第18回 JHPCN シンポジウム ／ 2026年度 公募型共同研究（研究計画）

オープンソース大規模言語モデルと検索拡張生成を活用した古代言語における引用・引喩の自動探知システムの開発

課題番号 jh261019

ICoMa
Intertextuality
Collation Machine
https://i-co-ma.vercel.app

研究代表者：宮川 創[1]／ 研究分担者：天野 恭子[2]・塚越 柚季[3]／ 研究協力者：京極 祐希[4]

[1]筑波大学 人文社会系（准教授）／ [2]京都大学 大学院文学研究科／ [3]東京大学 大学院人文社会系研究科／ [4]ライプツィヒ大学 南・中央アジア研究所（独）

Correspondence: miyagawa.so.kb@u.tsukuba.ac.jp ／ 研究実装：ICoMa v2.0（LLM+RAG）

1. 背景と目的

01

コンピュータ科学の語で言えば、本研究はノイズの多い多言語テキスト列の近似検索・アラインメント・根拠提示の問題である。入力は古代語文献のコーパス、出力は「どの箇所が、どの先行箇所を、どの程度再利用しているか」という候補リストである。

引用は、先行テキストの語句・句・文を比較的明示的に再利用する場合を指す。**引喩（allusion）**は、同じ語をそのまま使わず、場面・概念・比喩・語順の響きによって先行文献を想起させる、より曖昧な再利用である。

例：注釈書が聖典の一句をほぼ同じ形で繰り返すなら、文字列類似度で検出しやすい引用である。一方、「火を再び置く」「秩序を回復する」という儀礼的構図だけを借り、語形や表現を変える場合は、表層一致だけでは拾いにくい引喩になる。

ここが**テキストリユース探知**と**インターテクスチュアリティ**の交差点である。前者は候補を計算的に発見する探索問題、後者は候補が伝承・注釈・翻訳・権威づけの中で何を意味するかを解釈する人文学的問題である。本研究はICoMaを基盤に、LLM＋RAGで**意味的・引喩レベルの検出**へ拡張する。

既存成果：対象と手法

2. ヴェーダ3伝承の計算的校合

02

クリシュナ・ヤジュルヴェーダは、複数の学派（śākhā）が口承で伝えた**並行散文テクスト**として残る。本研究は3伝承（**マイトラーヤニー（MS）・カータカ（KS）・タイッティリーヤ（TS）**）の間テクスト性を、多アルゴリズムで定量化した。

この成果はACL 2026内のワークショップNLP4DHで発表採択され、国際会議プロシーディングズ論文としてACL Anthologyで公開予定である。

- **対象章：**アグニウパスターナ（聖火の日常礼拝）とプナルアーダーナ（聖火の再設置）の2儀礼章
- **テクスト形態：**表層形（plain）と見出し語化形（lemmatized）の2種
- **6実装・5独立手法：**Levenshtein距離・Character n-gram・Jaccard係数・Word n-gram・スクリプト対応正規化・Smith-Waterman局所整列（窓幅 n=3）。Character n-gramは現設定ではLevenshteinと同一結果のため、独立比較では5手法として扱う。

負の対照群としてMSの異質章（Darśapūrṇamāsaマントラ）を併置。

同一成果：結果

3. 主要な結果

03

① **一貫したランキング：**5つの独立手法・両形態が例外なく **MS-KS > KS-TS > MS-TS** を支持。文献学的通説を定量的に裏付け。

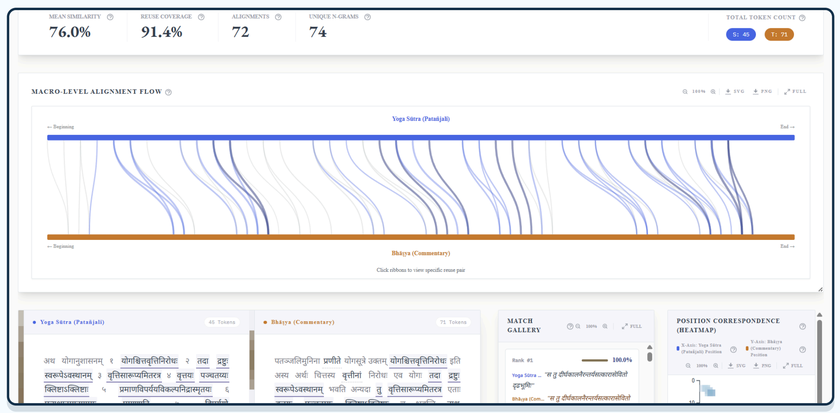
ペア	アグニウパスターナ	プナルアーダーナ
MS-KS	50.1%	93.5%
KS-TS	36.2%	20.7%
MS-TS	31.9%	20.5%

表：再利用カバレッジ（Levenshtein・表層形）。プナルアーダーナのMS-KSが突出。

② **2儀礼章の対照的プロファイル：**プナルアーダーナはMS-KSが**最大93.5%**と近似する一方TSとは乖離。アグニウパスターナは3ペアに広く中程度の類似が分布。文献間の類似は一定ではなく、儀礼章ごとに違いがあることが分かる。この儀礼章ごとの類似度の変化は、その章が成立した時代の違いを反映している可能性がある。アグニウパスターナMS-KSの50.1%は儀礼と思想の共有を表すが、プナルアーダーナMS-KSの93.5%は直接の借用関係を示すと考えられる。

4. ICoMa画面が示す再利用

04



サンスクリット語の全体画面：上段のmacro-level alignment flow、下段の本文ハイライト、match gallery、position correspondence heatmapを同時に表示し、候補の位置・順位・分布を一画面で検証する。

シリア語（Peshitta）の全体画面：右横書きの文字体系でも、リボン図、本文中の対応箇所、完全一致候補、ヒートマップを同一の読解手順で確認できる。

ICoMaは「候補一覧を見る道具」ではなく、候補の位置・文脈・順位・全体分布を同時に示す読解支援UIである。

5. テクスト史への含意

05

- ✔ **表層形（plain）の文字レベル解析が有効。**音を改変せず保存するヴェーダ文献の特性を捉え、意味解析（Word2Vec）と同等の結果を低コストで得る。
- ✔ プナルアーダーナ MS-KSの近接は、単なる共通祖型からの分岐でなく**直接的な借用関係**を示す強い数値的証拠。
- ✔ 類似度の差は、テクストが**固定化した時期の差**を反映する、という新たな成立史の説明枠組みを開く。

成果：Miyagawa, Amano, Tsukagoshi & Kyogoku。ACL 2026内NLP4DH採択、ACL Anthologyでプロシーディングズ論文公開予定。ICoMaと実験データは無償公開。

研究計画 ①

6. ICoMa v2.0：LLM＋RAG

06

課題：現行ICoMaは文字列の表層的一致に基づくため、語彙・語順が変化した引用や、直接の語の一致がない引喩の検出に限界がある。

本計画はLLMとRAGによる意味的類似性検出機能を追加し、3点を実現する：

- **意味理解に基づく引用検出：**表層一致でなく意味的類似性で、変形された引用を検出
- **引喩の自動検出：**RAGで辞書・文法書・概念知識を統合し、概念的類似性から引喩を検出
- **多言語汎用設計：**文字体系・文法に依存せず、コプト語、古代エジプト語、ギリシア語、ラテン語、古典中国語、古典アラビア語、楔形文字言語、古典シリア語、古典エチオピア語等へ拡張可能

ケーススタディ：コプト語文献（古代末期エジプト・約300万語）とヴェーダ語文献（古代インド・約300万語）。

1. 入力	2. 検索	3. 知識	4. 読解
TEI/XML・校訂本文・注釈を正規化	表層一致と埋め込み検索で候補生成	辞書・文法・概念情報をRAGで接続	候補・根拠・反例をICoMaへ返す

研究計画 ②

7. 検出対象・評価設計

07

ICoMa v2.0では、候補生成を**retrieval＋reranking＋explanation**のタスクとして設計する。モデルのスコアだけでなく、なぜ候補になったかをUIに戻す。

- **候補生成：**表層一致、埋め込み類似度、RAG検索を組み合わせて高再現率で抽出
- **順位づけ：**語彙・語順・屈折形の変化を許容し、意味的近接性と文脈でrerank
- **説明：**候補本文、類似箇所、関連語、参照した知識源をUI上で対応づける

直引用	変形引用	引喩
語句・句・文が高く一致。AIは文字列類似と位置対応で候補化。先行本文を権威として明示的に呼び込む。	屈折・語順・同義表現が変化。AIは埋め込み類似度とrerankingで拾う。文脈に合わせて先行本文を再構成する。	語の一致が弱く、場面・概念・比喩が響く。AIはRAGで辞書・概念知識を参照。読者が先行伝承を想起する関係を作る。

3種	F1	UI
一致・変形引用 引喩	目標値 > 0.75	候補と根拠を読解へ戻す

モデルの出力だけで結論にせず、候補・根拠・反例を研究者が比較できる形にする。

研究計画 ③

8. 研究計画（2026.4-2027.3）

08

1. **コーパス整備・評価（4-6月）**
   前処理（TEI/XML→Plaintext, Unicode正規化）、ベースライン実験、訓練データ作成、パイロット（F1>0.75）
2. **LLM微調整・RAG統合（7月-1月）**
   ByT5/mBERT/XLM-RoBERTa微調整、Sentence-BERTで25万件をベクトル化、FAISS索引、ハイパラ探索
3. **大規模評価・公開準備（2-3月）**
   全量推論、A/Bテスト、説明可能AI（BertViz, SHAP）、ICoMa v2.0 REST API統合（<5秒/クエリ）

評価指標：Precision・Recall・**F1 > 0.75**、MRR > 0.75、専門家による引喩検出の妥当性評価。

9. オープンサイエンス・意義

09

- 全成果物をオープンライセンスで公開：モデル＝Hugging Face（Apache2.0/MIT）、データ＝Zenodo（CC BY 4.0・DOI）、コード＝GitHub（MIT）
- 著作権フリーの古代語コーパスのみ使用（Coptic SCRIPTORIUM, KELLIA, GRETIL, Muktabodha）。個人情報は一切扱わない
- 公開時には国内外の研究者・学生へ一般公開し、研究・教育で再利用できる形に整備する

波及：古代ギリシア語・ラテン語・古典中国語・古典アラビア語・楔形文字言語など、文字体系を超えた間テクスト性研究への汎用基盤を構築する。

公開物と検証単位

UI：候補の位置、本文、順位、ヒートマップを同じ画面で確認できる形で公開。
データ：入力テキスト、正規化手順、候補リスト、除外例を対応づけて保存。
モデル：推論設定、評価指標、失敗例を記録し、再評価可能な形で共有。

ユニバーサルデザイン：色だけに依存せず、番号・罫線・見出しで構造を示す。十分な文字サイズ、左揃え本文、高コントラストの表・図キャプションで読みやすさを確保する。

研究利用	教育利用	再実験
候補を専門家が確認し、校訂・注釈・伝承史の議論へ接続する。	候補、根拠、判断ログを教材化し、古代語読解の訓練に使う。	入力、設定、失敗例を保存し、他言語・他コーパスで検証する。

F	A	I	R
DOI・索引・言語タグで発見可能にする。	Web UI・API・静的データでアクセス可能にする。	TEI/XML・Unicode正規化・語分割規則を明示する。	ライセンス、版、評価ログを添えて再利用可能にする。

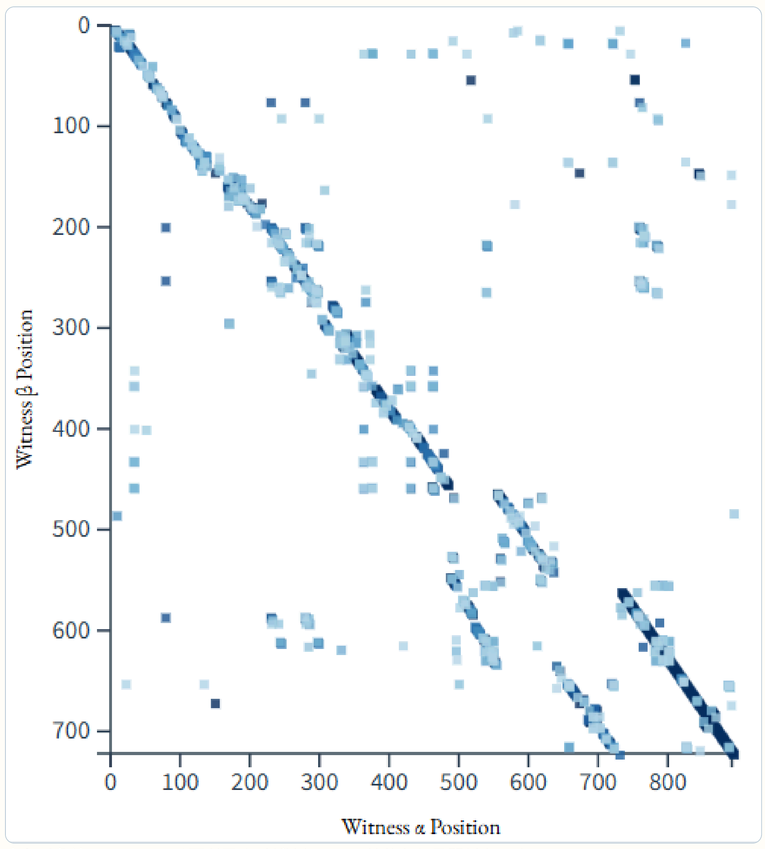
公開後の流れ：候補生成 → 専門家判断 → 失敗例の分類 → モデル再評価 → 多言語コーパスへの横展開、という循環をICoMa上で追跡できるようにする。

AI駆動デジタル・ヒューマニティーズへ

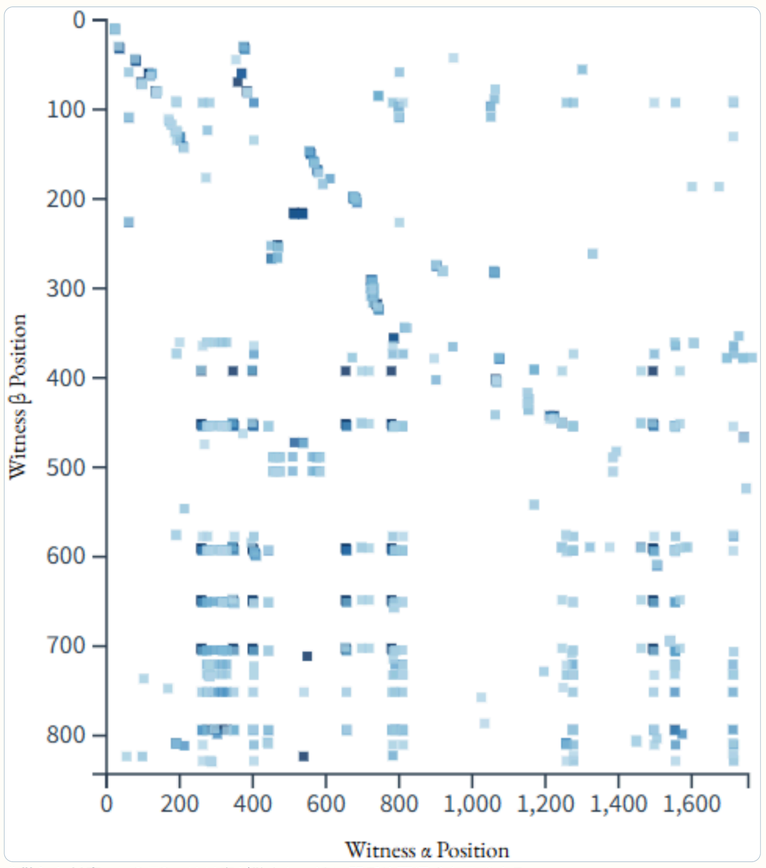
AIを「自動判定器」ではなく、古典文献の読解を拡張する共同作業基盤として位置づける。計算機が候補を広く拾い、人文学者が根拠・文脈・反例を検証する。

探索を高速化し、弱い引喩や遠い再利用を発見する。 | 判断理由をUIに残し、解釈の過程を共有可能にする。 | 多言語・多文字体系の古典資料を比較可能な研究データへ変える。

校合結果をどう読むか



密な対角線：MS-KSでは再利用箇所が連続し、直接的な伝承関係を示す。



分散した対応：MS-TSでは対応が散在し、遠い関係や再編成を示唆する。

結論：候補抽出を、読解可能な根拠へ

本研究は、古代語文献における引用・引喩を、近似検索・意味検索・RAGによって候補化し、最終判断に必要な本文・順位・根拠をICoMaのUIへ戻す。

既存成果
10文字体系対応のICoMaを公開。ヴェーダ3伝承の再利用を多手法で定量化し、ACL 2026内NLP4DHで発表採択。プロシーディングズ論文はACL Anthologyで公開予定。
表層形・見出し語化形・複数アルゴリズムを比較し、文献学的通説を再現可能な数値と可視化に変換した。

本申請
表層一致から、語彙・語順が変化した引用と直接一致しない引喩へ拡張。候補生成・順位づけ・説明を一体化する。
LLM、埋め込み検索、RAG、専門知識ベースを組み合わせ、候補本文・類似根拠・参照知識源を同じUIで提示する。

検証可能性
UI、入力データ、候補リスト、評価ログを対応づけて公開し追跡可能にする。モデル出力を人文学的判断へ戻す。
成功例だけでなく、失敗例・除外例・判断保留例も記録し、専門家評価と再実験に耐える研究基盤として整備する。

結論から次へ：多言語の引用・引喩検出へ広げる

本申請では、ICoMa v2.0をサンスクリット語・コプト語の事例に閉じず、コプト語、古代エジプト語、ギリシア語、ラテン語、古典中国語、古典アラビア語、楔形文字言語、シュメール語、アッカド語、ヒッタイト語、古典シリア語、古典エチオピア語へ広げる。文字体系・時代・ジャンルの異なる資料で実証することで、アルゴリズムが特定言語の癖ではなく、古典文献一般の再利用構造を捉えているかを検証する。

コプト語 | 古代エジプト語 | ギリシア語 | ラテン語 | 古典中国語 | 古典アラビア語 | 楔形文字言語 | シュメール語 | アッカド語 | ヒッタイト語 | 古典シリア語 | 古典エチオピア語

入力	検出	検証
聖典・注釈・翻訳・歴史叙述を、文字体系ごとに正規化して比較可能にする。	表層一致、意味検索、RAGを組み合わせ、直接引用から概念的引喩まで候補化する。	候補本文・根拠・失敗例をICoMa上で確認し、専門家評価と再実験に接続する。

公開・検証する成果物

多言語コーパス入力手順 | 正規化・分割ルール | 引用・引喩候補リスト | 失敗例・除外例ログ | 専門家評価プロトコル | ICoMa v2.0公開UI

評価の出口：候補の再現率、誤検出の型、言語別の得手不得手を比較し、次の多言語DH研究で再利用できる基準データへ整備する。

1. 近接伝承	2. 注釈関係	3. 翻訳関係	4. 広域比較
同一言語・同一ジャンル内の再利用を高精度に検証する。	聖典と注釈、本文と解釈の対応を候補化する。	言語を越えた引用・翻案・概念共有を比較する。	古代地中海・西アジア・南アジアの知の移動を可視化する。

本研究は JSPS 科研費 JP21KK0004 の助成を受けた