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　 ５ 月 ２０ 日現在

機関番号：１４５０１
研究種目：挑戦的研究（萌芽）
研究期間：2020～2023
課題番号：２０Ｋ２０６９９

研究課題名（和文）言語から見た日米マインドスケープ比較：データサイエンス志向型小説研究の試行

研究課題名（英文）Novels and Data Sciences: Mindscape Seen in Language

研究代表者
石川　慎一郎（Ishikawa, Shin'ichiro）

神戸大学・大学教育推進機構・教授

研究者番号：９０３２０９９４

交付決定額（研究期間全体）：（直接経費）　　4,600,000円

研究成果の概要（和文）：異言語資料を用いた計量的マインドスケープ比較研究の基盤整備を行うべく、(1)既存英米コーパスの小説モジュールと比較できるよう、三大文芸誌より日本語小説データを収集した。(2)言語学的解析を可能にする品詞タグ付けを実施し、専門家による修正を行った。(3)2種のAI翻訳で全日本語テキストに英訳付与を行った。(4)著作権処理を経て「6121JFIC」コーパスを公開した。(5)英語・日本語の複数テキストデータに対して自動で形態素解析を行い、統合頻度表を出力するEJWFTGを2023年度末に開発・公開した。同様の試みは過去に例が僅少で、新しい人文テキスト研究の可能性を広げる一助になると思われる。

研究成果の学術的意義や社会的意義
伝統的な文学研究や、各種のイメージ研究、マインドスケープ（心象風景）研究は、分析者の主観に依存する部分が大きく、研究手法の記録と管理、また、研究結果の再現性の保証という点で、制約が多かった。本研究課題では、まずもって、日本語小説を時系列的に収集した新しい言語資源として6121JFICを構築したことで、既存の英米の小説コーパスと対照研究を行う素地が確立された。さらには、EJWFTGの開発とリリースにより、文芸テキストに限らず、日本語・英語の様々な言語テキストをオンライン上で単語レベルで解析し、統合語彙表を自動生成する環境が構築された。これらは、テキスト研究のさらなる発展の一助となりうる。

研究成果の概要（英文）：In order to develop a foundation for quantitative comparative studies of mindscapes using different language materials, (1) we collected Japanese novel data from three major literary magazines so that we can compare them with existing English novel corpora, (2) conducted part-of-speech tagging to enable linguistic analysis, (3) added English translations using two types of AI translations, (4) released “6121JFIC” corpus after clearing copyright matters, and (5) released "EJWFTG", which automatically performs morphological analysis on multiple text data in English and Japanese and outputs an integrated frequency table. Our project is practically the first attempt in the past, and it is expected to help expand the possibilities of new text studies.

研究分野：コーパス言語学

キーワード：コーパス　日英対照分析　イメージ研究　マインドスケープ研究　自動語彙分析　計量的テキスト分析

様 式 C−19、F−19−1（共通）

1.研究開始当初の背景/Background
小説をはじめとする文芸テキストは、その国の母語話者が意識的・無意識的に継承する文化・風俗・精神・言語の表象である。小説は、まずもって、文学研究の対象であるが、「データサイエンス」の観点から見ると、多彩な文化研究の新しいリソースでもありうる。ランダムサンプリングされた日本語・英語の小説資料を比較することで、両国の精神風景（マインドスケープ）を計量的に論じることが可能になる。本研究では、米語コーパスの原型とされる Brown Corpus の「小説セクション」と比較可能な日本語小説コーパスを構築した上で、日本語・英語小説データの言語分析に基づく計量的マインドスケープ比較を実施する研究環境を構築し、もって、新しい観点に基づくテキスト研究の可能性を試行しようとするものである。

Fiction and other literary texts are representations of culture, customs, spirit, and language that the native speakers of a country consciously or unconsciously inherit. Fiction is, first and foremost, the object of literary research, but from the perspective of “data science,” it can also be a new resource for diverse cultural studies. A comparison of Japanese and English fiction makes it possible to quantitatively discuss the mental landscapes (mindscapes) of the two discourse communities. In this study, we aim to construct a research platform for quantitative analysis of mindscapes of different language communities by developing a new Japanese fiction corpus that can be compared with the “fiction section” of the Brown Corpus, a prototype of many English corpora. This project attempts to explore possibilities of new textual research based on a new perspective.

2.研究の目的/Aim
本科研プロジェクトでは、以下の 5 つの目標を掲げた。(1)既存英米コーパスの小説モジュールと比較できるよう、三大文芸誌より日本語小説データを収集する。(2)収集した小説データについて、言語学的解析を可能にする品詞タグ付けを実施する。(3)英米小説との直接比較が可能になるよう、収集した日本語小説データに英訳付与を行う。(4)著作権処理を経て収集したデータを公開する。(5)小説以外のテキストデータについても、同様の比較研究が実施できるよう、英語・日本語の複数テキストデータに対して自動で形態素解析を行い、統合頻度表を出力するシステムを開発する。

The following five goals were set for this project: (1) to collect Japanese fiction data from the three major literary magazines so that it can be compared with the fiction subset of the existing English corpora, (2) to perform part-of-speech tagging of the collected fiction data to enable linguistic analysis, (3) to add English translations to the collected data to enable direct comparison with British and American fictions, (4) to release the collected data as a corpus after processing copyright matters, and (5) to develop an online system that automatically performs morphological analysis on multiple text data in English and Japanese and produces an integrated word frequency table so that users can apply a similar comparative approach to a wider variety of text data.

3.研究の方法/Method
【コーパス開発】 このプロジェクトでは、1961 年、1971 年、1981 年、1991 年、2001 年、2011 年、2021 年の 7 つのデータポイントを定め、各年に発行された「文学界」（文藝春秋）、「群像」（講談社）、「新潮」（新潮社）に掲載された小説作品 217 本（各 5,000 字）を収集する。これにより、同年にデータを収集した Brown (1961, US), Frown (1992, US), Crown (2009, UK), LOB (1961, UK), FLOB (1991, UK), BE06 (2006, UK), CLOB (2009 , UK) などとの直接的な比較が可能になる。収集した日本語小説資料は、電子化して品詞タグ付けを行い、1961-2021 Japanese General Fiction Corpus (6121JFIC)としてリリースする。また、オンライン検索サイトを開発し、公開する。
【語彙頻度表作成システム開発】 日本語・英語を対象として、複数テキストに含まれるすべての語を形態素解析して、個々の語の頻度を調べ、それらを統合語彙表として出力する新しいオンラインシステムを開発し、リリースする。

[Corpus Development] This project will define seven data points (1961, 1971, 1981, 1991, 2001, 2011, and 2021) and collect 217 fictions (5,000 characters each) published in Bungakukai (Bungeishunju), Gunzo (Kodansha), and Shincho (Shinchosha) in the respective years. This will allow direct comparison with Brown (1961, US), Frown (1992, US), Crown (2009, UK), LOB (1961, UK), FLOB (1991, UK), BE06 (2006, UK), CLOB (2009 , UK) and others for which data were collected in the same year. The collected Japanese fiction materials will be digitized, part-of-speech tagged, and released as a new corpus (1961-2021 Japanese General Fiction Corpus/6121JFIC).

[Development of a Word Frequency Table Generator] A new online system will be developed that analyzes the frequencies of all words appearing in multiple Japanese and English texts and outputs the results as an integrated vocabulary list.

4. 研究成果/Outputs
(1) コーパスの構築と公開
2022 年 2 月度に、6121JFIC (V1.0) をオンラインで公開した。データセットは、石川研究室で構築しているコーパス統合検索サイトに統合され、多様な検索が可能になっている。また、2024 年 3 月には、形態素解析データを手作業で修正した v2.0 が公開された。

(1) Corpus Construction and its Release
6121JFIC (V1.0) was released online in February 2022. The dataset has been integrated into the corpus query system constructed by Dr. Ishikawa Lab, which enables a variety of corpus searches. In March 2024, v2.0 was released, in which morphological analysis data were re-examined and manually modified.

図 1 コーパス紹介サイト

図 2 終助詞「わね」を含む用例の抽出

図 3 英訳テキスト中の love の包含例（英訳検索モード）

(2) 語彙表作成ツールの開発と公開
English/Japanese Word Frequency Table Generator（EJWFTG）を開発し、2024 年 3 月に公開した。

(2) Development and Release of Vocaulary Analysis Tool
English/Japanese Word Frequency Table Generator（EJWFTG） was released in March 2024.

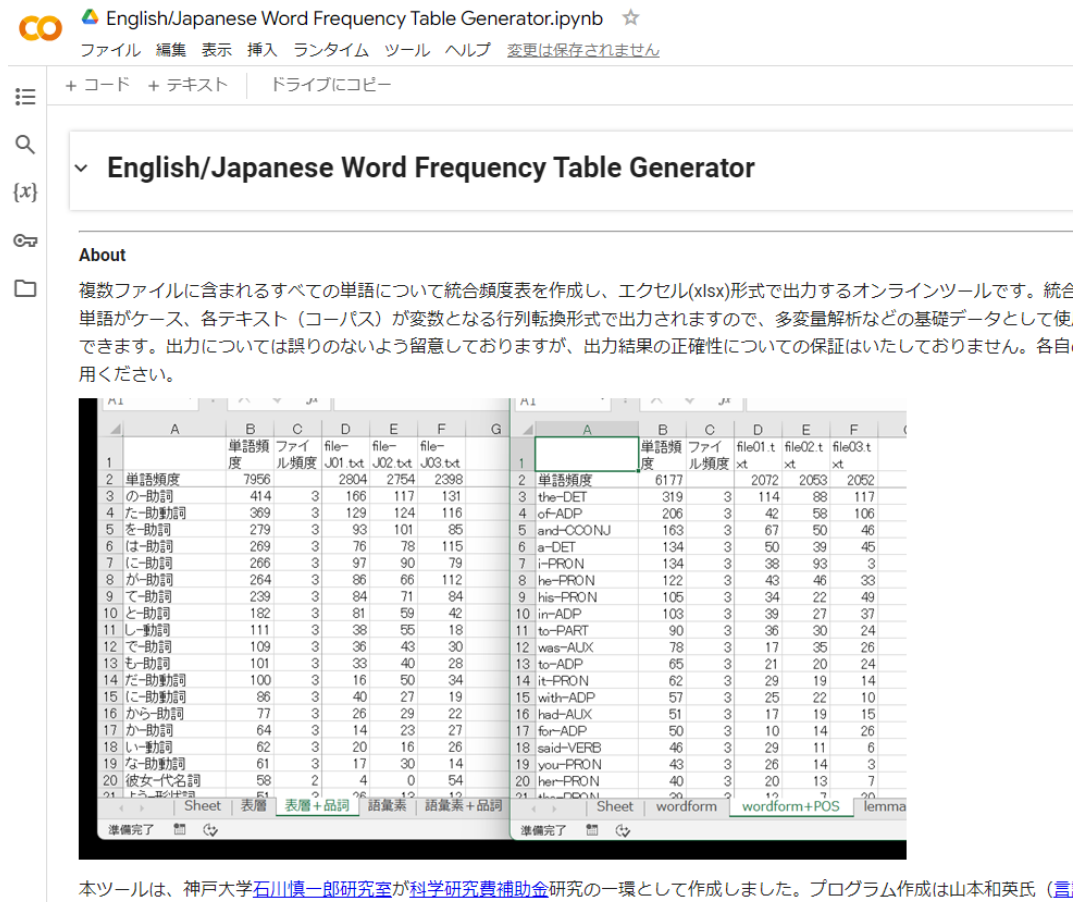

図 4 EJWFTG の画面

(3) 関連論文の公刊
6121JFIC 関連の論文として、「「1961-2021 日本語小説コーパス」の構築―日英小説対照研究の新しい可能性―」（石川、2021）、「時代変種と学習者変種の観点から考える日本語終助詞―時系列日本語小説コーパス「6121JFIC 」と国際日本語学習者コーパス「I-JAS」を用いた統合分析の試み」（石川、2022）ほか、また、EJWFTG 関連の論文として、「森を見ながら木を見る」コーパス研究の意義 : 複数テキストから統合語彙頻度表を作成する EJWFTG の開発」（石川、2024）ほかが刊行された。

(3) Publication
Regarding 6121JFIC Corpus, Ishikawa (2021), Ishikawa (2022), and other papers were published. Also, regarding EJWFTG, Ishikawa (2023) was published.

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　0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 石川 慎一郎	４．巻 469
２．論文標題 「森を見ながら木を見る」コーパス研究の意義 ： 複数テキストから統合語彙頻度表を作成するEJWFTGの開発	５．発行年 2024年
３．雑誌名 統計数理研究所共同研究リポート	６．最初と最後の頁 95～122
掲載論文のDOI（デジタルオブジェクト識別子） 10.24546/0100487709	査読の有無 無
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 横山 詔一、石川 慎一郎	４．巻 29
２．論文標題 オープンサイエンス時代の言語系研究と教育 ： プレプリントの公開をめぐって	５．発行年 2022年
３．雑誌名 言語・情報・テクスト ： 東京大学大学院総合文化研究科言語情報科学専攻紀要	６．最初と最後の頁 67～80
掲載論文のDOI（デジタルオブジェクト識別子） 10.15083/0002005966	査読の有無 無
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 石川慎一郎	４．巻 2021
２．論文標題 「1961-2021日本語小説コーパス」の構築―日英小説対照研究の新しい可能性―	５．発行年 2021年
３．雑誌名 英語コーパス学会大会予稿集	６．最初と最後の頁 7-12
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 －

１．著者名 石川 慎一郎	４．巻 456
２．論文標題 時代変種と学習者変種の観点から考える日本語終助詞 ： 時系列日本語小説コーパス「6121JFIC」と国際日本語学習者コーパス「I-JAS」を用いた統合分析の試み	５．発行年 2022年
３．雑誌名 統計数理研究所共同研究リポート	６．最初と最後の頁 73～88
掲載論文のDOI（デジタルオブジェクト識別子） 10.24546/81013070	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 －

1．著者名 石川慎一郎	4．巻 23
2．論文標題 習得研究の資料としての学習者コーパスの可能性と課題：計量研究におけるコーパスデータの制約性をめぐって	5．発行年 2020年
3．雑誌名 第二言語としての日本語の習得研究	6．最初と最後の頁 138-144
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 －

1．著者名 石川慎一郎	4．巻 54(1)
2．論文標題 計量的言語研究の現状と展望：検証型研究と探索型研究の界面	5．発行年 2021年
3．雑誌名 看護研究	6．最初と最後の頁 10-17
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10件（うち招待講演　7件／うち国際学会　3件）

1．発表者名 石川慎一郎
2．発表標題 コーパスと日本語教育：コーパスで変わる日本語の教授と学習
3．学会等名 上海大学外国語学院言語文化と文明系列講座No．106（招待講演）（国際学会）
4．発表年 2023年

1．発表者名 石川慎一郎
2．発表標題 日本語研究と計量的手法：言葉を数えることで見えてくること
3．学会等名 上海外国語大学日本語学講演会（招待講演）（国際学会）
4．発表年 2023年

１．発表者名
石川慎一郎

２．発表標題
データとAIの時代における日本語の研究と教育 ―変わるもの、変わらないもの―

３．学会等名
北京外国語大学主催日本語学講演会（招待講演）（国際学会）

４．発表年
2023年

１．発表者名
石川慎一郎

２．発表標題
「森を見ながら木を見る」コーパスデータ処理方法の提案―英語・日本語の複数テキストから形態素解析済み統合頻度表を自動作成するEJ-WFTGの開発―

３．学会等名
英語コーパス学会ESP研究会例会（招待講演）

４．発表年
2024年

１．発表者名
石川慎一郎

２．発表標題
「1961-2021日本語小説コーパス」の構築―日英小説対照研究の新しい可能性―

３．学会等名
英語コーパス学会第47回大会

４．発表年
2021年

１．発表者名
石川慎一郎

２．発表標題
日本語学習者データと日本語時系列データが出会うとき：I-JASと6121JFICの統合分析の試みー終助詞をめぐって－

３．学会等名
学習者コーパス研究会2022年2月例会

４．発表年
2022年

１．発表者名
石川慎一郎

２．発表標題
経年的日本語小説コーパス「6121JFIC」の開発と公開－日本語時系列分析の新しいリソースとして－

３．学会等名
統計数理研究所「計量的コーパス研究の展望2022」

４．発表年
2022年

１．発表者名
石川慎一郎

２．発表標題
MMRオープンフォーラム：計量的言語研究の現状と展望―検証型研究と探索型研究の界面―

３．学会等名
第6回日本混合研究法学会年次大会(JSMMR2020)シンポジウム（招待講演）

４．発表年
2020年

１．発表者名
石川慎一郎

２．発表標題
言語研究における有意性検定の今後の動向を考える

３．学会等名
学習者コーパス研究会第7回例会（招待講演）

４．発表年
2020年

１．発表者名
石川慎一郎

２．発表標題
多次元分析法（MD法）による学術論文の言語特性分析―コンピュータ工学系論文とコンピュータ援用言語学習系論文の比較―

３．学会等名
ESPシンポジウム2021「ジャンルとしての工学英語―理論と実践－」（招待講演）

４．発表年
2021年

〔図書〕　計3件

１．著者名 石川慎一郎／長谷部陽一郎／住吉誠	４．発行年 2020年
２．出版社 開拓社	５．総ページ数 273
３．書名 『コーパス研究の展望』	

１．著者名 石川慎一郎	４．発行年 2021年
２．出版社 ひつじ書房	５．総ページ数 277
３．書名 『ベーシックコーパス言語学』第２版	

１．著者名 【編】石川有香／【著】石川有香／Judy Noguchi／石川慎一郎／松田真希子／竹井智子／福永淳／小野義正	４．発行年 2021年
２．出版社 大学教育出版	５．総ページ数 232
３．書名 『ジャンルとしての工学英語ー理論と実践ー』	

〔産業財産権〕

〔その他〕

6121JFICウェブサイト
https://language.sakura.ne.jp/jfic/index.html

6121JFICコーパス検索リンクページ
http://language.sakura.ne.jp/onlinecorpus.html

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