

人文学におけるAI研究基盤：

古代エジプト語・コプト語などの 研究の実践と支援政策について

AI Research Infrastructure in Humanities: Practices from Ancient
Egyptian & Coptic Studies and Policy Recommendations



宮川 創 Dr.phil.

筑波大学 人文社会系 准教授

📅 2025年12月26日

🏛️ 文部科学省 人文学・社会科学特別委員会

E-Mail runa.uei@gmail.com / HP <https://somiyaagawa.com/>

SPEAKER PROFILE

宮川 創 So Miyagawa

筑波大学 人文社会系 / 西アジア文明研究センター
エジプト学部門 准教授

専門分野・学位

エジプト学

コプト学

デジタル人文学

計算言語学

 Dr. phil. (ゲッティンゲン大学, 2022年)

主な研究テーマ

- ▶ 古代エジプト語・コプト語の言語学的分析
- ▶ テキスト再利用 (Text Reuse) 分析・引用検出
- ▶ 低資源言語のためのコーパス構築と自然言語処理

その他の活動

アイヌ語・沖縄語・宮古語・奄美語・ヌビア語・タロコ語などの危機言語のためのツール開発
デジタル技術を活用した言語復興と継承支援



宮川が東海大学湘南キャンパスにおいて「東海大学古代エジプト及び中近東コレクション (AENET)」 (鈴木コレクション) のパピルス文献を調査している場面 (2025年)



Humanities × Technology

デジタルヒューマニティーズ
文理融合のアプローチ

古代エジプト語・コプト語

古代エジプト語

BC 3250 -

世界最古の記述言語の一つ。3つの主要な書記体系を持つ。

ヒエログリフ (聖刻文字) : 記念碑・神殿・墓・葬送文書

ヒエラティック (神官文字) : 行政・文学・日用など

デモティック (民衆文字) : 日常・法文書・勅令など

最古の文献は、紀元前3,250年頃。古代エジプト文明の主要言語。



ヒエログリフ

コプト語 (Coptic)

AD 300 - 2025

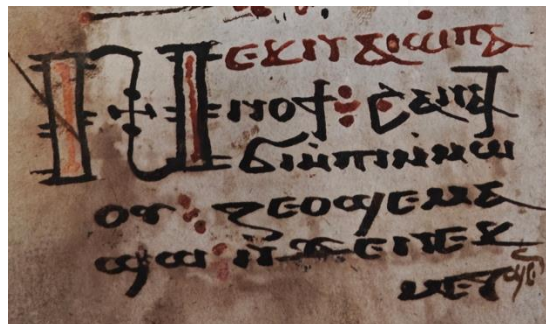
古代エジプト語の最終段階。ヘレニズム / キリスト教化の影響。

文字体系 : ギリシア文字 + デモティック由来の文字

主な資料 : 聖書翻訳、修道院文書、グノーシス文書

特徴 : 母音が表記される唯一のエジプト語段階

母語話者は17世紀に最後の記録。だが、現在でもコプト・キリスト教で用いられ、言語復興運動もある



コプト語写本

世界最長の書記期間を持つ言語

古代エジプト語の最終段階・コプト語は現在でもコプト・キリスト教で使用

▲ 研究の課題

資料の断片化 : 完全な形で残る文献は稀で、パピルスの断片をつなぎ合わせる作業が必要

方言の多様性 : サイド語、ボハイラ語など地域差が激しく、標準化が困難

研究者人口の少なさ : AI学習データ (コーパス) が不足する「低資源言語」

💡 なぜデジタル化・AIが必要か

📁 保存とアクセス

脆弱な原典をデジタル保全し、世界中で共有

🔍 解読の効率化

断片の自動接合や欠損文字の復元予測

🧠 大規模分析

手作業では不可能な全文献の横断的分析

研究活動におけるAIの位置づけ

従来の手法からデジタル・AI活用へのパラダイムシフト

DH・・・デジタルヒューマニティーズ
AI・・・人工知能
NLP・・・自然言語処理
LLM・・・大規模言語モデル
コーパス・・・分析ができるテキストデータの集合体



従来の研究方法

手作業による文献調査 紙の辞書やコンコーダンスを参照し、物理的な書籍から該当箇所を探し出す。

目視でのテキスト比較 異なる写本を並べて目視で照合。異同（異読）の記録に多大な時間を要する。

限定的なスケール 一人の研究者が扱える資料量に限りがあり、大規模な統計分析は困難。

Labor-intensive

Subjective

Small Data



DH + AIの導入

コーパス構築と検索 デジタル化されたテキストデータベースを瞬時に検索。関連文献を網羅的に収集。

NLPパイプライン処理 形態素解析、タグ付け、係うけ解析などを自動化。構造的な分析が可能に。

再現性とスケーラビリティ 分析手法（コード）を共有し、検証可能に。数千・数万の文書を一括処理。

Automated

Quantitative

Big Data

AI活用の具体例（1）：テキスト処理

📄 OCR / HTR

画像認識AI

手書き文字認識 (HTR) の進化

古代エジプト語の難解な手書き文字（ヒエラティック・デモティック）を高精度でデジタル化

TranskribusやeScriptorium等のツールを用いた低資源言語向けの学習モデル構築

🔗 形態素解析・タグ付け

自然言語処理

文法情報の自動付与

単語の区切りがない古代語テキストへの自動分かち書き
品詞タグ、見出し語（lemma）の自動付与による検索性の向上

📖 テキストリユース（再利用）検出

比較分析

引用・類似箇所を発見

聖書引用や著者間の影響関係を数千の文書から自動検出
「TRACER」や「ICOMA」等のツールを用いたn-gram分析

🗄️ コーパス言語学への展開

✔️ 大規模定量的分析

従来の手作業では不可能だった語彙統計やコロケーション分析が可能に。
言語変化の客観的証拠を提示。

✔️ 「遠読 (Distant Reading)」

個別のテキスト精読だけでなく、数千の文書をマクロな視点で俯瞰し、ジャンルや時代の傾向を把握。

🔄 ワークフローの変化

デジタル・エコシステム：スキャン画像 → HTRによるテキスト化 → 自動タグ付け → 検索可能データベースへの登録という一連の流れが確立。

Human-in-the-loop：AIの出力を専門家が修正し、それを再学習させることで精度が向上する循環モデル。

AI活用の具体例（2）：翻訳と解釈

機械翻訳モデルの構築

例：古代エジプト語・コプト語 ↔ 英語・日本語・その他現代語

低資源言語への対応：少量の対訳データを用いた転移学習モデルの開発

解釈の補助：文法的に複雑な箇所「下訳」として利用し、研究者の読解をサポート

次世代の語彙・文法検索

ベクトル検索の導入：単語の「意味的類似性」に基づく検索（例：「王」で検索して「ファラオ」もヒット）

RAG（検索拡張生成）：辞書などのベクトル化されたデータベースを参照しながら、文脈に応じた語義をAIが提案

文法パターンの抽出

構文構造の自動検出：特定の文法パターン（例：関係節、命令形）の使用傾向を網羅的に抽出

統計的分析：時代やジャンルによる文体変化を定量的に可視化

類似テキストの発見（Intertextuality）

引用・参照の同定

聖書からの引用や、他の修道院文書との類似箇所を自動検出

断片接合の示唆

離れ離れになったパピルス断片の内容的つながりを発見

研究プロセスの効率化と高度化

AI活用の具体例（3） 研究支援

論文執筆・多言語対応

執筆支援と構成提案：アイデア出し、章立てのブレインストーミング、論理展開のチェック。

文章校正とスタイル調整：学術英語としての自然さ、文法チェック、投稿規定へのフォーマット調整。

マルチリンガル研究発信：日本語・英語・ドイツ語間の翻訳補助により、国際的な研究発信を加速。 JPN ENG DEU



時間的コストの削減

定型的な作業（フォーマット調整、単純なコーディング、一次翻訳）をAIに任せることで、研究の本質的な「思考・解釈」に割ける時間が大幅に増加。

プログラミング・調査支援

DHツール開発の民主化：Python等のコード生成支援により、非エンジニア研究者でも分析スクリプトを作成可能。

データ処理の自動化：テキスト整形、データクレンジング、可視化コードの作成を対話型で実現。

文献調査の効率化：大量の先行研究の要約、重要ポイントの抽出、関連文献のリストアップ。



言語バリアの低減

非ネイティブ話者にとっての英語論文執筆の負担を軽減。コプト語やドイツ語など、多言語資料を扱う際のリファレンスとしても機能。



技術的ハードルの低下

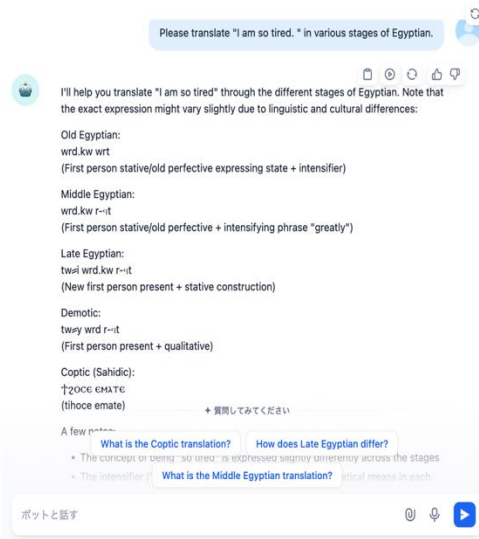
「やりたい分析はあるがコードが書けない」というデジタル人文学（DH）参入の壁を取り払い、より高度な定量分析を可能にする。

THOTH AI : 古代言語学習・研究支援

RAG-Enhanced Neural Machine Translation & Education System

📌 システム概要

Claude 4.5 Sonnetを基盤とし、Difyプラットフォーム上で構築。RAG（検索拡張生成）技術によりベクトル化された専門的な古代エジプト語・コプト語辞書データを参照し、幻覚を抑制した高精度な翻訳・解説を実現。



主な機能と成果



Text Analysis & OCR

古代エジプト語・コプト語のOCR（光学文字認識）と文法構造解析



SOTA Machine Translation

中エジプト語→英語翻訳において「State-of-the-Art（最高水準）」を達成



Educational Support

歴史的背景を含めた対話的な言語指導（Difyプラットフォーム活用）

So Miyagawa. 2025. [RAG-Enhanced Neural Machine Translation of Ancient Egyptian Text: A Case Study of THOTH AI](https://somiyagawa.com/thoth). In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 33–40, Albuquerque, USA. Association for Computational Linguistics.

🔧 技術スタック

Claude 4.5 Sonnet 基盤LLMとして採用。RAGによって古代語の複雑な文脈理解に優れた性能を発揮。

RAG x Dify 語源情報付きコプト語辞書(Burns et al.)等の専門DBをDify経由で参照し精度向上。

👥 6,500+

総ユーザー数

🏆 SOTA

翻訳精度(最高水準)

<https://somiyagawa.com/thoth>

“

「*pr-ḥb* (House of Life: 生命の家)のように、単語の意味だけでなく、エジプトの歴史的・文化的背景も同時に学べる点が画期的でした。」

— 大学4年生（ユーザーの声）



[Ancient Egyptian-Coptic AI](#)

[自作ツール/Tools](#)

[研究業績/CV](#)

[授業資料/CourseMaterials](#)

[リンク/Links](#)

[GitHub・SNS](#)

[古代オリエント博物館クローズアップ展](#)

[その他/Other](#)



Thoth: Ancient Egyptian AI Teacher



English

次をコプト語に翻訳して文法解釈をつけてください。「私は今、本を書いています」





Identify Language ▾



English



次をコプト語に翻訳して文法解釈をつけてください。「私は今、本を書いています。」



「私は今、本を書いています。」のコプト語翻訳

1. サイド方言 (Sahidic Coptic)

コプト文字：

ⲧⲉⲥⲁⲓ ⲙⲡⲭⲱⲙⲉ ⲧⲉⲛⲟⲩ

ローマ字転写：

ti-shai m-p-jōme tenou

逐語訳：

「私は書いている その本を 今」

詳細な文法解説 (サイド方言)

1. ⲧⲉⲥⲁⲓ (ti-shai)

形態素分析：

- ⲧ- = 現在進行形の主語接頭辞 1人称単数「私は～している」
- ⲉⲥⲁⲓ = 動詞不定形「書く」

Talk to Thoth: Ancient Egyptian AI Teacher





T H O T H . A I

Ancient Egyptian & Coptic Language Assistant

Unlock the secrets of Ancient Egyptian and Coptic languages through cutting-edge AI



Looking at this Sahidic Coptic text, I can identify a psalm quotation near the end:

Psalm Quotation Identified

NEKNAZITOOTE AN PE ENINOS N2WB MMNTAT2OTE

This is a reference to **Psalm 55:23 (54:23 LXX)**

Analysis:

Sahidic Coptic:

NEKNAZITOOTE AN PE ENINOS N2WB MMNTAT2OTE

Transliteration:

nek-na-hito-otk an pe e-ni-noč n-hōb m-mnt-at-hote

Translation:

"You will not give them over to the great matters of fearlessness"
or more idiomatically: "You will not deliver them to the pit of destruction"

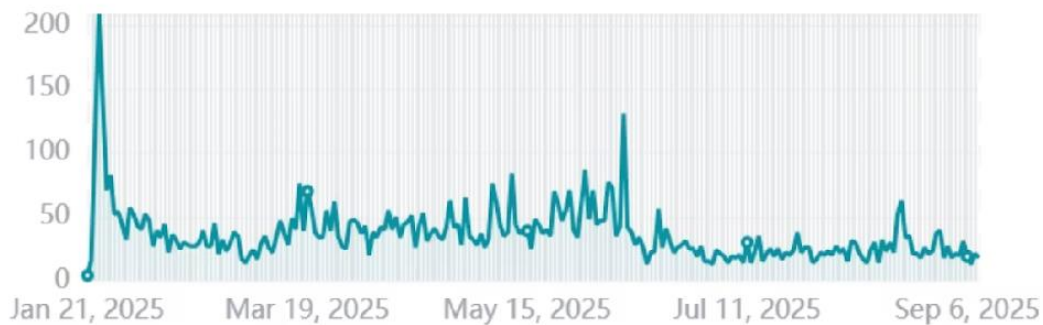
Modern Bohairic:

ΧΝΑΗΙΤΟΥ ΑΝ ΕΝΙΝΙΥ† Ν2ΩΒ ΝΙΕ ΤΜΕΤΑΤ2Ο†

Total Conversations ?

ALL TIME

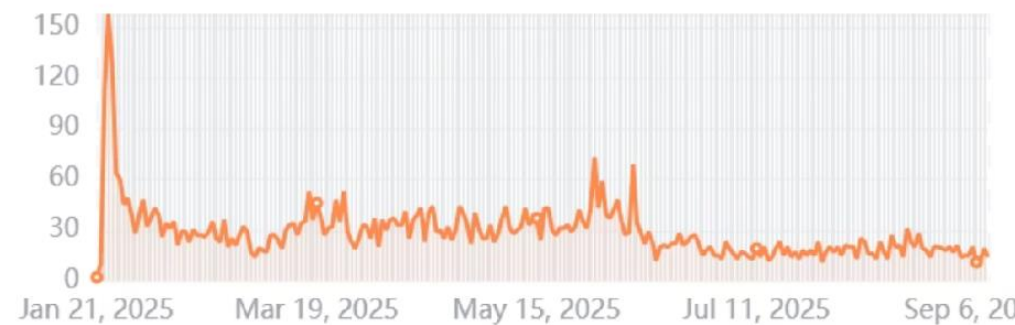
8,479



Active Users ?

ALL TIME

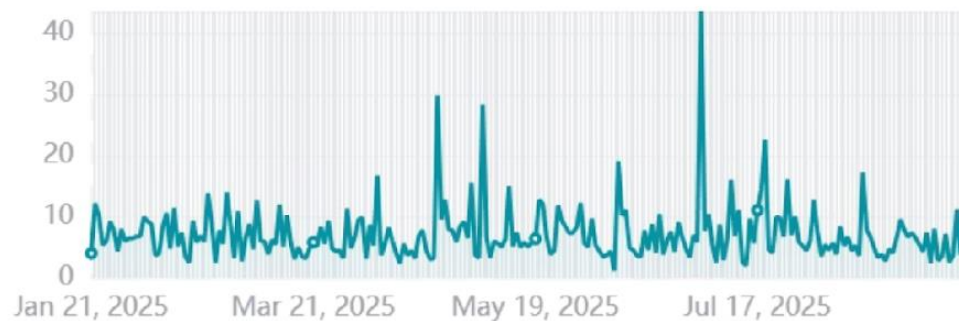
6,653



Avg. Session Interactions ?

ALL TIME

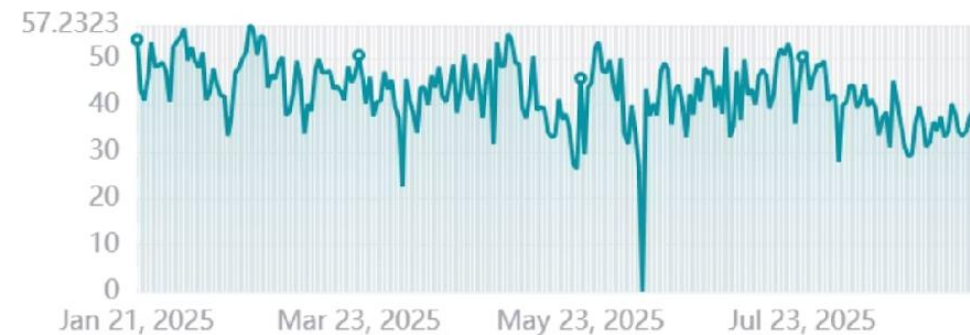
7.151



Token Output Speed ?

ALL TIME

42.745 Token/s



Mario Bishara氏によるTHOTH AIの支援によるグリム 童話のコプト語翻訳

†ⲱⲟⲙ† ⲛⲁⲥⲡⲓ

Οὐπαβολὴ ἡτε νισνηοῦ Γρίμ

Ἦεν †ῒελβετία νε οὔον οὔἔλλο ἡκοῖς ἐνεοὔονταϥ ἡναῦ ἡοὔωῃρι ἡναῦατϥ πε, ἀλλὰ νε οὔατῒητ πε οὔοῖ ἡπεϥῶχενχον ἡἔμι ἡῖλι. Τοτε πεϥῶτ αϥχος χε: "ῒωτεμ, παωῃρι, †ναωχενχον αν ἡ†σβω νακ κατα φῃ† ἔ†οὔωω. ῒω† ἔροκ ἔτεκωε ἔβολ ται, †νατῃκ ἔτοτϥ ἡοὔρεϥ†σβω εϥσωτπ ρινα ἡτεϥερπιαζιν ἡμοκ." Δὔοὔωρπ ἡπιἔελωρι ἔοὔβακι ἡωενμο, οὔοῖ αϥωωπῃ ἡατεν πιρεϥ†σβω ἡοὔρομπῃ τηρς. Μενενσα πιςνοῦ ἔτεμναῦ αϥτασῶο ἔπεϥῃ, οὔοῖ πεϥῶτ αϥωῃνῃ χε: "Παωῃρι, οὔ πε ἔτακἔμι ἔροϥ?" - "Παιωτ, αῖἔμι ἔ†αⲥⲡῃ ἡτε νιοὔῖωρ," αϥεροὔῶ. "Φῃοὔ† ναῃαι!" αϥωω ἔβολ ἡχε πεϥῶτ, "φαι πε ἔτακἔμι ἔροϥ τηρϥ? †ναοὔορπκ ἔκεβακι ρα κερϥ†σβω."





https://www.youtube.com/watch?v=aWr6tUKX-Q&list=RDWR6tUKX-Q&start_radio=1

研究基盤となる主要プロジェクト

コプト語テキストのデジタル化

Coptic SCRIPTORIUM

コプト語研究の中核的デジタル基盤

CORE
PLATFORM

オープンアクセスで利用可能な豊富なデジタルコーパス

標準化されたフォーマットによる検索・分析機能

Caroline T. SchroederとAmir Zeldesが率いる国際共同プロジェクト

Coptic NLP Service

自動言語解析パイプライン

NATURAL LANGUAGE
PROCESSING

トークン化、正規化、品詞タグ付け、見出し語化を自動処理

Web APIを通じた他のツールとの連携が可能

機械可読な形式での高度な言語データ生成

Coptic OCR

AUTOMATIC
RECOGNITION

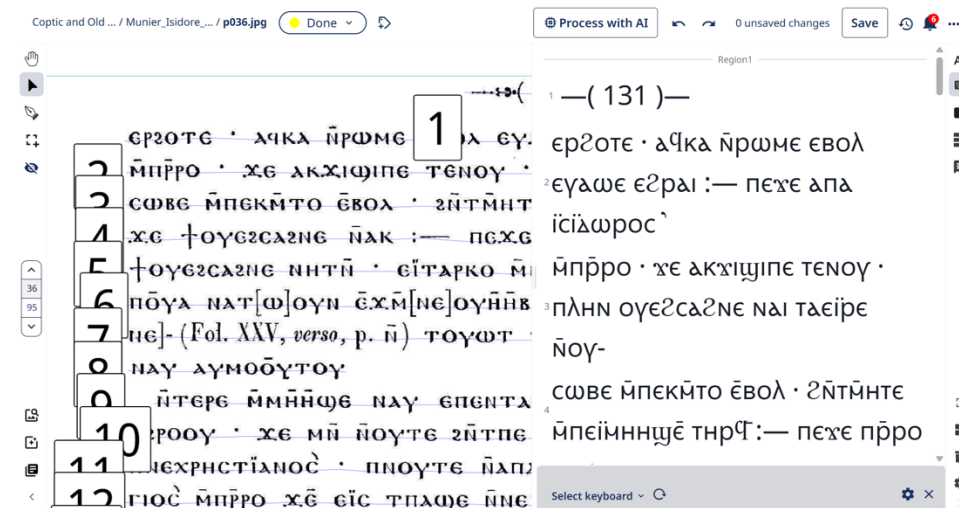
高精度な自動文字認識

ニューラルネットワークを用いたOCRモデルの開発

手書き写本や古い印刷物のデジタルテキスト化を加速

90%以上の認識精度を実現し、入力作業を大幅削減

処理対象となる実際のデータ（Transkribus上の訓練済みモデルを使用）



統合された研究ワークフロー

📄 画像データ (写本・版本)

✍️ Coptic OCR によるテキスト化

⚙️ NLP Service による自動解析

📖 Coptic SCRIPTORIUM (公開・検索)

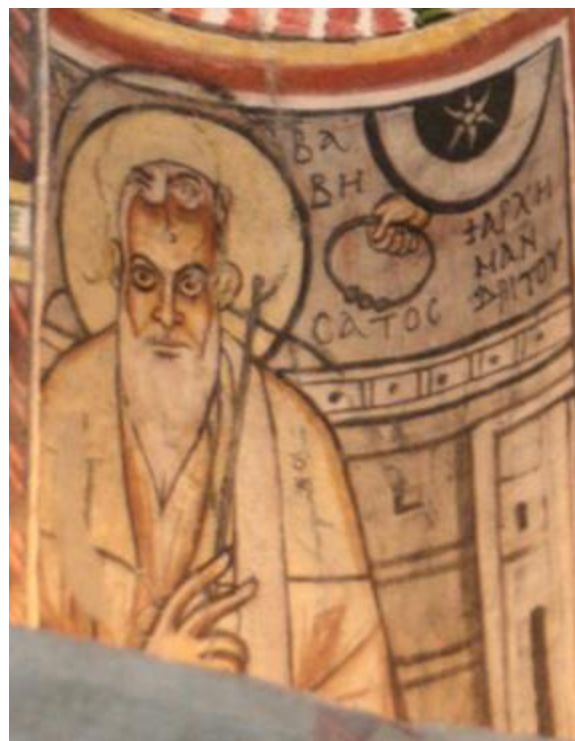
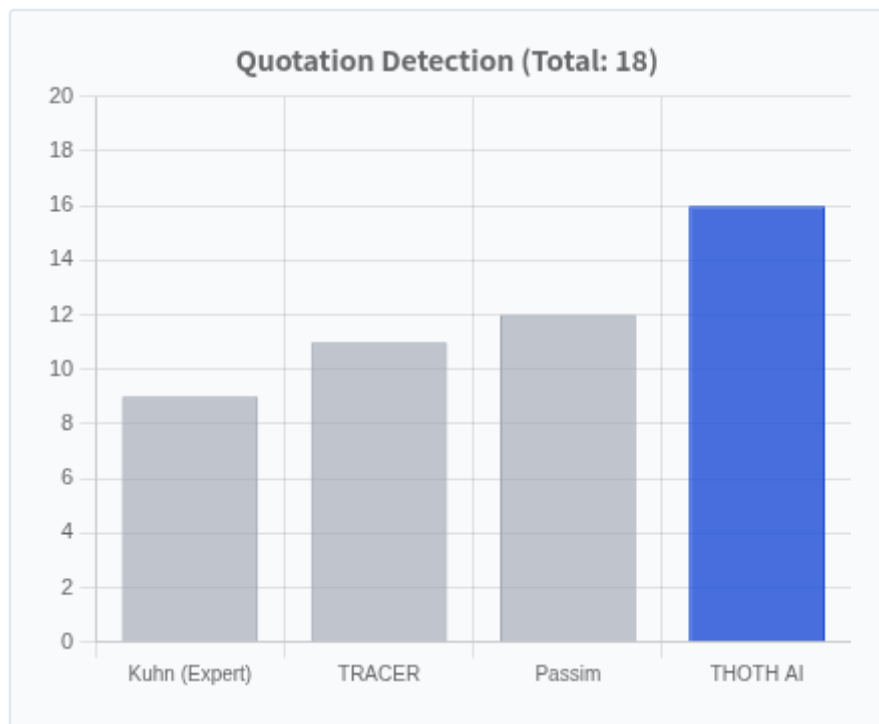
CASE STUDY

THOTH AIによるテキスト再利用検出

Comparing AI Performance in Coptic Psalm Quotation Detection

◎ プロジェクトの成果

紀元後5世紀のベーサによるコプト語のテキストにおける詩篇の引用（全18箇所）の検出率比較。THOTH AIは従来の手法や先行研究を上回る検出精度を達成し、AIと専門知識の融合の有効性を実証。



ベーサを描いたフレスコ画（赤修道院）

📊 定量的成果

16/18

THOTH AI 検出数（精度約89%）

+44%

対Kuhn版（先行研究）比の向上

💡 新発見の具体例

PS. 59:10 (NEWLY DETECTED)

"I will extend my shoe over Edom"

(ⲧ-ⲛⲁ-ⲥⲟⲟⲩⲧⲛ ⲙ-ⲡⲁ-ⲧⲟⲟⲩⲉ ⲉⲗⲛ-ⲧ-ⲁⲟⲩⲙⲁⲓⲁ)

※従来の手法では見逃されていた引用箇所をAIが特定



COLLABORATION MODEL

AIの提案を専門家が検証する「Human-in-the-loop」により、最高水準の成果を実現。

AI活用のハードルとその克服

💡 きっかけ：小規模PoCの成功体験から段階的拡張へ

✓ 4つの主要な障壁と実践的な解決アプローチ

データ不足（低資源言語）

- 課題：コプト語等の学習データが圧倒的に不足し、既存モデルの精度が出ない。

解決策：

- 既存資料（辞書・論文）の統合・正規化
- RAG技術による知識補完と合成データの活用

権利・倫理への不安

- 課題：学習データの著作権処理や、AI利用による倫理的懸念。

解決策：

- CCライセンス明示データの優先利用
- 利用規約テンプレートの整備と公開範囲の段階設定

品質管理（幻覚・誤訳）

- 課題：専門性の高い分野での「もっともらしい嘘（ハルシネーション）」のリスク。

解決策：

- 人間によるDouble-blind検証プロセスの導入
- BLEU/chrF等の客観的評価指標とエラーログの活用

スキルギャップと計算資源

- 課題：プログラミングスキルの不足と高価な計算資源の確保。

解決策：

- ショートコース（Python/NLP）とGitHubコード共有
- クラウドクレジット活用や学内GPUの共同利用

タロコ語（Truku Seediq）復興とAI活用

コプト語研究からの知見展開と台湾原住民言語への応用



AI対話型教育システム

消滅危機言語の学習支援チャットボットとの対話を通じた実践的学習。

若年層への訴求親しみやすいUIで学習ハードルを低減。



低資源言語の機械翻訳

双方向翻訳モデル タロコ語⇔中国語・日本語。転移学習応用。

方言差への対応地域差を吸収する多様性対応モデル。



コプト語研究との共通点

Methodology Transfer

「消滅危機」から「再生」へのモデル展開：コプト語の成功例（AI×コミュニティ）を水平展開。

汎用的復興フレームワーク：データ不足等の共通課題に対するソリューション提示。

REVITALIZATION



地理的コンテキスト

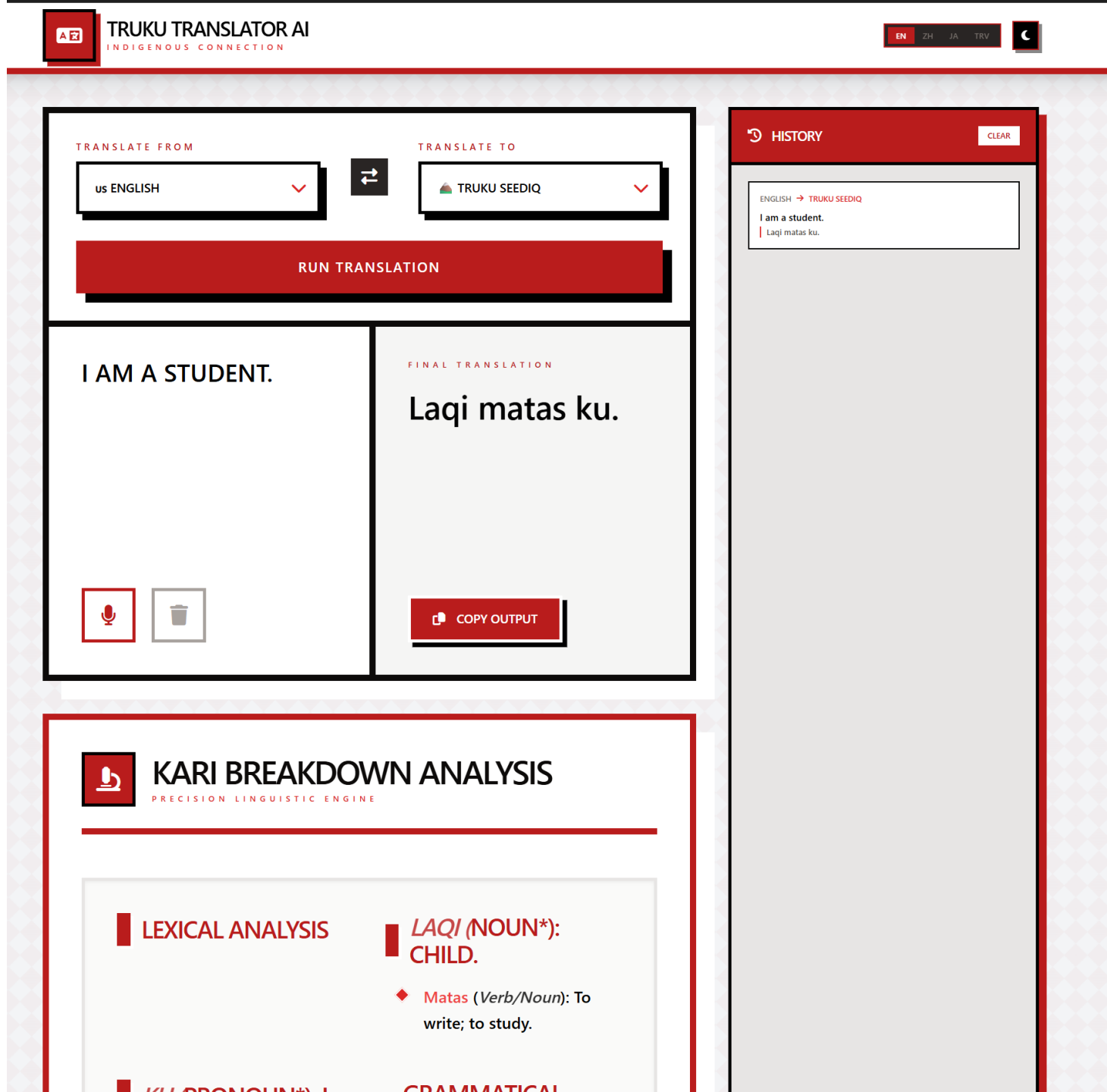


台湾原住民族の分布とタロコ族（Truku）の居住地域

タロコ語と現代語（日英中）の 自動相互翻訳アプリ

JSPS科研費 基盤研究 (A) 「生理指標付き自然談話コーパスに基づく危機言語の擬似動態保存と脳内処理過程の解明」

（研究代表者：東北大 小泉政利）
において発表者が開発





同科研の小泉・里・宮川の講演も設けられたタロコ語言語復興とAI応用の祭典（2025年2月）原視新聞網
「族語學習新亮點 AI建構太魯閣族語言模型 | 每日熱點新聞 | 原住民族電視台」（宮川開發のAI生成タロコ語歌曲やタロコ語などを話すAIアバターを披露）<https://www.youtube.com/watch?v=OLxb7AQbnkQ>

研究におけるAI使用のベストプラクティス



倫理的な利用

バイアスへの意識AI出力に含まれる文化的・歴史的偏見を批判的に評価し、無批判な再生産を防ぐ。

責任の所在AIは著者にはなり得ない。最終的な内容の正確性と独創性の責任は人間が負う。



透明性と再現性

使用ツールの明示論文内で使用したAIモデル名、バージョン、アクセス日時を正確に記載する。

プロンプトの公開生成に使用した指示（プロンプト）を付録やリポジトリで公開し、検証可能性を担保。



品質管理と検証

ハルシネーション対策AIが生成した事実、引用文献、翻訳結果は必ず一次資料で裏付けを取る（Fact Checking）。

Human-in-the-LoopAI任せにせず、専門知識を持つ人間が必ずプロセスに介入し監修を行う。



引用とデータ保護

入力データの保護未公開の研究データや個人情報を含む資料を、学習に利用されるAIに入力しない。

著作権の尊重学習元の権利関係と、生成物の利用規約を確認し、適切な引用を行う。

研究活動の変化：課題と懸念



AI依存と認知リスク

批判的思考の低下AIの回答を鵜呑みにし、原典確認を怠るリスク。

思考のアウトソーシング試行錯誤のプロセスが省略され、深い洞察が得にくくなる。

ブラックボックス化結論に至る推論過程が不透明なまま利用してしまう。



データの偏りと質

低資源言語の学習データ不足コプト語などのデータが少なく、AIの精度が安定しない。

英語中心主義バイアス英語圏の価値観や論理構成が、非英語圏の研究にも混入。

幻覚（ハルシネーション）もっともらしい嘘（架空の文献や誤った文法）の生成。



倫理的問題

著作権と透明性学習データの権利関係が不明瞭。AI利用の開示基準も過渡期。

剽窃との境界線AI生成テキストのそのままの利用は、オリジナリティを損なう。

再現性の欠如生成AIの出力はランダム性があり、科学的検証が困難な場合も。



デジタルデバイド

技術格差の拡大高度なプロンプトエンジニアリングやAPI利用ができる層との乖離。

コストの問題高性能な有料モデルやGPUリソースへのアクセス権の不平等。

インフラ依存特定企業のプラットフォームに研究基盤が依存する脆弱性。

研究者が備えるべき能力と、それを支えるインフラストラクチャ 必要なスキルと環境

学術リポジトリの例：
欧州のCERNによって運
営されるZenodo

zenodo



研究者に求められるスキル

データ整備・管理 (RDM) TEI/CSV/JSON形式への構造化、メタデータ付与、ライセンス理解と適切な管理。

NLP・OCR・MTの基礎 自然言語処理や機械翻訳の仕組みを理解し適切な評価指標で精度を判断する能力。

実装力と再現性 Python/Git/APIを活用したデータ処理、可視化、及び分析プロセスの再現性担保。

研究倫理とリテラシー バイアスの認識、情報セキュリティ、AI生成物の責任ある利用。

Data Literacy

Technical Skills

Ethics



必要な研究環境・インフラ

安全な計算基盤 学内または国レベルで提供される、セキュアなGPU環境やコンテナ実行基盤。

信頼できるデータリポジトリ DOI付与、標準メタデータ対応、長期保存が保証されたデータ共有プラットフォーム。

ドメイン資源とツール 専門分野に特化した辞書、語彙資源、モデルカード、データカードの整備。

ワークフロー支援 CI/CD、品質管理、監査ログなど、研究プロセスを効率化・自動化する仕組み。

Computing Power

Repository

Resources

まとめ：人文学研究の変革と支援

研究変革の実証

不可能が可能に：古代文字の欠損復元や大規模コプト語文献の全数調査など、従来の手法では到達困難だった成果を実現。

時間の劇的短縮：数年かかる翻刻・翻訳作業を数週間に短縮し、研究者が「解釈と議論」に集中できる環境を創出。



活用の鍵と波及効果

障壁の低下：ノーコードツールやチャットボット活用により、文系学生や市民研究者の参入障壁が大幅に低下。

国際連携の加速：データ標準化（TEIやIIIFやLOD）により、世界中の国際プロジェクトとの接続と共同開発による共同発展。



成功のための必須要素

インフラ整備：権利処理済みの高品質な学習データセットと、安全に利用できる計算資源の確保。

スキル教育：プロンプトエンジニアリングだけでなく、統計的評価や倫理的判断を含む体系的なカリキュラム。



具体的なアクション提言

パイロット支援：小規模でも成功事例を作るための「PoC（概念実証）ファンド」の創設。

AI開発：人文学の文脈を理解する、日本語などの英語以外の言語や、少数言語や消滅危機言語などに特化した学術用LLM（ソブリンAI）の構築推進。

共有知の形成：失敗事例も含めたナレッジベースと、分野を超えた相談窓口の設置。

