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

令和　２ 年　 ６ 月 １０ 日現在

機関番号：３２６８２
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ０２７４０

研究課題名（和文）日英語対訳コーパスからの統語的不一致情報抽出とその活用

研究課題名（英文）Extracting the information of syntactic divergence from a Japanese-English parallel corpus and applying the data

研究代表者

大矢　政徳（Oya, Masanori）

明治大学・国際日本学部・専任准教授

研究者番号：６０３１８７４８

交付決定額（研究期間全体）：（直接経費）　　1,100,000円

研究成果の概要（和文）：平成29年度においては、日本語と英語との間の翻訳可能性を一定の原理に沿って分類する方法についての客観的なデータに基づいた重要な示唆が得られた。
平成30年度においては、特定の国の出身者が産出した英文中で特定の依存関係タイプが比較的高頻度であった場合や、出身国の違いに関わらず高頻度であった依存関係タイプがコーパス分析の結果観察された。また、科学技術分野の英文では、複合名詞や、形容詞が名詞に依存している関係の頻度が、他のジャンルと比較して高いことなどが明らかとなった。
令和元年度においては、複数の文の統語的不一致の明示的記述、およびコーパスデータの統計的処理についての知見を深めることが可能になった。

研究成果の学術的意義や社会的意義
日英語間の統語的不一致を明示的に記述し、その結果得られた知見をより効率的な英語学習に応用する可能性を探る、という本研究の目的は、助成期間中に得られた研究成果により完全に達成されたとは言えないものの、今後もこの問題意識をもって研究を進めていくことは、当初の目的であった英語教育への応用可能性を拓くものである。本研究の結果得られた知見の英語教育への応用が実現するならば、現在は一部の語学熟達者によって占有されている暗示的知識を明示化し伝達可能とすることに繋がり、これが一般に語学学習の可能性を広げることが期待される。

研究成果の概要（英文）：In 2017, significant suggestions have been made based on objective data with respect to the procedure to categorize the possibilities between English and Japanese according to a set of principles.
In 2018, some instances have been observed in which certain dependency types are more frequent than others (1) in the texts produced by learners of English from certain countries or regions, or (2) in all the texts produced by learners of English regardless of their countries or regions.
In 2019, research has enabled us to describe the syntactic divergence among plural sentences in transparent manners and to deepen our understanding for statistical analyses of corpus data.

研究分野：依存文法、コーパス言語学

キーワード：依存文法　統語論　コーパス言語学　日英語間構造的不一致

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

本研究では、日英対訳コーパスに対して各収録文中の単語間の依存関係に関する情報及び各日英語対訳文ペアの統語依存構造木間の統語的不一致に関する情報を付与することによって日英対訳コーパスの応用可能性を拡大することを目的とする。統語的不一致とは、Mel'čuk & Wanner (2006)で提案された概念であり、二言語間の対訳文ペア間の統語依存構造間の不一致の分類である。日英語対訳コーパス中の対訳文ペアに対して統語的不一致に関する情報を付与することによって、その言語的資料としての利用価値をさらに高める。特に、より構造的不一致が多い対訳文ペアは、学習者にとっては学習困難なのではないかという予想を検証し、日英語間の統語的不一致という概念の外国語教育への応用可能性を探る。

１．１　依存文法

近年、統語論の分野にて依存文法(dependency grammar; Tesnière 1959, Debusmann and Kuhlmann2007, 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語構造を表現することを目指す文法理論の枠組みである。単文中の各二単語間の依存関係は、依存している単語(dependent)と依存されている単語(dependee)との意味的関係に応じてタイプ分けされる。例えば、単文 "David is studying."では、動詞 studying に David が依存しており、この二つの単語間の依存関係は「主語(subject)」としてタイプ分けされる。この二単語の依存関係は、de Marneffe, MacCartney and Manning (2006)では、"nsubj(studying-3,David-1)"のように表記される。

１．２　翻訳元単文と翻訳先単文との統語的不一致

統語的不一致(Syntactic mismatches)とは、Mel'čuk & Wanner (2006)で提案された概念であり、二言語間の対訳文ペア間の統語依存構造間の不一致の分類である。例えば、日本語の希望表現(desiderative)の「私はコーヒーが飲みたい」とその英語での翻訳文 "I want to drink coffee."とでは、単語間の依存関係に関して不一致が生じている。具体的には、日本語では系助詞「は」を伴う名詞「私は」はトピックとして動詞「飲みたい」に依存している一方、それに対応する英語の要素 "I"は主語として動詞"want"に依存している。そして、日本語では格助詞「が」を伴う名詞「コーヒーが」は主語として動詞「飲みたい」に依存しているが、それに対応する英語の要素"coffee"は目的語として動詞"drink"に依存している。さらに、日本語では「飲みたい」と動詞語幹と接尾要素とが組み合わされた単語で表現されている意味が、英語では"want to drink"と三つの単語で表現されている。Mel'čuk & Wanner (2006)では、Dorr (1993, 1994)を援用しながら、統語的不一致を五つに分類し、これらを機械翻訳の領域で統一的な様式で取り扱う方法が提案されている。彼らが提案している統語的不一致は、(i)統語項の文法役割の違い、(ii)依存関係の反転、(iii)単語の融合・分割、(iv)品詞交替、(v)機能語挿入・削除である。上記の例では、「私は」と I との間、そして「コーヒーが」と coffee との間に(i)文法役割の違いがあり、そして「飲みたい」と"want to drink"との間に(iii)単語の分割がある。全体として、上記の例では三つの統語的不一致があることになる。

１．３　日本語と英語との統語的不一致

Mel'čuk & Wanner (2006)では、主に英語と他のヨーロッパ諸言語との翻訳を例にとって統語的不一致の統一的取り扱いが論じられていたが、日本語と英語との統語的不一致については Mel'čuk &Wanner (2006; pp.110-111)では上述の希望表現について述べられているのみである。この点を鑑み、日本語と英語との統語的不一致を包括的に取り扱い、その知見を機械翻訳だけでなく英語教育の分野にも応用することは、当該分野に新しい研究可能性を拓くと期待される。

２．研究の目的

本研究では、日英対訳コーパスに対して各収録文中の単語間の依存関係に関する情報及び各日英語対訳文ペアの統語依存構造木間の統語的不一致に関する情報を付与することによって日英対訳コーパスの応用可能性を拡大することを目的とする。統語的不一致とは、Mel'čuk & Wanner (2006)で提案された概念であり、二言語間の対訳文ペア間の統語依存構造間の不一致の分類である。日英語対訳コーパス中の対訳文ペアに対して統語的不一致に関する情報を付与することによって、その言語的資料としての利用価値をさらに高める。特に、より構造的不一致が多い対訳文ペアは、学習者にとっては学習困難なのではないかという予想を検証し、日英語間の統語的不一致という概念の外国語教育への応用可能性を探る。

２．１　研究の第一目標：日英語翻訳ペア間の統語的不一致抽出

本研究の第一の目標は、Oya(2015)での日本語一語文とその英語対訳文との統語的不一致に関する知見を踏まえて、より多彩な日本語構文と、それに対応する英語との間で、(a)どのタイプの統語的不一致が高頻度で見いだされるか、そして(b)複数の統語的不一致がどのような組み合わせで生じているかを、より大量の、そして正確な日英語対訳コーパスのデータから探ることである。

２．２　研究の第二目標：日英語翻訳ペア間の統語的不一致と学習困難度との関連

本研究の第二の目標は、どの種類の統語的不一致が日本人英語学習者にとって学習が困難かを探ることである。本研究では、使用語彙の難易度など他の条件が同一であれば、統語的不一致を含まない日英語翻訳ペアは、何らかの統語的不一致を含む日英語翻訳ペアよりも、学習者にとって理解が容易であるという前提に立っている。日英語対訳コーパスから得られた日英語翻訳ペアには、統語的不一致の有無・その種類に関して様々な可能性があり得る。この可能性は、大まかに言えば、(a)全く統語的不一致を含まないペア、(b)複数の統語的不一致を含むペア、または(c)全く統語的には一致していないペアに分類される。直観的には学習困難度は(a)から(c)へと学習困難度が高まることが予想されるが、本研究では、実際に英語学習者を被験者としてこの直観を検証する。また、(b)の場合において、複数ある統語的不一致の中でどれが学習を困難にしているのかもあわせて検証する。

３．研究の方法

＜平成 29 年度の研究計画＞

＜日本語文とその英語対訳文の構文解析と統語的不一致抽出＞

平成 29 年度は、『日本語文型辞典』（グループ・ジャマシイ編著、2015）で取り上げられている日本語の文型とその英語対訳文のペアとの統語的不一致を抽出する。その際のデータ形式は、翻訳ペアの各文を構文解析して得られた統語依存構造のペアと、その間の統語的不一致に関する情報である。構文解析は、日本語単文に対しては KNP （黒橋、長尾 1997）を、英語に対しては Stanford Parser (De Marneffe at al. 2006)を用いることによって、人手による解析よりも速くかつ正確な構文解析結果を得る。

KNP の日本語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ため、Oya (2010)で論じた手法を使い、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を利用する。これによって、日本語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付与が自動で行われる。

＜平成 30 年度以降の研究計画＞

＜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、日英語対訳文間の統語的不一致が日本人の英語学習にどのような影響を及ぼし得るかを検証する。ここで、(a)高頻度な統語的不一致が必ずしも日本人英語学習者にとっては学習が困難ではない、あるいは(b)ある種の統語的不一致がひとつでもあると学習困難度が高まる、といった予想が立てられる。この予想を検証するために、統語的不一致の種類・数が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。例えば “David discovered the fact” と「デイヴィッドがその事実を発見した」のように英語の単語数と日本語の文節数が同一で、しかも統語的不一致がない例文、前述の"I want to drink coffee"と「私はコーヒーが飲みたい」のように統語的不一致が複数ある例文など、前年度までに抽出された統語的不一致のパターンに則した様々な日英語対訳ペアを、語彙的なレベルの違いが判断に影響しないよう工夫しながら創作する。そして、これらを解答とする日本語文英訳問題及び英語文和訳問題を学生に提示し、(a)高頻度な統語的不一致が多い日英語ペアについて正答率が高いか否か、そして(b)いずれかの統語的不一致タイプが含まれている場合には正答率が低いか、を測定する。

４．研究成果

〈平成 29 年度〉

『英語コーパス研究』第 24 号にて、"Syntactic Divergence Patterns among English Translations of Japanese One-Word Sentences in a Parallel Corpus"と題した論文を発表した。これは、日英パラレルコーパスに見られる日本語の一語文（単語一つから成る文)が対応する英文でどのように翻訳されているのかを、依存木の構造的不一致の観点から分類し、どのような不一致が頻出するかを計算し、日英翻訳における文脈情報の重要性を指摘した。この研究は、さらに多くの単語数を持つ日本語文がどのように英文へ翻訳され、その依存木間の構造的不一致の実態を知るという最終的目標を達成するにあたって、最も単純な構造を持つ日本語一語文から取り掛かるという初期的目標として位置づけられるものである。

さらに、英語コーパス学会第 43 回大会にて『日英対訳コーパス中の「～ことになる」構文とその英訳文間の構造的不一致』と題して研究発表を行った。これは、日本語の「～ことになる」構文が、対応する英文ではどのように翻訳されているのかを構造的不一致の観点から分類した結果を発表した。これは、日本語と英語との間で逐語訳的な対応関係を持たず一義的な翻訳が不可能な場合に、どのような翻訳可能性があり、それらの可能な翻訳文を恣意的にではなく一定の原理に沿って分類することが出来るのかについて、客観的なデータに基づいた示唆を与えるものであった。

〈平成 30 年度〉
Asia TEFL 2018 (University of Macau)にて、"Analysis of Learner-corpus Data Based on the Dependency-grammar Formalism"と題して研究発表を行った。これは、英語学習者による英語エッセイを出身国別・習熟度別に集約した The International Corpus Network of Asian Learners of English (The ICNALE)中の英文を構文解析した結果得られた依存木内の各依存関係タイプの使用頻度の違いを、学習者の出身国または地域別・学習者の習熟度別に集計し、その差異を検証したものであった。集計結果の一部を検証したところ、出身国別にみると特定の依存関係タイプが比較的高頻度であった場合や、出身国の違いに関わらず高頻度であった依存関係タイプも見られた。
さらに、PACLIC 2018 にて、"Utilization of Dependency Type per Sentence to Identify Differences among Genres of English Texts"と題して研究発表を行った。これは、1 センテンス当たりでの各依存関係タイプの使用頻度(Type per sentence, TPS)をひとつのメトリックとして、異なるジャンルの英文間でどのような依存タイプが異なる TPS 値を見せるか、あるいはジャンル横断的に高い/低い TPS 値を見せる依存タイプは何か、という観点でジャンル間の差異を明確化することを目的とした研究である。顕著な結果として、科学技術分野の英文では、複合名詞つまり名詞に別の名詞が依存している関係の頻度や、形容詞が名詞に依存している関係の頻度が、他のジャンルと比較して高いことなどが明らかとなり、これは先行研究での結果を跡付けるものであった。そして、当学会で研究成果を公表することで有益なフィードバックを得ることが出来た。

＜令和元年度＞
日英語のパラレルコーパス中の日英語翻訳ペア文から得られる言語学的情報の利用についての研究を進めた。具体的には、以下の論文を発表した。
1. "Structural divergence between root elements in English-Japanese translation pairs" Global Japanese Studies Review, Meiji University 12(1)
107 - 126 2020 年 3 月
2. "A corpus-based investigation of collexemes for active-passive alternation in the English part of an English-Japanese parallel corpus"
Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation 516 - 521 2019 年
論文 1 では、日英語翻訳ペア間の統語的不一致を統合的に記述する際の指針と、それに基づく規則ベースの自動翻訳過程の明示的記述を試みた。論文 2 では、日英語パラレルコーパス中の英語文をデータとして、個々の動詞が能動態として使用される傾向が強いのかあるいは受動態で使用される傾向が強いのかを、Fisher の正確確率検定を利用して統計的に解明することを試みた。これらの研究を通じて、(1) 複数の文の統語的不一致の明示的記述、および (2) コーパスデータの統計的処理についての知見を深めることが可能になった。これは今後の研究、特に令和 2 年度から科研費により助成されることになっている研究を進めるうえで重要な意義を持っている。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　1件／うちオープンアクセス　0件）

１．著者名 Masanori Oya	４．巻 32
２．論文標題 Utilization of Dependency Type per Sentence to Identify Differences among Genres of English Texts	５．発行年 2018年
３．雑誌名 Proceedings of PACLIC 32	６．最初と最後の頁 1-8
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Masanori Oya	４．巻 24
２．論文標題 Syntactic Divergence Patterns among English Translations of Japanese One-Word Sentences in a Parallel Corpus	５．発行年 2017年
３．雑誌名 英語コーパス研究	６．最初と最後の頁 19-40
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Masanori Oya	４．巻 33
２．論文標題 A corpus-based investigation of collexemes for active-passive alternation in the English part of an English-Japanese parallel corpus	５．発行年 2019年
３．雑誌名 Proceedings of PACLIC 33	６．最初と最後の頁 516 - 521
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Masanori Oya	４．巻 12 (1)
２．論文標題 Structural divergence between root elements in English-Japanese translation pairs	５．発行年 2020年
３．雑誌名 Global Japanese Studies Review, Meiji University	６．最初と最後の頁 107 - 126
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　2件）

１．発表者名
Masanori Oya

２．発表標題
Utilization of Dependency Type per Sentence to Identify Differences among Genres of English Texts

３．学会等名
PACLIC 32（国際学会）

４．発表年
2018年

１．発表者名
Masanori Oya

２．発表標題
Analysis of Learner-corpus Data Based on the Dependency-grammar Formalism

３．学会等名
Asia TEFL 2018（国際学会）

４．発表年
2018年

１．発表者名
大矢　政徳

２．発表標題
日英対訳コーパス中の「～ことになる」構文とその英訳文間の構造的不一致

３．学会等名
英語コーパス学会

４．発表年
2017年

１．発表者名
Masanori Oya

２．発表標題
Collostructional Analysis of English Future Tenses in a Learner Corpus

３．学会等名
The 24th Conference of Pan-Pacific Association of Applied Linguistics

４．発表年
2019年

１．発表者名 Masanori Oya
２．発表標題 A corpus-based investigation of collexemes for active-passive alternation in the English part of an English-Japanese parallel corpus
３．学会等名 PACLIC 33
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考