Desenvolvimento de um Classificador do Catálogo do Arquivo Histórico Ultramarino: Um Experimento com Processamento de Linguagem Natural e Inteligência Artificial Aplicado a Resumos Arquivísticos

Saulo Rogério Pacheco Rocha

https://doi.org/10.1590/SciELOPreprints.16461

Submetido em: 2026-06-09
Postado em: 2026-06-22 (versão 2)
(AAAA-MM-DD)

Justificativa da versão: Correção de uma imprecisão quantitativa no Resumo e no Abstract. A versão anterior afirmava erroneamente que o corpus era composto por '71.000 resumos', confundindo a dimensão da base de dados bruta original (71.000 linhas de texto) com o escopo do recorte metodológico. O texto foi retificado para esclarecer que o corpus final analisado é composto por 7.051 resumos documentais.

Desenvolvimento de um Classificador do Catálogo do Arquivo Histórico Ultramarino: Um Experimento com Processamento de Linguagem Natural e Inteligência Artificial Aplicado a Resumos Arquivísticos

Development of a Classifier for the Arquivo Histórico Ultramarino Catalog: An Experiment with Natural Language Processing and Artificial Intelligence Applied to Archival Summaries

Saulo Rogério Pacheco Rocha (UFSC)

https://orcid.org/0000-0003-3715-6706

Resumo

Este artigo descreve a arquitetura computacional e metodológica do projeto "Classificador AHU-Sul", voltado à construção de um corpus relacional e semanticamente anotado, composto por aproximadamente 7.051 verbetes de documentos do Arquivo Histórico Ultramarino (AHU) referentes ao Sul e Sudeste do Brasil (1737–1828), extraídos do Projeto Resgate Barão do Rio Branco. Para superar as limitações da busca lexical em massas de dados não estruturados, desenvolveu-se uma *pipeline* em Python que integra técnicas de higienização de metadados, engenharia reversa de códigos arquivísticos (padrão CRAV/DigitArq) e inferência sociolinguística baseada em Grandes Modelos de Linguagem (LLMs). Utilizando a API do modelo DeepSeek v3 sob restrições de zero-shot prompting, a ferramenta avalia os resumos para inferir categorias sociais, vetores de comunicação e a probabilidade de mediação por escrivães. A síntese dessa análise é quantificada no Score de Relevância Sociolinguística Potencial (SRSP), métrica inédita desenvolvida como indicador heurístico para apontar aos pesquisadores os manuscritos com maior propensão a abrigar inovações sintáticas do português brasileiro colonial. O trabalho detalha, ainda, o processo de vetorização semântica dos resumos para a implementação de um motor de busca híbrido (Ensemble Retrieval), que permite consultas tanto por termos lexicais específicos quanto por contextos semânticos amplos. Alinhado aos princípios da Ciência Aberta, o artigo apresenta o conjunto de dados completo, o código-fonte e uma interface interativa disponibilizados publicamente. O objetivo é demonstrar como a aliança entre Humanidades Digitais, Ciência de Dados e Processamento de Linguagem Natural pode otimizar substancialmente a seleção documental para pesquisas em Linguística Diacrônica e de Corpus.

Palavras-chave: Humanidades Digitais; Sociolinguística Histórica; Processamento de Linguagem Natural; Arquivo Histórico Ultramarino; Modelos de Linguagem.

Abstract

This article describes the computational and methodological architecture of the "AHU-South Classifier" project, aimed at constructing a relational and semantically annotated corpus from approximately 7,051 archival summaries from the Arquivo Histórico Ultramarino (AHU) concerning Southern and Southeastern Brazil (1737–1828), extracted from the *Projeto Resgate Barão do Rio Branco*. To overcome the limitations of lexical search in unstructured large-scale datasets, a Python pipeline was developed, integrating metadata cleaning techniques, archival code reverse-engineering (CRAV/DigitArq standard), and sociolinguistic inference based on Large Language Models (LLMs). Leveraging the DeepSeek v3 API under strict zero-shot prompting constraints, the tool evaluates archival summaries to infer social categories, communication vectors, and the likelihood of scribe mediation. This analysis is synthesized into the Score of Potential Sociolinguistic Relevance (SRSP), an unprecedented quantitative metric developed as a heuristic indicator to guide researchers toward manuscripts with a higher propensity for containing syntactic innovations of colonial Brazilian Portuguese. Furthermore, this study details the semantic vectorization process of the summaries to implement a hybrid search engine (Ensemble Retrieval), which enables both specific lexical term queries and broader semantic contextual searches. In alignment with Open Science principles, this work presents the complete dataset, source code, and a publicly available interactive search interface. The final objective is to demonstrate how the alliance between Digital Humanities, Data Science, and Natural Language Processing can substantially optimize documentary selection for research in Diachronic Linguistics and Corpus Linguistics.

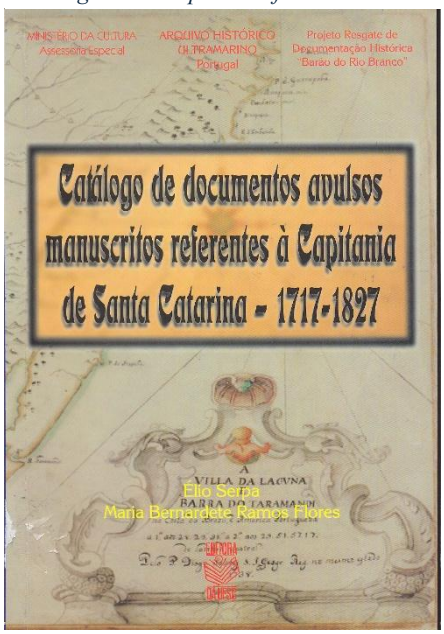
Keywords: Digital Humanities; Historical Sociolinguistics; Natural Language Processing; Arquivo Histórico Ultramarino; Large Language Models.

0. Introdução

A investigação da Mudança e da Variação Linguística no Português Brasileiro, que se sustente nos pressupostos de uma Linguística de Corpus (Mcenery & Hardie, 2012) voltada à realidade empírica da língua depende, inevitavelmente, do acesso às fontes documentais primárias para a composição de amostras. Para a macrorregião sul do Brasil[1], os acervos do Arquivo Histórico Ultramarino (AHU), disponibilizados pelo Projeto Resgate, constituem um repositório essencial da documentação administrativa do período colonial.

A organização do AHU compôs um levantamento exaustivo dos documentos em sua custódia, criando um breve resumo para cada um dos milhares de manuscritos.Esses verbetes, acompanhados de informações básicas (como a região à qual se referem, a cota arquivística e a tipologia documental) formam uma "capa" descritiva que intermedia o acesso ao original. O conjunto desses verbetes e metadados compõe o que é comumente chamado de "catálogo" do AHU.

Figura 1 - Capa do Referido Livro



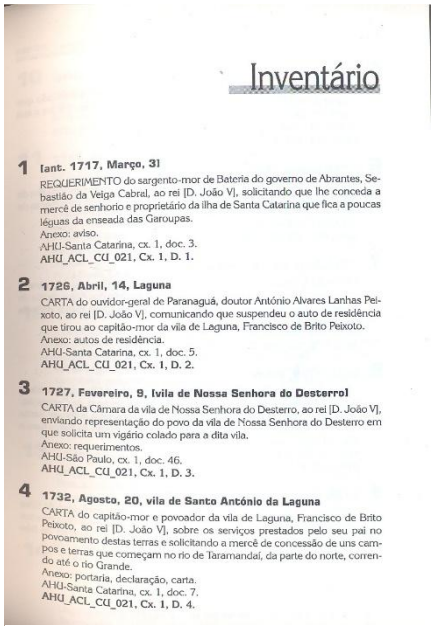
Fonte: Serpa & Ramos Flores (2000)

A partir dessa base, o Projeto Resgate Barão do Rio Branco coordenou, dentre outros projetos, um esforço monumental de publicações, microfilmagem e digitalização para compilar esses verbetes arquivísticos. O objetivo foi democratizar o acesso e divulgar o acervo e as fontes manuscritas referentes à nossa história colonial. Um exemplo desse esforço é a obra *Catálogo de documentos avulsos manuscritos referentes à Capitania de Santa Catarina (1717-1827)*, publicada por Élio Serpa e Maria Bernadete Ramos Flores (2000) pela Editora da UFSC. Nessas obras, publicadas de forma coordenada em todo o Brasil, os pesquisadores passaram acesso ao catálogo do arquivo e, num CD que vinha junto do filme, versões digitais dos fac-símiles microfilmados dos documentos. Atualmente, as versões digitalizadas dos fac-símiles microfilmados estão disponíveis no site institucional do Projeto Resgate/Biblioteca Luso

[1] Definida, para o escopo deste trabalho, com base na catalogação original do Arquivo Histórico Ultramarino. A varredura algorítmica processou um total de 7.051 manuscritos, distribuídos nas seguintes pastas (*folders*): Rio de Janeiro (5.781), São Paulo (375), Rio Grande do Sul (245), Nova Colônia do Sacramento (221), Montevidéu (169), Santa Catarina (157), Brasil Geral (42), Ultramar (28), Buenos Aires (17) e Paraguai (16).

Brasileira[2] e os catálogos foram em lista num endereço institucional do Governo Federal/Fundação Biblioteca Nacional[3].

Figura 2 Primeira página da descrição do inventário do referido Catálogo de SC



Inventário

1 **[ant. 1717, Março, 31]**
REQUERIMENTO do sargento-mor de Bateria do governo de Abrantes, Sebastião da Veiga Cabral, ao rei [D. João V], solicitando que lhe conceda a mercê de senhorio e proprietário da ilha de Santa Catarina que fica a poucas léguas da enseada das Garoupas.
Anexo: aviso.
AHU-Santa Catarina, cx. 1, doc. 3.
AHU_ACL_CU_021, Cx. 1, D. 1.

2 **1726, Abril, 14, Laguna**
CARTA do ouvidor-geral de Paranaguá, doutor António Alvares Lanhas Peixoto, ao rei [D. João V], comunicando que suspendeu o auto de residência que tirou ao capitão-mor da vila de Laguna, Francisco de Brito Peixoto.
Anexo: autos de residência.
AHU-Santa Catarina, cx. 1, doc. 5.
AHU_ACL_CU_021, Cx. 1, D. 2.

3 **1727, Fevereiro, 9, [vila de Nossa Senhora do Desterro]**
CARTA da Câmara da vila de Nossa Senhora do Desterro, ao rei [D. João V], enviando representação do povo da vila de Nossa Senhora do Desterro em que solicita um vigário colado para a dita vila.
Anexo: requerimentos.
AHU-São Paulo, cx. 1, doc. 46.
AHU_ACL_CU_021, Cx. 1, D. 3.

4 **1732, Agosto, 20, vila de Santo António da Laguna**
CARTA do capitão-mor e povoador da vila de Laguna, Francisco de Brito Peixoto, ao rei [D. João V], sobre os serviços prestados pelo seu pai no povoamento destas terras e solicitando a mercê de concessão de uns campos e terras que começam no rio de Taramandaí, da parte do norte, correndo até o rio Grande.
Anexo: portaria, declaração, carta.
AHU-Santa Catarina, cx. 1, doc. 7.
AHU_ACL_CU_021, Cx. 1, D. 4.

Fonte: Serpa & Ramos Flores (2000)

Embora os catálogos impressos venham servindo fundamentalmente aos pesquisadores, o acesso a tais descrições arquivísticas pode ser otimizado para potencializar a exploração de seu volume massivo de dados. O pré-processamento computacional e a implementação de mecanismos de busca mais sofisticados facilitam o acesso a uma gama mais ampla de documentos, rompendo as barreiras impostas pela consulta manual, livro a livro. Ademais, é preciso contornar as limitações dos motores de busca lexicais convencionais, que, por se restringirem a correspondências exatas de caracteres, se mostram insuficientes para mapear os conceitos semânticos complexos e as dinâmicas sociolinguísticas subjacentes presentes nas descrições.

Para lidar com esse volume documental de forma analítica, este projeto se baseou no paradigma da 'Leitura Distante' (*Distant Reading*), proposto por Franco Moretti (2013). Em oposição à tradicional 'leitura atenta' (*close reading*) (Moretti, 2013, p. 38-39), que se preocupa em de fato analisar o objeto textual e as minúcias de um manuscrito, a 'leitura distante' propõe o distanciamento intencional do texto individual para permitir a observação de um sistema inteiro. Este projeto, portanto, ao converter arquivos massivos em dados quantificáveis, buscou tornar possível enxergar a "topografia" do acervo, mapeando padrões de poder, repetições formulaicas e dinâmicas sociais que seriam invisíveis à escala do pesquisador humano (Moretti, 2013, p. 59) para após isso, guiar a pesquisa profunda aos documentos interessantes. No contexto do AHU, a aplicação computacional desse conceito não visa substituir a crítica paleográfica, diplomática ou Linguística Histórica, mas sim atuar como uma sondagem metodológica; uma varredura semântica capaz de processar os resumos para apontar, estatisticamente, onde residem os focos de interesse para a pesquisa linguística.

Se acrescenta a isso novas complexidades impostas pelo estudo da linguística histórica, que, diferentemente de outras áreas ocupadas com o conteúdo dos documentos ou

[2] Disponível em <https://resgate.bn.gov.br/docreader/docmulti.aspx?bib=resgate> acesso em 27 de maio de 2026.
[3] Disponível em < https://www.gov.br/bn/pt-br/central-de-conteudos/projeto-resgate/novos-instrumentos-de-pesquisa> acesso em 27 de maio de 2026.

com a crítica documental, precisa se preocupar com as categorias sociais dos autores que condicionam a produção linguística registrada no manuscrito. A detecção da emergência de traços linguísticos inovadores, especialmente sintáticos, esbarra no peso das Tradições Discursivas (TD) (Kabatek, 2006). Uma vez que as formulações diplomáticas e a norma metropolitana atuam como moldes conservadores que mascaram e atrasam o registro da mudança linguística, exige-se o uso de ferramentas analíticas capazes de cruzar dimensões históricas, sociais e textuais. Assim, os pesquisadores da área necessitam isolar documentos não apenas por localidade e data, mas por parâmetros finos como a classe social do remetente, a hierarquia da comunicação (o vetor direcional.

Ademais, se torna imprescindível detectar, ou no mínimo levar em consideração certa dubiedade, a provável presença ou ausência da mediação de um escritor na produção do documento. Conforme discutem Lose e Souza (2020), as edições de documentos históricos frequentemente falham em apontar as mãos que tecem o manuscrito, sem se atentar a diferença entre o sujeito intelectual que dita ou manda fazer a cópia e a pessoa que de fato risca o papel com a pena, chamado pelas autoras de *scriptor*. A identificação da distinção entre os *scriptores* e os autores intelectuais ou quem assina o documento é determinante, pois o linguista deve considerar cada mão como uma variável social e linguística. A atuação desse intermediário afeta diretamente o grau de monitoramento estilístico do texto, já que cada *scriptor* apresenta características peculiares que vão muito além da caligrafia, são sujeitos distintos que agem sobre a produção do texto.

Para superar essa barreira, apresento aqui a arquitetura computacional e metodológica que desenvolvi para projeto Classificador do Catálogo AHU-Sul. Trata-se da construção de uma esteira de processamento (*pipeline*) de extração, anotação semântica e indexação vetorial que converteu aproximadamente 71.000 linhas de verbetes arquivísticos não estruturados do AHU (extraídos da organização do Projeto Resgate) num banco de dados relacional (JSON) legível por máquina.

A inovação central desta proposta reside na integração experimental de Inteligência Artificial generativa, especificamente a API do modelo DeepSeek v3-chat. A interação com o modelo se deu buscando *guard rails* que ampliassem a confiabilidade no modelo, para atuar estritamente como um motor de inferência algorítmica focado em categorizar os remetentes, analisar o vetor de comunicação (e.g., *Bottom-Up*, *Top-Down*) e avaliar a probabilidade de interferência de escrivães locais. O produto final dessa inferência é a criação de uma métrica quantitativa, batizada de *Score de Relevância Sociolinguística Potencial (SRSP),* pensada para sinalizar heuristicamente aos linguistas quais cotas arquivísticas abrigam a maior probabilidade

de conter inovações sintáticas do português brasileiro na fonte primária. Além da geração estruturada desses dados, detalho neste trabalho o desenvolvimento de uma interface *web* de recuperação de informação com motores de busca próprios. Esta aplicação funde a tokenização lexical clássica (algoritmo BM25) com *embeddings* semânticos vetoriais, otimizando a triagem de documentos e facilitando a seleção e acesso dos manuscritos digitais hospedados nas plataformas contemporâneas do DigitArq e no projeto Resgate.

Ao detalhar as fases de processamento, sanitização estrutural e auditoria algorítmica, busco fornecer transparência metodológica sobre a construção deste novo *corpus*. O objetivo central é demonstrar como este experimento de triagem documental, ancorado no cruzamento entre as ferramentas das Humanidades Digitais e o rigor teórico da Linguística Histórica, pode transformar arquivos textuais outrora inertes em instrumentos consultáveis, dinâmicos e focados para a pesquisa em Linguística Histórica.

1. Tratamento de Dados: Extração e Arquitetura do Corpus Base

A arquitetura deste projeto foi desenvolvida em Python, concebida para atuar como uma esteira de processamento (*pipeline)*, operando pontualmente nos dados não estruturados e criando um banco de dados em formato JSON (*JavaScript Object Notation*). O ponto de partida metodológico consistiu num arquivo de texto bruto contendo aproximadamente 71.000 linhas extraídas da Central de Conteúdos do Projeto Resgate Barão do Rio Branco (Brasil, 2026). Este arquivo original continha as informações básicas e verbetes arquivísticos do Arquivo Histórico Ultramarino (AHU) referentes ao Brasil colonial, materiais estes compilados e disponibilizados ao público pelos esforços de indexação do Projeto Resgate. Cabe lembrar que, devido a uma limitação de viabilidade, me restringi a tratar os dados apenas da Região Sul do Brasil e as regiões circundantes, objetos de minha pesquisa, isto é: Santa Catarina, Rio Grande do Sul, Colônia do Sacramento e Montevidéu; São Paulo e Rio de Janeiro; Buenos Aires, Paraguai, Ultramar e Brasil Geral. Num cenário com mais tempo, recursos e colaboração, o objetivo seria ampliar o escopo deste projeto a todo o catálogo disponível e incluir as demais regiões do Império Colonial Português, especialmente a região Nordeste do Brasil.

O desafio dessa primeira etapa do projeto residia na alta entropia do formato textual original que, por mais organizado que fosse, era pouco legível por máquina. Para solucionar isso, planejei um sistema de operações para serializar a informação do AHU num único *array*[4]

[4] Um array (arranjo ou vetor) é uma estrutura de dados que armazena uma coleção de elementos organizados em formato de lista. Ele funciona como um grande "fichário catalográfico digital" unificado: trata-se de uma única lista sequencial e contínua, delimitada por colchetes [], na qual cada entrada é um documento individual. Cada um

de objetos no formato JSON (*JavaScript Object Notation)*. A escolha deste formato se justifica pela sua capacidade de suportar informações complexas e hierárquicas de maneira simples e pela possibilidade de aplicação duma tipagem estrita de dados (além das *strings* para as informações, utilizei valores booleanos para metadados lógicos). O objetivo primário foi alcançar uma estrutura na qual cada documento correspondesse a um objeto composto pelas mesmas chaves, cujo valor poderia ser acessado e manipulado automaticamente.

A primeira fase da operação de tratamento consistiu no *parsing* dos metadados básicos para a construção do JSON. O script iterou sequencialmente pelas linhas do documento de texto original, utilizando a biblioteca nativa de expressões regulares do Python (re) como mecanismo principal de ancoragem, padronização e extração de informações. O primeiro padrão de busca focou na ancoragem ####[5], que deixei no TXT para separar cada caixa/região do AHU, denotando a transição de informações estruturais na indexação original do arquivo. Após a criação da chave *folder*, outras expressões regulares de mesma natureza mapeavam e extraíam informações de maneira semelhante: o prefixo AHU_ACL para capturar os códigos de referência arquivística únicos para a chave *new_code*; palavras com três ou mais letras maiúsculas seguidas no início dos parágrafos indicavam a tipologia a ser extraída (para a chave *extracted_typology*); e o conteúdo concentrado numa linha contínua longa indicava a descrição (chave *description*). Essa extração sequencial transformou a massa textual bruta num objeto JSON minimamente organizado, consolidando a base estrutural sobre a qual as camadas analíticas de linguagem natural seriam construídas.

Durante esse processamento inicial, a formatação original dos dados apresentou sérios desafios em relação à temporalidade, exigindo intervenções em duas frentes. O primeiro foi o fenômeno das "datas fraturadas". Em centenas de entradas, o campo formal de datação do catálogo constava como "[Sem datação]" ou "S. d.", mas a informação cronológica real (ou um indício dela) havia sido fundida indevidamente ao início do texto de descrição do manuscrito durante a classificação analógica. Sabendo que as descrições arquivísticas do AHU tradicionalmente iniciam com a tipologia documental grafada em letras maiúsculas, se

desses documentos é chamado de "objeto" e fica delimitado por chaves {}, abrigando suas próprias 'gavetas' de informação (como data, remetente e tipologia). É essa estruturação que permite ao computador ler e processar os milhares de resumos arquivísticos de forma previsível e automatizada.
Ex.: Array 1 = [**Objeto 1 =** {"folder": "Montevidéu", "reference_code": "PT/AHU/CU/065/0001/00002", "extracted_typology": "CARTA", "document_id_and_date": "7964. 1787, março, 28, Montevidéu"}, **Objeto 2 =** {"folder": "Santa Catarina", "reference_code": "PT/AHU/CU/021/0001/00004", "extracted_typology": "CARTA", "document_id_and_date": "7255. 1732, Agosto, 20, vila de Santo Antônio da Laguna"}]

[5] Marcadores que inseri manualmente entre os registros extraídos do site do Projeto Resgate. Esses delimitadores serviram para identificar e agrupar as unidades por capitanias; por exemplo, o marcador #### separava a documentação de Santa Catarina da seção referente ao Rio Grande do Sul.

implementou uma expressão regular um pouco mais complexa, capaz de localizar a primeira sequência de três ou mais letras em caixa alta. Esse algoritmo avaliava o índice de posição dessa palavra: se houvesse texto precedendo a tipologia em *allcaps*, esse trecho inicial era recortado da descrição, colado separadamente duas linhas acima no texto base (a *Ground Truth* extraída do Projeto Resgate) e inserido como conteúdo da chave *document_id_and_date* do documento correspondente.

No entanto, a estruturação esbarrou em um segundo fenômeno de fragmentação temporal: os metadados órfãos. Como o acervo passou por múltiplos esforços de catalogação, muitas linhas contendo as identificações cronológicas primárias estavam estruturalmente separadas no arquivo bruto de seus respectivos blocos documentais. A situação se tornava ainda mais complexa ao constatar que a base lidava simultaneamente com três camadas evolutivas de cotas arquivísticas para o mesmo manuscrito: o código físico legado ("*old_code":* "AHU-Santa Catarina, cx. 8, doc. 14."), a cota usada pelo Projeto Resgate ("*new_code"*: "AHU_ACL_CU_021, Cx. 6, D. 390.") e a necessidade futura de estabelecer o padrão contemporâneo português ("*reference_code*": "PT/AHU/CU/021/0006/00390"). Para resgatar essas datas órfãs e unificá-las à chave principal, se desenvolveu outro script de cruzamento de dados ancorado na lógica de *Left-Join*. O algoritmo utilizou a cota intermediária (new_code) como uma 'âncora' para buscar a data correspondente no arquivo *Ground Truth* e uni-la novamente ao banco de dados JSON. Para mitigar o risco de associar a data à cota errada devido a inconsistências de digitação da época, o cruzamento adotou uma "Auditoria Indulgente" (*Forgiving Auditor*), aplicando uma função em Python (expressa sintaticamente como .strip().rstrip('. ,;').lower()) para eliminar espaços invisíveis e pontuações acidentais das chaves antes da comparação, garantindo o reengate preciso.

Uma vez que a tipologia documental exata (a primeira palavra em ALLCAPS) foi isolada, o desafio seguinte foi organizar a dispersão causada pela flexão de número das palavras. A língua portuguesa, como bem sabemos, admite múltiplas formas de pluralização, o que fazia com que, quando se extraía uma tipologia, documentos idênticos fossem classificados em categorias distintas por causa da variação morfológica ou ortográfica, (e.g.: "CARTA" e "CARTAS"). Para viabilizar o agrupamento quantitativo das tipologias, se aplicou um processo de Normalização Morfológica (*Custom Stemming*). O código foi instruído a substituir terminações "-ÕES" por "-ÃO", remover "-ES" de terminações "-RES" e eliminar o "-S" de maneira controlada (exigindo que a palavra resultante mantivesse um comprimento mínimo para evitar a corrupção de vocábulos curtos).

Como se pode imaginar, a normalização morfológica forçada a partir dessas regras puramente de grafia criou algumas monstruosidades (o plural "ORDENS" se tornando a anomalia "ORDEN", por exemplo), além de não agir sobre problemas decorrentes de ortografias arcaicas, como "TRESLADO" em vez de "TRASLADO". A solução aplicada foi a comparação das tipologias extraídas com uma lista teórica de controle, baseada na obra de Bellotto (2002), a partir de um motor de Lógica *Fuzzy*. Para os termos que não possuíam correspondência exata, o *script* invocava (por meio da a biblioteca difflib do Python**)** um algoritmo de distância de Levenshtein para calcular a similaridade matemática entre a palavra anômala e as categorias válidas, estabelecendo um limiar de corte (*cutoff*) de 80%. Caso a correspondência atingisse esse índice, o sistema realizava o "encaixe" (*snapping*) automático da palavra errônea para a sua forma teórica correta.

Contudo, a automação de correções ortográficas sobre o vocabulário histórico carrega perigos para a confiabilidade dos dados finais. Para garantir a integridade científica e evitar a corrupção de termos que fossem escolhas deliberadas dos arquivistas originais, o processo de normalização evitou rotinas de alteração cega. Esse processo culminou na criação de uma nova chave no JSON, denominada *typology_status*. O valor booleano desta chave controla a normalização da tipologia: as extrações que correspondiam em pelo menos 80% à lista de Bellotto recebiam o *status* de *true*, eram validadas manualmente por mim, e submetidas à normalização morfológica. Em contrapartida, os vocábulos não reconhecidos pela obra recebiam o *status* de *false* e eram isolados do algoritmo de correção, sendo preservados na base de dados exatamente como constavam no catálogo original. Esse rigor impediu que o sistema corrigisse indevidamente termos arcaicos ou descrições atípicas que fugiam ao padrão diplomático estrito.

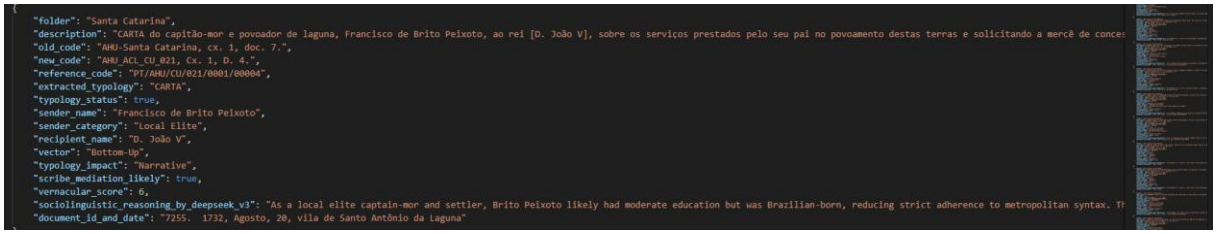
Ultrapassada a coesão interna dos dados, a consolidação estrutural da base exigiu uma intervenção de interoperabilidade externa. Os identificadores originais utilizados pelo Projeto Resgate ("*new_code"*: "AHU_ACL_CU_021, Cx. 6, D. 390."), embora úteis para o mapeamento e localização analógica, refletem um sistema arquivístico físico herdado, pautado na numeração de Caixas, Maços e Documentos. Para que a base de dados pudesse se comunicar com a infraestrutura digital do Arquivo Nacional da Torre do Tombo e de sua plataforma atual, o sistema CRAV (DigitArq), foi necessária, a aplicação de uma engenharia reversa sobre os códigos legados. Por meio de um novo bloco de expressões regulares, um novo *script* identificou e particionou as antigas cotas, isolando matematicamente os numerais referentes à série, à subsérie, ao número da caixa e à numeração sequencial do documento. Em seguida, submeteu-se esses componentes a uma operação algorítmica de preenchimento de zeros à

esquerda. Ao forçar que numerais isolados atinjam um comprimento métrico padronizado (transformando, por exemplo, o documento "390" em "00390", e o "*new_code"* em "reference_code": "PT/AHU/CU/021/0006/00390"), o algoritmo reconstruiu a cota exatamente como ela é compreendida pelos servidores de Portugal.

Ao conseguir criar os códigos de referência atuais de Portugal a partir das cotas utilizadas pelo Projeto Resgate foi possível converter dinamicamente os códigos de referências dos documentos do Catálogo em URIs (*Uniform Resource Identifiers*) de busca do sistema CRAV, tornou-se possível o roteamento de URLs. Assim, como descreverei nas próximas seções deste texto, foi possível criar a URL de busca "https://digitarq.arquivos.pt/search?query=PT%2FAHU%2FCU%2F021%2F0006%2F00390&isAdvancedSearch=false" a partir da cota "PT/AHU/CU/021/0006/00390" para cada um dos documentos organizados neste projeto, e dessa forma integrar este classificador diretamente ao sistema de controle de documentos do Arquivo Nacional da Torre do Tombo.

O arquivo JSON final contém, dentre outras informações:

Figura 3 Screenshot da interface do Antigravity mostrando o array correspondente ao documento PT/AHU/CU/021/0001/00004



Fonte: Elaborado pelo autor com base no conteúdo do Arquivo Histórico Ultramarino intermediado pelo Projeto Resgate.

2. Anotação Sociolinguística Automatizada e Inferência Baseada em LLM

Após a higienização estrutural, este projeto passou para sua fase analítica e de enriquecimento desses dados. A premissa central desta etapa é potencializar a busca dos pesquisadores pelo "vazamento vernáculo", conceito basilar para o estudo do português brasileiro histórico (Tarallo, 1990). Contudo, localizar a emergência da sintaxe coloquial impõe um paradoxo estrutural: o "vazamento" ocorre de forma microscópica e acidental na pena de *scriptores* menos letrados, mas o pesquisador dispõe de um oceano macroscópico de dezenas de milhares de manuscritos para investigar. É aqui que, novamente, se valendo da teoria da Leitura Distante (Moretti, 2013) se entrelaça aos postulados sociolinguísticos de Tarallo. Partindo da hipótese de que documentos produzidos por indivíduos sócio-geograficamente periféricos são objetos preferenciais para a quebra da barreira da norma metropolitana, esta etapa não busca o traço sintático em si, nem sequer tem acesso aos dados originais, mas as

condições sociais que o propiciam. Uma abordagem baseada puramente em correspondência lexical é incapaz de capturar as nuances de poder e letramento implícitas nos verbetes. Portanto, para operacionalizar a busca pelo vazamento vernáculo em escala arquivística, foi necessária a transição para uma análise semântica estruturada e distanciada.

Para operacionalizar essa tarefa em larga escala, processando todos os 7.051 documentos, integrei ao sistema, via API, o modelo DeepSeek (versão 3), um LLM que assumiu o papel de "motor de inferência". A arquitetura de interação com a IA foi estritamente confinada ao *zero-shot prompting* (comando sem exemplos prévios). Conforme descrevem Brown *et al.* (2020), o uso de *zero-shot* é, paradoxalmente, o método mais desafiador para a máquina, pois exige que ela compreenda a complexidade da tarefa se baseando exclusivamente na clareza das instruções. Contudo, a adoção dessa abordagem foi uma decisão metodológica e pragmática: a inclusão de exemplos metodológicos detalhados (*few-shot*) na chamada da API para cada milhares de resumos multiplicaria exponencialmente o consumo de *tokens* (a unidade de memória e custo computacional dos LLMs), o que tornaria o processamento do *corpus* inteiro financeiramente e computacionalmente inviável; e utilizar exemplos pequenos para todos os resultados tornaria o *output* enviesado, além de potencialmente fazer com que o modelo criasse relações espúrias entre os exemplos e casos extremos no *input*.

Apesar de ser o cenário mais exigente, o zero-shot apresenta a vantagem de evitar correlações espúrias entre exemplos dados e o texto de entrada, como descrevem os autores:

> Zero-Shot (0S) is the same as one-shot except that no demonstrations are allowed, and the model is only given a natural language instruction describing the task. This method provides maximum convenience, potential for robustness, and avoidance of spurious correlations (unless they occur very broadly across the large corpus of pre-training data), but is also the most challenging setting. In some cases it may even be difficult for humans to understand the format of the task without prior examples, so this setting is in some cases "unfairly hard". For example, if someone is asked to "make a table of world records for the 200m dash", this request can be ambiguous, as it may not be clear exactly what format the table should have or what should be included (and even with careful clarification, understanding precisely what is desired can be difficult). Nevertheless, for at least some settings zero-shot is closest to how humans perform tasks (Brown et al., 2020, p. 7)

Para mitigar a instabilidade inerente a esse cenário desafiador e garantir o rigor analítico, o modelo não operou de maneira conversacional ou generativa livre. Em vez disso, a parametrização do LLM foi contida por restrições inflexíveis: todos os verbetes foram processados segundo os mesmos parâmetros teóricos descritos num único *prompt*. Acima de tudo, para compensar a ausência de exemplos práticos, a exigência de uma saída de dados tipada atuou como um *guard rail* (mecanismo de contenção) estrutural vital. Instruí o modelo a ler

cada descrição arquivística e retornar sua análise exclusivamente através de geração restrita[6] dentro do formato JSON, preenchendo as chaves exatas exigidas pelo banco de dados. Essa formatação forçada impediu alucinações narrativas e garantiu que os *outputs* fossem reproduzíveis, padronizados e prontos para a indexação no motor de busca.

No interior dessa rotina de extração, a Inteligência Artificial foi instruída a avaliar as descrições textuais sob dimensões teóricas específicas. Primeiramente, o modelo realizava a classificação dos atores sociais envolvidos, categorizando o remetente e o destinatário do documento em estratos históricos definidos: Elite Metropolitana, Elite Local, Baixa Patente Militar, Plebeu, Marginalizado ou Clero. A partir dessa taxonomia, o algoritmo inferia um Vetor de Comunicação, estabelecendo a direcionalidade do poder, classificando a correspondência como *Bottom-Up* (da periferia colonial para o centro administrativo), *Top-Down* (da Coroa ou governadores colonais para os súditos) ou Horizontal (entre pares).

Em paralelo à dinâmica de poder, exigi que o modelo avaliasse a materialidade do texto a partir de uma inferência do Impacto da Tipologia. Esta etapa inferia o grau de monitoramento gramatical esperado para a espécie documental, distinguindo entre textos engessados por fórmulas da chancelaria (alvarás, provisões), narrativas semi-espontâneas (petições, requerimentos, cartas) e transcrições com traços de oralidade (devassas, interrogatórios).

O cruzamento dessas variáveis sociais e documentais permitiu a execução de uma das inferências centrais do classificador: a detecção da probabilidade de um s*criptor* intermediário (*scribe_mediation_likely*). Se apoiando na distinção filológica entre a atribuição de autoria intelectual e a realidade do *scriptor*, conforme destacada por Lose e Souza (2020), o LLM foi instruído a avaliar de forma booleana se a classe ou tipo social do remetente (indivíduos escravizados, indígenas, extratos populares ou militares de baixa patente) e o gênero textual do documento, historicamente, aumentaria a probabilidade da figura de um escrivão intermediando a escrita da carta. Essa inferência, contudo, é metodologicamente delicada, por mais que seja uma informação fundamental para a linguística diacrônica (uma vez que a escolaridade variável desses notários coloniais atuava como um filtro permeável, ora favorecendo a hipercorreção, ora permitindo o escape indeliberado da sintaxe local), a sua classificação de forma puramente heurística é perigosa. Não é possível atestar tal mediação de forma categórica sem um extenso estudo paleográfico e diplomático do manuscrito original.

[6] Prática conhecida como “Constrained Generation”, para mais informações leia Docherty (2025) disponível em <https://medium.com/@docherty/controlling-your-llm-deep-dive-into-constrained-generation-1e561c736a20> Acesso em 28/05/2026.

Por esse motivo, optei por restringir o uso dessa variável, removendo ela dos filtros do mecanismo de busca interativo. A avaliação algorítmica (que traduz o valor booleano *False* para "Pouco provável" e *True* para "Provável") foi disponibilizada ao usuário apenas no formato de dossiê exportável em PDF, servindo estritamente como um apontamento complementar para a avaliação qualitativa do pesquisador.

A síntese de todas essas extrações feitas pelo LLM culminou na métrica principal desta fase do projeto: o *Score de Relevância Sociolinguística Potencial* (SRSP). Ele se trata de uma variável quantitativa, parametrizada numa escala decimal de 0.0 a 1.0 (ou 0 a 10), que consolida o peso dessas inferências sociolinguísticas. O SRSP, aferido pela IA, pretende ser um índice da relevância sociolinguística potencial (por isso o nome) e traduzir numericamente a probabilidade de o documento físico original abrigar quaisquer características de inovação sintática próprias do português brasileiro. A ideia por trás dessa quantificação absoluta é criar um filtro único que permita o pesquisador remover os documentos mais provavelmente formulaicos do quadro de respostas e atingir os documentos que tem mais potencial de cumprir seu interesse sociolinguístico. Como um documento nunca será "completamente desprovido" de relevância sociolinguística, nem completamente relevante sociolinguisticamente, os valores totais de 0 e 10 nunca são utilizados em documento algum, a escala, na verdade, ocorre entre 1 e 9 organizada em em 3 níveis (que por si são divisíveis em 3 subníveis), entre 1 e 3 estão os documentos mais formulaicos e com menor SRSP, entre 4 e 6 estão os documentos intermediários, moderados, e entre 7 e 9 os documentos com maiores potencialidades de relevância sociolinguística.

Por fim, reconhecendo o risco de processos de inteligência artificial atuarem como "caixas-pretas" metodológicas, estabeleci uma pequena medida de rastreabilidade. O modelo, além de atribuir a nota numérica de SRSP, também foi obrigado a gerar o campo *sociolinguistic_reasoning_by_deepseek_v3*. Nesta chave, a IA redigia uma breve justificativa expondo o raciocínio linguístico e histórico que a levou a cruzar os dados do scriptor, do vetor e da tipologia para calcular aquele *score* específico. Essa abordagem de auditoria garantiu que cada decisão algorítmica tomada pelo classificador pudesse ser, pelo menos, qualitativamente revisada e validada pelo pesquisador. O prompt foi o seguinte:

> You are an expert in Historical Sociolinguistics, Paleography, and the Diachronic Syntax of Colonial Brazilian Portuguese. You are analyzing archival descriptions from the Southern Brazil region (Santa Catarina, Rio Grande do Sul, Uruguay, cities and regions

like Ilha de Santa Catarina, Laguna, Rio Grande, rivers like 'Araranguá', 'Tramandaí', 'Tubarão', etc.) between 1737 and 1828.
Your task is to extract historical metadata and assign a 'vernacular_score' (1 to 10). This score represents the probability that the ORIGINAL physical document contains colloquial syntax, linguistic innovation (like Brazilian clitic placement), or non-standard grammar, avoiding rigid metropolitan formulaicism.
EVALUATE THE FOLLOWING DIMENSIONS TO CALCULATE THE SCORE:
1. Typology & Degree of Monitoring:
- HIGHEST (8-10): Devassas, Testemunhos, Interrogatórios (Transcripts of semi-spontaneous speech with low grammatical monitoring).
- MODERATE (4-7): Requerimentos, Petições, Cartas Particulares (Highly formulaic openings/closings, but the central narrative often leaks vernacular syntax).
- LOWEST (1-3): Provisões, Decretos, Alvarás, Consultas, ofícios militares de alta patente (Rigid, heavily monitored Coimbra-standard syntax).

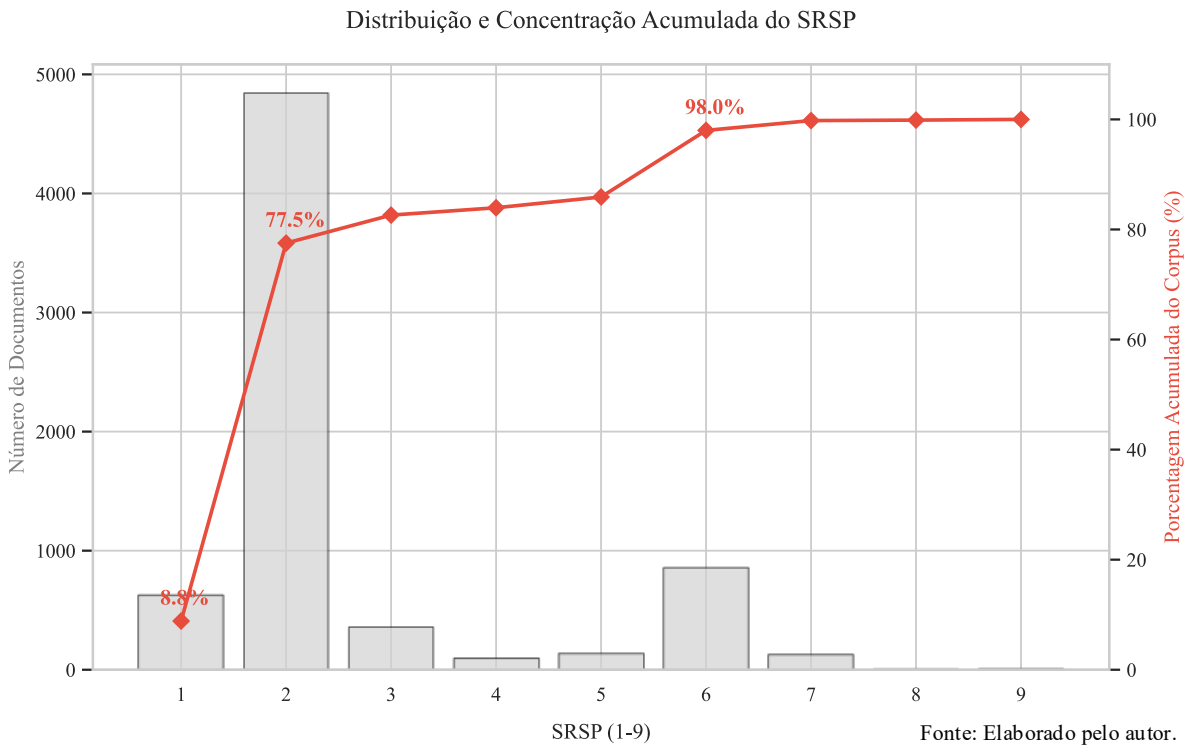
2. Social Rank, Gender & The "Scribe Filter" (scriptor/Amanuense):
- Marginalized actors (índios, escravizados, pardos) and commoners could not write. Their petitions were dictated to local scribes. This "Scribe Filter" results in a HIGH score (7-9) due to the local scribe's hypercorrection or failure to master metropolitan syntax.
- Women generally lacked formal grammatical schooling. Intersect gender with class: A poor widow scores VERY HIGH (8-10); an elite woman scores MODERATE (5-6).
- Local Elites (Capitães-mores, vereadores) score MODERATE (often Brazilian-born, less formal education than European peers).
- Metropolitan Elites (Ouvidores, Governadores, Bispos) score LOWEST (1-2) due to strict adherence to classical grammar.

3. Communication Vector & Geography:
- Bottom-Up (Periphery to Metropole): HIGH probability of vernacular features.
- Deep Periphery (Frontier camps, sertão, aldeias) scores higher than administrative capitals (Desterro, Rio Grande city).
- Top-Down (Metropole to Periphery): LOW probability.

You MUST respond strictly with a raw JSON object exactly in this format:
{
"sender_name": "Name or Title",
"sender_category": "Metropolitan Elite / Local Elite / Low Military / Commoner / Marginalized / Clergy / Unknown","recipient_name": "Name or Title",
"vector": "Bottom-Up / Top-Down / Horizontal / Unknown",
"typology_impact": "Highly Formulaic / Narrative / Speech Transcript",
"scribe_mediation_likely": true or false,
"vernacular_score": 8,
"sociolinguistic_reasoning_by_deepseek_v3": "A highly specific 3-sentence linguistic explanation evaluating the intersection of typology, the scribe filter, and the communication vector."
}

O total de classificações do modelo pode ser interpretado no gráfico 1 a seguir:

Gráfico 1



Fonte: Elaborado pelo autor.

A distribuição integral do *corpus* também oferece uma noção condizente com a realidade do acervo do AHU. A concentração de 68,67% da documentação no *Score* 2 espelha a natureza predominantemente burocrática da documentação administrativa colonial armazenada no AHU. Os dados indicam que o algoritmo foi capaz de mapear essa volumosa massa documental de "ruído administrativo", focada na norma metropolitana da chancelaria, reservando os índices mais altos para a minoria de documentos que foge ao padrão formulaico. Um caminho possível em relação a essa imensa concentração na categoria Score 2 seria tornar o prompt mais sensível a esse tipo de documento burocrático e ampliar o range classificatório dos documentos mais formulaicos nesses primeiros 3 scores (1, 2 e 3). Mas isso sacrificaria a compatibilidade deste método a arquivos de outra natureza que não administrativa, por isso decidi manter este método e aceitar a concentração natural dos documentos do AHU nos Scores mais baixos.

2.1. Validação Empírica das Inferências do LLM

Com o intuito de aferir a consistência interna da métrica gerada e mitigar os riscos de a inteligência artificial operar em descolamento da realidade arquivística, eu conduzi um experimento de validação empírica, já que a adoção de metodologias de anotação massiva baseadas em LLMs demanda etapas de aferição humana para que seus resultados possam ser

considerados cientificamente defensáveis. Cabe ressaltar que, devido ao escopo individual desta pesquisa, a classificação de controle da amostra foi realizada de forma unipessoal, servindo como uma prova de conceito para futuras auditorias com múltiplos anotadores. Para tanto, se desenvolveu um protocolo e um programa de Auditoria de Consistência Semântica.

Esse procedimento consistiu na extração de uma amostra aleatória correspondente a 10% do *corpus* tratado (705 documentos[7]), utilizando uma semente de randomização fixa (*seed*) para resguardar a reprodutibilidade metodológica do experimento[8]. Para evitar o viés de ancoragem durante a avaliação, se desenvolveu um *script* em Python operando sob uma interface cega: ela exibia exclusivamente a descrição original, a cota do documento e o ID com data, ocultando a inferência previamente atribuída pela máquina. A partir dessa leitura isolada, se atribuiu manualmente um *score* humano (de 1 a 9) a cada verbete, se baseando nas mesmas três dimensões teóricas fornecidas no *prompt* da inteligência artificial e, devo admitir, no meu viés subconsciente, dentre outras coisas que naturalmente compõem o julgamento humano. O *script* também registrou o tempo de reação cognitiva (em segundos) dispendido em cada avaliação. A Tabela 1 sintetiza as métricas de concordância extraídas dessa auditoria[9]:

Tabela 1: Resumo da Auditoria de Consistência e Distribuição do Corpus

Métricas de Precisão (10% do toral, N =705)	**Valores**
Erro Médio Absoluto[10]	1.13
Correlação de Pearson[11]	0.48
Viés Sistemático[12]	+0.59 p/ IA

[7] Para assegurar a confiabilidade dos scores humanos, foram selecionados na verdade 10% + 10 documentos, os 10 primeiros documentos a receberem score humano (meu) foram descartados nos cálculos finais comparativos, para evitar o viés enquanto me acostumava com o protocolo.

[8] Como se pode conferir no código auditool.py, a linha 108 "random.seed(42)" escolhe o "caminho de escolha aleatória" dos documentos a serem auditados, de forma que outro experimento replicar os resultados obtidos e se mantenha a natureza aleatória da amostragem de documentos.

[9] Todos os arquivos finais da auditoria foram inclusos na documentação dentro da pasta "auditoria".

[10] Métricas de precisão residual, como o Erro Médio Absoluto (e o Desvio Absoluto Médio), são recomendadas por Bruce et al., (2019, p. 40-41) para quantificar a dispersão de erros de forma linear e sem contra a influência de *outliers*. Cheguei a este resultado, como pode ser averiguado no código calculista.py na pasta de Auditoria do dataset, utilizando a combinação dos comandos de valor absoluto ".abs()" e média ".mean()" do NumPy/Pandas sobre os inteiros de vernacular_score e human_score.

[11] Métricas de Correlação, como Correlação de Pearson são recomendadas por Bruce et al., (2019, p. 59) para análises exploratórias deste tipo. Cheguei a este resultado, como também pode ser averiguado no código calculista.py na pasta de Auditoria do *dataset*, utilizando o comando ".corr()" do Pandas (Python) sobre os inteiros de vernacular_score e human_score.

[12]A análise de Viés Sistemático é recomendada pelo manual de Bruce et al., (2019, p. 75-77) para avaliar a ligação de um modelo e identificar inclinações estruturais acima ou abaixo do esperado em relação a um modelo controle (neste caso, os valores de human_score. Cheguei a este resultado, como também pode ser averiguado no código calculista.py, utilizando o comando de média ".mean()" do NumPy sobre a diferença linear simples (sem conversão modular).

Tempo Médio de Decisão Humana	4,63s / documento

Fonte: Elaborado pelo autor

A primeira implicação extraída desta auditoria se observa no valor do Erro Médio Absoluto (MAE). O valor de 1.13, situado numa escala de 1 a 9, indica que a IA apresenta um desvio heurístico contido. Os dados sugerem que o modelo não opera de forma estocasticamente aleatória, mantendo suas inferências em uma vizinhança semântica próxima ao meu julgamento. A Correlação de Pearson (0.48), considerada moderada, parece refletir a diferença epistemológica na natureza da leitura: enquanto o algoritmo pondera o cruzamento estatístico dos pesos atribuídos ao remetente, vetor e tipologia presentes no texto, o humano aciona sua intuição historiográfica extracorpus, preenchendo lacunas pragmáticas não explicitadas no resumo. Essa divergência moderada corrobora a premissa de que a IA atua não como um leitor histórico perfeito, mas como um extrator de padrões textuais não aleatórios ou alucinados.

Um dos indicativos metodológicos mais relevantes desse experimento de auditoria, contudo, é viés sistemático de +0.59. Este índice revela uma tendência do modelo a ser levemente mais “generoso” ou otimista na sua avaliação do que o escrutínio humano. Sob a ótica da Recuperação da Informação, este comportamento é considerado metodologicamente vantajoso. O viés positivo configura a ferramenta como um filtro de maior sensibilidade; para a pesquisa linguística, é preferível que o classificador sinalize documentos com um potencial superestimado (gerando falsos positivos que serão posteriormente filtrados pela leitura do pesquisador) do que atuar com rigor excessivo e descartar prematuramente manuscritos que poderiam abrigar traços sintáticos de interesse

Por fim, a auditoria registrou um tempo médio de 4.63 segundos para a tomada de decisão humana, sugerindo a viabilidade de metodologias de curadoria manual assistida, e revelou uma valiosa zona de incerteza analítica. O cálculo de erro discriminado apontou que a maior divergência média entre a máquina e o pesquisador ocorreu nas avaliações correspondentes ao *Score* 4 (MAE de 1.84). Essa nota representa a fronteira tipológica na qual o modelo parece oscilar entre a fórmula burocrática rígida e a emergência da narrativa. É precisamente neste limiar que a complexidade humana mais se afasta da probabilidade computacional, indicando que os documentos classificados no estrato intermediário podem configurar os terrenos mais profícuos para a análise qualitativa das Tradições Discursivas em transição. Diante desses resultados, se consolida a decisão de tratar a métrica gerada não como um atestado filológico final (mensurando a vernacularidade de fato), mas como um Score de Relevância Sociolinguística Potencial (SRSP). Essa ferramenta atua estritamente como a

operacionalização quantitativa da *Distant Reading* (Moretti, 2013). Ao comprovar estatisticamente que a esmagadora maioria do arquivo (68,67%) é composta por repetições de fórmulas chancelares (o *Score* 2), o algoritmo isola essa massa burocrática e cumpre, de certa forma, a promessa metodológica de Moretti: ele permite que o pesquisador contemple a imensidão normativa do sistema para, estrategicamente, desviar dela. O classificador poupa o historiador e o linguista de despenderem o inestimável tempo da leitura atenta (*close reading*) na norma metropolitana, direcionando o escrutínio humano rigoroso exclusivamente para aquele limiar estrutural periférico onde o vernáculo tem, de maneira aproximada, heurística, a maior probabilidade de romper a superfície do texto.

A título de exemplificação qualitativa, 3 dos 9 documentos com SRSP igual a 0.9 são os seguintes:

- **Pasta:** Brasil Geral
- **Código de Referência:** PT/AHU/CU/003/0042/03350
- **Descrição:** REQUERIMENTO do procurador e irmãos da Mesa da Irmandade de Nossa Senhora do Rosário dos Homens Pretos e Pardos, ereta no convento de Nossa Senhora da Graça de Lisboa, ao príncipe regente [D. João], solicitando que o irmão Joaquim dos Reis seja tirado da galera Sacramento para ser transportado para Maranhão ou Pernambuco como escravo, dado que se encontra preso apenas porque a sua Irmandade quis alforriá-lo pelo seu justo valor durante a vida do seu proprietário o capitão João Rodrigues da Costa da vila de Torres Vedras.
- **Justificativa do DeepSeek para o SRSP:** This petition from a Black and Pardo brotherhood represents marginalized actors dictating to a local scribe, creating a strong scribe filter likely to produce hypercorrection or vernacular leakage. The bottom-up vector from a peripheral religious group to the royal authority increases the probability of non-standard syntax. As a requerimento with a narrative core about manumission and legal injustice, it falls in the moderate typology range where formulaic openings give way to colloquial expression in the central plea.

- **Pasta:** Rio Grande do Sul
- **Código de Referência:** PT/AHU/CU/019/0003/00230
- **Descrição:** REPRESENTAÇÃO dos índios, naturais do Rio Grande de São Pedro, José Custódio, Lucas da Silva, João da Rosa, Pedro Antônio, Miguel Antônio dos Santos, Manuel dos Santos e Antônio José da Silva à rainha [D. Maria I], solicitando auxílio para

sobreviverem em Lisboa, onde se encontram depois de terem sido presos pelas tropas espanholas, levados para Montevidéu e, mais tarde, para Cádiz, e pedindo transporte para voltar para o Rio Grande de São Pedro.

- **Justificativa do DeepSeek para o SRSP:** *This representation by indigenous individuals from Rio Grande de São Pedro to the queen involves a dictated narrative of personal hardship, which typically bypasses rigid formulaic structures. The marginalized status of the senders necessitates scribal mediation, increasing the likelihood of local Brazilian Portuguese syntactic features due to the scribe's potential hypercorrection or incomplete mastery of metropolitan norms. The bottom-up communication vector from a frontier region to the metropole further amplifies vernacular intrusion, as such petitions often preserve colloquial speech patterns in their core narrative despite formal framing.*

- **Pasta:** Rio de Janeiro
- **Código de Referência:** PT/AHU/CU/017/0126/10118
- **Descrição:** REQUERIMENTO do índio e capitão-mor da aldeia de São Francisco Xavier de Tagoaí, estabelecida na Fazenda de Santa Cruz, no Rio de Janeiro, José Pires Tavares, à rainha [D. Maria I], solicitando que os índios, trazidos dos sertões da Lagoa dos Patos pelo padre José de Anchieta, possam permanecer nas suas terras da qual Manoel Joaquim, Manoel de Araújo Gomes e José Teixeira os tentam expulsar.
- **Justificativa do DeepSeek para o SRSP:** *This requerimento involves an indigenous captain-mor from a frontier aldeia, indicating high scribe mediation due to his marginalized status and likely limited formal literacy. The bottom-up vector from a peripheral settlement to the monarch increases vernacular probability, as local scribes often imperfectly standardize dictated narratives. The typology is narrative-heavy within a requerimento framework, allowing colloquial syntax to leak through despite formulaic elements.*

3. Vetorização Semântica e Implementação do Motor de Busca Híbrido

A consolidação desse conjunto de dados estruturado cumpriu o objetivo primário de higienização relacional e extração de inferências sociolinguísticas. Contudo, para que este *corpus* não se restringisse a um repositório inerte, busquei projetar, no limite das minhas capacidades, uma interface (*front-end*) de recuperação de informação sustentada por arquiteturas de aprendizado de máquina. A construção desta aplicação web, desenvolvida com

a biblioteca Streamlit em Python (Streamlit, 2026), adotou uma abordagem que supera conceitualmente as limitações da busca clássica por palavras-chave.

Para transformar os textos organizados do arquivo histórico num formato pesquisável de maneira mais livre, era necessária uma indexação semântica das informações. Reconhecendo que as descrições arquivísticas, se isoladas do seu contexto de produção, podem apresentar ambiguidades semânticas significativas, optei por não vetorizar a descrição bruta de forma isolada, em vez disso, busquei aplicar uma técnica de "enriquecimento de contexto" para a busca. Um novo algoritmo foi aplicado para extrair a tipologia documental, o nome do remetente, a localização e a descrição original do arquivista e transformar num novo bloco sintético, removendo todas as demais chaves de informação desnecessárias para a busca semântica. Este novo bloco textual foi processado pelo modelo de linguagem intfloat/multilingual-e5-large, desenvolvido por pesquisadores da Microsoft (Wang *et al*, 2024), para criar uma representação semântica legível por máquina. Através desse processo, chamado de embedding, cada documento histórico do corpus foi convertido no que pode ser chamado de "vetor semântico multidimensional" (*dense vector*). Isto é, essa representação vetorial criada pelo modelo de linguagem (que opera como um extrator de significados, e não como uma IA generativa como o DeepSeek) mapeia o sentido subjacente da combinação entre tipologia, autoria e descrição dos documentos e cria um grande mapa de relações semânticas, sobre o qual serão feitas as pesquisas na interface do classificador.

A ferramenta de pesquisa da interface web se baseia num motor de busca híbrido (um *Ensemble Retrieval*). O objetivo desse sistema é encontrar a correspondência exata de caracteres enquanto se aproximar da intenção por trás da consulta do pesquisador. Para isso, duas tecnologias de recuperação de informação diferentes são usadas em conjunto:

- **Motor Lexical:** O motor Okapi BM25 (Robertson & Walker, 1994) é o responsável pela recuperação rigorosa de entidades singulares presentes no banco de dados, como cotas arquivísticas específicas (e.g., "PT/AHU/CU/021/0006/00390"), topônimos locais e nomes próprios de capitães e oficiais. Conforme descrevem por Manning, Raghavan e Schütze (2009, p. 213-214), o BM25 supera os modelos tradicionais de frequência de termos ao incorporar uma algorítimo de filtro para o peso das palavras-chave, combinada com a normalização do comprimento do documento. Essa modelagem impede que resumos arquivísticos excessivamente longos dominem artificialmente o quadro de resultados, já que probabilisticamente eles têm mais chances de conter os termos

pesquisados, garantindo a equidade estatística na busca por correspondência tipográfica exata.

- **Motor Semântico:** Operando sobre os vetores gerados na etapa anterior, este motor projeta a consulta do usuário sobre o mesmo espaço matemático pré-processado dos documentos. O algoritmo recupera os textos através do cálculo da similaridade de cossenos, um método tradicional na área de Recuperação de Informação (Manning; Raghavan; Schütze, 2009, p. 291-292) e que também é recomendado pelos desenvolvedores da família de modelos E5 para tarefas de avaliação de afinidade semântica (Wang et al., 2024, p. 4). Esse procedimento permite localizar documentos que estão conceitualmente próximos à pesquisa, independentemente das palavras exatas utilizadas pelo usuário. Trata-se de um mecanismo semelhante às antigas pesquisas no Google (antes da visão geral gerada por IA), nas quais os resultados nem sempre continham as palavras idênticas ao que se digitava, mas se aproximavam de forma inteligente do significado pretendido pelo pesquisador.

O balanço entre esses dois motores opera de forma dinâmica por meio de uma combinação linear ponderada. Conforme os princípios de arquitetura de sistemas de busca detalhados por Manning, Raghavan e Schütze (2009, p. 146), a fusão de diferentes algoritmos de pontuação exige que a pontuação de cada query seja normalizada antes da soma, evitando que uma métrica sobressaia sobre a outra devido a amplitudes matemáticas distintas. Desse modo, o *script* que hibridiza os mecanismos de busca numa única lista realiza a normalização e, antes de efetuar a pesquisa, detecta a extensão da *query* submetida para recalibrar automaticamente as proporções de peso entre as buscas lexical e semântica e fazer o cálculo de ponderação que gera a lista de resultados do *app*.

Esse cálculo opera por meio da normalização e combinação condicional dos vetores de resultados. Primeiramente, os *scores* brutos de ambos os modelos são normalizados para uma escala comum entre 0 e 1: as pontuações do motor BM25 são divididas pelo seu valor máximo, enquanto a similaridade de cosseno gerada pelo modelo E5 sofre normalização min-max. Após a normalização, o algoritmo avalia a extensão da *query* (o seu número de tokens (n)). Para buscas curtas e nominais, em que o número de tokens seja menor que 2 ($n \leq 2$), se aplica um peso de 75% ao score lexical e 25% ao semântico. Para consultas mais extensas ou narrativas ($n > 2$), o sistema prioriza o significado abstrato e aumenta o peso dos resultados do motor semântico nos resultados finais, atribui 65% de peso à similaridade semântica (a

similaridade de cosseno) e 35% ao score lexical. O ranqueamento final do documento é dado pela equação de soma linear ponderada:

$$Pontuação_{Hibrida} = (Score_{\text{Lexical}} \times Peso_{\text{Lexical}}) + (Score_{\text{Semântico}} \times Peso_{\text{Semântico}})$$

Para além do mecanismo central de busca híbrida, a interface foi pensada para incluir um painel de controle lateral interativo, que oferece filtros de restrição baseada nos metadados e nas anotações geradas na etapa anterior pelo modelo de linguagem. Essa arquitetura permite que o pesquisador aplique recortes sociolinguísticos e geográficos sobre o *corpus*. Este painel é dividido em três eixos funcionais:

I **Perfis de Busca Predefinidos (Lentes Metodológicas):** Para otimizar a exploração do catálogo, existem atalhos que funcionam como "macros" de pesquisa. Ao selecionar lentes como *"Vozes Marginalizadas & História Social"* ou "*Máquina Administrativa (Top-Down)"*, o sistema calibra automaticamente os filtros sociolinguísticos da ferramenta para isolar os documentos que melhor representam esses recortes históricos. O usuário também pode optar pela lente padrão "*Busca Livre"*, que destrava todos os parâmetros para uma exploração personalizada.

II. **Filtro de Seções do AHU:** É uma ferramenta de controle de delimitação geográfica, por meio dela, o pesquisador pode restringir a varredura do motor de busca a capitanias ou regiões específicas do acervo (como Santa Catarina, São Paulo ou Rio Grande do Sul), isolando as buscas a apenas os documentos classificados pelo AHU àquela região.

III. **Filtros Sociolinguísticos:** É a interface que permite o controle de todas as informações preditivas criadas pela análise do DeepSeek, ela é composta por três controles:

- **Score de Relevância Sociolinguística Potencial (SRSP):** Um *slider* que permite ao usuário definir o intervalo numérico da probabilidade de inovação sintática do documento (de 0.0 a 1.0). O maior intervalo possível já vem predefinido, de forma que o pesquisador só terá sua pesquisa influenciada por essa métrica caso queira.
- **Direção da Comunicação:** Permite isolar o vetor de poder da correspondência (*Bottom-Up*, *Top-Down* ou *Horizontal*).
- **Perfil Social do Remetente:** Um filtro de múltipla escolha que cruza as categorias sociais históricas (como Plebeus, Elite Metropolitana ou Marginalizados).

A ideia dessas ferramentas é a de permitir, por meio da convergência delas, uma filtragem com diferentes graus de granularidade à pesquisa. É possível, por exemplo, que um linguista histórico filtre o *corpus* inteiro para exibir exclusivamente documentos originados em um vetor de comunicação *Bottom-Up*, cujos remetentes pertençam a extratos sociais marginalizados (*Commoner* ou *Marginalized*), restringindo os resultados da busca semântica apenas àqueles documentos que o LLM indicou possuir um alto Score (SRSP superior a 0.7). Ao mesmo tempo em que, um historiador investigando a ocupação militar no sul do Brasil cruze a busca semântica por conceitos como "deserção de tropas" ou "falta de soldos" com o filtro geográfico focado estritamente no "Rio Grande do Sul" e "Santa Catarina". Ao parametrizar a interface para exibir apenas documentos de um vetor de comunicação *Bottom-Up* cujos remetentes pertençam à categoria *Low Military* (Baixa Patente Militar), o pesquisador isola os relatos diretos da base e tem um dossiê documental a partir do qual começar a sua pesquisa.

Pensando nessa segunda ação que, por fim, foi implementada uma ferramenta de exportação de um dossiê documental contendo os filtros e os resultados de busca, utilizando a biblioteca de formatação nativa do Python, *FPDF*. Este dossiê contém a lista de quantos resultados de busca o usuário escolher, contendo todas as cotas arquivísticas, links de busca no Digitarq e descrição arquivística, além dos metadados sociolinguísticos criados pelo projeto, como as tipologias normalizadas e as justificativas do SRSP.

4. Barreiras Epistemológicas e Limites da Ferramenta

Antes de concluir, é necessário reconhecer as barreiras intrínsecas ao desenho metodológico do trabalho que acabei de apresentar. A aplicação de métodos computacionais sobre arquivos históricos não constitui, nem substitui, a crítica documental; pelo contrário, apresentam novas camadas de complexidade analítica que podem ser levadas em consideração pelos pesquisadores da área:

O primeiro limite reside na natureza do objeto processado. A divisão metodológica que estabeleci neste trabalho, entre as Seções 1, 2 e 3, deve ser muito clara. Uma etapa consistiu em ordenar os dados legados e disponibilizados publicamente pelo AHU e pelo Projeto Resgate (Seção 1) e processá-los para a busca lexical e semântica/vetorial via embeddings (Seção 3). Uma etapa completamente diferente, contudo, foi o experimento descrito na Seção 2, em que a anotação sociolinguística e a avaliação massiva dos dados foram realizadas utilizando Inteligência Artificial generativa.

Por mais que eu tenha aplicado diversos *guard rails* para obter o resultado mais rigoroso possível, desde o uso da API de um modelo de pesos abertos como o DeepSeek (com

processamento via *tokens* pagos) até exigências estritas de auditoria qualitativa, os resultados produzidos pela IA generativa neste trabalho, ainda assim, enfrentam quatro limites epistemológicos irredutíveis:

i. Diz respeito à natureza da fonte de entrada: as inferências baseiam-se estritamente no *proxy* arquivístico (os resumos textuais redigidos pelos técnicos do AHU), e não na transcrição paleográfica do manuscrito original. O algoritmo analisa a interpretação contemporânea do documento, sem alcançar a materialidade e as minúcias sintáticas diretas da fonte primária.

ii. Se refere à inauditabilidade inerente aos LLMs. Conforme um debate já antigo e conhecido, consolidado principalmente por pesquisadores como Bender et al. (2021), essas arquiteturas operam como "papagaios estocásticos[13]" (*stochastic parrots*), produzindo texto ao elencar sequências de palavras observadas num oceano de dados de treinamento, se guiando apenas por cálculos de probabilidade de combinação, sem qualquer acesso real ao significado. O argumento dos autores é o que, como nós, seres humanos, temos uma predisposição natural para interpretar atos comunicativos como portadores de sentido e intenção (vide os cachorros que "falam" com seus donos ao serem treinados a pressionar botões no chão, vídeos que alcançam milhões de visualizações globais nos *reels* do Instagram), a coerência gerada pela máquina é, muitas vezes, uma ilusão de compreensão. Essa opacidade arquitetônica, de irreplicabilidade, ou inauditabilidade dos cálculos que criaram o *ouput)* significa que as inferências da IA, por mais que sejam confinadas em um esquema estrito, derivam de "dados insondáveis" (*unfathomable training data*), que não garantem qualidade, diversidade e nem neutralidade. (cf. Bender *et al*., 2021, p.4)

iii. Para além da opacidade algorítmica, a utilização de um modelo treinado com o contingente textual da internet (obviamente produções contemporâneas, como é o caso do DeepSeek v3) introduz um claro *viés anacrônico*. Como os dados de treinamento da máquina refletem quase em sua totalidade as estruturas sociais, a demografia e a epistemologia dos séculos 20 e 21, a IA tende, invariavelmente, a projetar visões de mundo modernas sobre a realidade colonial. Na prática da anotação sociolinguística, isso significa que o modelo pode incorrer em generalizações precipitadas ao classificar certos

[13] O termo 'estocástico' se refere a sistemas ou processos cujo estado e evolução são determinados por cálculos probabilísticos e eventos aleatórios, em oposição a sistemas determinísticos. No contexto da IA, a metáfora do 'papagaio estocástico' (Bender *et al.,* 2021) diz respeito a ideia de que o modelo de linguagem apenas costura sequências de formas linguísticas se baseando na probabilidade estatística de coocorrência observada em seus dados de treinamento, sem possuir qualquer compreensão ou acesso ao significado "real" da linguagem.

atores sociais, reduzindo dinâmicas históricas complexas em categorias modernas. Se tem o risco constante, e na minha opinião um erro em que este projeto incorreu, de a rede neural subestimar o nível de letramento e a agência política de mulheres ou lideranças indígenas e afro-brasileiras do século 18 (como os capitães-mores de aldeia ou as lideranças das Irmandades de Homens Pretos), assumindo eles como sujeitos analfabetos ou totalmente alijados do letramento que tornaria sua escrita formulaica como a de um membro da alta elite ou administração colonial.

iv. Se refere à natureza unipessoal da validação empírica das categorizações. Embora as inferências geradas pela IA generativa tenham sido submetidas a uma auditoria estatística comparada com uma inferência humana (conforme detalhado na Seção 2.1), esta checagem foi conduzida exclusivamente por mim. A ausência de uma checagem estatística de concordância com mais anotadores, por uma banca de especialistas numa amostra de controle maior, significa que este experimento se propõe a ser uma ferramenta heurística inicial, que prioriza o escalonamento massivo. Se ele se provar relevante, trabalhos futuros poderão aplicar métricas de concordância amostral com múltiplos avaliadores para refinar e atestar numericamente a precisão objetiva deste classificador.

Como consequência dessas limitações, o experimento analítico da Seção 2 não atesta a existência material de inovações sintáticas, é exclusivamente um indicador heurístico e preditivo. Sua função é servir como um guia opcional para a amostragem organizada dos catálogos, apontando em quais documentos pode, probabilisticamente, residir o maior potencial para abrigar o "vazamento vernáculo" que quase todo linguista busca. Cabe ressaltar, ainda, que na interface web do aplicativo aqui descrito, todos os filtros sociolinguísticos que dependem das classificações do DeepSeek vêm desativados por padrão; a pesquisa só será influenciada por essas métricas caso o pesquisador ativamente decida utilizá-las como parte da busca.

5. Disponibilidade de Dados e Código-Fonte

A transparência metodológica e a reprodutibilidade científica são pilares deste trabalho. Todo o conjunto de dados, os scripts de processamento e a infraestrutura de busca aqui descritos foram disponibilizados publicamente para a comunidade acadêmica:

- **Conjunto de Dados e Código-Fonte:** O *dataset* estruturado (JSON), o índice semântico (pkl), o código da aplicação (para self-host) e toda a documentação de auditoria humana estão depositados no repositório Zenodo, sob o DOI: 10.5281/zenodo.18772667.

- **Interface de Consulta (Web App):** A ferramenta interativa, desenvolvida em Streamlit, pode ser acessada publicamente em: catalogo-ahu-sul.streamlit.app/.
- **GitHub:** O histórico de versões e a documentação técnica do código-fonte estão disponíveis em: github.com/saulorrp/catalogo-ahu-cone-sul.

Ao utilizar estes recursos, solicita-se a citação formal do repositório conforme o DOI indicado acima.

6. Considerações Finais

O desenvolvimento da arquitetura computacional apresentada neste trabalho demonstra que a Linguística Diacrônica tem muito a ganhar com o tratamento mais sofisticado de dados, que oferece um novo paradigma para a exploração de arquivos de grande extensão. O banco de dados resultante e sua respectiva interface não encerram a investigação historiográfica ou linguística; mais que isso, instrumentalizam novas bases tecnológicas robustas para dar suporte a trabalhos dessa natureza. Apesar das fronteiras epistemológicas já ressaltadas, a ferramenta entrega aos pesquisadores uma lente macroscópica inédita. Ela permite um acesso aos documentos coloniais do Brasil Meridional no AHU de forma até então inatingível exclusivamente pelos métodos filológicos tradicionais.

7. Referências

BELLOTTO, Heloísa Liberalli. Como fazer análise diplomática e análise tipológica de documento de arquivo. São Paulo: Arquivo do Estado; Imprensa Oficial do Estado, 2002.

BRASIL. Ministério da Cultura. Fundação Biblioteca Nacional. Projeto Resgate Barão do Rio Branco. Brasília, DF: MinC/FBN, [s.d.]. Disponível em: https://www.gov.br/bn/pt-br/central-de-conteudos/projeto-resgate.

BROWN, Tom B. et al. Language models are few-shot learners. Advances in neural information processing systems, v. 33, p. 1877-1901, 2020.

BRUCE, Peter; BRUCE, Andrew. Estatística prática para cientistas de dados: 50 conceitos essenciais. 1. ed. Rio de Janeiro: Alta Books, 2019.

KABATEK, Johannes. Tradições discursivas e mudança lingüística. In: LOBO, Tânia; RIBEIRO, Ilza; CARNEIRO, Zenaide; ALMEIDA, Norma (org.). Para a história do português brasileiro: volume VII: novos dados, novas análises, tomo II. Salvador: EDUFBA, 2006.

LOSE, Alícia Duhá; SOUZA, Arivaldo Sacramento de. Para uma filologia na pesquisa em linguística histórica. Letras, Santa Maria, v. 30, n. 60, p. 11-31, jan./jun. 2020. Disponível em: https://doi.org/10.5902/2176148542058.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. Introduction to Information Retrieval. Cambridge: Cambridge University Press, Online Edition 2009.

MCENERY, Tony; HARDIE, Andrew. Corpus linguistics: method, theory and practice. Cambridge: Cambridge University Press, 2012.

MORETTI, Franco. Distant Reading. Londres; Nova York: Verso, 2013.

ROBERTSON, Stephen; WALKER, Steve. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In: Annual International Acm Sigirconference On Research And Development In Information Retrieval, 17., 1994, Dublin. Proceedings [...]. Dublin: ACM, 1994. p. 232-241.
STREAMLIT. Streamlit Python Package. Versão 1.26.0. [S.l.]: Zenodo, 2026. DOI: 10.5281/zenodo.18421864. Disponível em: https://doi.org/10.5281/zenodo.18421864.
WANG, Liang et al. Multilingual E5 Text Embeddings: A Technical Report. arXiv preprint arXiv:2402.05672, 2024. Disponível em: https://doi.org/10.48550/arXiv.2402.05672.

8. Declaração de Fianciamento

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

9. Declaração de Conflitos de Interesses

O autor declara que a pesquisa foi conduzida na ausência de quaisquer relações comerciais ou financeiras que possam ser interpretadas como um potencial conflito de interesses.

10. Declaração de Disponibilidade de Dados de Pesquisa

Todos os dados relevantes gerados ou analisados durante este estudo estão incluídos neste artigo pré-publicado ou nos links links permanentes (doi) descritos neste trabalho.

11. Declaração de uso de IA

O autor declara que esta pesquisa configura um trabalho assistido por IA. Tecnologias de Inteligência Artificial, incluindo Grandes Modelos de Linguagem (LLMs) e Modelos de Vetorização Semântica, foram empregadas centralmente como instrumentos metodológicos para o processamento, anotação e classificação do corpus documental, conforme exaustivamente detalhado e auditado no manuscrito. Além disso, ferramentas baseadas em IA foram utilizadas de forma restrita para o refinamento formal do texto (revisão ortográfica, gramatical e aprimoramento estilístico) e auxílio na escrita de código. O autor declara que nenhuma IA generativa foi utilizada para a concepção intelectual, estruturação e reflexão teórica ou formulação de ideias e conclusões. A autoria do trabalho é integralmente humana, cabendo ao autor a total responsabilidade pela originalidade, precisão, integridade e validade científica desta obra.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.