様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１２　日現在

機関番号：　３２４１４
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３７０５００

研究課題名（和文）依存文法とグラフ理論に基づく日英語文の構造特性分析

研究課題名（英文）Analysis of structural characteristics of Japanese and English sentences based on Dependency Grammar and Graph theory

研究代表者

大矢　政徳（OYA, Masanori）

目白大学・外国語学部・准教授

研究者番号：６０３１８７４８

交付決定額（研究期間全体）：（直接経費）　　900,000円

研究成果の概要（和文）：平成26年度には、既存の日英語対訳コーパス中の日本語及びその英訳文の構文解析を進め、得られた依存木のグラフ中心性を比較し、英語文はそれに対応する日本語文よりも深く埋め込まれた構造を持っている傾向が数値的に示された。
　平成27年度には、日英語の依存木間の構造的不一致の概念を導入し、日英語対訳コーパス中の対訳文間で日本語の格助詞「が」を伴う名詞句が英語でどのように訳されているかに注目し、その不一致パターンを構造的不一致の概念に基づいて分類した。
　平成28年度には、日本人英語学習者が産出した英語テキストと、ネイティヴスピーカーが産出した英語テキストとで、頻出する依存関係に差異があることを指摘した。

研究成果の概要（英文）：In 2014, I have obtained the parse output (dependency trees) of the Japanese sentences and their English translations in a English-Japanese parallel corpus, and compared the graph centrality of these dependency trees. It has been shown as numerical data that English sentences tend to contain more embedded syntactic structure than their Japanese translations.
In 2015, I have applied the notion of structural divergence between the dependency trees of English sentences and those of their Japanese counterparts. I focused on the Japanese noun phrases with the case particle "-ga" which have been found in the same English-Japanese parallel corpus, and on how they are translated in their English counterparts. The divergence pattern have been categorized according to the notion of structural divergence.
In 2016, the frequent dependency types are found to be different between English texts written by Japanese learners of English and those written by native speakers of English.

研究分野：　コーパス言語学、依存文法

キーワード：　日英対訳コーパス　グラフ理論　グラフ中心性　構造的不一致

１．研究開始当初の背景

近年、統語論の分野にて依存文法(dependency grammar; Tesnière 1959, Debusmann and Kuhlmann 2007, Hudson 2010, etc.)が注目されている。これは従来の句構造文法とは異なり、句の存在を想定せずに単語間の依存関係によって統合構造を表現することを目指す文法理論の枠組みである。単語間の依存関係は主要部と従属部として定義付けされ、その関係は主要部と従属部の意味的関係に応じてタイプ分けされる。例えば、“Sarah is running.”では、動詞 running が主要部であり、Sarah はこの主要部に従属しており、この二つの単語間の依存関係は「主語」としてタイプ分けされる。

（１）有向非循環グラフ

このような文中の単語間の依存関係は、各単語を頂点とし、単語間の依存関係を主要部から従属部への向きが定義づけられた辺とする有向グラフ(directed graph)として表現することが可能である。この場合、ある単語から依存関係に沿って移動した場合その単語へと戻ることがないため、このグラフは非循環(acyclic)であるといえる。したがって、文中の単語間の依存関係をグラフで表現した場合、そのグラフは有向非循環グラフ(directed acyclic graph, DAG)である。

（２）次数中心性、近接中心性

さらに、ネットワーク分析の分野では、所与のグラフから特徴量を算出し、そのグラフの構造特性を数値的に把握するという手法が研究されている(de Nooy, Mrvar, Batagelj 2005)。そのような特徴量には、ある頂点に辺が集中している度合を示す次数中心性(degree centrality)、ある頂点から別の頂点へと辿っていく場合にどの程度の数の頂点を通るかを示す近接中心性(closeness centrality)などがある(Freeman 1979)。

（３）先行研究

Oya(2010b)では、英語単文を既存の構文解析アプリケーション(Stanford Parser; de Marneffe et al. 2006)で構文解析し、その構文解析結果からその単文の単語間依存構造の次数中心性と近接中心性を自作のコンピュータプログラムで計算する手法を紹介した。これによって、手作業ではなく計算機を用いることで、大量の単文の次数中心性と近接中心性を正確に算出することが可能になった。その結果、日本人英語学習者が書いた英語単文は、Studies in Second Language Acquisition 掲載の学術論文のアブストラクト英文と比較して、次数中心性が高く、近接中心性が低い傾向にあるという結果が得られた。

２．研究の目的

（１）研究の第一目標：英語統語構造特性の数値的把握

本研究の第一の目標は、日本人英語学習者が書いた英語文の統語構造特性を数値的に把握し、その結果を英語教育に応用する可能性を探ることである。本研究では、英語文の統語構造を依存文法の枠組みで捉え、その構造特性は単語間の依存関係を有向非循環グラフに見立てた場合のそのグラフの次数中心性と近接中心性として捉える。

（２）研究の第二目標：日英語統語構造類似度の数値的把握

本研究の第二の目標として、日本語文―英語文の翻訳ペアに属する文中の依存関係を有向非循環グラフと捉え、その構造の類似度を数値的に把握し、その類似度の高低が日英語間翻訳の困難度をどの程度反映しているかを検証する。日本語の統語構造解析については、Oya(2010a)を踏まえ、既存の日本語構文解析アプリケーション(KNP; 黒橋、長尾 2005)を使って得られた文節間の依存関係を利用する。Oya (2010a)では日本語の統語構造の最小単位は文節であると論じたが、本研究でもそれを踏襲する。本研究では、第一目標で述べた次数中心性と近接中心性、さらに頂点数と依存関係タイプの異なり数にも注目する。まず、日本語文―英語文ペアの中には、同一の意味を表しながらも英語と日本語とでは単語数が異なる場合が多い。例えば、“She would have seen that movie.”と、これに意味的に対応する「彼女はその映画を見ていたのかもしれない」とでは、それぞれの文中の依存関係を有向非循環グラフと捉えた場合、英語文では単語の数から頂点数は６、日本語文では文節の数から頂点数は４個である。この頂点数の差異は、これらの有向非循環グラフ間の類似度を低めていると考えられる。また、“I can speak English.”に意味的に対応する「私は英語が話せる」では、動詞“speak”の直接目的語である“English”に対応する「英語が」は、主語として動詞に依存している。このような依存関係タイプの違いも類似度を低めていると考えられる。

３．研究の方法

（１）平成 26 年度

①日英語対訳コーパス内の文構造解析と構造類似度算出

既存の日英語対訳コーパスに関して、英語文は Stanford Parser (De Marneffe at al. 2006)で、日本語文は KNP （黒橋、長尾 1997）で構文解析し、その出力結果をデータとして依存構造類似度算出を進める。

KNP の構文解析出力には依存関係のタイプ分けがされていないが、KNP の出力結果に依存関係タイプを付与するコンピュータプログラムはすでに作成してある。このプログラム作成に当たっては、Oya (2010)で論じた手法を使った。これは文節内部の主要部の品詞、文節内部の格助詞・係助詞の種類に応じて依存関係タイプを自動で行う手法である。日英語対訳コーパスとして、『Wikipedia 日英京都関連文書対訳コーパス』を選択する。これを選択する理由として、複数の翻訳者による正確な翻訳であるとされている点、そして

約 50 万文収録と大規模である点があげられる。
②日本語英語学習者作成英文の依存構造特徴量算出
学習進度の異なる日本人英語学習者が書いた英文データを入手し､各文を Stanford Parser で構文解析し、その解析データをもとに各文の構造特性算出を進める。英文データは、早稲田大学教育学部中野美知子教授の協力を仰ぎ、プレイスメントテストによってレベル分けされた授業を履修している学生が書いた英文をデータとして利用することを予定している。
③英語母語話者作成英文の依存構造特徴量算出
英語母語話者が作成した英文データとして、American National Corpus の下部コーパスである Manually Annotated Sub Corpus(Ide, Baker, Fellbaum, Fillmore, and Passonneau 2008)のタグなしデータを利用する。このコーパスは、19 の異なるジャンルから採択された約 50 万文の英文が収録されている。本研究では、その中から書き言葉である 13 のジャンルの英文を全て Stanford Parser で構文解析し、その解析データをもとに各ジャンルの構造特性傾向の算出を進める。パイロットスタディを Oya (2012)で発表しているが、新聞見出しやエッセイの小見出しなども一文として計算していたため、これらをどのように扱うかについて検討を加え、必要に応じて変更・削除する手順をとったうえで改めて解析するという手順を取る。
（２）平成２７年度以降
①日英語ペアの作成と日英・英日翻訳テスト
前年度までの成果を踏まえ、日英対訳コーパスから得られた日英語ペアの構造類似度に基づき、類似度の異なる日英語ペアを新たに作成し、日本人英語学習者に日本語文英訳と英語文和訳テストを課す。類似度の高いペアとは、例えば “Sarah runs”と「サラが走る」のように英語の単語数と日本語の文節数が同一で、しかも依存関係タイプが同一のものであり、これら二つの文は類似度が高いと言える。一方、日本語と英語とで(i) 次数中心性や近接中心性の差が大きいほど、(ii) 単語数の違いが大きくなるほど、そして(iii)使われている依存関係タイプが異なっているものが多いほど、類似度は低くなるものとする。類似度の高いものから低いものまで、様々な日英語ペアを多数創作する作業を経て、これらを解答とする日本語文英訳問題および英語文和訳問題を学生に提示し、類似度の高い日英語ペアについては正答率が高いか否かを測定する。

４．研究成果
（１）平成 26 年度
既存の日英語対訳コーパス中の日本語文及びその英訳文の構文解析は順調に進み、いくつかの学会および論文集でその解析結果に基づいた分析結果について発表した。特に、"A Study of Syntactic Typed-Dependency Trees for English and Japanese and Graph-centrality Measures"と題した博士学位請求論文（早稲田大学受理）は、依存文法とグラフ中心性に関してそれまでの私の研究をまとめたもので、英語と日本語の対訳ペアそれぞれの依存木のグラフ中心性を比較すると、英語文のほうが日本語文よりも度数中心性および近接中心性ともに小さい傾向にあり、つまり英語文はそれに対応する日本語文よりも深い構造を持つ傾向にあることが数値的に示された。
（２）平成 27 年度
日英語の依存機関の構造的不一致(Structural divergence)の概念を導入し、特に英語コーパス学会では「日英語パラレルコーパス中の対訳文間における格助詞「が」を伴う名詞句の依存関係の構造的不一致」と題した研究発表を行い、そこでは日英語パラレルコーパス中の日本語文で格助詞「が」が使われている単文が英語対訳文でどのように訳されているかに注目して、日英語対応パターンを構造的不一致の概念に基づいて分類した。
（３）平成 28 年度
"Dependency Types in Learner English and Authentic English"と題して環太平洋応用言語学会にて研究発表を行い、日本人英語学習者の産出した英語テキストと、ネイティヴスピーカーの産出した英語テキストとで、頻出する依存関係に差異があることを指摘した。また、平成 29 年度に『英語コーパス研究』にて発表が決定している"Syntactic Divergence Patterns among English Translation of Japanese One-Word Sentences in a Parallel Corpus"と題した研究論文では、日英語対訳コーパス中の日本語の一語文が英語ではどのように訳されているかに注目し、その対応パターンを構造的不一致の概念に基づいて分類した。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 4 件）
① Oya, Masanori. (in press) "Syntactic Divergence Patterns among English Translations of Japanese One-Word Sentences in a Parallel Corpus." 『英語コーパス研究』第 24 巻 査読有
②Oya, Masanori. 2015. "Japanese One-Word Sentences and their English Translations in a Parallel Corpus. 『目白大学人文学研究』第 11 巻、pp. 237-248. 査読有
③ Oya, Masanori. 2014. "Typed-dependency Tree Pairs of English and Japanese." 『目白大学人文学研究』第 10 巻、pp. 205-215.査読有
④Oya, Masanori. 2014. "An English-Japanese bilingual corpus-based comparison of their syntactic dependency structures". *Proceedings of the 19th Conference of Pan-Pacific Association*

of Applied Linguistics, pp 69-70. 査読有

〔学会発表〕（計 6 件）
①Oya, Masanori. "Dependency Types in Learner English and Authentic English." The 21st Conference of Pan-Pacific Association of Applied Linguistics. August 2016, Tamkang University, Taiwan.
②大矢政徳「日英語パラレルコーパス中の対訳文間における格助詞「が」を伴う名詞句の依存関係の構造的不一致」英語コーパス学会第 41 回大会　愛知大学（愛知県名古屋市）
③Oya, Masanori. "The Possibility of Building a Dependency-based Japanese-English Construction Dictionary." Asialex 2015. June 2015, The Hong Kong Polytechnic University, Hong Kong.
④ Oya, Masanori. "Centrality Measures of Sentences in an English-Japanese Parallel Corpus." PACLING 2015. May 2015, The Stones Hotel. (Bali, Indonesia).
⑤ Oya, Masanori. "An English-Japanese bilingual corpus-based comparison of their syntactic dependency structures." The 19th Conference of Pan-Pacific Association of Applied Linguistics. August 2014, Waseda University. (Shinjuku, Tokyo)
⑥大矢政徳「統語依存関係コーパスからの構造特性特徴量抽出」英語コーパス学会第 40 回大会　熊本学園大学（熊本県熊本市）

〔その他〕
博士学位請求論文
①Oya, Masanori. 2014. "A Study of Syntactic Typed-Dependency Trees for English and Japanese and Graph-Centrality Measures." （早稲田大学受理）

６．研究組織
(1)研究代表者
大矢　政徳（OYA, Masanori）
目白大学・外国語学部・准教授
　研究者番号：60318748