

令和 6 年 6 月 17 日現在

機関番号：32644

研究種目：挑戦的研究（萌芽）

研究期間：2021～2023

課題番号：21K18367

研究課題名（和文）再現性を担保した容認性判断のアーカイブの開発

研究課題名（英文）Developing a linguistic archive for repeatable acceptability judgments

研究代表者

成田 広樹（Narita, Hiroki）

東海大学・文学部・准教授

研究者番号：60609767

交付決定額（研究期間全体）：（直接経費） 4,900,000円

研究成果の概要（和文）：理論言語学における容認性判断を用いた研究手法は、データ収集の手法が確立しておらず、データの再現性や客観性の確保が大きな課題として残っている。本研究では、このような言語学版「再現性の危機」問題を乗り越え、分野横断的な理論提案と実証的検証のサイクルを実現するための共通基盤を創出するために、言語事実・理論データのオープンアクセス・データベースGrammarXiv(グラマカイク)を構築し、世界に向けてウェブ公開を行った（<https://grammarxiv.net>）。また、研究代表者の日英語の主語に関わる研究成果を中心にサンプルデータ登録を行い、GrammarXivを通じてデータ公開を行った。

研究成果の学術的意義や社会的意義

本研究は、オープンアクセスの、共同編集、柔軟なエントリ間リンク、および強力な横断検索が可能な、SNS機能付きのデータベースを開発公開する。- の特徴全てを備えたデータベースは未だ存在せず、世界に類を見ない革命的な成果となる。

本研究の提示する開かれた知見集積の場を提供することは、理論と実証の連携という観点からも、人間の洞察と計算機・ビッグデータ解析技術との連携という観点からも、諸科学で広く議論される「再現性の危機」問題に対する言語学分野独自の解決という観点からも、アカデミア成員の相互互恵的協同という観点からも、科学全体の今後の可能な発展のあり方を示唆する試金石/モデルケースとなる。

研究成果の概要（英文）：In theoretical linguistics, research methodologies using grammaticality/acceptability judgments face significant challenges such as the lack of established methods for data collection, which poses issues concerning the reproducibility (repeatability) and objectivity of data. In order to overcome this linguistic version of reproducibility crisis, we developed an open-access platform to foster cross-disciplinary interactions concerning theoretical proposals and empirical verification. Our endeavor has crystallized into the open-access database of linguistic facts and theoretical data, called GrammarXiv (Grammar Archive) and made it publicly available on the web (<https://grammarxiv.net>). Furthermore, we registered sample data, focusing on research findings related mainly to the subjects of Japanese and English by the principal investigator, and made these data publicly accessible through GrammarXiv.

研究分野：理論言語学

キーワード：容認性判断・文法性判断 理論言語学 実験言語学 オープンアクセスデータベース 主語性

1. 研究開始当初の背景

理論言語学は、理論を証拠付けるデータとして、例文に対して解釈の容認度を記したデータを積極的に活用する。例えば、「太郎は二郎に魚を食べさせた」という表現は容認されるが、助詞ニをヲに置き換えた「*太郎は二郎を魚を食べさせた」は容認不可であるという事実は、日本語文法研究にとって有用なデータとなる(例文先頭の*は容認不可であることを表す)。このような容認性判断を用いた研究手法は、言語理論を検証する上で示唆に富むデータを提供してくれるものの、そのデータ収集に関しては、未だ個人の主観的内省に頼らざるを得ない面が多く、データの再現性や客観性の確保が大きな課題として残っている。人文社会系諸科学で昨今広く議論されている「再現性の危機」の言語学版とも言える。事実、追試実験で統計的に有意な再現性が得られない事例も少なくないが(「*健が思い切り直美を頭を叩いた」を容認不可とする報告など; Linzen & Oseki 2018 等を参照)、現状、妥当性が疑わしい理論仮説や観察命題が十分に淘汰されない深刻な状況が続いている。

問題の根源は分野・理論の細分化によるコミュニケーション不全である。理論言語学及びその隣接分野の研究者は、様々な手法(計算機への実装、心理学的・脳科学的実験、言語習得研究)により、理論データの有効性を実証的に検証しようとしているが、個々の研究者集団は個別の流儀や方法論に囚われ、その外で行われた批判検証には必ずしも十分に注意を払っていない。こうした状況を打開し、透明性の高い、多元的かつ合目的なデータの蓄積・参照・検証を実現するため、分野の垣根を超えた組織的努力がこれまで以上に必要とされる状況であった。

2. 研究の目的

本研究は、言語事実・理論データのオープンアクセス・データベース^{グラマカイク}GrammarXivを構築し、理論集積の系統的可視化及び多元的評価を実現することで、言語学版「再現性の危機」問題を乗り越え、分野横断的な理論提案と実証的検証のサイクルを実現するための共通基盤を創出する。特に、図1に示す①～⑤の機能を実装したオープンアクセスプラットフォームを新たに研究コミュニティ全体へ提供することを目指す。また、⑦多種多様な分野理論からの関心の高さ、⑧データの豊富さ、および⑨応募者自身のこれまでの研究活動との関連性の高さの3つの要件を満たす「主語(性)」をサンプルトピックに選定し、応募者自身の継続課題と紐付けて計画を実施することで、言語理論そのものに対する貢献とデータベースの表現力の実地的検証の両方を射程に研究を遂行する。

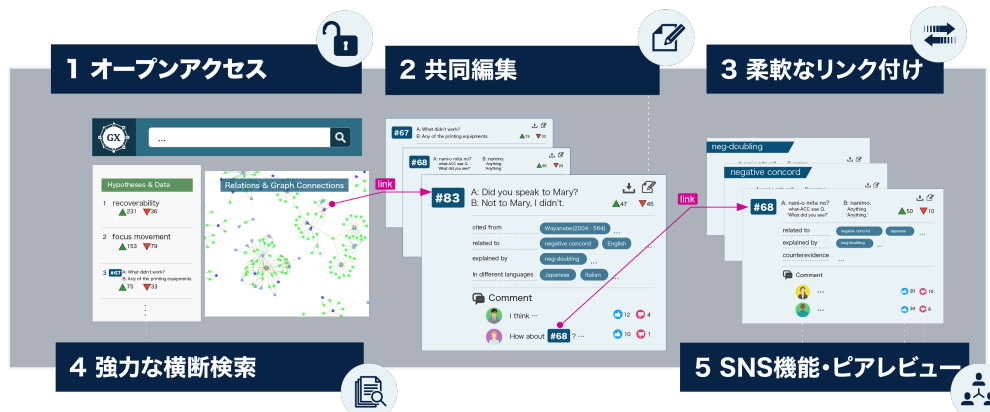


図1: GrammarXiv仕様(イメージ図)と5つの要件

3. 研究の方法

[3-a 研究体制]

本研究は言語研究における容認性判断データと理論仮説を一元的に集約するデータベースを構築するという野心的なものであり、分野内の実験観察・仮説命題に関する深い理論的洞察を持つ言語学者と、データベース構築の豊かなノウハウ・技術を持つ開発体制(一部業務外部委託)との緊密な連携のもとで研究を推進した。開発においては、理論言語学研究に計算言語学の手法を持ち込む研究や言語資源構築に豊富な経験を持つ窪田(研究計画開始当初より研究協力者; 2022年度より研究分担者として参加)が主となりプロトタイプシステムを構築し、成田(研究代表)と協働してアーカイブのデータエントリ登録内容の策定、およびグラフデータベース・ユーザーインターフェースの設計・開発・管理を担当した。また、情報技術開発に高度な知見を持つ松林が綿密に設計の監修を行うことで、人文社会系諸学に不足する工学的・計算機科学的知見を現場レベルで指揮することとなった。データ入力作業には成田が総指揮を採る入力チームが担当した。また、集積したデータの一部に対する再現実験の計画立案を大関が担当した。言語研究諸分野における多様な研究手法や価値付けの在り方を十分に担保しながら作業を進めることで、分野融合的・相互補完的なデータベース設計を実現する。

[3-b 研究計画]

(A)データベース開発に関して、以下3期を各年度に対応させて遂行した。

- (A1) プロトタイプ設計期(プロトタイプ開発、入力ユーザインターフェース検証)
 - (A2) α版設計期(入力仕様・閲覧性の検証、バグ検証、モニターユーザーへの公開)
 - (A3) β版設計期(SNS機能検証期、期間内でのβ版公開、利用者拡大に向けた準備)
- 正式版公開、コミュニティへの普及活動は、今後の継続課題となった。

また、(B)データベースエントリ拡充について、以下の目標を設定し、概ね計画通りに研究を遂行することができた。

- (B1) 著作権利用に同意・入力に協力してくれるモニターユーザー募集、入力マニュアル整備
- (B2-3) 研究協力者およびモニターユーザー有志による GrammarXivers 研究会の発足、月毎(公開/限定)ウェビナー開催。研究補助バイト雇用による組織的入力作業の拡充

さらに、(C)応募者(成田)自身の研究について、下記の研究を推進した。

- (C1) Narita & Fukui (2022) (以下 NF)の完成・出版、応募者自身による研究サンプル登録作業→自らの過去の研究成果の整理集約
- (C2) NFの推進する新たな統辞理論 Symmetrizing Syntax を巡る理論-仮説-データを網羅的にサンプル登録を行った。また、主語性の統語的・意味的基準の統一分析において積み残されていた主語性基準及び主語(島)制約を巡る日英語データの再検討を行った。さらに、GrammarXiv を用いたシステムデモの学会発表に着手し、特に Symmetrizing Syntax の内容について扱う集中講義を東北大学にて行った。
- (C3) 言語学分野の主要学会でのシステムデモ発表を行った。また、日本語の主語性を弁別するテストとしてこれまで広く用いられてきた再帰代名詞「自分」の先行研究を渉猟し、GrammarXiv 成果を論文として投稿・出版した。最後に、研究課題の総括として、GrammarXiv を中心テーマとした論集の出版企画を成田・窪田を主編集担当として計画し、投稿し、受理された(2026年に出版予定)。

	(1) 2021(R3)年度	(2) 2022(R4)年度	(3) 2023(R5)年度
A 設計開発	<ul style="list-style-type: none"> 入力UI設計 <p style="text-align: center;">プロトタイプ完成</p>	<ul style="list-style-type: none"> 事実-仮説-理論の多層的関係の表現 閲覧性の検証 <p style="text-align: center;">α版公開</p>	<ul style="list-style-type: none"> 利用者拡大準備 SNS機能検討 <p style="text-align: center;">β版公開</p>
B エン트리 拡充	<ul style="list-style-type: none"> モニターユーザー募集 入カマニュアル整備 	<ul style="list-style-type: none"> 研究協力者およびモニターユーザー有志によるGrammarXivers研究会の発足、月毎(公開/限定)ウェビナー開催 研究補助バイト雇用→組織的入力作業の拡充 	
C 応募者 自身の研究	<ul style="list-style-type: none"> Narita & Fukui (2022;NF)完成 自らの過去の研究成果の整理集約 	<ul style="list-style-type: none"> GrammarXivを用いたシステムデモ発表 NFの提示する新たな研究枠組み Symmetrizing Syntaxを推進する集中講義 	<ul style="list-style-type: none"> 日本語再帰代名詞「自分」を巡る論文投稿・出版 GrammarXiv論集計画・投稿

図 2: 研究計画概要

4. 研究成果

[4-a システム開発] 3年度を通じ、概ね研究計画通りに研究を進めることができた。

2021(R3) 年度のプロトタイプシステム (ver0.2.0) では、あくまでデータ構造は基本的なリレーショナル DB 構造にとどめながら、研究トピック・理論枠組み・仮説・容認性判断を含むデータ・およびそれらの間の関係エン트리など、様々なエントリタイプの基本プロパティの洗い出しを行った。

2022 (R4) 年度に開発した α 版 (ver.0.4.0; 最終的なウェブリリースは 2023 年 9 月) においては、グラフデータベース Neo4j に基づく関係グラフ描写の基本的なシステム要件を策定した。また、数式エディタ LaTeX と同等の数式表現能力を備え、かつ言語学論文によくつかわれるアノテーションを簡易に記述するための独自の記法を開発した(GramTeX と命名)。また、プロトタイピングで検討を重ねてきた検索結果のグラフ表示機能を新たに本システムに追加し、表示ユーザーインターフェイスのデザイン調整を行った。さらに、専門の弁護士に法律相談を行い、著作権法、個人情報保護法等への法令遵守の観点からデータベースの利用規約およびプライバシーポリシーの策定を行った。また、研究期間内の法令遵守における課題の洗い出しを行った。

2023 (R5) 年度に開発したウェブ β 版 (ver.0.6.0; 2024 年 1 月公開) においては、容認性判断データを含め、GrammarXiv エントリそれぞれに対するユーザーの真偽判断を集約してアンケートを取るための survey 機能を開発・実装した。また、複数の法律専門家と相談をしながら、ウェブ世界公開へむけて利用規約・プライバシーポリシーの最終調整を行った。また、エントリそれぞれに「Japanese-native」(ユーザーエントリ用、日本語の母語話者であるという認定) や「Repeatability Confirmed」(データエントリ用、再現性が確認されているという認定) などの様々な認定をあてがうための「バッジシステム」の簡易版を開発した。SNS 機能の検討には当初想定した以上の時間がかかったが、翌年度以降に順次公開する予定である。

[4-b データサンプルの登録]

2021 年度は、プロトタイプシステム開発と並行しつつ、成田(研究代表者)および窪田(研究協力者; 2022 年度より研究分担者追加)の研究実績を中心にサンプルデータを入力した(ver.0.2.0; サンプルデータ 6778 件)。

2022 年度は、α 版の策定とともにサンプルエントリ登録も順次進め、報告書作成時の 2023/5/17 時点で登録エントリ総数は 23,418 件となった。

2023 年度は、ウェブ β 版の公開とともにテストユーザーによる使用実例も含めてエントリ数が増大した。報告書作成時の 2023/6/17 時点で登録エントリ総数は 38,109 件となった。

[4-c 学会発表等におけるシステムデモ]

2021 年度には、日本言語学会第・言語系学会連合共同開催公開特別シンポジウム『データベースをつくる・つかう:課題と展望』を成田が主となり企画し、大関(研究分担者)が同シンポジウムにて理論言語学データベースに求められる諸条件について研究発表を行った。

2022 年度には、言語処理学会第 28 回年次大会 併設ワークショップ JED2022.「日本語における評価用データセットの 構築と利用性の向上」分科会に参加し、瀧田健介・窪田悠介・成田広樹の 3 名共同研究「GrammarXiv:言語理論・事実のインタラクティブ・データベースの開発」を発表した。

2023 年度には、津田塾大学の国際ワークショップで招待講演を行い、GrammarXiv チュートリアルを世界に向けて発信した(成田・瀧田・窪田)。また、木村博子氏(研究協力者)との共同研究をいくつかの学会にて発表することができた(WCCFL41, および津田塾大学国際ワークショップ)。

[4-d 論文等の著作物]

2021 年度には、成田(研究代表者)が数年かけて取り組んできた福井直樹氏(研究協力者)との共同研究を Routledge より書籍の形で出版することができた(Narita and Fukui 2022 *Symmetrizing Syntax*)。

2022 年度には、研究協力者の木村博子氏、および酒井邦嘉氏の研究室との共同研究を進め、翌年度の出版業績につながる成果を得た。

2023 年度には、日本語の再帰代名詞「自分」を巡る諸問題を概説し、「自分」を伴う容認性判断データの適切な扱いに関わる議論を展開する慶應義塾大学言語文化研究所の紀要論文を執筆した(成田・小林・竹内・小町)。また、本論考の成果を GrammarXiv 上に登録・公開した。また、酒井邦嘉研究室との共同研究(Umejima et al. 2023)および木村博子氏との共同研究(木村・成田 2023、Kimura and Narita 2024a,b)をいくつかの論文の形で出版することができた。

また、これらの諸成果に加えて、本研究計画の期間全体の研究成果をまとめる論集を組む計画を立案している。研究代表者の成田と研究分担者の窪田で共編とし、本研究の研究分担者、研究協力者等で分担執筆する書籍の執筆企画を作成し、出版機関に投稿・受理された。論文集は 2025 年度に出版する予定で各研究分担者、研究協力者と共同して、2023 年度後期より執筆にあたっている。

5. 主な発表論文等

〔雑誌論文〕 計5件（うち査読付論文 4件 / うち国際共著 0件 / うちオープンアクセス 2件）

1. 著者名 Keita Umejima, Isso Nakamura, Naoki Fukui, Mihoko Zushi, Hiroki Narita, and Kuniyoshi L. Sakai	4. 巻 14
2. 論文標題 Differential networks for processing structural dependencies in human language: linguistic capacity vs. memory-based ordering	5. 発行年 2023年
3. 雑誌名 Frontiers in Psychology	6. 最初と最後の頁 1153871
掲載論文のDOI (デジタルオブジェクト識別子) 10.3389/fpsyg.2023.1153871	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 成田広樹・小林亮一朗・竹内士瑛伊・小町将之	4. 巻 55
2. 論文標題 照応表現「自分」をめぐる諸問題: GrammarXivを用いた論点整理と将来的展望	5. 発行年 2024年
3. 雑誌名 慶應義塾大学言語文化研究所紀要	6. 最初と最後の頁 89-113
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 木村博子・成田広樹	4. 巻 1
2. 論文標題 語と句の相違からみる省略・削除の再考	5. 発行年 2023年
3. 雑誌名 大関洋平・漆原朗子(編)『分散形態論の新展開』開拓社	6. 最初と最後の頁 187-211
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 Kimura, Hiroko, and Hiroki Narita	4. 巻 41
2. 論文標題 Immobile Remnants of Japanese Why Stripping	5. 発行年 2024年
3. 雑誌名 Proceedings of the 41st West Coast Conference on Formal Linguistics (WCCFL41)	6. 最初と最後の頁 1-12
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 Kimura, Hiroko, and Hiroki Narita	4. 巻 41
2. 論文標題 Substitutive Why Stripping: A Case of Japanese	5. 発行年 2024年
3. 雑誌名 Proceedings of the 41st West Coast Conference on Formal Linguistics (WCCFL41)	6. 最初と最後の頁 1-12
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 無
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9件 (うち招待講演 3件 / うち国際学会 4件)

1. 発表者名 Kimura, Hiroko, and Hiroki Narita
2. 発表標題 The In Situ Approach to Clausal Elliptical Constructions in Japanese
3. 学会等名 Current Issues in Comparative Syntax 2: Boundaries of Ellipsis Mismatch, Tsuda University [Hybrid Workshop], September 2, 2023. (招待講演) (国際学会)
4. 発表年 2023年

1. 発表者名 Narita, Hiroki, Yusuke Kubota, and Kensuke Takita
2. 発表標題 Introducing GrammarXiv: An Open Access Graph Based Database for Linguists' Truths (with Special Focus on Ellipsis)
3. 学会等名 Current Issues in Comparative Syntax 2: Boundaries of Ellipsis Mismatch, Tsuda University [Hybrid Workshop], September 2, 2023. (招待講演) (国際学会)
4. 発表年 2023年

1. 発表者名 Kimura, Hiroko, and Hiroki Narita
2. 発表標題 Immobile Remnants of Japanese Why Stripping
3. 学会等名 41st West Coast Conference on Formal Linguistics (WCCFL41) (国際学会)
4. 発表年 2023年

1. 発表者名 Kimura, Hiroko, and Hiroki Narita
2. 発表標題 Substitutive Why Stripping: A Case of Japanese
3. 学会等名 41st West Coast Conference on Formal Linguistics (WCCFL41) (国際学会)
4. 発表年 2023年

1. 発表者名 瀧田健介・窪田悠介・成田広樹
2. 発表標題 GrammarXiv:言語理論・事実のインタラクティブ・データベースの開発
3. 学会等名 言語処理学会第28回年次大会併設ワークショップJED2022.「日本語における評価用データセットの構築と利用性の向上」分科会. Evidence based Linguistics Workshop. 国立国語研究所. 2022年9月7日. (招待講演)
4. 発表年 2022年

1. 発表者名 染谷大河, 大関洋平
2. 発表標題 日本語版CoLAの構築
3. 学会等名 言語処理学会 第28回年次大会 (1872-1877)
4. 発表年 2022年

1. 発表者名 染谷大河, 大関洋平
2. 発表標題 日本語版CoLAの構築の舞台裏
3. 学会等名 言語処理学会 ワークショップ「日本語における評価用データセットの構築と利用性の向上」
4. 発表年 2022年

1. 発表者名 染谷大河, 進藤裕之, 大関洋平
2. 発表標題 情報抽出技術を用いたJCoLAの拡張に向けて
3. 学会等名 言語処理学会 第28回年次大会, 290-295
4. 発表年 2022年

1. 発表者名 石月由紀子, 栗林樹生, 松林優一郎, 大関洋平
2. 発表標題 情報量に基づく日本語項省略の分析
3. 学会等名 言語処理学会 第28回年次大会, 442-447
4. 発表年 2022年

〔図書〕 計1件

1. 著者名 Hiroki Narita and Naoki Fukui	4. 発行年 2022年
2. 出版社 Routledge	5. 総ページ数 344
3. 書名 Symmetrizing Syntax: Merge, Minimality, and Equilibria	

〔産業財産権〕

〔その他〕

GrammarXiv https://grammarxiv.net/ SUGHT (Narita Lab) https://sought-naritalab.com/
--

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
研究分担者	大関 洋平 (Oseki Yohei) (10821994)	東京大学・大学院総合文化研究科・講師 (12601)	
研究分担者	松林 優一郎 (Matsubayashi Yuichiroh) (20582901)	東北大学・教育学研究科・准教授 (11301)	
研究分担者	窪田 悠介 (Kuboya Yusuke) (60745149)	大学共同利用機関法人人間文化研究機構国立国語研究所・研究系・准教授 (62618)	

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