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　５ 年　 ５ 月 ２９ 日現在

機関番号：１７１０２
研究種目：若手研究
研究期間：2019～2022
課題番号：１９Ｋ１３１８０

研究課題名（和文）アノテーション付き大規模通言語コーパスを利用した言語変種についての計量的研究

研究課題名（英文）A Quantitative Study of Linguistic Varieties Using a Large Annotated Corpus

研究代表者
伊藤　薫（Kaoru, Ito）

九州大学・言語文化研究院・助教

研究者番号：３０７６９３９４

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本研究課題の成果として、日本語Universal Dependencies (UD)ツリーバンクのアノテーション（情報付与）に関する考察や提言を行った。加えて、UDツリーバンクをコンピュータ上で読み込むためのツールであるChaKi.NETを改良し、ChaKi.NET liteとして公開した。ツールの改良により、インターフェイスを備えていない言語データの使用に習熟していない言語学者にとっても危機言語のデータを容易にアクセスできるようにした。

研究成果の学術的意義や社会的意義
本研究課題において作成したツールは、危機言語に関するデータ作成も活発なUniversal Dependencies (UD)プロジェクトにおいて生み出される情報へのアクセスを容易にする。UDプロジェクトは元々情報系分野のプロジェクトであり、利用者はプログラミングに精通していることが想定されている。しかし、UDツリーバンクには他の形式では公開されていない危機言語などのデータなども含まれており、本ツール開発により活用の裾野を増やし、言語学分野のデジタル・トランスフォーメーションに貢献した。

研究成果の概要（英文）：Through this research project, recommendations were made on annotating the Japanese Universal Dependencies (UD) treebanks. In addition, we improved "ChaKi.NET", a tool for processing UD treebanks on computers, and released it as "ChaKi.NET lite". This tool makes data on endangered languages more accessible to linguists who are not proficient in using language data without an interface.

研究分野：言語学

キーワード：Universal Dependencies　言語変種　コーパス　言語資源

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

研究の構想段階では、今日まで続く第三次人工知能ブームにより学習用データの開発が活発化していた。自然言語に関するデータも増加しており、多言語に渡って共通した基準で（通言語的）文法的な情報が付与（アノテーション）される Universal Dependencies (UD)と呼ばれる枠組みのコーパスが開発されていた。UD は人工知能開発向けのデータであり言語学者にとって扱い難い点もあるが、これまでにない様々な言語変種（言語の使用目的、話者などによる個別言語内の変異）についてのコーパスが開発されており、これらを言語学研究に活用する利点は大きいと見込んでいた。そこで本研究では、(I)UD を利用した言語変種についての計量的研究と、(II)別分野で開発されたデータを言語学研究に最大限利活用するための方法、の 2 点を柱とした研究を計画した。

２．研究の目的

本研究の目的は、(I) 言語変種の特徴の解明という言語そのものについて説明するという目的と、(II) 爆発的に増加する「きれい」ではない電子的な言語資料をいかに活用するかを示すという言語学の研究手法開発という目的の２つである。(I)については、UD ツリーバンクという新たな資料をもとに、これまでまとまった資料が存在しなかった言語変種についての特徴を明らかにすることが目的であり、(II)については言語学の資料としての UD の価値を評価することや、その活用方法を提案することが目的となる。

本研究の学術的独自性と創造性は、上記(II)に比重が置かれている。つまり、事実の正確な記述や理論的整合性を追究する科学の立場から、実用性の高い手法の追求という工学的な関心に基づいて作成された膨大なデータを利活用するための方法を探る点にある。つまり、科学的目的と工学的目的は一致しない場合があるが、関連する分野であれば資源を共有し得るため、そのための方法を探る価値がある。

３．研究の方法

NLP のために作られた UD ツリーバンクは必ずしも言語学者にとって使いやすいフォーマットとは言えないため、それらを言語学の研究に活用するための戦略が必要となる。そこで、本研究では次のような戦略を取り、言語学における UD ツリーバンク普及を図った。

(A)本研究ではツリーバンクに含まれる情報が言語学のために作成されたコーパスのアノテーションとどのように異なるかを量的に調査したり、開発途上である日本語 UD ツリーバンクのアノテーションガイドラインについて言語学的観点から提言を行う。具体的な手法としては、言語学向けに作成されたコーパスからタグのコンバージョン（変換）によって作成された UD ツリーバンクを比較した。比較対象としたのは日本語 UD のアノテーション仕様の中でも比較的議論の余地の少ない品詞タグである。UD Japanese-BCCWJ というツリーバンクを対象とし、国語研単単位品詞と UPOS と呼ばれる UD で規定された品詞の頻度分布を比較、主成分分析によってレジスター別の文書の分布を観察した。また、発展途上の日本語 UD のアノテーション仕様に対する提言として、言語類型論の研究を参考にしながら述語並列に関して提言を行った。

(B)UD ツリーバンクは情報学向けに開発されたため、研究開始当初はプログラミング技術をもたないユーザにとって情報抽出は容易でなかった。言語学者による UD ツリーバンク利用促進のため、指標を簡単な操作で抽出できるようなツールの開発を行い、ソフトウェアを公開することを目標とした。ソフトウェア開発に当たっては、UD ツリーバンクのフォーマットである CoNLL-U 形式のファイルを読み込むことのできる既存のコーパスツールを UD 向けに改良し、ツールに不慣れなユーザでも使用しやすい軽量版を開発することとした。

４．研究成果

まず、伊藤(2020)で行った研究では、品詞の分布の差異について文書を単位として頻度分布に対し主成分分析を行った結果、図 1 のような分布が得られた。グラフ上の各点は出版書籍(PB)、雑誌(PM)、新聞(PN)、白書 (OW)、Yahoo!知恵袋(OC)、Yahoo!ブログ(OY)のいずれかのレジスターに属する文書 1 つに対応し、色によってどのレジスターに属するか示されている。また、原点から伸びる矢印は各主成分についての主成分負荷量を表す。図中の楕円は ggbiplot パッケージによって描画された各レジスターに対応する正規確率楕円である。両品詞体系の違いついては、名詞の細分化や助動詞に関するものが顕著である。一般的に名詞として分類される語は今回対象とした 2 つの品詞体系でさらに細かく分類されている。具体的には、短単位では接頭辞と接尾辞が名詞とは別の品詞として立てられており、UPOS では一般的に名詞に該当する語が NOUN, PROPN, NUM の 3 つに細分化されている一方で、短単位では単に名詞としてまとめられている。短単位に関して名詞、接頭辞、接尾辞の主成分負荷量を見ると、大きさは異なるもののいずれも主に第 1 主成分の負荷が高くなっている。UPOS に関して主成分負荷量を見ると NOUN が第 1, 2 主成分ともに一定の負荷があるのに対し、PROPN, NUM に関しては主に第 2 主成分に関して負荷

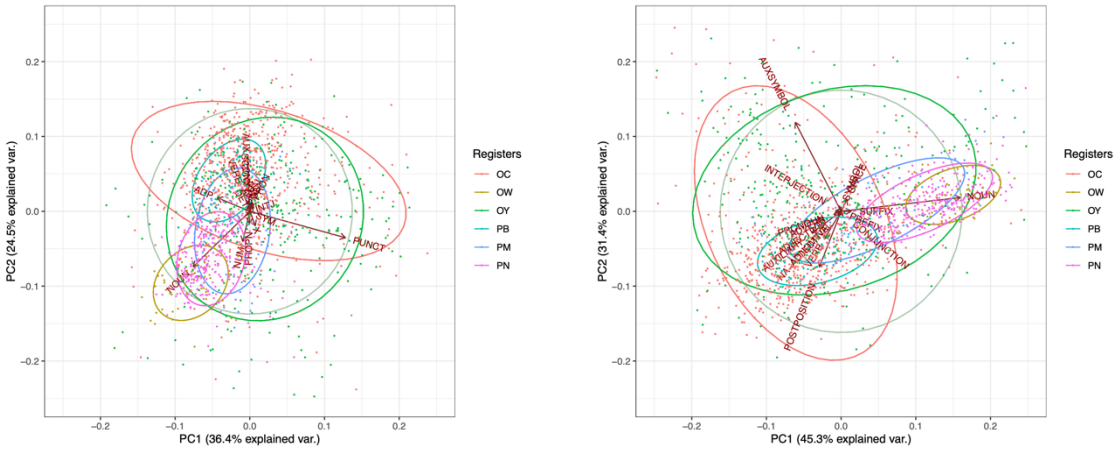

図 1UPOS 頻度（左）と短単位品詞頻度（右）を特徴量としたレジスター別の文書分布

に対し、短単位では第 2 主成分までの負荷は 0.2 程度とさほど大きくないという細かな違いはが高くなっている。また、助動詞に関しては UPOS で名詞と並んで第 2 主成分の負荷が大きいの見られる。全体的な結果を見ると、UPOS を用いても国語研単単位を用いても、少なくとも OW, PN, PM, PB の広がり方など、広域的な文書の分布に関しては大きな影響を与えないと考えられる。

次に、日本語 UD における並列アノテーションに関して問題視されていた述語並列表現の扱いに対する提言(伊藤 2021)について述べる。UD は通言語的に共通した品詞タグを付与する設計になっているが、個別言語に当てはめるときには類型論的な考察が必要になる。UD において述語並列は等位接続を表す *conj* タグを用いて表される。パラレルコーパスを用いて英語と日本語の対応を調査した結果、英語では and, but, or、もしくは(as well) as の 4 種が抽出されたのに対し、日本語では 37 語がこれらに対応していた。日本語このように並列表現が多様であるのに加え、並列表現の日本語訳として多く見られた動詞のテ形は言語類型論で converb(副詞的な従属接続標示することを主な機能とする不定動詞)とみなされることがあることも指摘した。このような観点から、日本語 UD における並列表現のアノテーションでは、意味的観点もしくは統語的観点のどちらを重視するか、元コーパスからの変換のしやすさ、UD 全体との整合性などを考慮して修正するべきことを提言した。

最後に、UD 向けコーパスツール ChaKi.NET lite の開発(伊藤・森田 2023)について述べる。UD は言語学、特に少数言語研究や類型論分野で有用なコーパスだが、自然言語処理分野で開発されたコーパスであるため、利用者がプログラミングスキルを持つことを想定している。このため自分でスクリプトを書くことができれば必要な情報を抽出できるが、そうでない場合は何らかのツールを利用する必要がある。そこで、本研究では既存の ChaKi.NET を活用し、UD ツリーバンクの利用障壁を下げ利用しやすくした。既存の ChaKi.NET は習熟した利用者にとっては高機能で使いやすいものの、高機能ゆえに操作が複雑であるため、新規ユーザにとって学習コストが高い状況となっている。こうした状況をふまえ、本研究では CoNLL-U の読み込みやタグ，依存構造木表示、依存構造情報の検索など、UD コーパスの利用に適した基本的機能を備える ChaKi.NET を改良することで、言語学や関連分野の研究を促進するツールの開発を目標として以下の方針で開発を行った。

- インターフェイスの改良により直感的に利用可能であること
- ChaKi.NET のデータベースと互換性があること
- 言語学研究に適した機能を持つこと
- コーパスを利用して言語学研究を行うユーザを主な対象とすること
- UD ツリーバンクの操作に適していること

更新箇所は主に、ツリーバンク一括読み込み機能の追加、画面レイアウト・ボタン配置の改善、パネル表示内容の UD への特化である。既存の ChaKi.NET は汎用的な用途でカスタマイズ性が高いため、UD の読み込みでは不要な機能が表示されていたり、ボタン配置がツリーバンク検索の操作フローを考えると分かりづらい位置に置かれたりしていた。しかし、これらの改良により、新規ユーザでも直感的にコーパスデータの検索が可能になったと思われる。改良後の画面レイアウトに操作フローの説明を加えたものを図 2 に示す。

完成したツールは https://github.com/chakidev/chakinet-lite で公開されている。

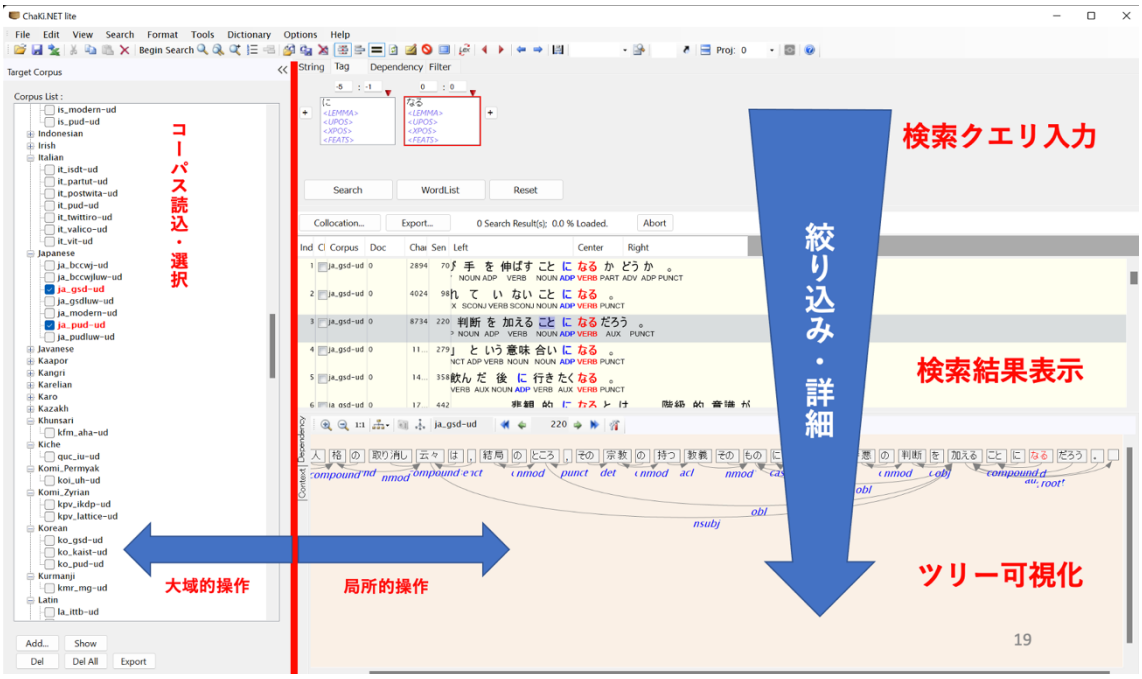

図 2 ChaKi.NET lite のデザイン

近年は沖縄語 UD ツリーバンク（宮川 他 2023）が構築されるなど国内外で危機言語の UD ツリーバンクが活発に公開されている。本研究で開発したツールはこのようなデータへのアクセスを容易にすることにより言語学のデジタル・トランスフォーメーションに貢献できると思われる。

５．主な発表論文等

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

１．著者名 伊藤薫・森田敏生	４．巻 25
２．論文標題 ChaKi.NET liteの開発	５．発行年 2023年
３．雑誌名 国立国語研究所論集	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 ―

〔学会発表〕　計4件（うち招待講演　1件／うち国際学会　0件）

１．発表者名 伊藤薫・ 森田敏生
２．発表標題 ChaKi.NET liteの開発 Universal Dependenciesコーパスの利用を見据えた ChaKi.NETユーザインターフェイスの改良
３．学会等名 Evidence-based Linguistics Workshop 2022
４．発表年 2022年

１．発表者名 伊藤 薫
２．発表標題 Universal Dependencies における述語並列記述の展望
３．学会等名 言語処理学会第27回年次大会
４．発表年 2021年

１．発表者名 伊藤　薫
２．発表標題 Universal Dependenciesコーパスを利用したレジスター研究の試み
３．学会等名 言語処理学会第26回年次大会(NLP2020)
４．発表年 2020年

１．発表者名
伊藤　薫

２．発表標題
Universal Dependencies に基づく言語学研究の射程

３．学会等名
Universal Dependencies シンポジウム（招待講演）

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
2023年公開の論文「ChaKi.NET liteの開発」にまとめられているコーパスツール"ChaKi.NET lite"は下記ページで公開されている。
https://github.com/chakidev/chakinet-lite

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