は単文、複文、そして重文の三種類に分類する考え方（橋本，1939）が一般的であるが、本研究では、文の種類をタイプ I とタイプ II の二種類に分類して分析を行う（表 1）。なお、ここでいう節とは「主語・述語の関係を含んで文の一部となる連文節」（山口ほか，2001）を指し、文節とは異なる。また、「明日、東京に行く。」のように主語が省略されたものも節として扱う。

表 1　本研究で採用した文の種類（構造上の分類）

| 橋本(1939)による文の種類 | | | 本研究での文の種類 | |
|---|---|---|---|---|
| 種類 | 条件 | | 種類 | 条件 |
| 単文 | 主語・述語の関係が一回しか成立しない文 | → | タイプ I | 節が一組ある文 |
| 複文 | 主語・述語の関係が成り立っている文の中に、更に別の主語・述語の関係が認められる部分（従属節）を含む文 | → | タイプ II | 節が二組以上ある文 |
| 重文 | 二つ以上の節が対等の資格で並列されて構成された文 | → | | |

### 3.4 使用するデータ

本研究で使用する作文データは、関東圏私立女子大学人間生活学部に在籍する日本人大学生の一回生 68 名に対して収集された。大学生の英語習熟度はヨーロッパ言語共通参照枠（CEFR）の A1（Breakthrough）であり、初級レベル学習者に相当する。

使用する作文データの執筆条件として、日本語作文では、プロンプト A :「大学生にとってアルバイトをすることは重要である。」、プロンプト B :「国内では全てのレストランで喫煙が完全に禁止されるべきだ。」のどちらかを選択し、文字数は 400 字、制限時間は 30 – 40 分、辞書の使用は不可とした。また、英語作文では日本語作文で選択したプロンプトを翻訳し、制限時間は 50 – 60 分、辞書の使用は不可とした。

日本語作文データの収集方法は、国立国語研究所の作文対訳 DB のように日本語学習者に目標言語である日本語作文を要求し、その作文を執筆者に母語で翻訳させ原言語のデータを採取するという手順を踏むと、学習者は問題に直面すると誤りを犯しそうな構造や語彙を避けるコミュニケーション・ストラテジーをとる（Tarone, 1980）可能性がある。この可能性を排除するため、本研究では日本人英語学習者にまず原言語である日本語による作文を要求し、その後、その日本語作文を目標言語である英語に翻訳するという順序で作業を課した。留意点として、日本語作文を書く際には後に英語に翻訳することを告知していない。

表 2　分析用作文データ一覧

| | | | |
|---|---|---|---|
| A | 日本人英語学習者の主語がある日本語作文 | 原言語 | 100 文 |
| B | 日本人英語学習者の主語がない日本語作文 | 原言語 | 100 文 |
| C | 日本人英語学習者の主語がある日本語作文に対応する英語作文 | 目標言語 | 100 文 |
| D | 日本人英語学習者の主語がない日本語作文に対応する英語作文 | 目標言語 | 100 文 |

今回の分析では、データとして収集した日本語作文・英語作文は各 110 エッセイ、計 1499 文あり、そのうち、主語が存在する日本語作文 100 文とそれに対応する英語作文 100 文、主語が存在しない日本語作文 100 文とそれに対応する英語作文 100 文の合計四種類 400 文を任意に取り出して分析に使用した。作文データの一覧は上表に示した通りである。

### 3.5 分析の手順

分析はリサーチクエスチョンに従って行う。

RQ1 については、まず日本語作文 A および日本語作文 B から人手判定により、それぞれ主語および述語を抽出する。次に日本語構文解析器 Cabocha を利用し、前述の抽出条件に従って選定した主語候補および述語候補を抽出する。人手判定の結果と自動抽出の結果の比較は本研究で設定した文の種類（I および II）ごとに行い、それぞれどの程度一致しているか確認する。

RQ2 についても、RQ1 と同様に英語作文 C および英語作文 D から人手判定により主語および述語動詞を抽出する。次に英語構文解析器 Stanford Parser を利用し、前述の抽出条件に従って選定した主語候補および述語動詞候補を抽出する。人手判定の結果と自動抽出の結果の比較は本研究で設定した文の種類（I および II）ごとに行い、それぞれどの程度一致しているか確認する。

## IV　結果と考察

### 4.1 RQ1　日本語作文の主語・述語判定

#### 4.1.1 RQ1　結果

表 3、4 は、それぞれ主語がある日本語作文 100 文 A と主語がない日本語作文 100 文 B について、主語候補と述語候補の抽出結果をまとめたものである。

表 3、4 では構文解析器 Cabocha を使用し、自動抽出によって得た主語候補および述語候補と人手判定による主語および述語を比較して集計した文の数を示している。ここでいう一致とは、自動抽出の候補と人手判定で抽出した主語および述語が全く同じ場合の完全一致と一部分だけ（文字単位ではなく文節単位で判定）が同じ場合の部分一致を指す。また、一致する部分がない場合は不一致とする。一文に人手判定による主語および述語は一つだけであるが、構文解析器による自動抽出では主語候補に限って複数抽出される場合があるため、主語のある日本語作文は抽出された主語候補が、ない、一つ、そして複数の三分類（0, 1, 複数）、主語のない日本語作文は抽出された主語候補が、ない、と一つ以上の二分類（0, 1 以上）としている。なお、傍線（—）は該当する項目がないことを示す。

表 3 より、主語がある日本語作文 A では節が一組あるタイプ I と節が二組以上あるタイプ II の比率がおおよそ 1 対 1 であることがわかる。主語候補と述語候補の抽出結果については個別に見ていく。まず主語候補について、構文解析器による自動抽出と人手判定による結果でもっとも厳しい基準である完全一致の割合は、タイプ I、タイプ II それぞれ 63.3%（31 / 49）、62.7%（32 / 51）、合わせて 64.0%であった。次に述語候補について、同様の基準で自動抽出と人手判定による結果が完全一致した割合をみると、それぞれ 87.8%（43 / 49）、84.3（43 / 51）、合わせて 86.0%であった。また主語候補と述語候補の抽出結果について併せてみると、タイプ I の完全一致の割合は 59.2%（29 / 49）、タイプ II の完全一致の割合は 54.9%（28 / 51）であった。

表 3　日本語作文 A における主語候補・述語候補の分類結果と各分類の件数

| | | 主語候補 | | | | 述語候補 | | | | 主語候補・述語候補 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 文の種類 | | 抽出タイプ | | マッチタイプ | | 抽出タイプ | | マッチタイプ | | マッチタイプ | |
| I | 49 | 1 | 38 | 完全一致 | 31 | 1 | 49 | 完全一致 | 43 | 完全一致 | 29 |
| | | | | 部分一致 | 4 | | | 部分一致 | 6 | | |
| | | | | 不一致 | 3 | | | 不一致 | 0 | | |
| | | 複数 | 10 | 完全一致 | 10 | — | — | — | — | | |
| | | | | 部分一致 | 0 | | | — | — | | |
| | | | | 不一致 | 0 | | | — | — | | |
| | | 0 | 1 | — | — | 0 | 0 | — | — | | |
| II | 51 | 1 | 37 | 完全一致 | 32 | 1 | 51 | 完全一致 | 43 | 完全一致 | 28 |
| | | | | 部分一致 | 3 | | | 部分一致 | 8 | | |
| | | | | 不一致 | 2 | | | 不一致 | 0 | | |
| | | 複数 | 8 | 完全一致 | 8 | — | — | — | — | | |
| | | | | 部分一致 | 0 | | | — | — | | |
| | | | | 不一致 | 0 | | | — | — | | |
| | | 0 | 6 | — | — | 0 | 0 | — | — | | |

表 4　日本語作文 B における主語候補・述語候補の分類結果と各分類の件数

| | | 主語候補 | | 述語候補 | | | | 主語候補・述語候補 | |
|---|---|---|---|---|---|---|---|---|---|
| 文の種類 | | 抽出タイプ | | 抽出タイプ | | マッチタイプ | | マッチタイプ | |
| I | 13 | 1 以上 | 7 | 1 | 13 | 完全一致 | 8 | 完全一致 | 6 |
| | | | | | | 部分一致 | 5 | | |
| | | | | | | 不一致 | 0 | | |
| | | 0 | 6 | 0 | 0 | — | — | | |
| II | 87 | 1 以上 | 18 | 1 | 87 | 完全一致 | 75 | 完全一致 | 66 |
| | | | | | | 部分一致 | 12 | | |
| | | | | | | 不一致 | 0 | | |
| | | 0 | 69 | 0 | 0 | — | — | | |

次に、表 4 より、主語がない日本語作文 B では、節が一組あるタイプ I と節が二組以上あるタイプ II の比率がおおよそ 1 対 9 であることがわかる。主語候補と述語候補の抽出結果について個別に見ていく。まず主語候補について、構文解析器による自動抽出と人手判定による結果は、それぞれ 46.2%（6 / 13）、79.3%（69 / 87）、合わせて 75.0%であった。

次に述語候補について、もっとも厳しい基準である完全に一致した割合をみると、61.5%（8 / 13）、86.2%（75 / 87）、合わせて 83.0%であった。また主語候補と述語候補の抽出結果について併せてみると、タイプ I の完全一致の割合は 46.2%（6 / 13）、75.9%（66 / 87）であった。

#### 4.1.2 RQ1　考察

以上の結果から、構文解析器の抽出の正確さは、主語のある日本語作文では文の種類によってほとんど影響を受けないこと、そして主語のない日本語作文では文の種類、特に節が一組みである文において正確さが低下することが明らかとなった。

主語がある日本語作文 A では、自動抽出された述語候補は 86.0%の正確さであったのに対し、主語候補は 54.9%の正確さにとどまった。主な原因は、主語の候補を複数抽出している文が 18 文（A では 10 文，B では 8 文）も存在していることによる。この 18 文は全て、複数の主語候補の中に人手判定と同じ主語が包含されており、構文解析器の主語抽出結果に何らかの条件を付加することで主語の候補を減らし、正確さを高めることができると考えられる。具体的には、主語候補の中に「最近では」、「レストランには」、「（〜の）時は」などのいわゆる副詞のように使用されている語が含まれており、それらを抽出しないような条件を付与することが考えられる。また、節が二組以上あり対等の資格で並列されて構成された文、いわゆる重文において、主語候補の中からどのようにして正しい複数の主語を抽出するのか、ということも今後の改善すべき課題と言える。

主語がない日本語作文 B においては、述語候補は 83.0%、主語候補は 75.3%の正確さで自動抽出できることが明らかとなった。改善すべき点としては、誤って主語として候補を抽出した 25 文中 10 文において、主語と述語の判定基準で示した「格助詞「が」が対象として使用される場合（ガ格名詞句）には主語とは認めないこととする。」という条件に当てはまるものを主語候補として抽出しており、それらを抽出しないように条件を付加することでさらに正確さを高めることが可能と考えられる。

以上のことより、本研究で収集したような日本語作文に対して、構文解析器を直接利用し、基本文構造である主語や述語を抽出することの妥当性は高いと考えられる。

### 4.2 RQ2　英語作文の主語・述語動詞判定

#### 4.2.1 RQ2　結果

表 5、6 は、それぞれ主語がある日本語作文に対応する英語作文 100 文 C と主語がない日本語作文に対応する英語作文 100 文 D について、主語候補と述語動詞候補の抽出結果をまとめたものである。

表 5、6 は構文解析器 Stanford Parser を使用し、自動抽出によって得た主語候補および述語動詞候補と人手判定による主語および述語動詞を比較して集計した文の数を示している。RQ1 と同様に、ここでいう一致とは、自動抽出の候補と人手判定で抽出した主語および述語が全く同じ場合の完全一致と一部分だけ（文字単位ではなく単語単位で判定）が同じ場合の部分一致を指す。また、一致する部分がない場合は不一致とする。英語作文の構文解析器による自動抽出では複数の語が抽出されないように基準を定めたため、英語作文は抽

出された候補が、ないと一つの二分類（0，1）としている。なお、傍線（—）は該当する項目がないことを示す。

表 5　英語作文 C における主語候補・述語動詞候補の分類結果と各分類の件数

| | | 主語候補 | | | | 述語候補 | | | | 主語候補・述語候補 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 文の種類 | | 抽出タイプ | | マッチタイプ | | 抽出タイプ | | マッチタイプ | | マッチタイプ | |
| I | 71 | 1 | 63 | 完全一致 | 58 | 1 | 65 | 完全一致 | 55 | 完全一致 | 51 |
| | | | | 部分一致 | 2 | | | 部分一致 | 5 | | |
| | | | | 不一致 | 3 | | | 不一致 | 5 | | |
| | | 0 | 8 | — | — | 0 | 6 | — | — | | |
| II | 29 | 1 | 29 | 完全一致 | 25 | 1 | 29 | 完全一致 | 23 | 完全一致 | 22 |
| | | | | 部分一致 | 0 | | | 部分一致 | 1 | | |
| | | | | 不一致 | 4 | | | 不一致 | 5 | | |
| | | 0 | 0 | — | — | 0 | 0 | — | — | | |

表 6　英語作文 D における主語候補・述語動詞候補の分類結果と各分類の件数

| | | 主語候補 | | | | 述語候補 | | | | 主語候補・述語候補 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 文の種類 | | 抽出タイプ | | マッチタイプ | | 抽出タイプ | | マッチタイプ | | マッチタイプ | |
| I | 40 | 1 | 37 | 完全一致 | 36 | 1 | 40 | 完全一致 | 29 | 完全一致 | 28 |
| | | | | 部分一致 | 0 | | | 部分一致 | 3 | | |
| | | | | 不一致 | 1 | | | 不一致 | 8 | | |
| | | 0 | 3 | — | — | 0 | 0 | — | — | | |
| II | 60 | 1 | 57 | 完全一致 | 54 | 1 | 60 | 完全一致 | 48 | 完全一致 | 48 |
| | | | | 部分一致 | 0 | | | 部分一致 | 3 | | |
| | | | | 不一致 | 3 | | | 不一致 | 9 | | |
| | | 0 | 3 | — | — | 0 | 0 | — | — | | |

表 5 より、主語がある日本語作文に対応する英語作文 C では節が一組あるタイプ I と節が二組以上あるタイプ II の比率がおおよそ 7 対 3 であることがわかる。主語候補と述語動詞候補の抽出結果について個別に見ていく、まず主語候補について、構文解析器による自動抽出と人手判定による結果でもっとも厳しい基準である完全一致の割合は、それぞれ 81.7%（58 / 71）、86.2%（25 / 29）、合わせて 83.0%であった。次に述語動詞候補について、同様の基準で自動抽出と人手判定による結果が完全一致した割合をみると、それぞれ 77.5%（55 / 71）、79.3%（23 / 29）、合わせて 78.0%であった。また主語候補と述語動詞候補の

抽出結果について合わせてみると、タイプ I の完全一致の割合は 71.8%（51 / 71）、タイプ II の完全一致の割合は 75.9%（22 / 29）であった。

次に、表 6 より、主語がない日本語作文に対応する英語作文 D では、節が一組あるタイプ I と節が二組以上あるタイプ II の比率がおおよそ 2 対 3 であることがわかる。まず主語候補の抽出結果について個別に見ていく。まず構文解析器による自動抽出と人手判定による結果をもっとも厳しい基準である完全に一致した割合をみると、それぞれ 90.0%（36 / 40）、90.0%（54 / 60）、合わせて 90.0%であった。次に述語動詞候補について、同様の基準で自動抽出と人手判定による結果が完全一致した割合をみると、それぞれ 72.5%（29 / 40）、80.0%（48 / 60）、合わせて 77.0%であった。また主語候補と述語動詞候補の抽出結果について合わせてみると、タイプ I の完全一致の割合は 70.0%（28 / 40）、タイプ II の完全一致の割合は 80.0%（48 / 60）であった。

#### 4.2.2 RQ2　考察

以上の結果より、主語がある日本語作文に対応する英語作文 C では、自動抽出された主語候補は 83.0%、述語動詞候補は 78.0%の正確さであることが判明した。主語候補と比較して述語動詞候補の正確さが低い原因として、学習者の述語動詞の誤用の多さが考えられる。

誤りの例（原文ママ）

・Restaurants is came to many people.
・That do not coming guest at smoker.
・I had be given five thousand moneys once a month by high school student.
・My father smoking cigarette.
・There are don't like human in the future.
・My grandfather is die of fast because he is smoker.

また、複数の動詞を同時に使用、あるいは不適切な形態で使用などの理由により構文解析器では正常に述語動詞候補を抽出できていない。このことは、述語動詞を抽出できないことから述語動詞部分に誤用があると推定できるが、誤用の種類まで推定することは不可能であり、この点に関しては今後の課題と言える。

最後に、主語がない日本語作文に対する英語作文 D では、自動抽出された主語候補は 90.0%の正確さ、述語動詞候補は 77.0%の正確さであり、主語がある日本語作文に対応する英語作文 C と同様、高い正確性を示していることが明らかとなった。

以上のことより、本研究で収集したような日本語作文に対応する誤りを含む英語作文に対して構文解析器を直接利用し、基本文構造である主語や述語動詞をする抽出することの妥当性は高いと考えられる。

## V　おわりに

本研究では、初級英語学習者の産出した日本語と英語の各作文から人手によって抽出した主語および述語（動詞）と日英構文解析器（日本語 : Cabocha，英語 : Stanford Parser）を利用して自動抽出した主語および述語（動詞）を比較し、その正確さについて観察を行った。以下、RQ ごとに得られた結果をまとめる。

RQ1 では、主語がある日本語作文から人手によって抽出した主語および述語と、構文解析器を利用して自動抽出した主語候補および述語候補を比較したところ、主語は 54.9%、述語は 86.0%の正確さで抽出できること、主語がない日本語作文では主語は 75.3%、述語は 83.0%の正確さで抽出できることが明らかとなった。

RQ2 では、主語がある日本語作文に対応する英語作文から人手によって抽出した主語および述語動詞と、構文解析器を利用して自動抽出した主語候補および述語動詞候補を比較したところ、主語は 83.0%、述語は 78.0%の正確さで抽出できること、主語がない日本語作文では主語は 90.0%、述語は 77.0%の正確さで抽出できることが明らかとなった。

本研究では、学習者の誤用検出のために既存の構文解析器を利用して学習者作文から主語および述語（動詞）を抽出することは可能であることが確認できた。さらに、主語および述語（動詞）抽出の正確さを高めるためには、データの前処理として行う英語作文の文の種類（単文、重文、複文）に応じた分類をより厳密に行う必要があることも確認できた。

依存構造情報に独自の抽出条件を適用して、主語と述語（動詞）どの程度正確に抽出できるのか検証した後、主語のある日本語作文 100 文（A）とそれに対応する英語作文 100 文（C）、主語のない日本語作文 100 文（B）とそれに対応する英語作文 100 文（D）の合計 400 文から主語および述語（動詞）を抽出し、それぞれを手動で比較した。その結果、主語のある日本語作文に対応する英語作文では 43 文、主語のない日本語作文に対応する英語作文では 34 文の global error が確認できた。これらの比較を自動で行うことは今後の課題である。

このように構文解析器を利用して日本語と英語の主語と述語（動詞）をそれぞれ比較することにより、教師が学習者の global error を容易に見つけることが可能となる。しかしながら、日本語と英語では言語構造が異なるため、単純な基本文構造の比較のみでは対応できない文も存在する。そこで、日本語を文型に分類し、それに対応する英語の文型を多く生成することにより、より正確性の高い誤用検出が可能だと考えられる。

（神戸大学国際文化学研究科博士後期課程）

## 参照文献

坂内昌徳・佐々木裕美（2005）．「第二言語における主語と動詞の「一致」の知識：日本人英語学習者のデータから」．『研究紀要』，*45*，101-109.

花岡洋輝・増田勝也・植松すみれ・美馬秀樹（2012）．「日本語助詞「と」コーパス構築」ポスターセッション．言語処理学会第 18 回年次大会．広島市立大学.

橋本進吉（1939）．『改制新文典別記』．東京：冨山房.

Klein, D. and Manning, C. D. (2003). Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423-430.

工藤洋路（2009）．「英語ライティング能力のレベルが異なる学習者の Global Error の特徴に関する研究」．『ARCLE REVIEW』，*3*，110-121.

水品江里子・麻柄啓一（2007）．「英文の主語把握の誤りとその修正」．『教育心理学研究』，*55*(4)，573-583.

水谷修・加藤清方・佐久間勝彦・佐々木倫子・西原鈴子・仁田義雄（2005）．『新版日本語教育事典』．日本語教育学会，東京：大修館書店.

永田亮（2013）．「構文解析を必要としない主語動詞一致誤り検出手法」．『電子情報通信学会論文誌．D，情報・システム』，*96*(5)，1346-1355.

野地美幸（2008）．「L2 英語における目的語格標示: 日本人英語学習者の発話コーパス研究」．『上越教育大学研究紀要』, *27*，173-180.

Schwartz, B. D. (1998). The second language instinct. *Lingua*, *106*(1-4), 133-160.

Schwartz, B. D., & Sprouse, R. (1994). Word order and nominative case in nonnative language acquisition: a longitudinal study of (L1 Turkish) German interlanguage. *Language acquisition studies in generative grammar*, *31*(4), 71-89.

Schwartz, B. D., & Sprouse, R. (1996). L2 cognitive states and the full transfer/full access model. Second Language Research 12, 40-72.

鈴木一行（2014）．『日本語文法事典』．日本語文法学会，東京：大修館書店.

Taku Kudo, Yuji Matsumoto (2003). Fast Methods for Kernel-Based Text Analysis, ACL 2003 in Sapporo, Japan.

Tarone, E. (1980). Communication strategies, foreigner talk, and repair in interlanguage. *Language learning*, *30*, 417-431.

時枝誠記（1950）．『日本文法口語篇』．東京：岩波書店.

Trent, N. (2012). The challenge of English sentence subjects (shugo) to Japanese Leaners. *The 2012 Pan-SIG Proceedings*, 187-195.

山口明穂・秋本守英（2001）．『日本語文法大辞典』．東京：明治書院.

山内真理・内田充美（2011）．「日本人英語学習者の中間言語にみられる L1 の痕跡」．『千葉商大紀要』，*49*(1)，43-56.