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　６ 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：2020～2023
課題番号：２０Ｋ００５８３

研究課題名（和文）依存文法の枠組を利用して複数文間の相互関係を可視化した英語コーパス構築

研究課題名（英文）Development of an English corpus with visualized interrelationships among sentences in the framework of dependency grammar

研究代表者
大矢　政徳（Oya, Masanori）

明治大学・国際日本学部・専任教授

研究者番号：６０３１８７４８

交付決定額（研究期間全体）：（直接経費）　　800,000円

研究成果の概要（和文）：日英対訳コーパスを依存文法の枠組みで構文解析した結果をもとにして各文の構造的特性と日英語間の構造上の一般的な差異に注目する過程で、依存距離へと研究テーマが移行し、これに注目した研究発表を行い、各種学術雑誌に論文を投稿した。特に、依存距離の頻度分布が言語横断的に特定の分布に一致するという知見や、英語学習者が算出した英文の平均依存距離がその学習到達度によって異なる、といった当初の研究テーマを超えた知見が得られた。最終年度においては、従来注目されてこなかった、平均よりはるかに長い依存距離をもつ依存関係が、日英語間で大きな差異を見せることを発見し、新しい研究テーマの開拓につながったことは特筆に値する。

研究成果の学術的意義や社会的意義
日本語と英語の統語構造を依存構造の枠組みで明示し、その構造特性を依存距離という数値で表現しかつ比較対照することによって、これら2言語の構造的差異を従来より客観的に把握することが可能になる。さらに、日本人英語学習者が産出する英文の平均依存距離が学習者の熟達度に応じて変化するといった知見は今後の英語教育への応用可能性を秘めている。

研究成果の概要（英文）：In this study, the results of syntactic analysis using a dependency grammar framework on a Japanese-English parallel corpus focused on the structural characteristics of each sentence and the general structural differences between Japanese and English. During this process, the research theme shifted to dependency distances, leading to presentations at various academic venues and papers published in several scholarly journals. Notably, it is found that the frequency distribution of dependency distances aligns with a specific distribution cross-linguistically, and that the average dependency distance in English sentences produced by learners varies according to their proficiency level. Particularly significant in the final year was the discovery of dependency relationships with much longer than average dependency distances, which exhibited significant differences between Japanese and English, paving the way for new research themes.

研究分野：言語学

キーワード：依存文法　日英対訳コーパス　依存距離

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

機械学習の理論的洗練とハードウェアの性能向上の結果、現在の自然言語処理研究は隆盛を極め、応用可能性も広がっている。しかしながら、人工知能の分野ですでに Pearl (2018)が指摘しているのと同様に、自然言語処理においてもその内実は必ずしも言語学的知見を十分に取り入れたものであるとは限らない。例えばニューラルネットワークを活用した機械翻訳では、実際にどのような計算過程を経て翻訳元文から翻訳先文が得られているのかは、研究者自身にとってもいわばブラックボックスの中であるのが実情である。このようなブラックボックス化は、システムの精度や適用範囲の向上に課題を残すのみならず、結局のところ現在においてもコンピュータは言語学の知見を基礎として自然言語を処理してはいないことを意味する。この点を踏まえると、現在の自然言語処理研究と並んで、言語学的に動機づけられた自然言語処理の可能性を探ることには学術的意義があると考えられる。

２．研究の目的

本研究は、言語学的に動機づけられた自然言語処理研究の試みの一つとして、依存文法(Dependency Grammar)の枠組を利用して複数の文の相互関係を可視化した英語コーパス構築と、これを利用した自然言語推論システム構築を目的とする。依存文法を二つの文をまたいだ単語間の関係も対象とするように拡大し、自然言語推論(natural language inference, NLI)で利用されている前提文—仮定文ペア中の単語間の相互関係を可視化することで、前提文—仮定文ペアの間にある含意関係、矛盾関係、そして中立関係が各ペア中の単語間の相互関係によってどのように表現されるかを検証する。その結果を利用して、前提文と仮定文の間に含意関係、矛盾関係または中立関係のどれが成立するかを、単語間の関係から自動で判定するシステムの構築を目指す。

３．研究の方法

【第一段階：既存の NLI 英語コーパスへの文間単語関係タグ付け】MultiNLI に含まれている前提文—仮定文ペア中の単語間の関係を明示化するタグを人手で付与する。この作業を通じて、entailment, contradiction, そして neutral のそれぞれの含意関係にある場合にどのような文間単語関係が見いだされるのかを見極めることが期待される。現時点では MultiNLI に含まれた前提文—仮定文ペアは Penn Treebank 様式で構文解析されているが、本研究ではこれらの英文を Stanford Dependency Parser で依存文法の枠組で構文解析し、前提文の依存木中のどの単語が仮定文の依存木中のどの単語とどのような関係にあるかを明示的にタグ付けする。

【第二段階： NLI 英語コーパスへの文間単語関係自動タグ付けと含意関係自動判定】第一段階で得られた文間単語関係タグ付き前提文—仮定文ペアコーパス作成から得られた知見を応用して、文間単語関係がタグ付けされていない自然言語推論英語コーパス中の前提文—仮定文ペアに対して当該タグを自動で付与するコンピュータプログラムを作成し、その結果をもとに各ペアの含意関係の自動判定を試みる。その結果を機械学習の手法を使った含意関係自動判定結果と比較し、さらなる判定精度の向上を図る。

４．研究成果

2020 年度

1. "Structural divergence between root elements in English-Japanese translation pairs"（Global Japanese Studies Review, Meiji University 12(1), 107- 126）では、英語と日本語との翻訳ペア文中の単語の依存関係を依存木で表現し、その構造的不一致について論じた。

2.　"Analysis and Quantification of the Network of Lexical Connectedness Within a Text Based on Metrics Used in Network Analysis” (Global Japanese Studies Review, Meiji University 13(1), 15 - 38)では、英文テキスト中の文を頂点とし、文間の意味的関連を辺としたネットワークととらえ、これが文と文との関係つまり文脈の構造を表現しているとし、その構造特性の数値化を提案した。

2021 年度

1. “Developing a Japanese-English Parallel Corpus of "Japan, the Beautiful and Myself" by Yasunari Kawabata”(Global Japanese Studies Review 14(1), 13-26)は、川端康成のノーベル賞受賞スピーチ「美しい日本の私」のオリジナル日本語版とその英訳版との対訳コーパス構築と、主要な依存関係の依存距離を日英語間で比較した研究であり、本研究課題である文章中の文同士の関係を明示化するためのコーパスデータとして利用する予定であった。

2. “Three Types of Average Dependency Distances of Sentences in a Multilingual Parallel Corpus”(Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation (PACLIC 35) 652-661) は、当該年度に PACLIC35 で行った発表のプロシーディングスであり、平均依存距離の算出方法として、多言語パラレルコーパス中の（1）言語ごとに、(2）文ペアごとに、そして（3）依存タイプごとにそれぞれ平均依存距離を算出する手法を提案し、文ペアごとに算出された平均依存タイプが言語のカテゴリ分けに寄与しうることを解明した。
3. “Syntactic Similarity of the Sentences in a Multi-Lingual Parallel Corpus Based on the Euclidean Distance of Their Dependency Trees”(Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation (PACLIC 34) 225-233)は、前年度に PACLIC 34 で行った研究発表のプロシーディングスであり、多言語パラレルコーパスをデータとし、文の統語的類似性をユークリッド距離として計算する手法を提案した。

2022 年度
1. "The Relevance of Dependency Distances in the Study of L2 Production" (The 41st Thailand TESOL Conference Proceedings)は、前年度に Thai TESOL で発表した内容のプロシーディングスとして発行されたものであり、異なる熟達度の英語学習者が産出したテキストの文中の単語の依存距離の確率分布が、熟達度の違いに関わらず Zipf-Alekseev 分布に一致するという先行研究の知見を、より多くの学習者の産出した英語エッセイのコーパスをデータとして検証した研究である。
2. "Differences of Mean Dependency Distances of English Essays Written by Learners of Different Proficiency Levels"(Glottometrics vol. 53,24-41)では、前論文の結果を踏襲しつつも、平均依存距離を尺度として異なる熟達度の英語学習者が産出した英文を比較すると、熟達度の違いによって平均依存距離は異なり、さらに異なる依存タイプの平均依存距離に注目すると、依存タイプによって熟達度の違いで平均依存距離が異なるものもあれば、依存タイプの違いに関わらず平均依存距離が変わらない場合もあるという結果が得られた。
3. "Propositional Idea Density of a Japanese Text and its English Translation in a Parallel Corpus "(Global Japanese Studies Review, Meiji University 15(1), 97-105)では、先行する研究で注目されている概念密度に着目し、川端康成のノーベル文学賞受賞スピーチとその英訳を依存文法の枠組みで構文解析した日英対訳コーパスをデータとしてそれぞれの概念密度を計算したところ言語間で概念密度には有意な差がみられないという結果が得られた。

2023 年度
1. “Low-Frequency Long-Distance Dependencies as “Long Tails”” (Proceedings of 37th Pacific Asia Conference on Language, Information and Computing (PACLIC) 37, 89-95)は、当該年度に PACLIC37 で行った発表のプロシーディングスである。従来注目されてこなかった、平均よりはるかに長い依存距離をもつ依存関係の依存タイプごとの頻度が、日英語間で大きな差異を見せることを発見し、新しい研究テーマの開拓につながった。

５．主な発表論文等

〔雑誌論文〕　計8件（うち査読付論文　8件／うち国際共著　0件／うちオープンアクセス　5件）

１．著者名 Masanori Oya	４．巻 15
２．論文標題 Propositional Idea Density of a Japanese Text and its English Translation in a Parallel Corpus	５．発行年 2023年
３．雑誌名 Global Japanese Studies Review	６．最初と最後の頁 97-105
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Masanori Oya	４．巻 53
２．論文標題 Differences of Mean Dependency Distances of English Essays Written by Learners of Different Proficiency Levels	５．発行年 2023年
３．雑誌名 Glottometrics	６．最初と最後の頁 24-41
掲載論文のDOI（デジタルオブジェクト識別子） 10.53482/2022_53_400	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Masanori Oya	４．巻 -
２．論文標題 The Relevance of Dependency Distances in the Study of L2 Production	５．発行年 2022年
３．雑誌名 The 41st Thailand TESOL Conference Proceedings	６．最初と最後の頁 -
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Masanori Oya	４．巻 14 (1)
２．論文標題 Developing a Japanese-English Parallel Corpus of "Japan, the Beautiful and Myself" by Yasunari Kawabata	５．発行年 2022年
３．雑誌名 Global Japanese Studies Review	６．最初と最後の頁 1-26
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 35
2．論文標題 Three Types of Average Dependency Distances of Sentences in a Multilingual Parallel Corpus	5．発行年 2021年
3．雑誌名 Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation (PACLIC 35)	6．最初と最後の頁 652-661
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 オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Masanori Oya	4．巻 34
2．論文標題 Syntactic Similarity of the Sentences in a Multi-Lingual Parallel Corpus Based on the Euclidean Distance of Their Dependency Trees	5．発行年 2021年
3．雑誌名 Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation (PACLIC 34)	6．最初と最後の頁 225-233
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 オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Masanori Oya	4．巻 13
2．論文標題 Analysis and Quantification of the Network of Lexical Connectedness Within a Text Based on Metrics Used in Network Analysis	5．発行年 2021年
3．雑誌名 Global Japanese Studies Review	6．最初と最後の頁 15 - 38
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
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	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3件（うち招待講演　1件／うち国際学会　3件）

1．発表者名 Masanori Oya
2．発表標題 The Relevance of Dependency Distances in the Study of L2 Production
3．学会等名 The 41st Thailand TESOL International (Virtual) Conference（招待講演）（国際学会）
4．発表年 2022年

1．発表者名 Masanori Oya
2．発表標題 Three Types of Average Dependency Distances of Sentences in a Multilingual Parallel Corpus
3．学会等名 Pacific Asia Conference on Language, Information and Computation (PACLIC) 35（国際学会）
4．発表年 2021年

1．発表者名 Masanori Oya
2．発表標題 Syntactic similarity of the sentences in a multi-lingual parallel corpus based on the Euclidean distance of their dependency trees
3．学会等名 Pacific Asia Conference on Language, Information and Computation (PACLIC) 34（国際学会）
4．発表年 2020年

〔図書〕　計1件

1．著者名 大矢政徳	4．発行年 2022年
2．出版社 開拓社	5．総ページ数 240
3．書名 依存文法概説	

〔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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関