G-12

生成 AI を活用した消滅危機言語の方言地図作成支援手法：アイヌ語を事例とした Web アプリの開発

Creating Dialect Maps for Endangered Languages using Generative AI: A Web Application Case Study on the Ainu Language

春日 勇人†

Hayato Kasuga

1. はじめに

消滅危機言語における方言の地理的分布を視覚化する言語地図の作成は言語の多様性を記録しその歴史的変遷を分析する上で重要な研究手法である．近年のコンピュータ技術の発展はこの分野に大きな変革をもたらした．先行研究においては，専用ソフトウェアの開発，汎用的な地理情報システム（GIS）の活用，さらには機械学習を用いたデータ分類など，地図作成の効率化と分析の高度化を目指す多様なアプローチが試みられてきた．例えば GIS を用いることで言語の分布を地形や人口密度といった言語外情報と重ね合わせ，その背景にある社会的・地理的要因を考察する研究が行われている[1]．また機械学習を導入し，語形の音韻的特徴から方言を分類する試みも報告されている[2][3].

これらの先行研究が主に「分析手法の高度化」に焦点を当ててきたのに対し，本研究は異なる視点からアプローチする．それは大規模言語モデル（LLM）を方言地図の開発に活用することにより，研究者が直面する「データの前処理から成果物の生成に至る一連のワークフロー」そのものを支援し，自動化する Web アプリケーションを構築するという試みである．本稿では筆者が開発した「アイヌ語方言地図 SVG ジェネレーター」を事例として先行研究との比較を通じて本研究手法の新規性と独自性を明らかにする．

LLM との対話を通じてアプリケーションを開発するというプロセス自体の革新性，そして生データから多様な形式の成果物までをシームレスに生成する統合的な機能が従来の研究手法といかに異なるかを論じる．

2. 先行研究

2.1 コンピュータを用いた言語地図作成の変遷と展望

方言の地理的分布を視覚化する言語地図は，言語の歴史的変遷や伝播の様相を解明するための重要な資料である．かつて，その作成は地図に手作業でスタンプを押印するといった膨大な労力を要するものであったが，コンピュータ技術の発展は言語地図の作成手法に大きな変革をもたらした．初期にはメインフレームコンピュータを用いた作図システムが開発され，やがてパーソナルコンピュータの普及に伴い，研究者自身が開発した専用ソフトウェアや汎用作図ソフトウェアが利用されるようになった．

近年では GIS の導入により単なる地図作成の効率化に留まらず，言語情報と地形，交通網，人口密度といった多様な地理空間情報とを重ね合わせて分析する，より高度な言語地理学的研究が可能となった．さらに機械学習技術の応用は研究者の主観を排した客観的な語形の分類や言語地図が示す分布パターンの類型化といった新たな研究領域を切り開きつつある．

コンピュータを用いた言語地図作成の技術的変遷を概観し，データベースの構築から GIS による空間分析，そして機械学習の応用へと至る各段階での成果と課題を詳述する．これにより現代の言語地理学が持つ可能性と今後の展望を明らかにする．

2.2 言語地図作成の電子化とデータベースの構築

コンピュータによる言語地図作成の第一歩は，調査で得られた言語データの電子化である．日本では国立国語研究所が実施した大規模調査である「日本言語地図」（LAJ）や「方言文法全国地図」（GAJ）が，その主要な対象となった．これらのプロジェクトで収集された膨大な紙媒体の調査カードは手作業で入力され，コンピュータで扱えるデータベースとして整備された．特に GAJ では編纂当初からコンピュータ処理を見据え，データがフロッピーディスクに保存されていた[4]．これらのデータベースは，後に Excel 形式やテキストファイルとして Web 上で公開され，多くの研究者が利用できるようになった．

電子化されたデータを用いて言語地図を描画するため様々なツールが開発されてきた．古くは荻野綱男による GLAPS，PC 時代になってからは福嶋秩子が開発した SEAL（System of Exhibition and Analysis of Linguistic Data）などが知られる．SEAL は，白地図の作成からデータ入力，描画，集計までを一貫して行えるソフトウェアであり，複数の言語地図の重ね合わせも可能である[5].

一方で汎用の作図ソフトウェアを応用する試みも行われた．大西拓一郎は，Adobe Illustrator のプラグインとして機能する LMS（Language Map System）を開発した．これは，調査地点番号と記号コードを記述した CSV ファイルを読み込み，Illustrator 上の白地図の対応する位置に自動で記号を配置するシステムである[6]．この方法は Illustrator の持つ高い描画品質を活用しつつ，地図作成の自動化を実現するものである．

これらの取り組みにより言語地図の作成は格段に効率化され，データの再分析や異なる仮説に基づく地図の再描画が容易になった．コンピュータ上で扱うデータは単なる資料ではなく，検索や抽出が可能な「言語データベース」としての性格を強め，後の GIS や機械学習による分析の基盤を築いたのである．

2.3 GIS の導入と多角的な空間分析

言語地理学研究における大きな進展は GIS の導入によってもたらされた．GIS とは位置情報を持つ多様なデータ（空間データ）を統合的に管理，加工，可視化し，高度な

†京都芸術大学，Kyoto University of the Arts

空間分析を可能にする技術である．言語研究において GIS を利用する最大の利点は，言語情報と，地形，河川，交通網，行政区画，人口密度といった言語外の地理情報を共通のプラットフォーム上で結びつけ，その相関関係を客観的に分析できる点にある[5].

GIS の最も基本的な分析手法は異なる主題図のレイヤを重ね合わせる「オーバーレイ」である．大西（2007）は GIS を活用し，従来の方言周圏論に代表される「配列モデル」だけでは解明が困難であった言語事象の分析を試みた．例えば待遇表現の分布を人口密度や家族構成（核家族世帯の割合，1 世帯あたりの人数）のデータと重ね合わせることで共通語形「オ書キニナル」が人口密度の高い大都市圏に現れやすい傾向や，父親への尊敬語の使用が世帯規模の小さい「小家族制」の社会構造と関連している可能性を指摘した[1].

また林（2016）は LAJ の「かたつむり」のデータを秋田県内の地図にプロットし，地形や交通網のデータと重ね合わせることで特定の語形の広がりが平野部に限定されたり国道に沿って分布したりする様相を視覚的に示した．これにより山などの地理的障害や道という伝播経路が言語分布に与える影響を具体的に考察できることが示された[4]．さらに峪口有香子・平井松午・岸江信介（2013）は瀬戸内海地域を対象に過去の言語地図データと現代の追跡調査データを GIS 上で比較し言語の経年変化を分析する研究を行っている[7].

GIS を言語研究で活用するためには技術的な課題も存在する．特に，LAJ や GAJ で用いられている「調査地点番号」という独自の位置情報システムは，そのままでは GIS で利用できない．これは，日本の地図をメッシュ状に分割し，6 桁または 8 桁の数字で地点をコード化するものである．この地点番号から，基準点（北緯29度，東経122度30分）を基に緯度・経度を算出する計算式が考案・公開されており，これにより過去の膨大な方言データを GIS 上で活用する道が開かれた[4][8]．現在では，この計算に基づいて緯度・経度情報が付与されたデータセットが提供されており研究者の利便性は向上している．

このように GIS は言語地図を単なる分布図から，多様な情報と結びついた分析ツールへと進化させ，言語地理学を言語史の解明という伝統的な目的に留まらせず，より広く学際的な研究分野へと解放する大きな力となっている．

2.4 機械学習の応用と新たな展開

近年，GIS による分析からさらに一歩進んで，機械学習技術を言語地図作成に応用する研究が登場している．これは従来の専門家による主観的な分類や解釈を補完しデータ駆動型の客観的な分析を目指す試みである．

近藤（2023）は方言形の分類を自動化する手法を提案した．この研究では，まず方言語形をローマ字表記し，その語形を構成する音素（アルファベット）の出現頻度をベクトル化する「Bag of Characters (BoC)」という手法を用いた．例えば「DENDENMUSI」という語形は各アルファベットの出現数を数え，26次元の頻度ベクトルとして表現される．この手法で得られた語形ごとのベクトルを，主成分分析（PCA）で次元圧縮し，K-Means 法でクラスタリングすることで，音韻構成の類似性に基づいて語形を自動的にグループ分けする．この結果を用いて地図上のアイコンの色を決定することで従来は研究者の判断に委ねられていた煩雑な分類作業の自動化を実現した[2].

さらに近藤と持橋（2024）はこのアプローチを発展させ語形そのものの情報に加え地理的な分布状況をベクトル化する手法を提案した．この研究では，各語形の採取地点の緯度・経度を 1 次元に圧縮し，その線上に語形を並べたものを一種のテキストと見なした．そして単語埋め込みモデルである FastText を用いて，この「テキスト」から語形のベクトル表現を学習させる．この方法により地域的に隣接して現れる語形は，ベクトル空間上でも近い位置に配置されることになる．得られた語形ベクトルを地図ごとに平均化することで，その言語地図全体が持つ分布の「模様」を一つの特徴ベクトルとして表現し，それを階層的クラスタリングにかけることで言語地図そのものを分類することを試みた．

この分析の結果，語の性質によって分布パターンが異なることが明らかになった．例えば，「本」や「甘い」といった漢語や基本語は，西日本や九州などにまとまって分布する「ブロック状」の地図を形成する傾向があった．一方で「行こうよ」や「先生だろう」といった文法要素を含む表現は特定の地域にまとまらず，まだらに分布する傾向が見られた[3]．この研究は個々の語形を分類するだけでなく言語地図が示す分布パターン自体を類型化しそのパターンと語の素性との関係を探るという全く新しい研究の可能性を示した点で画期的である．

2.5 課題と展望

コンピュータ技術の導入は言語地理学に大きな革新をもたらしたが，同時に新たな課題も浮き彫りになっている．

第一に，データの共有と統合の問題がある．これまで，多くの言語地図やデータベースが個々の研究者や機関によって作成されてきたが，そのフォーマットは統一されておらず，異なる調査データを結合したり比較したりすることは容易ではない．調査票や文字コード，データ形式などの標準化を進め，国境を越えた広域言語地図の作成も視野に入れたデータの共有・統合が今後の課題となる．

第二に，分析に必要な基盤データの整備である．言語分布を歴史的に，あるいは社会的に考察するためには，調査当時の行政境界，古道や旧街道，過去の人口統計といった歴史的な地理空間データが必要不可欠である．しかし，こうしたデータは GIS で容易に利用できる形式で整備されているとは限らず，研究者自身が作成する必要があるのが現状である．

第三にツールの普及と操作性の課題が挙げられる．GIS ソフトは高機能である反面，操作が複雑であり，言語学の研究者が使いこなすには習熟を要する．また論文に掲載するための描画品質の点では，専用の作図ソフトに利便性を感じる研究者も少なくない．研究者が手軽に高度な分析を行えるような環境づくりが求められる．

第四に研究成果の公開のあり方である．データベースや言語地図を専門家だけでなく教育現場や一般市民にも役立つ形でいかに提供していくかという視点が重要である[9]．Web 上でインタラクティブに地図を操作できるシステムや音声・画像といったマルチメディア情報とリンクしたデータベースの公開などが今後の方向性として考えられる．

これらの課題を克服し，GIS や機械学習といった新たな手法を駆使することで，言語地理学は大きな可能性を秘めている．大西が繰り返し指摘するようにその目的を伝統的な「言語史の解明」という枠に限定する必要はない．言語の地理的・空間的あり方を基盤とし，言語とそれを使用す

る人間の社会，文化，歴史，地理との関わりを多角的に解明する，ダイナミックで学際的な学問分野として発展していくことが期待される．コンピュータ技術は，そのための極めて強力なツールとなるだろう．

3. 提案手法

3.1 Web ツールの開発過程

アイヌ語の方言分布を視覚化する Web ツール「アイヌ語方言地図 SVG ジェネレーター」の開発過程を，その技術的課題と解決策の観点から時系列に沿って記述する．このツールは初期の手動操作を基本とする仕様から段階的な機能改善を経て，データ処理の自動化，外部システム連携，バッチ処理といった機能を実装するに至った．

本開発プロジェクトの特筆すべき点は，その全工程が Google によって開発された大規模言語モデル（LLM）である Gemini 2.5 Pro との対話を通じて行われたことである．

この開発手法を背景として，利用者のニーズに応じて機能が実装され，システムの利便性と安定性が向上していった過程を解説する．

3.2 開発環境と手法：LLM を活用した対話型アプリ開発

本ツールの開発は従来の人間がコードを直接記述する開発手法とは異なり，自然言語による指示（プロンプト）とそれに応答して生成されるコードを基盤とする，対話型の開発プロセスを採用している．以下に，その開発環境と手法について説明する．

3.3 コード生成の主体：大規模言語モデル「Gemini」

開発における実際のコード生成は，Google が開発したマルチモーダル大規模言語モデル「Gemini 2.5 Pro」を用いた．

このモデルは自然言語（本件では日本語）で記述された曖昧さを含む要求を解釈し，それを具体的なプログラムコードに変換する能力を持つ．本プロジェクトでは，HTML，CSS，JavaScript という Web アプリケーションを構成する複数の言語を横断して，単一のファイルとして機能するコードの生成，修正，リファクタリング（構造改善）といったソフトウェア開発における一連の知的作業を Gemini 2.5 Pro が実行した．

3.4 対話型開発環境としての「Canvas」

開発プロセスは「Canvas」と呼ばれる対話型のインターフェース上で行われた．これは利用者（人間）と Gemini モデルとの間で継続的な文脈を保持しながら共同作業を行うための環境として機能する．開発サイクルは以下に示す反復的なプロセスで進行した．

1. 要求提示：利用者が「シンボルの描画位置がずれている」や「Lua コードを生成する機能を追加したい」といった要求を自然言語で提示する．
2. コード生成：Gemini は要求を解釈し，アプリケーション全体のソースコードを生成・修正して提示する．
3. 実装と検証：利用者は生成されたコードを実際にブラウザで動作させて機能や修正内容を検証する．
4. フィードバック：検証結果に基づき，新たな課題（例：「バッチ処理時にファイルが上書きされる」）や追加の要望を，多くの場合，動作中のコード全体を再度提示しながらフィードバックとして入力する．

この「要求→生成→検証→フィードバック」という密なループを繰り返すことで，アプリケーションは段階的に洗練されていった．Canvas はアプリケーションの全状態（コード）を文脈として保持する共有空間として機能し，Gemini が複雑な修正や機能追加を全体の整合性を保ちながら行うことを可能にした．次章で分析する開発履歴は，この対話型開発環境におけるやり取りの記録そのものである．

3.5 先行研究との比較：LLM 対話開発とワークフロー統合

本研究が提示する「アイヌ語方言地図 SVG ジェネレーター」の最大の新規性は開発プロセスそのものの革新性とそれによって実現された研究ワークフロー全体の統合的支援という二点に集約される．

第一に開発手法において本研究は Google の大規模言語モデル「Gemini 2.5 Pro」との自然言語による対話を通じて，アプリケーションの設計，コーディング，修正，リファクタリングといった全工程を実行した．これは研究者自身がプログラミングを行う（例：福嶋の SEAL 開発），あるいは既存の汎用ソフトウェア（例：大西の Illustrator プラグイン，林の GIS 活用）を利用するといった従来の開発パラダイムとは根本的に異なる．この対話型開発はプログラミングの専門家でない言語研究者であっても自身のニーズに基づいた高度に専門的なツールを迅速に開発・改良することを可能にする新しい開発モデルを提示するものである．

第二に機能面において，本ツールは従来であれば複数のソフトウェアや手作業を介して行われていた研究プロセスを単一の Web アプリケーション上で全てを実行する点に独自性がある．先行研究では，データの電子化，語形の分類，地図の描画，そして成果の公開は，それぞれ独立したステップとして扱われることが多かった．しかし，本ツールは注釈などが混在する生の言語データ（TSV 形式）の入力から①データの自動整形，②ユニークな語形の自動分類と ID 割り当て，③SVG 形式での方言地図の生成，④外部プラットフォーム（Wikipedia, Wiktionary）で利用可能な形式でのデータ出力までを一連の自動化されたフローとして統合している．特に出力結果の公表を直接支援する具体的な外部連携コードの生成機能は，先行研究には見られない本研究ならではの実践的な機能である．

3.6 先行研究にはない本研究の取り組み

本研究で実装された機能の多くは先行研究では着手されていなかった研究者の実務的な負担を軽減し研究プロセスを加速させるためのものである．

1. 生データの自動処理と分類

先行研究では地図化の前段階として，研究者が手作業でデータをクレンジングし語形を分類・整理することが一般的であった．本ツールは括弧内の注釈や多様な区切り文字を含む生の言語データを正規表現によって自動的に解析・整形し，純粋な単語データのみを抽出する．さらに抽出された単語の出現頻度を計算しユニークな単語に一意の ID を自動で割り当てることで，従来は多大な時間と労力を要した分類作業を完全に自動化した．

2. 外部連携コードの自動生成と応用的実装

本ツールは，単に地図を画像として出力するだけでなく，その基となった分類データを再利用可能な形式で出力する機能を持つ．特に知識共有プラットフォームで広く利用される Wikipedia 形式の表や，Wiktionary で方言情報を記述するための Lua モジュールをワンクリックで自動生成できる点は本研究の大きな特徴である．さらに，Lua モジュール生成時にはアイヌ語の音韻規則に基づき，長音や末尾子音

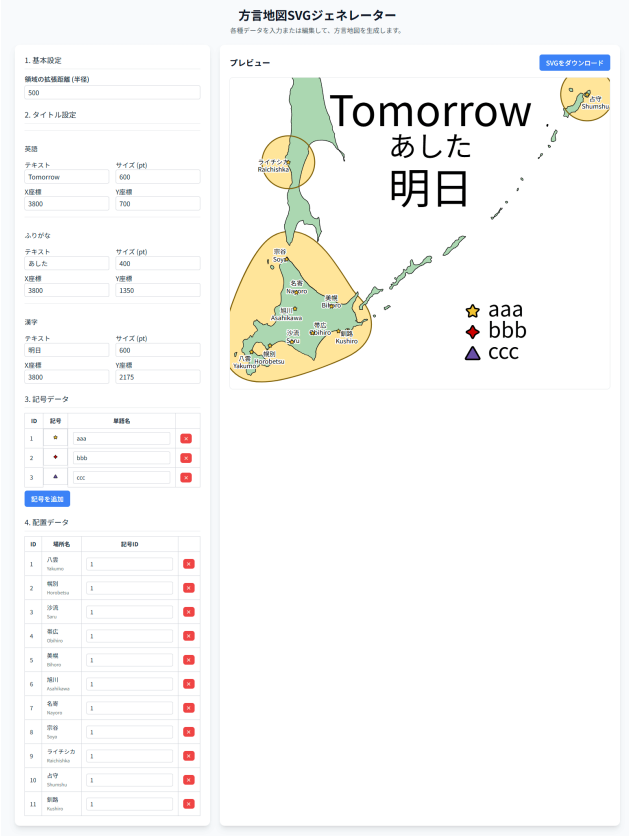

図 1　開発初期の Web アプリケーション

を適切に処理しながらローマ字表記をカタカナに変換するロジックが組み込まれており，単なるデータ変換に留まらない言語学的知見に基づいた応用的な実装がなされている．

3.　堅牢なバッチ処理機能

研究では多数のデータセットを同時に扱う場面が頻繁に生じる．本ツールは複数の TSV ファイルをドラッグ＆ドロップで一括処理し，生成された多数の SVG 地図と Lua モジュールをファイル名の衝突を自動で回避しながらサブフォルダに整理し，単一の ZIP ファイルとしてダウンロードできるバッチ処理機能を実装している．このような大規模データ処理に対応した堅牢なシステムは，個別の地図作成を主眼としていた先行研究のツールには見られない．

4.　汎用フレームワークとしての設計と実証

本ツールは，特定の言語や地域情報（ベースマップ SVG，地点座標データ）を「データ層」としてデータ処理や描画を行う中核的な機能を「ロジック層」として明確に分離する設計を採用している．このアーキテクチャによりデータ層を差し替えるだけで，最小限の労力で他の言語（本研究ではニヴフ語への応用を実証）の方言地図ジェネレーターを迅速に展開できる．これは本ツールが個別具体的なアプリケーションであると同時に多様な言語研究に応用可能な「フレームワーク」として機能することを示しており，ソフトウェアの再利用性と拡張性という点で先行研究にはない明確な設計思想が存在する．

3.7　先行研究で扱われているが本研究では未対応の領域

一方で，本研究のツールは研究ワークフローの自動化と効率化に特化しているため，先行研究で重点的に扱われてきた高度な分析機能の一部は実装していない．

1.　地理空間情報との高度なオーバーレイ分析

林や大西らによる研究では，GIS の強力な機能を活用し方言の分布と，標高，河川，人口密度，世帯構成，生産年齢人口比率といった多様な非言語的地理情報とを重ね合わせ，その相関関係を統計的・視覚的に分析している．これにより，言語変化の背景にある地理的・社会的要因を深く考察することが可能となっている．本研究のツールはあくまで背景地図上にシンボルを正確に配置する「視覚化」に主眼を置いており，GIS ソフトウェアが持つような高度な空間解析機能や，外部の地理空間データと動的に連携する機能は有していない．

2.　機械学習による客観的な語形分類

近藤らの研究は語形そのものが持つ音韻的・形態的特徴や，地理的な共起情報に着目し，BoW（Bag of Words）や FastText といった手法で語形をベクトル化し，クラスタリングによって客観的に分類・分析することを目的としている．このアプローチは，語形間の「類似度」や「距離」をデータから自動的に算出し，新たな方言区画の可能性を探るなど，分析そのものに新規性がある．対照的に，本研究の「自動分類」機能は，ユニークな単語をリストアップし ID を振るという処理であり，語形の類似性に基づく分類ではない．これは分析の前処理を効率化するための「ユーティリティ機能」であり，近藤らの研究が目指す「データ駆動型の分析」とは目的と手法が異なる．

4.　アイヌ語方言地図 SVG ジェネレーターの開発

4.1　開発史と技術的進化に関する詳細考察

本章は言語データを視覚的な地図情報へと変換する Web アプリケーション「アイヌ語方言地図 SVG ジェネレーター」の開発過程を，その技術的課題の特定と解決策の実装という観点から時系列に沿って詳細に分析・考察するものである．このツールの開発履歴は単一機能の単純なスクリプトから，データ処理の自動化や堅牢なバッチ処理能力を備えた多機能な研究支援ツールへと進化する，典型的なソフトウェア開発の成熟過程を示している．

初期段階における手作業中心の煩雑なワークフローは利用者の負担を増大させ実用性を著しく損なうものであった．

この初期の課題を起点とし，描画精度の向上，ユーザーインターフェースの洗練，データ処理の自動化，そしてシステムの堅牢性確保という各段階を経て，どのように高度な機能として昇華させていったかを，開発履歴から説明する．専門分野に特化したツールの開発における課題発見から実装，そして改善に至る一連のプロセスモデルを提示することを目的とする．

4.2　初期段階における深刻な課題とその分析

開発の黎明期（2025 年 6 月 11 日時点の記録）において本ツールは単一パネルで構成された，ごく基本的な機能のみを持つプロトタイプであった．開発初期の Web アプリケーションのスクリーンショットを図 1 に示す．その機能は利用者が提供した情報を基に地図上にシンボルを描画するという一点に集約されていたが，実用的なアプリケーションとして運用するには，複数の深刻な課題を内包していた．

第一にデータ入力プロセスにおける極端な非効率性が挙げられる．地図のタイトル，凡例に表示する各シンボルの定義，そして各地名に対応する単語データの配置といった描画に必要なすべての情報が，それぞれ独立したテーブル形式のフォームへの手動入力を前提としていた．これは少

数のデータを扱う試用段階では機能するものの，実際の研究で想定される数十，数百のデータポイントを処理する際には膨大な時間と労力を要求し，入力ミスを誘発する温床ともなり得る極めて脆弱なワークフローであった．

第二にシステムの安定性を下げる構造的欠陥が存在した．地図の基盤となる SVG 画像を外部サーバーである Wikimedia から HTTP 経由で動的に読み込む設計はネットワークの接続状況や相手方サーバーの応答性に完全に依存するものであった．このため読み込みのタイムアウトや失敗が頻発し，ツールの基本的な動作すら保証されないという信頼性に著しい問題を抱えていた．

第三にそして最も致命的であったのがシンボル描画ロジックの不具合である．SVG の座標系や viewBox の解釈，スケーリング処理に内在するバグにより，配置されたシンボルが意図した座標から大きく逸脱したり不適切なサイズで描画されたりする事象が確認された．これは地図情報の正確性を下げるような欠陥であった．

これらに加え，凡例の位置やサイズが固定されており利用者が調整できないといった柔軟性の欠如や地図上のテキストに縁取り処理が施されておらず背景の複雑な地形と重なると著しく視認性が低下するなど，ユーザー体験における配慮の不足も散見された．これらの課題は本ツールが単なる技術的試作から実用的な研究ツールへと変わるためにまず乗り越えなければならない最初の障壁であった．

4.3 UI/UX の改善と描画精度の抜本的向上

初期段階で露呈した複数の課題を克服すべく，直後（同 2025 年 6 月 11 日中の更新）に行われた改良ではツールの根幹である描画機能の信頼性確保と利用者にとっての使いやすさ，すなわち UI/UX の向上に主眼が置かれた．

最大の技術的進展は描画エンジンの心臓部である JavaScript 関数 createSymbol の全面的再設計である．従来は不完全であったシンボルの座標およびサイズ計算ロジックを抜本的に見直し，個々のシンボルが持つ固有の基準サイズ（baseArtworkSize）を基に地図上（80px 四方）と凡例（240px 四方）でそれぞれ要求される表示サイズへと正確にスケーリングし，かつ指定された座標の中心に寸分の狂いなく配置するための正規化と変換のプロセスが導入された．これにより多種多様なデザインのシンボルを一貫したルールで，かつ高い信頼性をもって描画することが可能となりツールの描画精度は向上した．

システムの安定性に関する問題に対しては外部サーバーへの依存という根本原因を削除するアプローチが採られた．具体的には地図のベースとなる SVG コードを HTML ファイル内に直接埋め込むことで外部ネットワークへの接続を不要とし読み込み失敗のリスクを完全に払拭した．これはオフライン環境での利用可能性をも担保する堅牢な設計への転換を意味する．

UI/UX の観点からは煩雑であった入力フォームの簡素化が図られた．特にシンボル定義テーブルからは色や補足情報といった冗長な項目が削除され，利用者は「単語名」の入力に集中できるようになった．加えて入力されたシンボルのプレビューが表示されるようになり，直感的な操作性が向上した．また視認性の低かった地図上のテキストには 25pt の白い縁取り（stroke）を付与するスタイルが適用された．これにより文字が背景の複雑な地図デザインから視覚的に分離され，あらゆる状況下での可読性が向上した．

これらの改良によりツールは「正しく，安定して動作する」というアプリケーションとしての最低限の要件を満たすに至った．しかしながら大量の言語データを手作業で整形しツールが要求する形式へと変換しなければならないという研究ワークフローにおける最大のボトルネックは依然として解決されておらず，次なる改良の段階へと課題は持ち越されることとなった．

4.4 データ処理の自動化と機能拡張

2025 年 6 月 12 日から 25 日にかけて，本ツールの機能は大きく改善した．単なる手動の描画ツールから言語データの整形・分類・変換という一連の研究プロセスを自動化する高度なデータ処理支援プラットフォームへと変貌したのである．この期間における技術的飛躍の中核をなすのが「データ自動入力」機能の実装である．

この新機能は利用者のワークフローを大きく変えるものである．タブ区切り形式（TSV）で記述された注釈や異表記が混在する生の言語データをテキストエリアに貼り付けるとツール内部で以下の多段階処理が自動的に実行される．

1. データ解析と整形: extractAndCleanWords 関数が括弧内の注釈や「~」「〜」といった様々な区切り文字を識別し正規表現を用いて的確に処理し分析対象となる純粋な単語データのみを抽出する．これは現実の言語資料が持つノイズへの実践的な対応能力を示す．

2. 自動分類と ID 割り当て: generateClassificationTable 関数が抽出された全単語の出現頻度を計算し，ユニークな単語ごとに一意の ID を割り当てる．これにより，これまで研究者が手作業で行っていた地道で時間のかかる分類作業が完全に自動化された．

3. 配置用データの生成（Placement Data Generation）: 元データ内の各地名に紐づいていた生の単語を上記で割り当てられた ID に置換しツール内部の描画エンジンが直接解釈できる形式の配置データを自動生成する．

この一連の自動化プロセスは，研究者がデータの前処理に費やしていた膨大な時間を解放し，本来注力すべき分析や考察に集中することを可能にした．

さらにツールの応用範囲を拡大するため外部プラットフォームとの連携機能が追加された．生成された分類データを基に知識的な情報共有の場で広く利用される Wikipedia 形式の表をワンクリックで生成する機能はデータの公表を支援する．また，言語記述の専門的プラットフォームである Wiktionary のための Lua モジュールを自動生成する機能が追加された．これにはアイヌ語のローマ字表記をカタカナへと変換するロジック（getKatakanaConversion 関数）が内包されており，単なるデータ変換に留まらない言語学的知見に基づいた処理を実現している．

これらの機能に加え，複数のデータセット（TSV ファイル）をドラッグ＆ドロップで一括処理し，生成された多数の SVG 地図と Lua モジュールを単一の ZIP ファイルとしてダウンロードできるバッチ処理機能，そして UI の多言語対応（日本語，英語，中国語，韓国語，ロシア語）が実装されたことで，本ツールは多様な利用者と大規模なデータに対応可能な汎用性と拡張性を兼ね備えた本格的な研究ツールとなった．

4.5 最終調整とシステムの堅牢性確保に向けた改良

2025 年 7 月上旬，一連の機能拡張を終えたツールは品質を最高水準にまで高めるための最終調整期間に入った．こ

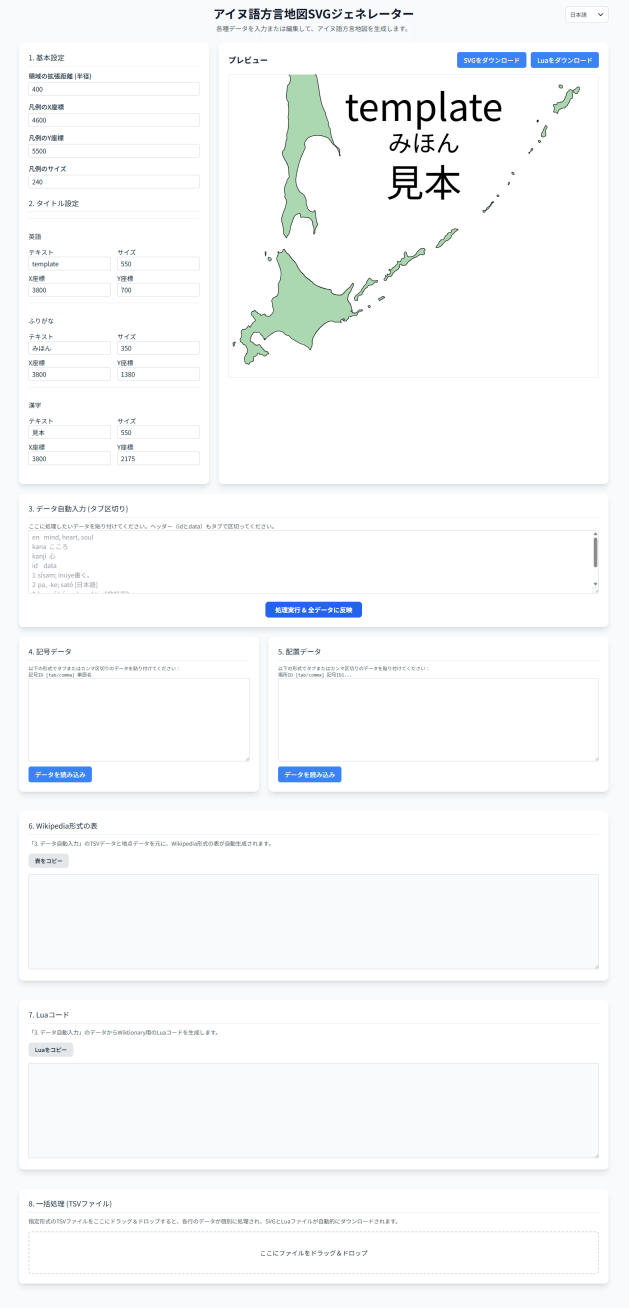

図２　完成した Web アプリケーション

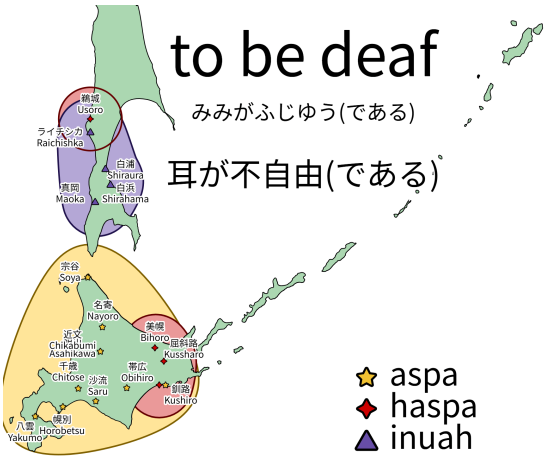

図３　アイヌ語の「耳が不自由」を表す語の地理的分布

の段階では既存機能の精密なバグ修正と多様な利用方法を想定したシステムの堅牢性の確保に焦点が当てられた．

第一に高度な機能である Lua コード生成機能に残存していた言語学的な正確性に関わるバグの修正が行われた．具体的にはカタカナ変換ロジックにおいて，kaa のような連続する母音と ka'a のような分節された母音を区別した長音処理や，n, y, w といった末尾子音の扱いに見られた誤りが是正された．完成した Web アプリケーションのスクリーンショットを図２に示す．

図３は「耳が不自由(である)」を意味する語彙の地理的分布を示した言語地図である．図３のデータは知里・和田（1943）[10]，知里（1954）[11]，服部（1964）[12]，中川裕（1995）[13]，釧路アイヌ語の会（2021）[14]に基づく．本図から，北海道東部およびサハリン中部には「haspa」が，北海道には「aspa」がそれぞれ分布していることが確認できる．このような言語地理学的な分布の可視化は，日本書紀にみられる「粛慎（あしはせ）」をアイヌ語で「言葉の通じない集団」と解し語源を「aspa say」（耳が不自由+群）に求める説について，その妥当性を検討する上で一つの視点を提供する．これは説の当否を直接証明するものではないが，議論における実証的な参考資料となり得るだろう．

第二にバッチ処理機能の実用性を飛躍的に向上させるための堅牢性の強化が図られた．従来の仕様では複数のデータセットから同名のファイル（例えば異なる地図に対してそれぞれ module.lua）が生成された場合，後から生成されたファイルが先に生成されたものを上書きしてしまうという致命的な欠陥があった．この問題に対し，生成されるファイルを ZIP アーカイブ内で SVG と Lua というサブフォルダに自動的に振り分ける構造を導入．さらに同名のファイルが生成される際には Map.svg，Map2.svg のようにファイル名の末尾に自動で連番を付与する衝突回避方法を実装した．これにより，いかなる規模のバッチ処理においてもデータの損失が発生しない信頼性の高いシステムが完成した．

最後にコード全体の品質向上のための変更が実施された．冗長なコードを置き換えるなど，コードの可読性，保守性，そして実行パフォーマンスの向上が図られた．これは将来的な機能追加や仕様変更にも迅速かつ柔軟に対応できる持続可能な開発基盤を確立するための重要な改良である．

4.6　おわりに

本稿で詳述した「アイヌ語方言地図 SVG ジェネレーター」の開発は，利用者の抱える課題を基点として，いかにして洗練され進化を遂げていくかを示す事例となった．

その試みはまず「正確に描画する」というツールの根源的価値を技術的に確立することから始まった．次に手動入力という障壁を「データ処理の自動化」という解決策によって克服し利用者を煩雑な作業から解放した．そして外部連携やバッチ処理といった機能拡張により，その応用範囲を専門的な研究領域へと広げた．最終的には細やかなバグ修正とシステムの堅牢化を通じて，誰がどのような状況で使っても安定して動作する信頼性のあるツールとなった．

この開発プロセスは単なる機能の逐次的な追加ではない．それは利用者のワークフローを深く洞察し，技術的な課題を特定し，そして最適な解決策を実装するという体系的かつ反復的な問題解決の連続であった．かくして，このジェネレーターは単なる地図作成ツールという初期の実装を変えて，アイヌ語方言研究のプロセスそのものを支援し加速させるプラットフォームとして完成したのである．

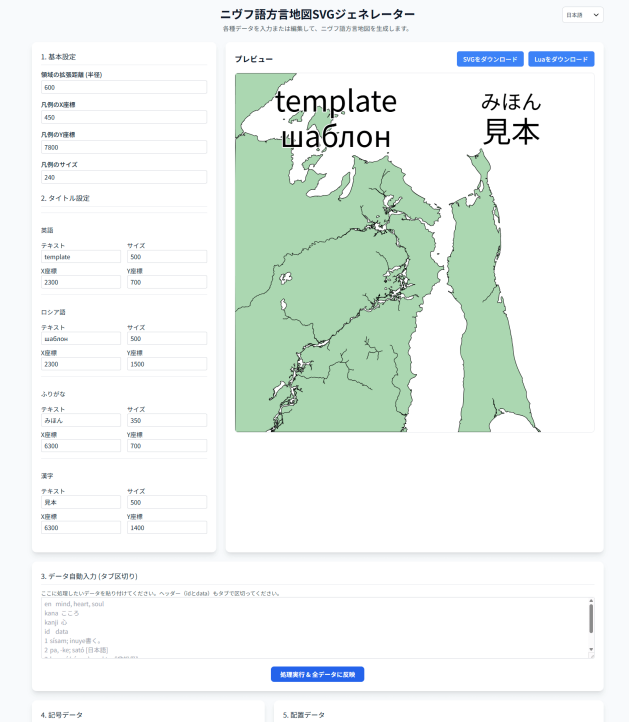

図４　ニヴフ語版の Web アプリケーション

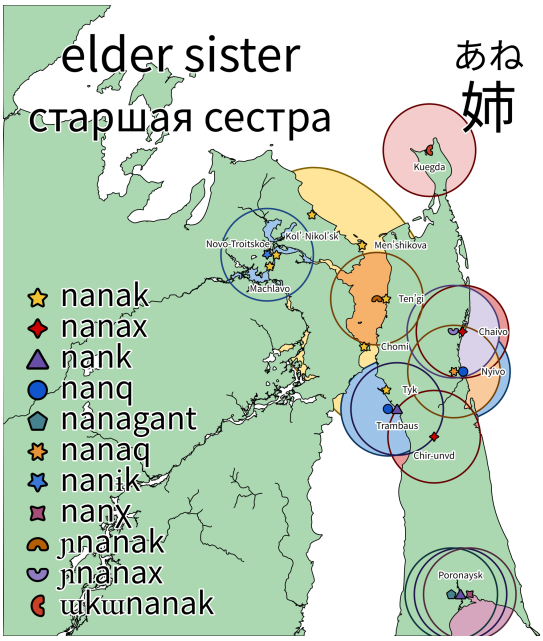

図５　ニヴフ語の「姉」を表す語の地理的分布

5. システムの汎用性と他言語への展開

5.1　ニヴフ語への応用事例

これまで詳述してきた「アイヌ語方言地図 SVG ジェネレーター」は，その開発過程において特定のデータセットから独立した再利用可能なロジックを持つモジュール性の高いシステムとして設計された．本章では，このシステムの汎用性と拡張性を実証する事例として，新たに作成された「ニヴフ語方言地図 SVG ジェネレーター」を取り上げる．

このニヴフ語版アプリケーションはアイヌ語版の基本構造を流用し，ごくわずかな変更を加えるだけで実現された．この事例を説明することにより本ツールのアーキテクチャが単一言語のための専用ツールではなく，多様な言語の方言地図作成に応用可能なフレームワークとして機能し得ることを明らかにする．

5.2　アプリケーションの基本構造の分離

本アプリケーションの構造は，大きく分けて「データ層」と「ロジック層」の二つに分離することが可能である．この設計が他言語への展開を容易にする上での鍵となった．

- データ層（Data Layer）

この層は特定の言語や地域に固有の情報を格納する部分である．具体的には以下の要素が含まれる．

1. ベースマップ SVG：地図の背景となる SVG 画像データ（アイヌ語版では北海道・サハリン南部・千島列島，ニヴフ語版ではサハリン北部・アムール川下流域）．
2. 地点情報データ：地図上にプロットされる地点の定義（ID，地名，SVG 上の XY 座標）を格納する JavaScript オブジェクト．
3. UI 上のテキスト：アプリケーションのタイトルや説明文など，利用者に表示される言語固有のテキスト．
4. 初期サンプルデータ：データ自動入力機能で使われる初期状態で表示されている言語データ．

- ロジック層（Logic Layer）

この層はデータ層から提供された情報を受け取り，処理を実行するエンジン部分である．言語や地域に依存せず汎用的に機能する．

1. データ処理エンジン：入力された生データを分析し分類・ID 割り当てを行う一連の機能（handleProcessData 関数など）．
2. SVG 描画エンジン：地点情報とシンボル定義に基づき地図，エリア，凡例を含む最終的な SVG を動的に生成する機能（generateSvg 関数など）．
3. 外部連携コード生成部：Wikipedia 形式の表や Wiktionary 用の Lua モジュールを生成する機能．

5.3　ニヴフ語版とアイヌ語版ジェネレーターの相違点

ニヴフ語版の作成は前述のロジック層にはほとんど変更を加えず，データ層の情報を差し替えるだけでほとんど完了した．具体的な変更点は以下の通りである．

1. ベースマップの変更

- アイヌ語版：北海道・サハリン南部・千島列島を中心とした SVG 地図．
- ニヴフ語版：ニヴフ語の話域であるサハリン北部・アムール川下流域を示す SVG 地図に差し替えられた．

2. 地点情報の変更

- アイヌ語版：沙流や美幌など，アイヌ語が記録された地点の座標データが定義されていた．
- ニヴフ語版：Chir-unvd や Nyivo など，ニヴフ語の諸方言が話されるサハリン・大陸の地点に対応する新たな地名と座標データに完全に置き換えられた．

3. UI テキストとサンプルデータの変更

アプリケーションのタイトルが「アイヌ語〜」から「ニヴフ語〜」に変更されたほか，説明文やサンプルデータもニヴフ語研究に即したものに更新された．

これらの相違点はほぼすべて「データ層」に属するものであり，データ処理，描画，外部連携といった「ロジック層」のプログラムコードはアイヌ語版とニヴフ語版でほとんど同一である．ニヴフ語版の Web アプリケーションのスクリーンショットを図４に示す．図５はニヴフ語で「姉」を意味する語彙の地理的分布を示した言語地図である．図

５のデータは高橋盛孝（1942）[15]，中川裕・佐藤知己・斉藤君子（1993）[16]，白石英才・丹菊逸治（2013）[17]，白石英才・丹菊逸治（2014）[18]，白石英才・丹菊逸治（2015）[19]に基づく．

5.4 考察：フレームワークとしての汎用性

ニヴフ語版アプリケーションをほぼデータ層の差し替えのみという最小限の作業で作成できた事実は本システムの設計が持つ高い汎用性と拡張性を明確に示している．

このことから本ツールは「アイヌ語方言地図ジェネレーター」という個別具体的なアプリケーションであると同時に，「特定地点に紐づく言語データを分類し，地図上に可視化するためのフレームワーク」であると結論付けられる．中核となるロジック層は，入力される地名や単語がどの言語に属するかを問わない．ただ構造化された「地点データ」と「言語データ」を受け取り，定められたアルゴリズムに従って処理を実行するだけである．

したがって，ベースとなる地図 SVG とそれに対応する地点座標データを用意しさえすれば，理論上は世界中のあらゆる言語の方言・バリエーションの分布図作成に応用することが可能である．この事例は一度確立したロジックを再利用し異なるデータセットに適用することで，開発コストを大幅に抑制しながら類似のアプリケーションを迅速に展開できることを実証したと言える．

6. おわりに

本稿では，生成 AI との対話型開発を通じて構築した「アイヌ語方言地図 SVG ジェネレーター」を事例に，その手法と機能が先行研究といかに異なるかを比較分析した．本研究の最大の貢献は，LLM を開発の協働パートナーと位置づけることで，従来の開発パラダイムを刷新し，言語研究者が自身のニーズに即した高機能なツールを迅速に構築できる可能性を示した点にある．

機能面では，データの投入から整形，自動分類，地図描画，多様な形式での成果物出力まで，研究のワークフローを一気通貫で支援する統合的なプラットフォームを実現した．これは，個別の分析手法の高度化を目指した先行研究とは一線を画し，「研究プロセス全体の効率化と加速」という実務的な課題に正面から応えるものである．また，データ層とロジック層を分離した汎用的なフレームワーク設計は，一度確立した技術を他言語へ迅速に展開できることを実証し，今後の研究支援ツール開発における持続可能なモデルを提示した．

一方で，GIS を用いた高度な空間分析や，機械学習による客観的なデータ分類といった領域は，本研究では未対応であり，今後の課題として残されている．将来的には，本ツールが持つワークフロー自動化機能と，先行研究で培われた高度な分析手法とを連携させることで，より強力で包括的な研究支援環境を構築することが期待される．本研究は，方言地図作成という具体的な応用を通じて，生成 AI がデジタル・ヒューマニティーズ分野における研究のあり方そのものを変革しうる，新しいパラダイムの到来を告げる一例となったと言えよう．

生成したコードは GitHub にて公開している[20].

謝辞

英語での表記や生成したコードの公開方法においては岡山大学の Stephen West 氏にご助言を頂いたことを深謝する．アイヌ語とニヴフ語の語彙資料の情報提供においては白鳥詩織氏にご教示を頂いたことを深謝する．

参考文献

[1] 大西拓一郎：方言分布の解明に向けて―原点に帰る言語地理学―，日本語科学，Vol.21，pp.125-142（2007）．

[2] 近藤泰弘：単語音素のベクトル化による言語地図作成，言語処理学会第 29 回年次大会 発表論文集，pp.2381-2385（2023）．

[3] 近藤泰弘, 持橋大地：語形の分布状況のベクトル化による言語地図の分類方法，言語処理学会 第 30 回年次大会 発表論文集，pp.1301-1305（2024）．

[4] 林良雄：方言研究における GIS の活用に関する一考察, 秋田大学教育文化学部紀要 人文科学・社会科学自然科学, No.71，pp.89-96（2016）．

[5] 池田潤：GIS と言語研究，一般言語学論叢，No.9，pp.1-10（2006）．

[6] 大西拓一郎：地図作りのデモンストレーション―コンピュータで方言地図を作る―，国立国語研究所 第 13 回ことばフォーラム，pp.1-8（2003）．

[7] 峪口有香子, 平井松午, 岸江信介：地域言語に関する GIS 分析 ―瀬戸内海地域をフィールドとして―，人文地理学会大会 研究発表要旨 2013 年度大会，pp.126-127（2013）．

[8] 大西拓一郎：方言分布・言語地図データベース―時空間情報を持つ言語データ―，第 23 回公開シンポジウム「人文科学とデータベース」発表論文集，pp.43-50（2017）．

[9] 福嶋秩子：世界の言語地図作成・活用状況に見る言語地理学の現状と課題，日本語科学，Vol.23，pp.5-15（2008）．

[10] 知里真志保・和田文治郎：樺太アイヌ語に於ける人体関係名彙，樺太庁博物館報告，Vol.5，No.1，pp.39-80（1943）．

[11] 知里真志保：分類アイヌ語辞典 第 3 巻 (人間篇)，日本常民文化研究所（1954）．

[12] 服部四郎編：アイヌ語方言辞典，岩波書店（1964）．

[13] 中川裕：アイヌ語千歳方言辞典，草風館（1995）．

[14] 釧路アイヌ語の会編：釧路地方のアイヌ語語彙集，藤田印刷エクセレントブックス（2021）．

[15] 高橋盛孝：樺太ギリヤク語，朝日新聞社（1942）．

[16] 中川裕・佐藤知己・斉藤君子：サハリンにおけるニヴフ語基礎語彙の地域差，サハリンの少数民族 文部省科学研究費補助金研究成果報告書，pp.209-254（1993）．

[17] 白石英才・丹菊逸治：ニヴフ語アムール方言の基礎語彙，北方言語研究，No.3，pp.201-212（2013）．

[18] 白石英才・丹菊逸治：ニヴフ語アムール方言の基礎語彙 2，北方言語研究，No.4，pp.173-184（2014）．

[19] 白石英才・丹菊逸治：ニヴフ語アムール方言の基礎語彙 3，北方言語研究，No.5，pp.215-226（2015）．

[20]“言語地図ジェネレーターLinguistic Atlas Generator”. https://github.com/AsPJT/LinguisticAtlasGenerator,（参照 2025-07-21）．