

AIを活用した低資源言語のコーパス構築

古代エジプト語・コプト語を中心に

みやがわ そう
宮川 創

筑波大学 人文社会系 言語コミュニケーション学域 准教授

筑波大学 西アジア文明研究センター エジプト学部門 准教授

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 共同研究員

目次

- ① 低資源言語とは？
- ② AI 駆動型コーパス構築
- ③ 実例紹介
 - 古代エジプト語
 - 間テキスト性研究
- ④ 最新研究事例：THOTH AI
- ⑤ その他の低資源言語への AI 応用
- ⑥ 今後の展望とまとめ



Figure: ルクソール博物館・トトメス3世のヒエログリフのレリーフ (宮川撮影)

目次

- 1 低資源言語とは？
- 2 AI 駆動型コーパス構築
- 3 実例紹介
 - 古代エジプト語
 - 間テキスト性研究
- 4 最新研究事例：THOTH AI
- 5 その他の低資源言語へのAI応用
- 6 今後の展望とまとめ

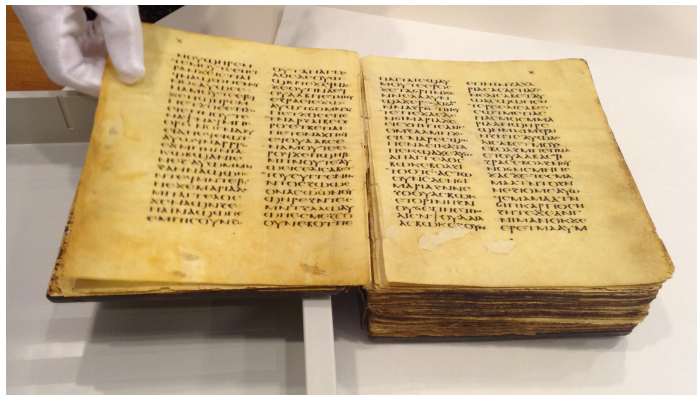


Figure: バルセロナ・イエズス会文書館所蔵・コプト語「バルセロナ福音書写本」(宮川撮影)

低資源言語とは？

デジタル時代の言語格差

低資源言語の定義

- デジタルテキスト資源が少ない言語（少資源言語とも）
- 言語モデルに学習させるテキストが少ないため、AI や大規模言語モデルの開発が困難
- 数多くの消滅危機言語と消滅してしまった言語（注：ラテン語や古典ギリシャ語など消滅したが資源の数が多い言語を除く）

現状の課題

- 限定的なデジタルコーパスと専門家不足
- （複雑な文字体系を持つ文献言語に限ると）文字体系のコンピュータでの処理が不十分

よく AI の学習データとなっている **Common Crawl** における言語分布

CC-MAIN-2025-21
(1 タイル $\approx$ 1%)



言語トップ 10

英語 (eng)	43.788%	フランス語 (fra)	4.254%
ロシア語 (rus)	5.892%	不明	2.921%
ドイツ語 (deu)	5.582%	イタリア語 (ita)	2.365%
中国語 (zho)	4.857%	ポルトガル語 (por)	2.341%
日本語 (jpn)	4.850%	その他 (約 150 言語)	18.440%
スペイン語 (spa)	4.711%		

消滅危機言語の現状

世界と日本の言語多様性の危機

言語の消滅とは

母語話者・学習者がいなくなること

約 7,000 言語のうち、約 2,500 言語が消滅の危機

日本の消滅危機言語（UNESCO 認定）

【極めて深刻】

アイヌ語

【重大な危機】

八重山語、与那国語

【危険】

八丈語、奄美語、国頭語、沖縄語、宮古語

世界の言語状況

● 消滅した言語

タスマニア諸語、エフタル語、ピクト語など

● 古典語として存続

ラテン語、シュメール語、アッカド語など

● 消滅したが復興した言語

ヘブライ語、コーンウォール語、マン語など

● 復興がかなり成功した言語

ハワイ語、マオリ語など

● 消滅危機にあるが復興中の言語

アイヌ語、琉球諸語、ノビイン語など

● 一度消滅してしまったが復興中の言語

コプト語など

目次

- 1 低資源言語とは？
- 2 AI 駆動型コーパス構築
- 3 実例紹介
 - 古代エジプト語
 - 間テキスト性研究
- 4 最新研究事例：THOTH AI
- 5 その他の低資源言語への AI 応用
- 6 今後の展望とまとめ

output/Luke_71_1L/010001.bin.png

7667PFXECC

2662P.E.X.EC6

output/Luke_71_1L/010002.bin.png

ΖΝΟΥΜΑΝΔΑ

ՀՈՍՄԱՆՃԱ

output/Luke_71_1L/010003.bin.png

ΙΕ·ΝΉΘΥΛΕΙΙΕ

ΙΕΝΤΟΥΔΕΠΕ

output/Luke_71_1L/010004.bin.png

ΧΑΥΝΑΥΧΕΜΑ

Figure: 2016 年に Kirill Bulert と宮川によって強化学習されたコプト語の OCR 自動翻刻モデルの結果

AIを活用した構築アプローチ: AI開発とコーパス構築の好循環

専門家による検証

専門家がデータの質を確保



コーパスの増大

質の良いコーパス
データの獲得



AI駆動型 コーパス構築

AIによる効率化

改善されたAIが教師
データ構築を効率化



学習量の増加

AIモデルが質の高い教
師データを学習



目次

- ① 低資源言語とは？
- ② AI 駆動型コーパス構築
- ③ 実例紹介
 - 古代エジプト語
 - 間テキスト性研究
- ④ 最新研究事例：THOTH AI
- ⑤ その他の低資源言語への AI 応用
- ⑥ 今後の展望とまとめ



Figure: アビュドスのセティ 1 世葬祭殿の内部のレリーフ (宮川撮影)

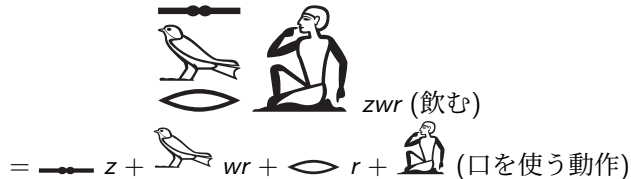
5000 年以上に渡って文字で書かれた古代エジプト語の世界

時代区分

- 紀元前 3200 年頃～紀元後 21 世紀
- 古エジプト語→中エジプト語→新エジプト語→民衆文字エジプト語→コプト語

文字体系

- ヒエログリフ：聖刻文字
(石碑・神殿・パピルス)
- ヒエラティック：神官文字
(パピルス・土器片など)
- デモティック：民衆文字
(パピルス・土器片・碑文など)
- コプト文字：ギリシャ文字＋いくつかの民衆文字



古代エジプト文字 (コプト文字以外) の特徴

表音文字は子音しか書けない / 単語の意味カテゴリーを表す限定符を置く / 振り仮名のような音声補字 など

コプト語とは

言語復興運動もなされている、古代エジプト語の最終段階

特徴

- 紀元 2-3 世紀頃に文字体系が成立
- ギリシャ文字 24 字＋デモティック由来 6-9 字
- キリスト教（コプト正教会）の典礼言語
- 初期キリスト教、マニ教などの資料が多数
- 日常語としての最後の記録は 17 世紀ころだが、19 世紀から言語復興運動が続いている

重要性

- 母音を表記→古代エジプト語の音韻再建の鍵
- シャンポリオンがコプト語を用いてヒエログリフ解読の基礎を築いて以降、コプト語が古代エジプト語解読の最大の鍵になった



京都府木津川市の「聖母マリア・聖マルコ・コプト正教会」の典礼では、コプト語・ギリシャ語・アラビア語・英語・日本語が用いられる（宮川撮影）

なぜ「低資源」なのか

データ不足の現実

- 日常話者の不在：新たな言語データが増えない
- 資料の限定性：残された文献は断片的
- デジタル化の遅れ：多くが未整理の写本・碑文
- 専門家不足：学習者・研究者が少ない

AI が学習するための

大規模データセットの構築が困難



宮川が大英図書館で第 25 王朝時代の擬古的なヒエログリフ文が刻まれたシャバカストーンについて調査をしている場面 (2014 年)

研究の課題①古典語の場合

資料の散在と未デジタル化

現状

- 世界中の博物館・図書館に分散保管
- 現地での閲覧許可が必要
- 手作業での転写・翻刻

実体験

発表者も大英図書館などで未刊行コプト語写本を調査した際、撮影不可のため、隅々まで手描きで写し取る必要があった



宮川がバルセロナのイエズス会研究施設にてパラウ・リベス・パピルス文書群を調査している場面（2014年）

研究の課題②

技術的な困難

文字体系の複雑さ

- ヒエログリフ：表音文字・表語文字・限定符の組み合わせ ⇒ 複雑 & 字の多さ
- 矩形中に縦横などに文字をはめこむ「矩形書法」
- Unicode 対応は不完全（ヒエラティック・デモティックは未対応）

専門知識の要求

- 高度な古代エジプト語の知識
- コンピュータ技術の両方が必要
- 両方に通じた人材は限定的

宮川が留学して博士課程を取ったドイツのゲッティンゲン大学エジプト学・コプト学専攻では、コプト語を含む古代エジプト語の各時代の専門教育が行われたが、このような環境は世界でも珍しい



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)

実例紹介：古代エジプト語とコプト語のコーパスと辞書

TLA & Coptic SCRIPTORIUM

<https://thesaurus-linguae-aegyptiae.de/>

Thesaurus Linguae Aegyptiae (TLA)

- 世界最大の古代エジプト語デジタルコーパス
- ベルリン・ブランデンブルク科学アカデミーのプロジェクト
- 各国研究機関の数十年にわたる協力の結晶
- 古エジプト語から民衆文字エジプト語までをカバー

TLA Thesaurus Linguae Aegyptiae v2.0

Search results: 1 - 20 of 1014 sentences with occurrences.
✂ (Possibilities to sort the results list will be added in a future version of the TLA web application.)

1 2 3 ... 99 100 101

ntr	ʿ	wʿs.n	=f	{hr}	rn	=f
substantive_masc (unspecified)	adjective Adj.sg	verb_3-4t SC.n.act.ngem.3sgm	personal_pronoun (unspecified)	preposition (unspecified)	substantive_masc Noun.sg.stpr.3sgm	personal_pronoun (unspecified)
N.ass	ADJ.sg	V.10M-ACT-INT.3TTP	-3sg.m	PREP	N.ass.3sgm	-3sg.m
Gott	groß, erhaben	ausprechen	[Suffix Pron. sg.3.m]	[Präposition]	Name	[Suffix Pron. sg.3.m]

n	ʿs.t	wr.t-hk3{t}	.
preposition (unspecified)	gods_name (unspecified)	epith_god (unspecified)	
PREP	DIVN	DIVN	
zu (jmd.)	Isis	die Zauberreiche (Götterinnen)	

Der große Gott, er offenbarte seinen Namen der Isis, der Zaubermächtigen.

Magische Texte Magische Papyri Sammelhandschriften des Neuen Reiches pTurin CGT 54051 = pTurin Cat. 1993
Recto 2,11-5,5: Spruch gegen Skorpiongift mit Legende von Isis und Re rto 2,14

Dating (time frame): 13. Dynastie

Main author(s): Katharina Stegbauer; further editor(s): Ägyptisches Wörterbuch, Lutz Popko, Florence Langermann, Anja Weber (latest changes to text: 10/18/2022)

Go to/quote sentence

Hieroglyphs
Switch annotation/block view
Word class
Gram. tagging BTS
Linguistic Glossing
Token translation
Sentence token ID
Translation
Dating of Text
Text editor

実例紹介：古代エジプト語とコプト語のコーパスと辞書

TLA & Coptic SCRIPTORIUM

Coptic SCRIPTORIUM <https://copticSCRIPTORIUM.org/> 全米人文科学基金 (NEH)

- コプト語最大のデジタルコーパス
- 徹底的な言語学的な語釈・文献学的な注釈
- 豊富な検索機能・統計解析機能
- データはすべてオープンソース
- ⇒ 宮川は2014年からコーパス開発に加わる(旧約聖書、修道文学など)

The screenshot displays the Coptic SCRIPTORIUM web interface. On the left, a search bar contains the query "norm="nyte"". Below it, a list of matches is shown, with "besa.letters" selected. The main area displays the search results for the selected match, showing the text "29 matches in 5 documents". The right side of the interface shows a detailed view of the selected match, including the text "29 matches in 5 documents" and a table of linguistic annotations. The table has columns for various linguistic features, including "cb_n", "lang", "lb_n", "lemma", "morph", "norm", "norm_group", "orig", "orig_group", "p_n", "pb_xml_id", "pos", "func", "entity", "translation", and "chapter_n". The table contains data for the selected match, showing the text "29 matches in 5 documents" and the corresponding linguistic annotations.

深掘り事例：間テキスト性 / テクストリユース研究

隠れた引用・引喩など、TRACER によるテキストリユース探知

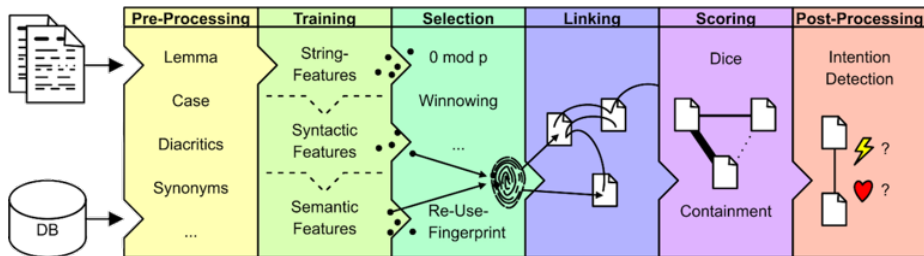
コーパス構築と間テキスト性解析が宮川のドイツ研究振興協会 (DFG) 研究員としての仕事 & 博士論文

課題

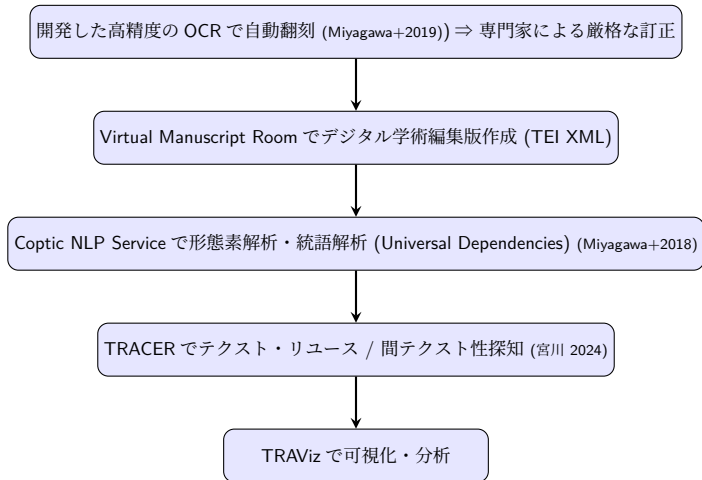
- 古代の文献では引用が明示されないことが多い
- 語形・語順の変化や巧みな言い換えにより、一瞥するだけでは発見は困難

TRACER の役割

コンピュータ支援により、引用・引喩・パラフレーズなどを自動検出し、人文学研究を加速させる



TRACER のワークフロー



Coptic WordNet (Slaughter+ 2019, Miyagawa+ 2025a) で同義語・類義語を考慮し、より高い精度の探知を実現。

TRACERによる新発見：コプト語修道書簡と旧約聖書詩編の間テキスト分析

先行研究で見逃された旧約聖書・詩編からの引用 (専門家がキリスト教学の基準で認めたもの)

対象文献

新規発見数

シェヌーテ『第6カノン』

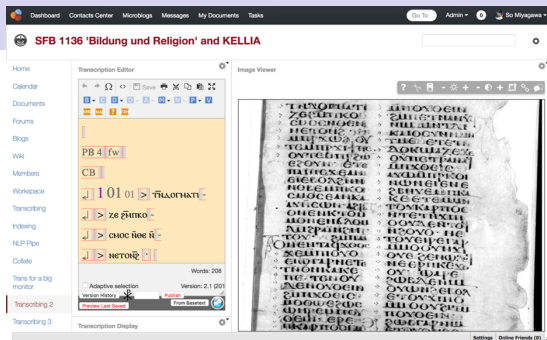
14件

ベーサ『書簡と説教集』

6件

発見のポイント

- 引用導入句がない
- 語彙や語順が変更されている (Miyagawa 2022)
- WordNet による同義語・類義語の考慮が有効 (Miyagawa+ 2025a)



⇒ 京都大学・天野恭子准教授（インド古典学）の監督のもと、東京大学・塚越柚季講師、ライプチヒ大学・京極祐希 RA とともに、この手法をヴェーダ・サンスクリット語に応用 (Miyagawa+ 2025b)

目次

- ① 低資源言語とは？
- ② AI 駆動型コーパス構築
- ③ 実例紹介
 - 古代エジプト語
 - 間テキスト性研究
- ④ 最新研究事例：THOTH AI
- ⑤ その他の低資源言語への AI 応用
- ⑥ 今後の展望とまとめ

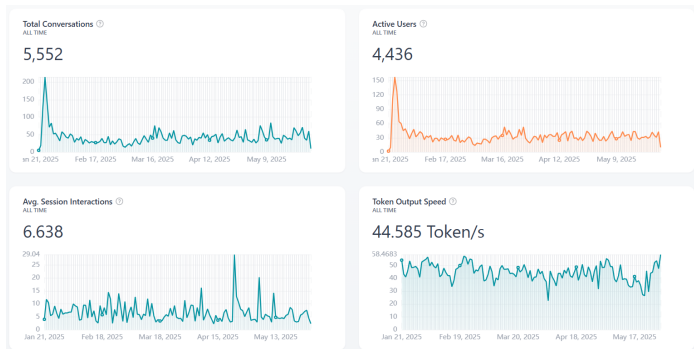


Figure: THOTH AI の 2025 年 2 月から 5 月までのユーザ数など

最新研究：RAG による学習支援

どの LLM が古代語タスクに最適か？

研究目的

検索拡張生成（RAG）技術を活用し、専門知識データベースを参照して古代エジプト語の翻訳や質問応答を行う AI 学習支援システム「THOTH AI」を開発・評価する

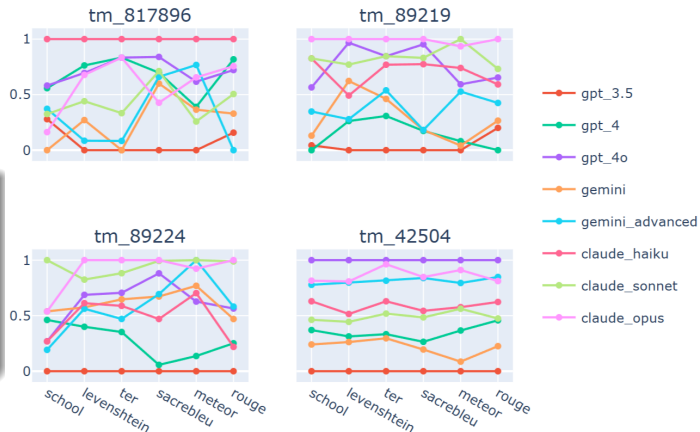


Figure: パイロット研究：Wannaz & Miyagawa 2024: 469 によるコプト語での諸大規模言語モデルの性能検証

評価対象

- GPT-4
- Claude 3.5 Sonnet
- Google Gemini 2.0 Pro
- その他最新モデル

評価タスク

中エジプト語（翻字）→英語翻訳
「トトメス1世戴冠ステラ」を使用

- 第18王朝トトメス1世の戴冠を告げる勅令
- 王名と称号の宣言
- 南方領土の行政官トゥリへの勅令

評価指標

- **BLEU**: n-gram 一致率
- **SacreBLEU**: 標準化版 BLEU
- **METEOR**: 意味的合致度
- **ROUGE**: 語句重複度
- **chrF**: 文字レベル F 値

評価方法

Nederhof による英語訳を参照訳として、各 LLM にゼロショットで翻訳を命令

実験結果

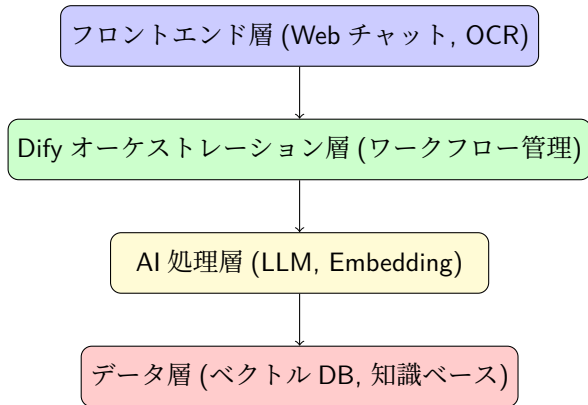
Model	BLEU	SacreBLEU	ROUGE-1	ROUGE-L	chrF	METEOR
Nederhof (Reference)	1.000	100.000	1.000	1.000	100.000	1.000
THOTH AI (RAG)	0.354	35.431	0.730	0.680	61.052	0.650
Claude 3.5 Sonnet	0.325	32.457	0.717	0.652	58.064	0.640
Gemini 2.0 Pro Exp.	0.288	28.772	0.714	0.681	59.937	0.4881
Gemini 2.0 Flash Exp.	0.256	25.590	0.697	0.664	59.298	0.4871
DeepSeek R1	0.222	22.163	0.675	0.4887	53.735	0.4810
GPT o1 Pro mode	0.216	21.598	0.625	0.4834	53.167	0.4809
GPT 4o	0.196	19.615	0.4881	0.484	51.109	0.417
GPT o3 Mini High	0.034	3.413	0.375	0.233	35.148	0.249

主要な発見

Claude 3.5 Sonnet と Gemini 2.0 Pro が高い性能を示したが、Claude 3.5 Sonnet にベクトル化した辞書を与えた THOTH AI が最高性能 (Miyagawa 2025)

システム構築：THOTH AI

Dify プラットフォームと 4 層アーキテクチャ



What is "Hi, my name is So Miyagawa, nice to meet you?" in every stage of Ancient Egyptian?



I'll help you translate this greeting through the major stages of Egyptian. Note that some concepts like "nice to meet you" didn't have exact equivalents, so we'll use culturally appropriate greetings.

Old Egyptian (c. 2700-2200 BCE):

ink rn=i So Miyagawa, nfr m*ḥ*=i tw
(A more formal, archaic construction)

Middle Egyptian (c. 2000-1350 BCE):

ink So Miyagawa rn=i, nfr m*ḥ*=i tw
(The classical form)

Late Egyptian (c. 1350-700 BCE):

p*ḥ*y=i rn So Miyagawa, nfr p*ḥ*y=i m*ḥ* tw
(More analytical structure with articles)

Demotic (c. 700 BCE-400 CE):

p*ḥ*y=i rn So Miyagawa, nfr p*ḥ*y=n nw r=i
(More colloquial form)

Coptic (Sahidic):

ⲡⲁⲣⲁⲛ ⲛⲉ ⲥⲟ ⲙⲓⲛⲁⲓⲛⲁⲱⲁ, ⲡⲣⲁⲱⲉ ⲉⲡⲣⲓⲛⲁⲱ ⲉⲣⲟⲕ

Coptic (Bohairic):

ⲡⲁⲣⲁⲛ ⲛⲉ ⲥⲟ ⲙⲓⲛⲁⲓⲛⲁⲱⲁ, ⲡⲣⲁⲱⲓ ⲉⲡⲣⲓⲛⲁⲱ ⲉⲣⲟⲕ

Note: The name "So Miyagawa" is kept as is since it's a modern name. In actual ancient usage, Egyptian names would have been used. The greeting concept evolves from more formal constructions in Old Egyptian to more analytical forms in later stages.

目次

- ① 低資源言語とは？
- ② AI 駆動型コーパス構築
- ③ 事例紹介
 - 古代エジプト語
 - 間テキスト性研究
- ④ 最新研究事例：THOTH AI
- ⑤ その他の低資源言語への AI 応用
- ⑥ 今後の展望とまとめ



Figure: 「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」原視新聞網 TITV News の報道より (Youtube; <https://www.youtube.com/watch?v=0Lxb7AQbnkQ>)

低資源言語への応用事例：沖縄語・アイヌ語・タロコ語

沖縄語：語彙・表記体系データ開発

- 多様な表記法の混在
- 標準日本語にない音素の表記法の不統一
- Miyagawa & Carlino 2023: 表記体系 DB 化
- OMOLO プロジェクトにおける辞書開発 (Miyagawa+ 2023, 宮川 2023a, 宮川 2023b)

アイヌ語：機械翻訳モデル開発

- 複統合語的特徴への対応
- 10 万文対のパラレルコーパス構築
- 転移学習による双方向翻訳 (Miyagawa 2024)
- 様々な方言に対応 (Igarashi & Miyagawa 2024)

タロコ語 (Truku Seediq)：AI 言語復興

- 台湾原住民諸語の 1 つ
- FALCOHN プロジェクトによる AI 活用プロジェクト
- 東北大学・小泉政利教授、(研究代表者)、沖縄国際大学・里麻奈美教授、東北学院大・那須川訓也教授と大沼仁美准教授、宮川創 (AI 担当)
- 現地協力者：東華大学・Apay Ai-yu Tracy Tang 教授、Chan 牧師ほかタロコ族の皆様
- HeyGen AI アバター教師の開発
- 音声認識・合成・機械翻訳の統合システム
- コミュニティ参加型研究 (CBPR)
- 2025 年 2 月：AI 言語復興イベント開催

AI技術の統合的アプローチ：古代語から現代危機言語まで

共通技術基盤

- 諸大規模言語モデル (LLM) 活用：転移学習や RAG
- **RAG 技術**：専門知識データベースとの統合
- **マルチモーダル処理**：テキスト・音声・画像の統合
- **Universal Dependencies**：言語横断的統語情報付きコーパス開発 (宮川+ 2023)

言語別カスタマイズ

- 古代エジプト語・コプト語：OCR 強化
- **タロコ語**：世代間変異への対応 (VOS → SVO)
- **沖縄語**：表記ゆれへの柔軟な対応

タロコ語プロジェクトの革新性

- **データ収集**：31 冊以上の書籍、10 以上のウェブサイト
- **音声技術**：Whisper ベースの音声認識・合成
- **教育応用**：SUNO AI 音楽生成、対話型学習
- **文化保存**：語りのデジタルアーカイブ化

自動文字認識→コーパス構築効率化

- 与論語 (ユンヌフトゥバ) の昔話文献を NDL OCR で自動翻刻したものをベースにコーパス化・分析 (宮川 forthcoming)
- 古ヌビア語の辞書 (ヌビア文字、コプト文字、アラビア文字、ギリシャ文字、ラテン文字が混在) を Transkribus を用いて訓練・自動翻刻 (Miyagawa 2023)

目次

- ① 低資源言語とは？
- ② AI 駆動型コーパス構築
- ③ 実例紹介
 - 古代エジプト語
 - 間テキスト性研究
- ④ 最新研究事例：THOTH AI
- ⑤ その他の低資源言語への AI 応用
- ⑥ 今後の展望とまとめ



Figure: 筑波大学筑波キャンパス中地区 (宮川撮影)

学際的協働と社会実装

チーム科学から地域社会へ

国際的研究ネットワーク

- 古代語：TLA（ベルリン）、Coptic SCRIPTORIUM（米国）
- 日琉語族：国立国語研究所、琉球大学、鹿児島大学等
- 台湾原住民語：国立東華大学、地域の文化協会

コミュニティ中心アプローチ

- データ主権の尊重と段階的アクセス管理
- 地域主導の優先順位設定
- 双方向の知識移転と能力構築
- 長期的パートナーシップの構築

教育・文化的インパクト

- デジタル教材：年齢別カスタマイズ（子供向けアニメーション、若者向けアバター）
- オンライン学習：THOTH AI、インタラクティブ文法ツール
- 文化継承：伝統的知識のデジタル保存
- 社会啓発：メディアを通じた言語多様性の発信

持続可能性の確保

- オープンソース化による技術共有
- 地域人材の育成プログラム
- 政府・民間との協働体制

研究成果と今後の展望: 2025 年の成果から 2030 年への道筋

主要成果 (2024-2025)

技術的成果

- Claude/Gemini の優位性実証
- RAG 技術の有効性検証
- 統合プラットフォーム構築

言語別成果

- エジプト語: THOTH AI
- アイヌ語: 翻訳モデル開発
- タロコ語: AI アバター実装

社会的成果

- コミュニティ連携確立
- 教育プログラム開始
- 国際協力体制構築

今後の研究計画 (2025-2030)

短期目標 (2025-2027)

- ヒエラティックとデモティックの OCR 開発
- マルチモーダル RAG (図像+テキスト統合)
- 世代別教材の効果測定と改善
- 他の台湾原住民語への技術展開

長期目標 (2028-2030)

- 統合プラットフォームの完成と公開
- 100 年スパンのデジタルアーカイブ構築
- AI による言語変化予測モデル開発
- グローバル危機言語ネットワーク形成

ご清聴ありがとうございました

消滅の危機にある言語の保存と継承は 人類共通の課題です

私たちの先端技術が、過去の知恵と未来の希望を繋ぐ
架け橋となることを願っています

2025 年 7 月 8 日 宮川 創

お問い合わせ：miyagawa.so.kb@u.tsukuba.ac.jp

ホームページ：<https://somiyaagawa.com/>

謝辞

本研究は JSPS 科研費 JP15J05370、JP24K03232、JP24K00074、JP25K00399、JP24H00085、JP23H00007、JP23K21993、JP22H00721、JP21KK0004、JP21K18376、JP21K00537、JP20K21975、JP19H01265、JP19H05589 の助成を受けたものです。

また、以下の研究プロジェクトからも支援を受けています：

- 人間文化研究機構 共同研究プロジェクト「学術知デジタルライブラリの構築」（2022 年度～）共創先導プロジェクト 共創促進研究
- 国立国語研究所 共同研究プロジェクト「異分野融合による総合書物学の拡張的研究」国語研ユニット「古辞書類に基づく語彙資源の拡張と語彙・表記の史的変遷」（2022 年度～）広領域連携型基幹研究
- 国立国語研究所 共同研究プロジェクト「消滅危機言語の保存研究」（2022 年度～）機関拠点型基幹研究「開かれた言語資源による日本語の実証的・応用的研究」基幹型
- 大学共同利用機関情報・システム研究機構データサイエンス共同利用基盤施設 2022-2024 年度公募型共同研究「ROIS-DS-JOINT」「日琉諸語の言語類型アトラス LAJaR の開発と分析」

2025 年

- Miyagawa, So. 2025. “RAG-Enhanced Neural Machine Translation of Ancient Egyptian Text: A Case Study of THOTH AI.” In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities 2025*, edited by Mika Härmäläinen et al., 33–40. Albuquerque: Association for Computational Linguistics.
- Miyagawa, So, Luis Morgado da Costa, Laura Slaughter, and Heike Behlmer. 2025a. “Automatic Detection of Coptic Text Reuse: Applying Coptic Wordnet to Intertextuality Studies in Selected Coptic Monastic Writings.” In *Proceedings of 13th International Global WordNet Conference (GWC2025)*.
- Miyagawa, So, Yuki Kyogoku, Yuzuki Tsukagoshi, and Kyoko Amano. 2025b. “Computational Pathways to Intertextuality of the Ancient Indian Literature: A Multi-Method Analysis of the Maitrya and Khaka Sahits.” *Journal of Data Mining and Digital Humanities* 15085: 1–12.

2024 年

- Igarashi, Ryo, and So Miyagawa. 2024. “Enhancing Neural Machine Translation for Ainu-Japanese: A Comprehensive Study on the Impact of Domain and Dialect Integration.” In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, edited by Mika Härmäläinen et al., 413–422. Miami: Association for Computational Linguistics.
- Miyagawa, So. 2024. “Ainu – Japanese Bi-directional Neural Machine Translation: A Step Towards Linguistic Preservation of Ainu, An Under-Resourced Indigenous Language in Japan.” *Journal of Data Mining & Digital Humanities* 13151: 1–10.
- Wannaz, Audric-Charles, and So Miyagawa. 2024. “Assessing Large Language Models in Translating Coptic and Ancient Greek Ostraca.” In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, edited by Mika Härmäläinen et al., 463–471. Miami: Association for Computational Linguistics.
- 宮川創. 2024. 「間テキスト性のコーパスデータ駆動型研究: テクストリユース探知プログラム TRACER の 古代末期キリスト教コプト語書簡への応用」『情報処理学会論文誌』65(2): 305–321.

2023 年

- Miyagawa, So. 2023. “Digitizing Old Nubian Dictionary: Optical Character Recognition for Multi-Lingual and Multi-Script Text From Medieval Africa.” In *Proceedings of 7th IEEE Congress on Information Science and Technology (CiSt), Agadir - Essaouira, Morocco, 2023*, 587–592.
- Miyagawa, So, and Salvatore Carlino. 2023. “Database of Writing Systems and Orthographies for Okinawan Language: Toward Preservation of Okinawan Linguistic Cultural Heritage.” In *Proceedings of JADH conference, vol.2023*, 33–35. Tokyo: JADH Program Committee and Local Organizing Committee.
- Miyagawa, So, Kanji Kato, Miho Zlazli, Salvatore Carlino, and Seira Machida. 2023. “Building Okinawan Lexicon Resource for Language Reclamation/Revitalization and Natural Language Processing Tasks such as Universal Dependencies Treebanking.” In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, edited by Nikolai Ilinykh et al., 86–91. Stroudsburg, PA: Association for Computational Linguistics.

2019 年

- Miyagawa, So, Kirill Bulert, Marco Büchler, and Heike Behlmer. 2019. “Optical character recognition of typeset Coptic text with neural networks.” *Digital Scholarship in the Humanities* 34(Suppl. 1): i135–i141.
- Slaughter, Laura, Luis Morgado Da Costa, So Miyagawa, Marco Büchler, Amir Zeldes, and Heike Behlmer. 2019. “The Making of Coptic Wordnet.” In *Proceedings of the Tenth Global Wordnet Conference*, 166–175.

2018 年

- Miyagawa, So, Amir Zeldes, Marco Büchler, Heike Behlmer, and Troy Griffiths. 2018. “Building linguistically and intertextually tagged Coptic corpora with open source tools.” In *Proceedings of the 8th Conference of Japanese Association for Digital Humanities*, 139–141.