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 ３０ 日現在

機関番号：１４４０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ００７４２

研究課題名（和文）自然言語処理CUIアプリケーションの汎用GUIコーパスツールへの組み込み

研究課題名（英文）Incorporation of a Natural Language Processing CUI Application into a General-Purpose GUI Corpus Tool

研究代表者

今尾　康裕（Yasuhiro, Imao）

大阪大学・大学院人文学研究科（言語文化学専攻）・准教授

研究者番号：５０６０９３７８

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：本研究は、CUIの自然言語処理アプリケーションをGUIコーパス分析アプリケーションであるCasualConcから利用できるようにして、自然言語処理の研究成果を手軽に言語研究へ応用する橋渡しをすることを目指したもので、依存文法による文法タグ付け処理ができるStanford CoreNLPとトピックモデル分析ができるMALLETとの連携機能をCasualConcに組み込んだ。また、利用を促進するためにワークショップを行うとともに、オンラインチュートリアルなどの作成を行った。

研究成果の学術的意義や社会的意義
本研究の成果であるCasualConcと自然言語処理アプリケーションの連携は、コマンドラインでの入力やスクリプト言語での処理など、CUIアプリケーションの利用促進を妨げる要素を取り除き、処理の設定や結果の扱いを容易にすることで、技術的に不利な立場にある研究者が結果の解釈に注力できる環境を整えて、これらを利用した言語分析の更なる応用への道を開き、言語研究自体の発展に寄与することができるであろうと考える。

研究成果の概要（英文）：The goal of this research was to make CUI natural language processing application programs accessible from the GUI corpus analysis application CasualConc. This would bring the products of natural language processing research to language studies. The dependency grammar tagging function of Stanford CoreNLP and the topic modeling analysis function of MALLET were incorporated into CasualConc. Additionally, to promote its use, workshops were given and online tutorials were created.

研究分野：ツール開発

キーワード：コーパスツール開発　自然言語処理　GUIアプリケーション

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

1990 年代からコンピュータの急速な発達により広まったコーパス分析は、ローカルのコンピュータで使用できるアプリケーションから、ウェブ上でのツールにその中心が移ったが、大規模なコーパスではなく、小規模なコーパスやデータを公開できないコーパスなどの分析は、ローカルでのアプリケーションを利用することが中心となっている。ただし、そのようなアプリケーションは、文字列の処理を中心として、KWIC や頻度集計にとどまるものがほとんどであった。そのような状況下、Mac 用汎用コーパス分析アプリケーションである CasualConc を開発し、作成した頻度表を元に統計環境 R と連携させて、多変量分析を含めたデータを視覚化する機能を追加した。

自然言語処理関連では、研究成果を基に CUI での UNIX アプリケーションが公開されていたが、使用するにはコマンドラインやスクリプトが書ける必要があり、多くの研究者には手の届かないものであった。そこで、GUI で操作できる汎用コーパス分析アプリケーションである CasualConc と連携させることで、手軽に自然言語処理研究の成果を利用できる環境を提供することを思いついた。

２．研究の目的

本研究では、自然言語処理を行うための UNIX アプリケーションとの連携機能を CasualConc に組み込み、テキスト処理を行うだけでなく、処理した結果を利用して更なる分析へと繋げるために、これらのアプリケーションの結果を読み込んで、CasualConc の既存の機能を利用してデータを視覚化するなど、新たな分析への道を開くことを目的とした。それと同時に、チュートリアルやワークショップなどを通じて、実際に研究者が分析に利用することを促進することを目指した。

３．研究の方法

(1) 本研究開始時にすでに着手していたものも含め、申請時点で候補に入れていたアプリケーションを中心に、自然言語処理関連の専門書籍や論文、さらには、他の研究者らが公開している自然言語処理アプリケーション利用のチュートリアルなどを参照するとともに、本務校における同僚研究者や大学院生などとの交流などから、CasualConc に連携させる UNIZ CUI アプリケーションを選定した。

(2) Stanford CoreNLP を利用した依存文法に基づく文法タグ付けと MALLET を利用したトピックモデル分析の機能を組み込むことを決定し、まずは Stanford CoreNLP で文法タグ付けをするためのアプリケーションを開発し、CoreNLP での分析結果をもとにデータベースファイルを作成して、それを読み込んで結果を表示させる機能を CasualConc に追加した。MALLET については、テスト用に、MALLET を利用してトピックモデルの分析をし、その結果を読み込んでデータを視覚化する機能を持ったアプリケーションを開発し、トピックモデルでの研究を行う研究者の協力を得て、機能の追加・改善を行なったのち、CasualConc に組み込んだ。

(3) ウェブチュートリアルを作成し、機能を追加した時点で学会発表を行い、ワークショップなどで CasualConc の利用を促進した。

４．研究成果

本研究の助成初年度は、5 月にオンライン開催された外国語メディア教育学会関西支部の地区大会において、CasualConc と関連アプリケーションを利用したコーパス分析のワークショップを行なった。ワークショップ自体は、スライドショーを書き出したもので、テキスト収集・処理を含めたコーパス作成から CasualConc でのコーパス分析をその手順を含めて紹介した。また、このワークショップスライドに連携させた CasualConc などの操作ビデオを作成し、ウェブチュートリアルの形で、合計 34 本の操作ビデオクリップを YouTube 上に公開した。現在も、外国語メディア教育学会関西支部のウェブサイトを通してワークショップのスライドショーおよびビデオチュートリアルへのアクセスが可能である。

アプリケーション開発面では、さまざまな観点から考慮した結果、申請時点で候補に入れていた UNIX アプリケーションとの連携を優先して行うことに決定した。優先順位としては、すでに試験的な開発を開始していた Stanford CoreNLP を利用した文法検索機能を最優先として、次に

MALLET でのトピックモデル分析機能を追加することにした。

初年度の開発は、Stanford CoreNLP による依存文法解析を利用して文法タグ付け処理を行ったデータを基に SQLite データベースを作成する機能と、文法タグを利用して依存関係にあるコロケーションを検索する機能を中心に行なった。データベース作成に関しては、文法タグ情報とともに品詞タグやレマなどの情報をそれぞれ別のテーブルに格納する構造にした。GUI に関しては、すでにプロトタイプを作成していたため、文法タグを指定して検索するモード、英語用にあらかじめ文法タグを組み合わせて検索できるプリセットを作成し、それを指定することで検索するモードを用意した。結果の出力は、コロケーションの頻度を集計するものと、検索したコロケーションを含む文を表示するものを用意した。この GUI のオプションに関しては、研究者が分析に使用することを最終的な目的としているため、学習者などが簡易的に検索できることよりも、研究者向けに検索の柔軟性を重視したものにした。

データベースのテーブル構造については、複数のテーブル構造を持つデータベースファイルを作成し、検索処理における処理時間と使用メモリを参考に、最適なデータベースのテーブル構造を選択した。また、検索・集計処理においても、さまざまな SQL 検索パターンとインデックスの作成の組み合わせなどを試し、最適化を行なった。この最適化に関しては、継続して行なっており、最新版のリリースにおいても更なる改良を行なっている。

図 1 文法プリセット検索・コロケーション頻度集計

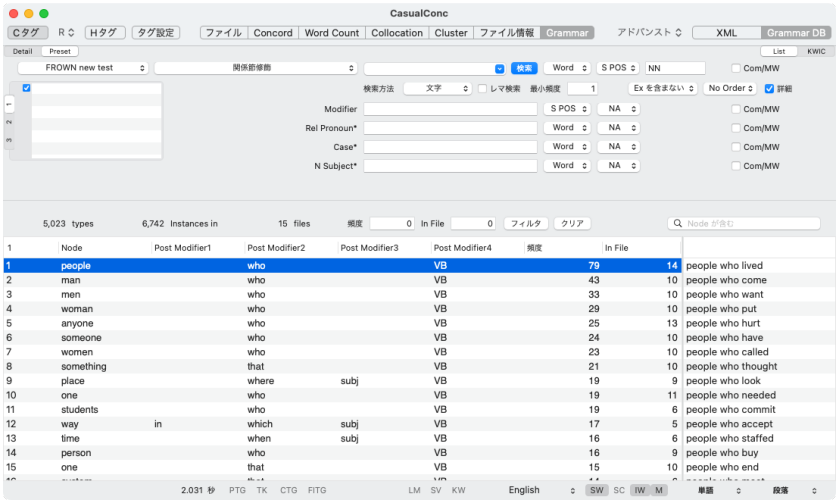
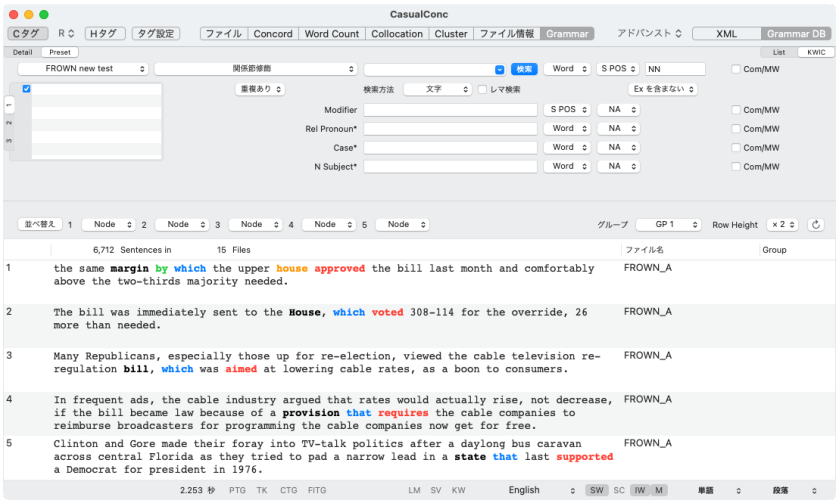


図 2 文法プリセット検索・KWIC 表示



このデータベースを使った分析は、文法検索の機能だけではなく、KWIC 検索、単語・n-gram リスト作成、Collocation 検索などの CasualConc の既存のツールでの処理も可能とし、文字列検索においてはレマ・品詞タグを利用した検索も可能にし、大規模なコーパスにおいて処理に時間とメモリを消費する n-gram リストの作成などをあらかじめ作成しデータベースファイルに保存して読み込む機能を追加した。これにより、文法検索のためのテキストデータベー

ス、つまりコーパスを多面的に分析することを可能とした。

この文法検索機能を備えたバージョンは、2022 年春に公開し、同年の英語コーパス学会で新機能について発表し、2023 年の年明けに行なったワークショップでこの機能についても紹介し、普及に努めた。

2023 年度からは、トピックモデル分析が行える UNIX アプリケーションの MALLET との連携について調査をはじめ、どのような分析が行われているか、データの視覚化にはどのようなグラフが用いられているかを探り、CasualConc で MALLET の処理や結果をどのように扱うかについて検討を行った。これと並行して、word2vec などの「単語埋め込み (word embedding)」アプリケーションとの連携も検討したが、CasualConc からのアクセスが困難であることが判明したため保留とし、MALLET との連携に注力することにした。

2022 年度後半からは、MALLET を GUI で扱うためのテストアプリケーションである CasualMallet の開発に着手した。開発の方向性としては、MALLET のトピックモデルの機能だけに特化して連携させ、分析のオプションを GUI で設定することを可能とするとともに、MALLET で処理するためのテキストデータの下処理を行う機能や、MALLET での分析結果出力を読み込んで統計処理をした上で、結果データの視覚化を行う機能を備えたものとして開発した。CasualConc は、各種機能をモジュール化して組み込むことができるような設計になっているため、テストアプリケーションの機能の一部は、CasualConc に組み込むモジュールに改変できる設計とした。2022 年末に限定的に CasualMallet を公開し、トピックモデル分析を行っている研究者や本務校での研究会などでフィードバックを得て、CasualMallet の修正を行うとともに、CasualConc に組み込む機能について検討した。

2023 年度は、引き続き CasualMallet の機能を CasualConc へと組み込んで、既存のツール・機能との連携について検討するとともに、上記の文法検索機能の強化を行った。2023 年度後半には、CasualMallet の機能の組み込みと CasualConc の既存機能との連携について方針を固めて、組み込みを開始した。開発自体の大枠はほぼ 2023 年度内に終了したが、機能の確認やマニュアルの整備などが間に合わず、年度内に公開することが叶わなかった。しかしながら、本報告書執筆中の 5 月下旬には、MALLET での処理と結果の表示および結果の統計処理、既存の機能との連携を追加したバージョンと対応するマニュアルを公開した。

機能的には、処理するテキストファイルを含むフォルダを指定して MALLET の設定を行い、MALLET でトピックモデル分析処理を行った結果を SQLite データベースファイルに保存して、そこからデータを読み込んで表示させる方式をとった。また、MALLET で処理するテキストファイルの下処理を独立したツールとして組み込み、CasualConc に登録されているコーパスファイルの処理を行うことや、CasualConc のタグフィルタ機能などを使った処理を行うことを可能にした。保存した MALLET の出力結果は、CasualMallet と同様によく行われる統計処理・視覚化は簡易的に行えるようにするとともに、CasualConc の R との連携機能と連携して処理させることも可能にした。

図 3 CasualConc での MALLET 処理設定

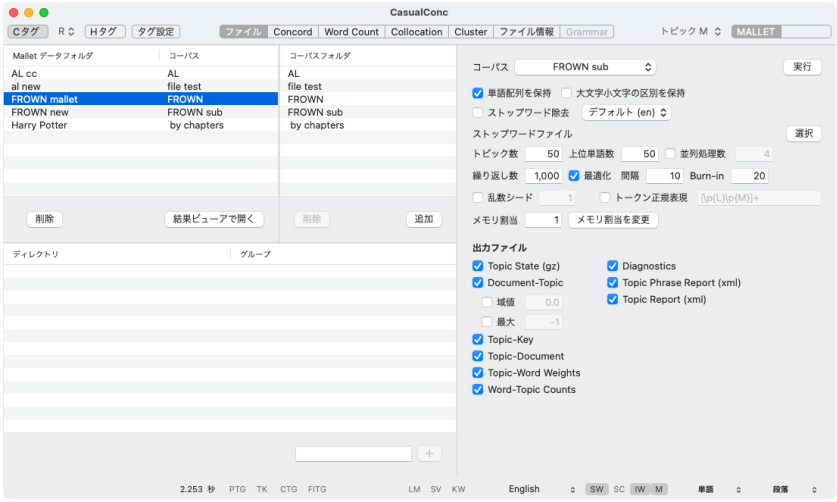


図 4 MALLET の出力（結果）表示

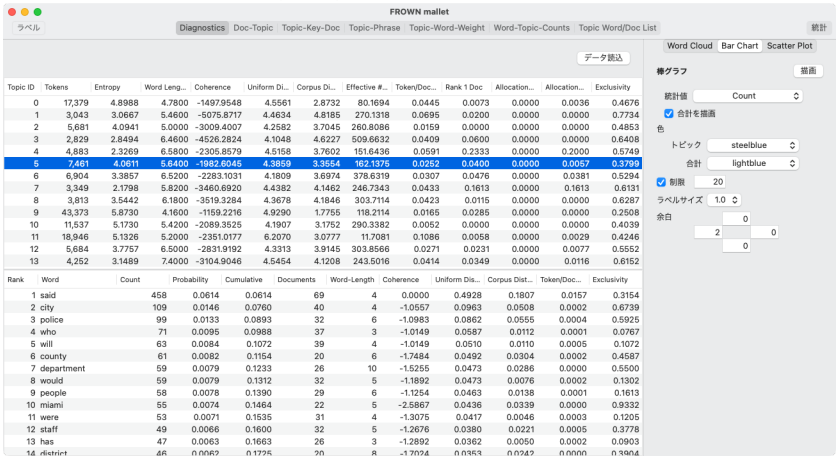
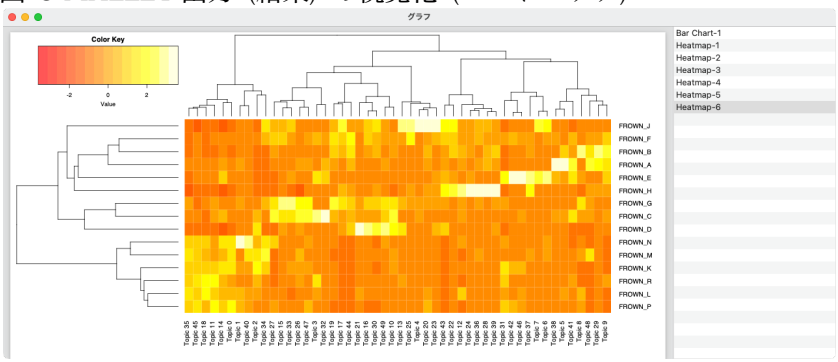


図 5 MALLET 出力（結果）の視覚化（ヒートマップ）



本研究の助成受付期間中には、コロナ禍や CasualConc を動かす環境である Apple 社の Mac において、CPU の変更という大きな変化があり、これらへの対応に時間を要したため、研究計画に含めた、新機能を使った応用研究や、iPad への基本的なテキスト処理機能の移植、ウェブチュートリアルの充実などは達成できなかったが、2024 年度からの基盤研究 C の助成を受ける研究において、CasualConc の機能の充実や iPad への移植などを行うことになっている。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 今尾康裕
２．発表標題 CasualConc 3.0 - Universal Dependency タグを利用した文法検索の試み
３．学会等名 英語コーパス学会
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

CasualMallet https://sites.google.com/site/casualconc/other-applications/casualmallet CasualConc（アプリケーションサイト） https://sites.google.com/site/casualconcj/ CasualConc Blog https://casualconc.blogspot.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

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