

Estado da publicação: O preprint não foi publicado em outro meio.

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: 2026-10-29 (versão 4)

(AAAA-MM-DD)

Justificativa da versão: Reformulação e atualização para a descrição de uma nova etapa do projeto.

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 Catálogo AHU-Brasil, uma ferramenta de triagem documental que reúne, num banco de dados estruturado e pesquisável, cerca de 167 mil verbetes de documentos do Arquivo Histórico Ultramarino (AHU) referentes ao Brasil colonial. O projeto começou como o “Classificador AHU-Sul”, restrito a 7.051 verbetes do Sul e Sudeste do Brasil extraídos do Projeto Resgate Barão do Rio Branco, e foi depois ampliado a todas as séries do Conselho Ultramarino relativas ao Brasil, a partir dos catálogos publicados pelo próprio arquivo, no Arquivo Científico Tropical Digital (ACTD), e pelo Projeto Resgate. Para superar as limitações da busca lexical em massas de dados não estruturados, foi desenvolvida uma pipeline em Python que integra a coleta automatizada dos catálogos, a higienização de metadados, a padronização e a conversão das cotas arquivísticas (padrão CRAV/DigitArq), o controle de qualidade dos verbetes e a ligação dos documentos aos seus registros no DigitArq. O trabalho detalha, ainda, o processo de vetorização semântica dos resumos para a implementação de um motor de busca híbrido, que permite consultas tanto por termos lexicais específicos quanto por contextos semânticos amplos. Como descrição complementar, o artigo apresenta um experimento de anotação sociolinguística feito com o modelo DeepSeek v3 sobre os 7.051 verbetes da coleção AHU-Sul, que infere categorias sociais, vetores de comunicação e a probabilidade de mediação por escritas, sintetizados no Score de Relevância Sociolinguística Potencial (SRSP), um indicador heurístico para apontar os manuscritos com maior propensão a abrigar inovações sintáticas do português brasileiro colonial. Alinhado aos princípios da Ciência Aberta, o trabalho disponibiliza publicamente o conjunto de dados completo, o código-fonte e uma interface interativa. 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 Catálogo AHU-Brasil, a documentary triage tool that gathers, in a structured and searchable database, about 167,000 archival summaries of documents from the Arquivo Histórico Ultramarino (AHU) concerning colonial Brazil. The project began as the "AHU-South Classifier", restricted to 7,051 summaries concerning Southern and Southeastern Brazil extracted from the Projeto Resgate Barão do Rio Branco, and was later expanded to every series of the Overseas Council concerning Brazil, based on the catalogues published by the archive itself, in the Arquivo Científico Tropical Digital (ACTD), and by the Projeto Resgate. To overcome the limitations of lexical search in unstructured large-scale datasets, a Python pipeline was developed, integrating the automated collection of the catalogues, metadata cleaning, the standardization and conversion of archival call numbers (CRAV/DigitArq standard), quality

control of the summaries, and the linking of the documents to their DigitArq records. This study also details the semantic vectorization of the summaries to implement a hybrid search engine, which enables both specific lexical term queries and broader semantic contextual searches. As a complementary description, it presents a sociolinguistic annotation experiment carried out with the DeepSeek v3 model on the 7,051 summaries of the AHU-South collection, which infers social categories, communication vectors, and the likelihood of scribe mediation, synthesized into the Score of Potential Sociolinguistic Relevance (SRSP), a heuristic indicator to guide researchers toward manuscripts with a higher propensity for containing syntactic innovations of colonial Brazilian Portuguese. 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.

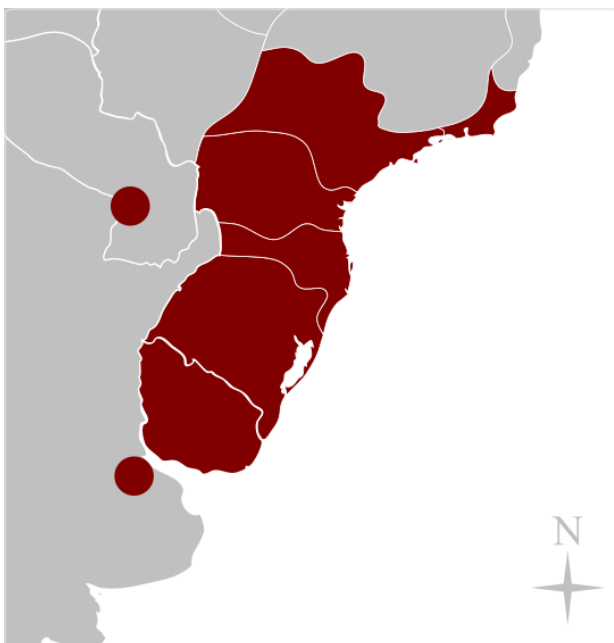
0. Introdução

A investigação da mudança e da variação linguística no português brasileiro, quando se sustenta 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 o Brasil colonial como um todo, e em especial para a macrorregião sul do Brasil¹, os acervos do Arquivo Histórico Ultramarino (AHU), disponibilizados pelo Projeto Resgate, constituem um repositório essencial da documentação administrativa do período.

O AHU realizou um levantamento exaustivo dos documentos sob sua custódia, criando um breve resumo para cada um dos milhares de manuscritos. Esses verbetes, acompanhados de informações básicas (como a região a que se referem, a cota arquivística e a tipologia documental), formam uma “capa” descritiva que intermedeia o acesso ao original. O conjunto desses verbetes e metadados compõe o que se costuma chamar de “catálogo” do AHU.

A partir dessa base, o Projeto Resgate Barão do Rio Branco coordenou, entre outras iniciativas, um esforço monumental de publicação, microfilmagem e digitalização desses 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 o catálogo *dos documentos avulsos referentes à Capitania de Santa Catarina (1717-1827)*, publicado pela Editora da UFSC e hoje também disponível no Arquivo Científico Tropical Digital (ACTD, s.d.). Por meio dessas obras,

Figura 1 – Representação da “macrorregião sul” descrita

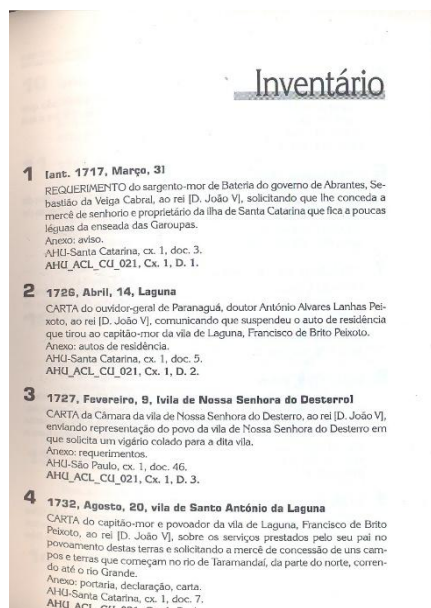


Fonte: Elaborado pelo autor a partir de vetores originais da Wikimedia Commons)

¹ Definida, para o escopo deste trabalho, com base na catalogação original do Arquivo Histórico Ultramarino. Na primeira etapa deste trabalho (a coleção AHU-Sul), 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).

publicadas de forma coordenada em todo o Brasil, os pesquisadores passaram a ter acesso ao catálogo do arquivo e, num CD que acompanhava a publicação, às versões digitais dos fac-símiles microfilmados dos documentos. Atualmente, essas versões digitalizadas estão disponíveis no site institucional do Projeto Resgate/Biblioteca Luso-Brasileira², e os catálogos estão listados num endereço institucional do Governo Federal/Fundação Biblioteca Nacional³.

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



Fonte: Serpa & Ramos Flores (2000)

Embora os catálogos impressos continuem a servir aos pesquisadores, o acesso a essas descrições arquivísticas pode ser otimizado para explorar melhor o seu enorme volume de dados. O pré-processamento computacional e 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. Além disso, 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

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 ocupa de analisar de fato 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. Ao converter arquivos massivos em dados quantificáveis, este projeto buscou, portanto, tornar visível a “topografia” do acervo, mapeando padrões de poder, repetições formulaicas e dinâmicas sociais que seriam invisíveis na escala do pesquisador humano (Moretti, 2013, p. 59), para então guiar a pesquisa aprofundada até os documentos de interesse. No contexto do AHU, a aplicação computacional desse conceito não pretende substituir a crítica paleográfica, a diplomática ou a Linguística Histórica, mas atuar como uma

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.

sondagem metodológica: uma varredura semântica capaz de processar os resumos para apontar, estatisticamente, onde estão os focos de interesse para a pesquisa linguística.

A isso se somam as complexidades próprias do estudo da linguística histórica, que, diferentemente de outras áreas ocupadas com o conteúdo dos documentos ou com a crítica documental, precisa considerar 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, se torna necessário o uso de ferramentas analíticas capazes de cruzar dimensões históricas, sociais e textuais. Assim, os pesquisadores da área precisam isolar documentos não apenas por localidade e data, mas também por parâmetros finos, como a classe social do remetente e a hierarquia da comunicação (o vetor direcional).

Além disso, se torna imprescindível detectar, ou ao menos considerar com a devida cautela, a provável presença ou ausência de um escrevente intermediário na produção do documento. Conforme discutem Lose e Souza (2020), as edições de documentos históricos frequentemente deixam de apontar as mãos que tecem o manuscrito, sem atentar para 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, chamada pelas autoras de *scriptor*. Distinguir os *scriptores* dos autores intelectuais, ou de 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 o Catálogo AHU-Brasil: uma esteira de processamento (*pipeline*) de coleta, extração, padronização e indexação vetorial que converteu cerca de 167 mil verbetes arquivísticos não estruturados do AHU referentes ao Brasil, publicados pelo próprio arquivo e pelo Projeto Resgate, num banco de dados estruturado legível por máquina. O projeto nasceu, contudo, de forma mais modesta, como o Classificador do Catálogo AHU-Sul, que tratava apenas dos 7.051 verbetes referentes ao Sul do Brasil e às regiões circundantes, extraídos de aproximadamente 71.000 linhas de texto do Projeto Resgate. Foi sobre essa primeira coleção que os procedimentos aqui descritos foram desenvolvidos e

testados; só parte deles (a coleta, a padronização das cotas e a busca) foi depois estendida ao restante do catálogo.

Sobre esse banco, desenvolvi uma interface web de recuperação de informação com motores de busca próprios, que funde a busca lexical clássica (algoritmo BM25) com *embeddings* semânticos vetoriais, otimizando a triagem de documentos e facilitando a seleção e o acesso aos manuscritos digitais hospedados nas plataformas contemporâneas do DigitArq e do Projeto Resgate. Como experimento complementar, restrito aos 7.051 documentos da coleção AHU-Sul, integrei também a Inteligência Artificial generativa, especificamente a API do modelo DeepSeek v3-chat. A interação com o modelo foi cercada de *guard rails* que ampliassem a sua confiabilidade, para que ele atuasse estritamente como um motor de inferência algorítmica voltado a 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 uma métrica quantitativa, batizada de Score de Relevância Sociolinguística Potencial (SRSP), pensada para sinalizar heurísticamente aos linguistas quais cotas arquivísticas abrigam a maior probabilidade de conter inovações sintáticas do português brasileiro na fonte primária.

Ao detalhar as fases de processamento, sanitização estrutural e auditoria algorítmica, busco oferecer 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 até então inertes em instrumentos consultáveis, dinâmicos e direcionados à pesquisa na área.

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

O banco de dados foi construído em duas etapas. A primeira, que serviu de piloto e que descrevo a seguir, tratou apenas dos 7.051 verbetes referentes ao Sul do Brasil e às regiões circundantes. A segunda, descrita na Seção 1.1, ampliou o banco a todo o catálogo do AHU referente ao Brasil, chegando a cerca de 167 mil verbetes, mas com um tratamento mais enxuto e inteiramente automatizado: a coleta dos catálogos, a padronização e a conversão das cotas, a extração da tipologia e o controle de qualidade. A revisão manual das tipologias e as camadas de análise descritas neste artigo continuam restritas aos 7.051 verbetes da primeira etapa.

A arquitetura deste projeto foi desenvolvida em Python e concebida como uma esteira de processamento (*pipeline*) que opera sobre os dados não estruturados e cria um

banco de dados em formato JSON (*JavaScript Object Notation*). O ponto de partida metodológico foi um arquivo de texto bruto com aproximadamente 71.000 linhas extraídas da Central de Conteúdos do Projeto Resgate Barão do Rio Branco (Brasil, [s.d.]). Esse arquivo continha as informações básicas e os verbetes arquivísticos do Arquivo Histórico Ultramarino (AHU) referentes ao Brasil colonial, compilados e disponibilizados ao público pelos esforços de indexação do Projeto Resgate. Cabe lembrar que, por uma questão de viabilidade, me restringi inicialmente aos dados da Região Sul do Brasil e das regiões circundantes, objetos da minha pesquisa: Santa Catarina, Rio Grande do Sul, Colônia do Sacramento e Montevideú; 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 do projeto a todo o catálogo disponível e incluir as demais regiões do Império Colonial Português, especialmente a região Nordeste do Brasil. Posteriormente, ao menos para o Brasil, essa ampliação foi realizada, como descrevo na Seção 1.1.

O desafio dessa primeira etapa estava na alta entropia do formato textual original, que, por mais organizado que fosse, era pouco legível por máquina. Para resolver isso, planejei uma sequência de operações para serializar a informação do AHU num único *array*⁴ de objetos JSON. A escolha desse formato se justifica pela sua capacidade de representar informações complexas e hierárquicas de maneira simples e pela possibilidade de aplicar uma tipagem estrita aos dados (além de *strings* para as informações, utilizei valores booleanos para os metadados lógicos). O objetivo primário era chegar a uma estrutura em que cada documento correspondesse a um objeto composto pelas mesmas chaves, cujos valores pudessem ser acessados e manipulados automaticamente.

A primeira fase do tratamento consistiu no *parsing* dos metadados básicos para a construção do JSON. O script percorreu sequencialmente as linhas do documento de texto original, usando a biblioteca nativa de expressões regulares do Python (re) como mecanismo principal de ancoragem, padronização e extração de informações. O primeiro padrão de busca

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: é uma única lista sequencial e contínua, delimitada por colchetes [], na qual cada entrada é um documento individual. Cada um 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": "Montevideú", "reference_code": "PT/AHU/CU/065/0001/00002", "extracted_typology": "CARTA", "document_id_and_date": "7964. 1787, março, 28, Montevideú"}, **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"}]

se concentrou na marcação #####⁵, que deixei no TXT para separar cada caixa/região do AHU e que indicava a transição entre blocos estruturais na indexação original do arquivo. Depois da criação da chave *folder*, outras expressões regulares da mesma natureza mapeavam e extraíam as demais informações: o prefixo AHU_ACL indicava 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 (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, a base estrutural sobre a qual seriam construídas as camadas analíticas de linguagem natural.

Durante esse processamento inicial, a formatação original dos dados apresentou sérios desafios em relação à temporalidade, o que exigiu intervenções em duas frentes. A primeira 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. Como as descrições arquivísticas do AHU tradicionalmente começam pela tipologia documental grafada em letras maiúsculas, se implementou uma expressão regular um pouco mais complexa, capaz de localizar a primeira sequência de três ou mais letras em caixa alta. O algoritmo avaliava a posição dessa palavra: se houvesse texto antes da tipologia em caixa alta, 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 num segundo fenômeno de fragmentação temporal: os metadados órfãos. Como o acervo passou por vários esforços de catalogação, muitas linhas com as identificações cronológicas primárias estavam separadas, no arquivo bruto, dos seus respectivos blocos documentais. A situação se tornava ainda mais complexa porque a base lidava simultaneamente com três camadas 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 o padrão contemporâneo português, que ainda seria preciso estabelecer (“*reference_code*”:

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

“PT/AHU/CU/021/0006/00390”). Para resgatar essas datas órfãs e as unir à chave principal, se desenvolveu outro script de cruzamento de dados, baseado na lógica de *Left-Join*. O algoritmo usou a cota intermediária (*new_code*) como ‘âncora’ para buscar a data correspondente no arquivo *Ground Truth* e a reintegrar ao banco de dados JSON. Para reduzir o risco de associar a data à cota errada por inconsistências de digitação, o cruzamento adotou uma “Auditoria Indulgente” (*Forgiving Auditor*), aplicando uma função em Python (expressa como `.strip().rstrip('.,').lower()`) que elimina espaços invisíveis e pontuações acidentais das chaves antes da comparação, garantindo o reengate preciso.

Uma vez isolada a tipologia documental exata (a primeira palavra em ALLCAPS), o desafio seguinte foi organizar a dispersão causada pela flexão de número. A língua portuguesa, como bem sabemos, admite múltiplas formas de pluralização, o que fazia com que documentos da mesma tipologia 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 o “-ES” de terminações em “-RES” e eliminar o “-S” final 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 por essas regras puramente gráficas criou algumas monstruosidades (o plural “ORDENS” se tornou a anomalia “ORDEN”, por exemplo), além de não resolver problemas decorrentes de ortografias arcaicas, como “TRESLADO” em vez de “TRASLADO”. A solução foi comparar as tipologias extraídas com uma lista teórica de controle, baseada na obra de Bellotto (2002), por meio de um motor de Lógica *Fuzzy*. Para os termos sem correspondência exata, o *script* invocava (por meio da biblioteca *diff*lib do Python) um algoritmo de similaridade entre sequências de caracteres, aparentado à distância de Levenshtein, para calcular a proximidade entre a palavra anômala e as categorias válidas, com um limiar de corte (*cutoff*) de 80%. Quando a correspondência atingia esse índice, o sistema fazia o “encaixe” (*snapping*) automático da palavra errônea na sua forma teórica correta.

Contudo, a automação de correções ortográficas sobre o vocabulário histórico traz riscos para a confiabilidade dos dados finais. Para garantir a integridade científica e evitar a corrupção de termos que fossem escolhas deliberadas dos arquivistas, o processo de normalização evitou rotinas de alteração cega e culminou na criação de uma nova chave no JSON, denominada *typology_status*. O valor booleano dessa chave controla a normalização da

tipologia: as extrações que correspondiam em pelo menos 80% à lista de Bellotto recebiam o *status true*, eram validadas manualmente por mim e submetidas à normalização morfológica. Já os vocábulos não reconhecidos na obra recebiam o *status false*, ficavam fora do algoritmo de correção e eram preservados na base de dados exatamente como constavam no catálogo original. Esse cuidado impediu que o sistema corrigisse indevidamente termos arcaicos ou descrições atípicas que fugiam ao padrão diplomático estrito.

Resolvida a coesão interna dos dados, a consolidação estrutural da base exigiu uma intervenção de interoperabilidade externa. Os identificadores usados pelo Projeto Resgate (*“new_code”*: “AHU_ACL_CU_021, Cx. 6, D. 390.”), embora úteis para a localização física, refletem um sistema arquivístico 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 com a sua plataforma atual, o sistema CRAV (DigitArq), foi necessária uma engenharia reversa sobre esses códigos. Por meio de um novo bloco de expressões regulares, um *script* identificou e particionou as antigas cotas, isolando os numerais referentes à série, à subsérie, à caixa e à numeração sequencial do documento. Em seguida, esses componentes passaram por uma operação de preenchimento de zeros à esquerda. Ao forçar que cada numeral atinja um comprimento 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 portugueses.

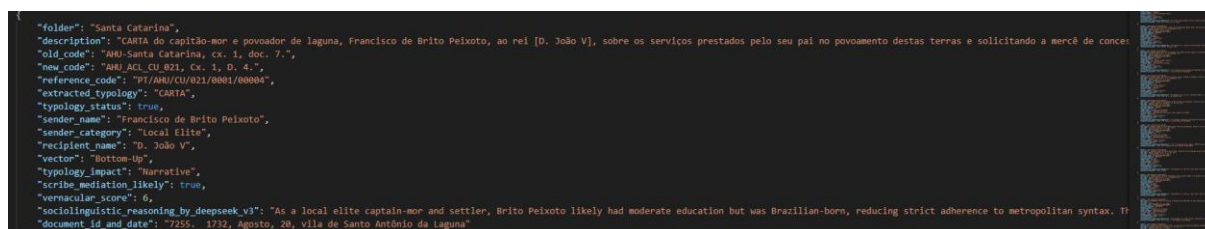
Com os códigos de referência atuais de Portugal gerados a partir das cotas do Projeto Resgate, foi possível transformar os códigos de referência dos documentos do catálogo em URIs (*Uniform Resource Identifiers*) de busca do sistema CRAV. Assim, foi possível criar, num primeiro momento, a URL de busca [“https://digitarq.arquivos.pt/search?query=PT%2FAHU%2FCU%2F021%2F0006%2F00390&isAdvancedSearch=false”](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 a ferramenta diretamente ao sistema de controle de documentos do Arquivo Nacional da Torre do Tombo.

Mais tarde, já com o corpus ampliado descrito na Seção 1.1, ao conferir esses links contra o próprio DigitArq, percebi que a busca pela referência nem sempre encontrava o documento, sobretudo na série do Rio de Janeiro organizada por Castro e Almeida, em que o DigitArq descreve vários documentos num único registro, na forma de um intervalo. Para resolver isso sem adivinhar, listei, pela interface pública do DigitArq, todos os registros das caixas citadas no corpus (cerca de 152 mil) e passei a ligar cada documento diretamente ao

registro que o contém. Dos cerca de 157 mil documentos que têm código de referência do DigitArq (os cerca de 10 mil restantes, quase todos da Série Reino e dos Códices, não o têm), apenas 221 não têm registro correspondente; nesses casos, o link leva à caixa em que o documento está guardado.

Ao fim dessa primeira etapa, o arquivo JSON continha, 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.

1.1. Ampliação do Corpus: o Catálogo AHU-Brasil

Depois da primeira versão deste trabalho, foi possível realizar aquilo que, acima, apontei como um objetivo para um cenário com mais tempo e recursos: ampliar o escopo do projeto a todo o catálogo do AHU referente ao Brasil. A ferramenta, agora chamada Catálogo AHU-Brasil, reúne cerca de 167 mil documentos de todas as capitanias e séries do Conselho Ultramarino relativas ao Brasil. Essa ampliação, contudo, se deu em duas camadas com estatutos diferentes, e é importante deixar essa distinção clara. A primeira é a coleção AHU-Sul, os 7.051 documentos descritos ao longo deste artigo, que continuam sendo os únicos a receber as camadas de análise (as inferências sociolinguísticas do DeepSeek, a tipologia revisada e a identificação das pessoas, com o local de nascimento de parte delas). A segunda é a coleção AHU-Brasil, com os cerca de 160 mil documentos restantes, que traz o catálogo do arquivo tal como foi publicado: o resumo, a data, o local, a série, a cota e os links para as fontes, além da tipologia, extraída automaticamente pelas mesmas regras de normalização descritas acima, mas sem a validação manual. Quando um documento está nas duas, ele aparece uma única vez, na versão da AHU-Sul, que é a mais rica. Na versão atual, o banco deixou também de ser um único arquivo JSON: as duas coleções e cada camada de análise ficam em arquivos separados, em formato Parquet, cada um com a sua proveniência.

A identificação das pessoas da coleção AHU-Sul foi feita numa etapa própria, posterior à anotação do DeepSeek, em cinco passos. Primeiro, um script separou cada remetente registrado pelo DeepSeek em pessoas, títulos, cargos, coletivos e instituições, sem

quebrar os sobrenomes compostos. Depois, as diferentes grafias de uma mesma pessoa foram reunidas e os homônimos de épocas diferentes foram separados, o que resultou em 1.359 pessoas. Para cada uma, montei um dossiê com as variantes do nome, os títulos, os cargos, o período de atividade, os locais de onde escreveu, os principais destinatários e trechos dos próprios resumos do AHU. Em seguida, cada dossiê foi enviado ao modelo sonar-pro, da Perplexity, que pesquisa na web, com a instrução de identificar a pessoa apenas quando o nome, o período e ao menos um elemento do contexto coincidisse com a fonte, de evitar homônimos de outras épocas e parentes com o mesmo nome ou título, e de tratar “não encontrado” como uma resposta correta. Quando identificava a pessoa, o modelo devia informar o local de nascimento, a página consultada e o trecho literal da fonte que o registra. A confiança de cada resposta não é a declarada pelo modelo: ela é calculada por regras, que conferem o nome, a cronologia (ninguém pode ter nascido depois do primeiro documento que assinou), se a página citada está entre as que a busca de fato consultou, se o trecho menciona o local e, quando possível, se o Wikidata confirma o dado. Só os nascimentos com confiança alta ou média, de 253 das 1.359 pessoas, aparecem na ferramenta, e os resultados dessa etapa estão no repositório (camadas entidades_v2 e biografias_v2).

Para montar essa segunda coleção, a fonte principal deixou de ser o texto do Projeto Resgate. Durante o trabalho, descobri que o próprio AHU publica os catálogos de cada série do Conselho Ultramarino em PDF com texto nativo, no Arquivo Científico Tropical Digital (ACTD, s.d.), do antigo Instituto de Investigação Científica Tropical, hoje parte da Universidade de Lisboa. São 35 séries relativas ao Brasil, e é delas que vêm o resumo, a data e a cota de quase todos os documentos. As páginas do Projeto Resgate no site da Biblioteca Nacional continuaram sendo usadas para três funções: fornecer o link de cada documento no site do Resgate, conferir o texto do ACTD e cobrir aquilo que o ACTD não tem (alguns Códices, parte da Série Reino e Sergipe). Os catálogos temáticos do Resgate (sobre mulheres, escravidão, irmandades e povos indígenas, entre outros) serviram para marcar o tema dos documentos no banco da coleta (marcação que ainda não aparece na ferramenta) e só entraram como fonte quando a cota não aparecia em nenhum outro catálogo.

A coleta desse material foi automatizada. Um rastreador (*crawler*) percorreu as páginas do Projeto Resgate e registrou todos os seus endereços, e leitores (*parsers*) específicos para cada formato de catálogo separaram os verbetes e extraíram deles as mesmas informações descritas acima. Todas as páginas e PDFs foram guardados exatamente como vieram, antes de qualquer processamento, para que a leitura possa ser refeita e conferida por

qualquer pessoa. Os servidores foram consultados devagar, e nenhuma área restrita foi acessada.

Alguns catálogos (Bahia, São Paulo, Minas Gerais e Pará) só existiam, no Resgate, como imagens das páginas dos livros impressos, e outros (Pernambuco, Maranhão e Colônia do Sacramento) como PDFs escaneados com uma camada de texto pouco confiável. Para esses volumes, montei uma esteira de transcrição automática com o modelo Gemini Flash, instruído a transcrever literalmente, sem corrigir a ortografia nem desdobrar as abreviaturas. As páginas que o modelo se recusou a transcrever passaram antes pelo Tesseract, um programa de OCR tradicional, e o Gemini corrigiu esse texto olhando a imagem. Quando o texto nativo do ACTD foi encontrado, contudo, ele passou a ser a fonte, já que o texto publicado pelo próprio arquivo é mais confiável que qualquer transcrição automática. A transcrição passou, então, a servir apenas de conferência: comparando, pela cota, cerca de 61 mil verbetes transcritos com o mesmo verbete no ACTD, o texto coincide em 97% dos casos. Nenhum resumo do catálogo foi escrito ou reescrito por IA.

Juntar fontes diferentes exigiu, mais uma vez, lidar com as cotas. Como cada catálogo as escreve de um jeito (com ou sem a sigla ACL, com zeros a mais, com a pontuação trocada ou com um ponto separando o milhar), a engenharia reversa descrita acima se transformou num conversor de cotas, que reconhece essas variantes, reduz cada cota a uma chave comum (série, caixa e documento) e, a partir dela, gera as três formas: a cota do Resgate, o código de referência do DigitArq e, nas séries em que isso é possível, a cota antiga. A mesma etapa retirou das cotas o cabeçalho de página que, na extração, havia grudado em cerca de 3 mil delas, e refez as poucas referências que uma versão anterior do conversor havia montado com a série errada. Foi por essa chave que as fontes foram unidas e que 7.031 dos 7.051 documentos da AHU-Sul foram reencontrados na nova coleção. Cada documento recebeu um identificador estável, que se mantém caso a coleta seja refeita, e carrega consigo a sua proveniência: a citação da fonte e, quando existem, o PDF e a página do ACTD de onde veio e a página correspondente no Resgate.

Por fim, optei por não “consertar” por adivinhação os problemas das fontes. Cerca de 849 cotas aparecem repetidas e 310 verbetes não têm cota reconhecível; a sua correção exigiria inferir um número a partir da sequência do catálogo, e preferi os manter como estão na fonte, acompanhados de um aviso. Um script de controle de qualidade marcou, sem alterar o texto, os verbetes suspeitos (sem data, com cota repetida, fora da sequência numérica, com resumo vazio ou longo demais, ou com texto divergente entre o ACTD e o Resgate): são 4.424 dos 167.324 verbetes coletados, cerca de 2,6%. Nos cerca de 49 mil casos em que o

ACTD e o Resgate descrevem a mesma cota, os textos são praticamente idênticos. O passo a passo completo da ampliação, com os números e as decisões tomadas, está documentado no repositório do projeto (ver Seção 5).

Pelo mesmo motivo, os resumos do catálogo não são corrigidos. Erros que já vêm da fonte, como a abreviatura “SS.” desdobrada duas vezes (“SãoSão”), letras triplicadas (“Barreto”) ou preposições coladas ao nome (“deTávora”), ficam como estão, e a ferramenta os marca com um *[sic]*, para que o pesquisador saiba que vieram assim.

Está em andamento, ainda, uma extração mais leve, feita com o modelo Gemini 3.1 Flash-Lite, do remetente, do destinatário, das pessoas e lugares mencionados e do assunto de cada um dos 167 mil resumos. Diferentemente das inferências do DeepSeek, aqui se pede ao modelo apenas o que está escrito no resumo, sem interpretação, com exceção do assunto. Essas variáveis serão identificadas como produzidas pelo Gemini e mantidas separadas das do DeepSeek. Numa primeira comparação, com 600 documentos da coleção AHU-Sul, o remetente extraído pelo Gemini coincidiu com o já anotado em 97,5% dos casos, e o destinatário em 87,5%; nos desacordos examinados, o Gemini costumava estar certo, já que o DeepSeek, por vezes, inferia um destinatário que o resumo não menciona. Essas variáveis só entrarão na ferramenta quando a extração do corpus inteiro estiver concluída.

2. Vetorização Semântica e Implementação do Motor de Busca Híbrido

A consolidação desse conjunto de dados estruturado cumpriu o objetivo primário de higienização estrutural. 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 aplicação web, desenvolvida com a biblioteca Streamlit em Python (Streamlit, 2026), adota uma abordagem que supera as limitações da busca clássica por palavras-chave.

Para tornar os textos organizados do arquivo histórico pesquisáveis de maneira mais livre, era necessária uma indexação semântica das informações. Como as descrições arquivísticas, isoladas do seu contexto de produção, podem apresentar ambiguidades semânticas significativas, optei por não vetorizar a descrição bruta de forma isolada e apliquei uma técnica de “enriquecimento de contexto”. Um novo algoritmo extraiu a tipologia documental, o nome do remetente, a localização e a descrição original do arquivista e os reuniu num novo bloco sintético, deixando de fora as demais chaves, desnecessárias para a busca semântica. Nos documentos da coleção AHU-Brasil, que não têm remetente anotado, o bloco traz apenas a tipologia, a pasta e a descrição. Esse bloco textual foi processado pelo

modelo de linguagem intfloat/multilingual-e5-small, desenvolvido por pesquisadores da Microsoft (Wang *et al.*, 2024), para criar uma representação semântica legível por máquina. Na primeira versão da ferramenta usei a versão maior desse mesmo modelo (multilingual-e5-large); a troca pela versão menor foi necessária para que os vetores dos 167 mil documentos coubessem na memória de um servidor modesto. Ainda assim, esse modelo não cabe no servidor gratuito que hospeda a versão pública da ferramenta. Nela, a busca por sentido é feita por um modelo estático, bem mais leve (o potion-multilingual-128M, da biblioteca Model2Vec, de Tulkens e van Dongen, 2024), que guarda um vetor fixo para cada pedaço de palavra em vez de calcular o sentido de cada frase inteira. É uma busca menos fina, mas que permite manter a pesquisa semântica disponível a todos. Por meio desse processo, chamado de embedding, cada documento do corpus foi convertido no que se pode chamar de “vetor semântico multidimensional” (*dense vector*). Essa representação vetorial, criada por um modelo que opera como um extrator de significados, e não como uma IA generativa como o DeepSeek, mapeia o sentido subjacente à combinação entre tipologia, autoria (quando anotada) e descrição dos documentos e cria um grande mapa de relações semânticas, sobre o qual são feitas as pesquisas na interface.

A ferramenta de pesquisa da interface se baseia num motor de busca híbrido (*Ensemble Retrieval*), cujo objetivo é encontrar correspondências exatas de caracteres e, ao mesmo tempo, se aproximar da intenção por trás da consulta do pesquisador. Para isso, duas tecnologias diferentes de recuperação de informação são usadas em conjunto:

- **Motor Lexical:** O motor Okapi BM25 (Robertson & Walker, 1994) é 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 Manning, Raghavan e Schütze (2009, seção 11.4.3), o BM25 supera os modelos tradicionais de frequência de termos ao incorporar um mecanismo de saturação do peso das palavras-chave (um termo repetido muitas vezes deixa de somar indefinidamente), combinado com a normalização pelo comprimento do documento. Essa modelagem impede que resumos excessivamente longos, que probabilisticamente têm mais chances de conter os termos pesquisados, dominem artificialmente o quadro de resultados, garantindo equidade estatística na busca por correspondência exata.
- **Motor Semântico:** Operando sobre os vetores gerados na etapa anterior, este motor projeta a consulta do usuário no mesmo espaço matemático em que estão os

documentos. O algoritmo recupera os textos pelo cálculo da similaridade de cossenos, um método tradicional na área de Recuperação de Informação (Manning; Raghavan; Schütze, 2009, seção 6.3), também 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 conceitualmente próximos à pesquisa, independentemente das palavras exatas usadas pelo usuário. É um mecanismo semelhante ao das antigas pesquisas no Google (antes dos resumos gerados por IA), nas quais os resultados nem sempre continham as palavras exatas que se digitavam, mas se aproximavam de forma inteligente do significado pretendido pelo pesquisador.

Na versão atual da ferramenta, a busca passou a combinar três caminhos, e não mais dois. O primeiro é uma rota exclusiva para cotas: quando o pesquisador digita uma cota, em qualquer uma das três formas usadas pelo catálogo (o código de referência do DigitArq, a cota do Resgate ou a cota antiga), o sistema a reconhece e leva diretamente ao documento correspondente, sem passar pelos outros motores; se a cota vier no meio de uma frase, esse documento aparece primeiro e o resto da consulta segue a busca normal. Os outros dois caminhos são os motores lexical e semântico descritos acima. O motor lexical, além disso, passou a dar pesos diferentes a cada campo do documento: uma palavra encontrada no nome de uma pessoa, no destinatário (campos que, por enquanto, só a coleção AHU-Sul tem) ou na tipologia vale mais do que a mesma palavra encontrada na descrição, e menos ainda no nome da pasta. Para evitar, ainda, que um único termo frequente dominasse os resultados (uma busca por “ordem terceira de São Francisco” devolvia, por exemplo, ordens régias e pessoas chamadas Francisco), o motor passou a favorecer os documentos que contêm todos os termos da consulta, sobretudo os mais raros, e a dar um bônus quando os termos aparecem lado a lado, na mesma ordem em que foram digitados.

O balanço entre os dois motores também mudou. Na primeira versão, as pontuações de cada motor eram normalizadas e somadas numa combinação linear ponderada. Esse método se mostrou frágil, porque o modelo semântico tende a atribuir similaridades muito próximas a quase todos os documentos, e a normalização acabava distorcendo as diferenças entre eles. Por isso, substituí essa soma por uma fusão por posição (*Reciprocal Rank Fusion*, ou RRF), proposta por Cormack, Clarke e Büttcher (2009): em vez de somar as notas, o sistema considera a posição que cada documento ocupa na lista de cada motor, de modo que um documento bem colocado nos dois sobe para o topo. A extensão da consulta continua sendo levada em conta: em buscas curtas, de até duas palavras, o motor lexical pesa

mais; em buscas mais longas e narrativas, o semântico. A pontuação final de cada documento é dada pela soma, em cada motor, do peso daquele motor dividido pela posição do documento na sua lista, acrescida de uma constante (60) que suaviza a diferença entre as primeiras posições:

$$\text{Pontuação}(d) = \sum \text{peso_motor} \div (60 + \text{posição_motor}(d))$$

Para além do mecanismo central de busca híbrida, a interface se divide em duas seções: o Catálogo, com todos os documentos, e a coleção AHU-Sul enriquecida, em que ficam as variáveis que só essa coleção tem. Nas duas, um painel de controle lateral oferece filtros baseados nos metadados e, na seção enriquecida, também nas pessoas identificadas e nas anotações geradas pelo modelo de linguagem (descritas na Seção 3). Essa arquitetura permite que o pesquisador aplique recortes sociolinguísticos e geográficos sobre o *corpus*. Os controles se dividem em três eixos funcionais:

- I. **Busca e Filtros Gerais:** Valem para todo o catálogo. Junto da caixa de busca, o pesquisador escolhe o modo da pesquisa (automático, por palavras exatas ou conceitual) e o seu rigor, isto é, o quanto um documento pode se afastar do sentido da consulta para continuar na lista; um trecho entre aspas exige a expressão exata. No painel lateral, é possível restringir o período, a tipologia e a coleção, e incluir ou não os documentos sem data.
- II. **Filtro de Pastas e Regiões:** É uma ferramenta de delimitação geográfica: por meio dela, o pesquisador pode restringir a varredura do motor de busca a capitânias ou regiões específicas do acervo (como Santa Catarina, São Paulo ou Rio Grande do Sul), limitando a busca aos documentos que o AHU classificou naquela região. Com a ampliação do corpus, esse filtro passou a abranger todas as séries relativas ao Brasil.
- III. **Filtros da Coleção Enriquecida:** Só aparecem na seção AHU-Sul enriquecida, porque dependem das camadas que só essa coleção tem. Permitem buscar pelo remetente (pelo nome principal, pelas variantes e pelos títulos), pelo destinatário e pela origem do remetente (o país atual e a cidade de nascimento, exibidos apenas quando a identificação tem confiança alta ou média). Numa gaveta à parte, que começa desligada, ficam as informações preditivas criadas pela análise do DeepSeek, em três controles:
 - **Score de Relevância Sociolinguística Potencial (SRSP):** Uma caixa que, quando marcada, abre um *slider* para definir o intervalo numérico da

probabilidade de inovação sintática do documento (de 1 a 9). Desmarcada, como vem por padrão, a pesquisa não é influenciada por essa métrica.

- **Vetor de Comunicação:** Permite isolar o vetor de poder da correspondência (*Bottom-Up*, *Top-Down* ou *Horizontal*).
- **Categoria Social do Remetente:** Um filtro de múltipla escolha com as categorias sociais históricas atribuídas pelo modelo (como *Commoner*, *Metropolitan Elite* ou *Marginalized*).

A ideia dessas ferramentas é permitir, pela combinação entre elas, uma filtragem com diferentes graus de granularidade. É possível, por exemplo, que um linguista histórico, na seção AHU-Sul enriquecida, filtre o *corpus* para exibir apenas documentos de vetor de comunicação *Bottom-Up*, cujos remetentes pertençam a estratos sociais marginalizados (*Commoner* ou *Marginalized*), restringindo os resultados da busca semântica aos documentos que o LLM indicou como de SRSP alto (igual ou superior a 7). Do mesmo modo, um historiador que investigue a ocupação militar no sul do Brasil pode cruzar 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 em “Santa Catarina”. Ao parametrizar a interface para exibir apenas documentos de vetor *Bottom-Up* cujos remetentes pertençam à categoria *Low Military* (Baixa Patente Militar), o pesquisador isola os relatos vindos da base e obtém um conjunto de documentos a partir do qual começar a sua pesquisa.

Pensando nesse segundo uso, a ferramenta permite, por fim, marcar documentos ao longo de várias buscas e exportar essa seleção (ou uma lista de cotas colada pelo usuário) num PDF gerado com a biblioteca FPDF, do Python, pensado para a consulta no próprio arquivo: uma tabela com o número da linha, o ano, as cotas, o resumo completo, o link para o DigitArq e uma caixa para marcar à mão, que facilita conferir e fotografar os documentos. A seção enriquecida traz, ainda, uma aba de pessoas, com a página de cada remetente identificado; uma aba de amostras, que sorteia amostras diacrônicas (os documentos de uma mesma pessoa ao longo do tempo) ou estratificadas (por década, categoria social, tipologia, pasta ou país de nascimento), sempre com uma semente visível, para que o sorteio possa ser refeito; e um painel diatópico, que conta os documentos e as pessoas pelo país de nascimento do remetente. No Catálogo, há também um conversor que traduz uma cota entre as três formas descritas na Seção 1.1.

3. Experimento Complementar: Anotação Sociolinguística Automatizada e Inferência Baseada em LLM na Coleção AHU-Sul

Esta seção descreve um experimento complementar, que não faz parte do tratamento do catálogo inteiro e se restringe aos 7.051 documentos da coleção AHU-Sul. Diferentemente das etapas anteriores, que apenas organizam e tornam pesquisável a informação já publicada pelo AHU e pelo Projeto Resgate, aqui os resumos foram avaliados por um modelo de Inteligência Artificial generativa, e os resultados devem ser lidos como estimativas heurísticas. Na interface, essas variáveis só aparecem nos documentos da AHU-Sul, e os filtros que dependem delas vêm desligados por padrão.

Após a higienização estrutural, o projeto passou à sua fase analítica, de enriquecimento dos dados. A premissa central desta etapa é potencializar a busca dos pesquisadores pelo “vazamento vernáculo”, conceito basilar para o estudo do português brasileiro histórico. 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 a teoria da Leitura Distante (Moretti, 2013), novamente, se entrelaça aos postulados da Sociolinguística. Partindo da hipótese de que documentos produzidos por indivíduos sociogeograficamente 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 manuscritos 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 interação com a IA foi estritamente confinada ao *zero-shot prompting* (comando sem exemplos prévios). Conforme descrevem Brown *et al.* (2020), o *zero-shot* é, paradoxalmente, o cenário mais desafiador para a máquina, pois exige que ela compreenda a complexidade da tarefa apenas a partir da clareza das instruções. A adoção dessa abordagem, contudo, foi uma decisão metodológica e pragmática: incluir exemplos detalhados (*few-shot*) na chamada da API para cada um dos milhares de resumos multiplicaria o consumo de *tokens* (a unidade de memória e custo computacional dos LLMs), o que tornaria o processamento da coleção inteira financeira e computacionalmente inviável; e usar exemplos curtos para todos os resumos tornaria o *output* enviesado, além de poder levar o modelo a criar 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 a devolver a sua análise exclusivamente por meio de geração restrita⁶ em formato JSON, preenchendo as chaves exatas exigidas pelo banco de dados. Essa formatação forçada conteve digressões e alucinações narrativas e garantiu que os *outputs* fossem padronizados, reproduzíveis 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 classificava os atores sociais envolvidos, identificando o remetente e o destinatário do documento e classificando o remetente 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, que estabelece a direção do poder, classificando a correspondência como *Bottom-Up* (da periferia colonial para o centro administrativo), *Top-Down* (da Coroa ou dos governadores coloniais para os súditos) ou Horizontal (entre pares).

Em paralelo à dinâmica de poder, exigei 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

⁶ Prática conhecida como “Constrained Generation” (geração restrita).

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 *scriptor* intermediário (*scribe_mediation_likely*). Apoiado na distinção filológica entre a atribuição de autoria intelectual e a realidade do *scriptor*, destacada por Lose e Souza (2020), o LLM foi instruído a avaliar, de forma booleana, se a classe ou o tipo social do remetente (indivíduos escravizados, indígenas, estratos populares ou militares de baixa patente) e o gênero textual do documento aumentavam, historicamente, a probabilidade de um escrivão intermediar a escrita do documento. 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 involuntário da sintaxe local), a sua classificação puramente heurística é arriscada. Não é possível atestar essa mediação de forma categórica sem um extenso estudo paleográfico e diplomático do manuscrito original. Por esse motivo, optei por restringir o uso dessa variável e a retirei dos filtros do mecanismo de busca interativo. Na primeira versão da ferramenta, a avaliação algorítmica (que traduzia o valor booleano *False* para “Pouco provável” e *True* para “Provável”) aparecia apenas num dossiê exportável em PDF, como um apontamento complementar para a avaliação qualitativa do pesquisador. Na versão atual, esse dossiê deixou de existir e a variável saiu também da ficha dos documentos: ela continua guardada nos dados, mas a interface apenas explica o que ela é, com a ressalva feita acima.

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 é uma variável quantitativa, numa escala de 1 a 10, que consolida o peso dessas inferências sociolinguísticas. O SRSP, aferido pela IA, pretende ser um índice da relevância sociolinguística potencial (daí o nome) e traduzir numericamente a probabilidade de o documento físico original abrigar características de inovação sintática próprias do português brasileiro. A ideia por trás dessa quantificação é criar um filtro único que permita ao pesquisador remover do quadro de respostas os documentos mais provavelmente formulaicos e chegar aos documentos com mais potencial para o seu interesse sociolinguístico. Na prática, o modelo não atribuiu a nota 10 a documento algum, o que provavelmente se deve ao próprio prompt, cujas instruções situam os casos mais altos entre 7 e 9 e cujo modelo de resposta traz

a nota 8. A escala efetiva vai, assim, de 1 a 9, e foi essa a escala adotada na ferramenta e na auditoria humana. Para a leitura dos resultados, organizei-a em 3 níveis (cada um deles divisível em 3 subníveis), que não coincidem exatamente com as faixas tipológicas indicadas no prompt. Entre 1 e 3 estão os documentos mais formulaicos, com menor SRSP; entre 4 e 6, os documentos intermediários; e entre 7 e 9, os documentos com maior potencial 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; na interface, essa justificativa aparece na ficha de cada documento da seção AHU-Sul enriquecida. 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:

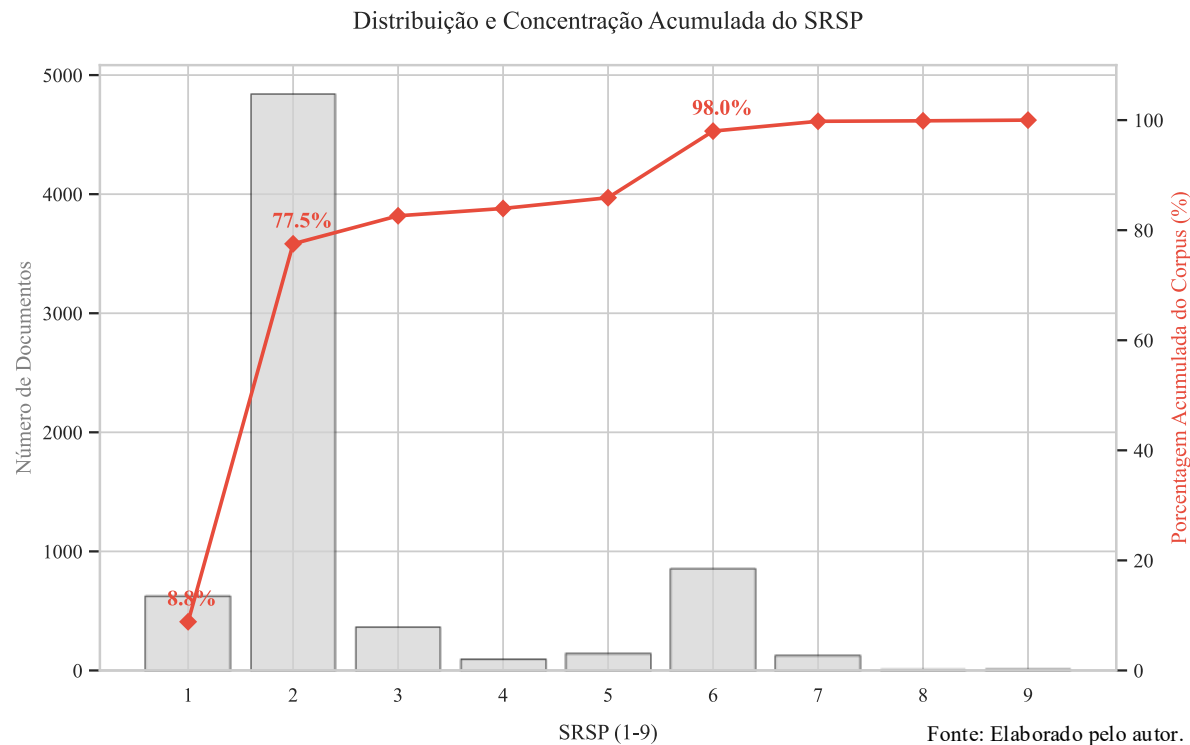
- Bottom-Up (Periphery to Metropole): HIGH probability of vernacular features.
- Deep Periphery (Frontier camps, sertão, aldeias) scores higher than administrative capitals (Desterro, Rio Grande city).
- Top-Down (Metropole to Periphery): LOW probability.

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."
}
```

A distribuição das classificações feitas pelo modelo aparece no Gráfico 1, a seguir:

Gráfico 1



A distribuição integral da coleção AHU-Sul também oferece uma noção condizente com a realidade do acervo do AHU. A concentração de 68,67% dos seus documentos no *Score 2* espelha a natureza predominantemente burocrática da documentação administrativa colonial guardada no AHU. Os dados indicam que o algoritmo foi capaz de mapear essa volumosa massa de “ruído administrativo”, presa à norma metropolitana da chancelaria, reservando os índices mais altos para a minoria de documentos que foge ao padrão formulaico. Um caminho possível diante dessa enorme concentração no *Score 2* seria tornar o prompt mais sensível a esse tipo de documento burocrático e ampliar a faixa de classificação dos documentos mais formulaicos nos três primeiros scores (1, 2 e 3). Isso, porém, sacrificaria a compatibilidade do método com arquivos de natureza não administrativa; por isso, decidi manter o método e aceitar a concentração natural dos documentos do AHU nos scores mais baixos.

3.1. Validação Empírica das Inferências do LLM

Com o intuito de aferir a consistência interna da métrica e mitigar o risco de a inteligência artificial operar descolada da realidade arquivística, conduzi um experimento de validação empírica, já que metodologias de anotação massiva baseadas em LLMs exigem etapas de aferição humana para que os seus resultados possam ser considerados cientificamente defensáveis. Cabe ressaltar que, devido ao escopo individual desta pesquisa, a classificação de controle da amostra foi feita por uma única pessoa, servindo como prova de conceito para futuras auditorias com múltiplos anotadores. Para isso, desenvolvi um protocolo e um programa de Auditoria de Consistência Semântica.

Esse procedimento consistiu na extração de uma amostra aleatória correspondente a 10% da coleção AHU-Sul (705 documentos⁷), com uma semente de randomização fixa (*seed*) para resguardar a reprodutibilidade do experimento⁸. Para evitar o viés de ancoragem durante a avaliação, desenvolvi um *script* em Python com uma interface cega: ela exibia apenas a descrição original, a cota do documento e o ID com data, ocultando a inferência previamente atribuída pela máquina. A partir dessa leitura isolada, atribuí manualmente uma nota humana (de 1 a 9) a cada verbete, com base nas mesmas três dimensões teóricas fornecidas no *prompt* da inteligência artificial e, devo admitir, no meu viés subconsciente,

⁷ A auditoria foi feita sobre 10% da coleção mais 10 documentos. O arquivo final da auditoria tem 716 avaliações (714 documentos distintos), e as métricas foram calculadas pelo script `calculista.py` sobre todas elas.

⁸ Como se pode conferir no código `auditoool.py`, a linha 108 “`random.seed(42)`” escolhe o “caminho de escolha aleatória” dos documentos a serem auditados, o que permite que outro experimento replique os resultados obtidos sem perder a natureza aleatória da amostragem.

entre outras coisas que naturalmente compõem o julgamento humano. O *script* também registrou o tempo de reação cognitiva (em segundos) despendido em cada avaliação. A Tabela 1 sintetiza as métricas de concordância extraídas dessa auditoria⁹:

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

Métricas de Precisão (10% da coleção AHU-Sul, N = 716)	Valores
Erro Médio Absoluto ¹⁰	1.13
Correlação de Pearson ¹¹	0.48
Viés Sistemático ¹²	+0.59 p/ IA
Tempo Médio de Decisão Humana	4,63s / documento

Fonte: Elaborado pelo autor

A primeira implicação extraída desta auditoria está no valor do Erro Médio Absoluto (MAE). O valor de 1.13, numa escala de 1 a 9, precisa ser lido com cautela. Como a maior parte das notas, humanas e da IA, se concentra nos níveis mais baixos, um classificador trivial que desse a nota 2 a todos os documentos teria, nesta mesma amostra, um MAE de 0.97, menor que o do modelo. O MAE, sozinho, não mostra que o modelo acerta; o que o distingue desse classificador trivial é a ordenação dos documentos, medida pela correlação. Como o SRSP é uma escala ordinal e muito assimétrica, calculei também, com os mesmos dados da auditoria, duas medidas mais adequadas a esse tipo de escala: a correlação de postos de Spearman (0.37) e o kappa ponderado quadrático (0.44), ambas indicativas de uma concordância de fraca a moderada. 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, ao vetor e à tipologia

⁹ Todos os arquivos finais da auditoria foram incluídos na documentação dentro da pasta “auditoria”.

¹⁰ O Erro Médio Absoluto é a média das diferenças absolutas entre a nota da IA e a nota humana, média(|IA – humano|); por usar o valor absoluto, e não o quadrado, ele mede o erro de forma linear, sem amplificar a influência de *outliers* (sobre métricas de erro e de variabilidade, ver Bruce e Bruce, 2019). 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 “np.abs()” e média “np.mean()”, do NumPy, sobre os inteiros de vernacular_score e human_score.

¹¹ Métricas de correlação, como a Correlação de Pearson, são recomendadas por Bruce e Bruce (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 O Viés Sistemático é a média das diferenças com sinal entre a nota da IA e a nota humana, média(IA – humano), e indica se o modelo tende a dar notas acima ou abaixo das de uma referência (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).

presentes no texto, o humano usa da sua intuição historiográfica extracorpore, 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 que não são aleatórios nem alucinados.

Um dos indicativos metodológicos mais relevantes desse experimento de auditoria, contudo, é o viés sistemático de +0.59. Esse 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, esse comportamento é 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 depois filtrados pela leitura do pesquisador) do que atue com rigor excessivo e descarte 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, o que sugere a viabilidade de metodologias de curadoria manual assistida, e revelou uma valiosa zona de incerteza analítica. O cálculo do erro por faixa 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 em que o modelo parece oscilar entre a fórmula burocrática rígida e a emergência da narrativa. É precisamente nesse limiar que a complexidade humana mais se afasta da probabilidade computacional, o que indica que os documentos do estrato intermediário podem ser 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 (que mediria a vernacularidade de fato), mas como um *Score* de Relevância Sociolinguística Potencial (SRSP). Essa ferramenta atua estritamente como uma operacionalização quantitativa da *Distant Reading* (Moretti, 2013). Ao mostrar estatisticamente que a esmagadora maioria da coleção AHU-Sul (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 despender o precioso tempo da leitura atenta (*close reading*) na norma metropolitana, direcionando o escrutínio humano rigoroso para aquele limiar estrutural periférico em que o vernáculo tem, de maneira aproximada e heurística, a maior probabilidade de romper a superfície do texto.

A título de exemplificação qualitativa, 3 dos 9 documentos com SRSP igual a 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 Montevideu 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 Tagoai, 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.*

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 a arquivos históricos não constitui nem substitui a crítica documental; pelo contrário, acrescenta novas camadas de complexidade analítica, que os pesquisadores da área precisam levar em consideração:

O primeiro limite está na natureza do objeto processado. A divisão metodológica que estabeleci neste trabalho, entre as Seções 1, 2 e 3, deve ficar muito clara. Uma etapa consistiu em ordenar os dados legados e disponibilizados publicamente pelo AHU e pelo Projeto Resgate (Seção 1) e em processar esses dados para a busca lexical e semântica via embeddings (Seção 2). Uma etapa completamente diferente, contudo, foi o experimento descrito na Seção 3, em que a anotação sociolinguística e a avaliação massiva dos dados foram feitas com 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 por *tokens* pagos) até exigências estritas de auditoria qualitativa, os resultados produzidos pela IA generativa neste trabalho enfrentam, ainda assim, quatro limites epistemológicos irredutíveis:

- i. Diz respeito à natureza da fonte de entrada: as inferências se baseiam estritamente no *proxy* arquivístico (os resumos 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 da fonte primária.
- ii. Diz respeito à inauditabilidade inerente aos LLMs. Conforme um debate já conhecido, consolidado principalmente por Bender et al. (2021), essas arquiteturas operam como “papagaios estocásticos¹³” (*stochastic parrots*): produzem texto encadeando sequências de palavras observadas num oceano de dados de treinamento, guiadas apenas por cálculos de probabilidade de combinação, sem qualquer acesso real ao significado. O argumento dos autores é 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, em vídeos que alcançam milhões de visualizações nos *reels* do Instagram), a coerência gerada pela máquina é, muitas vezes, uma ilusão de compreensão. Essa opacidade arquitetônica (a impossibilidade de replicar ou auditar os cálculos que produziram o *output*) significa que as inferências da IA, por mais que sejam confinadas num esquema estrito, derivam de “dados insondáveis” (*unfathomable training data*), que não garantem qualidade, diversidade nem neutralidade (cf. Bender *et al.*, 2021, p. 613).
- iii. Para além da opacidade algorítmica, o uso de um modelo treinado com o contingente textual da internet (produções obviamente contemporâneas, como é o caso do DeepSeek v3) introduz um claro *viés anacrônico*. Como os dados de treinamento da máquina refletem quase inteiramente as estruturas sociais, a demografia e a epistemologia dos séculos XX e XXI, 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 atores sociais, reduzindo dinâmicas históricas complexas a categorias modernas. Há o risco constante

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 à ideia de que o modelo de linguagem apenas costura sequências de formas linguísticas com base na probabilidade estatística de coocorrência observada em seus dados de treinamento, sem possuir qualquer compreensão ou acesso ao significado “real” da linguagem.

(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 de lideranças indígenas e afro-brasileiras do século XVIII (como os capitães-mores de aldeia ou as lideranças das Irmandades de Homens Pretos), tratando essas pessoas como analfabetas ou totalmente alijadas do letramento que tornaria a sua escrita tão formulaica quanto a de um membro da alta elite ou da administração colonial.

- iv. Diz respeito à natureza unipessoal da validação empírica das categorizações. Embora as inferências geradas pela IA generativa tenham sido submetidas a uma auditoria estatística, comparadas a uma avaliação humana (conforme detalhado na Seção 3.1), essa checagem foi conduzida exclusivamente por mim. Sem uma verificação de concordância com mais anotadores, por uma banca de especialistas e numa amostra de controle maior, este experimento se propõe apenas como uma ferramenta heurística inicial, que prioriza a escala. Se ele se mostrar relevante, trabalhos futuros poderão aplicar métricas de concordância com múltiplos avaliadores para refinar e medir com mais precisão o desempenho deste classificador.

Como consequência dessas limitações, o experimento analítico da Seção 3 não atesta a existência material de inovações sintáticas: é apenas um indicador heurístico e preditivo. A sua função é servir como um guia opcional para a amostragem organizada dos catálogos, apontando em quais documentos pode residir, probabilisticamente, o maior potencial para abrigar o “vazamento vernáculo” que quase todo linguista busca. Cabe ressaltar, ainda, que na interface web aqui descrita todos os filtros sociolinguísticos que dependem das classificações do DeepSeek vêm desativados por padrão; a pesquisa só é influenciada por essas métricas se o pesquisador decidir, por conta própria, recorrer a elas.

5. Disponibilidade de Dados e Código-Fonte

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 núcleo atualizado do projeto (o corpus com as duas coleções e as suas camadas de análise, os vetores semânticos, o código da aplicação, para self-host, e os scripts de coleta e de controle de qualidade) e toda a documentação de auditoria humana estão depositados no repositório Zenodo, sob o DOI: <https://doi.org/10.5281/zenodo.18772666>. A primeira versão do depósito, restrita à coleção AHU-Sul, fica no mesmo registro. Os arquivos depositados no Zenodo estão

sob embargo até 25 de fevereiro de 2027, data prevista para a publicação do artigo metodológico e da tese de doutorado a que este projeto se vincula; até lá, o código e os dados estão abertos no GitHub, indicado abaixo.

- **Interface de Consulta (Web App):** A ferramenta interativa, desenvolvida em Streamlit, pode ser acessada publicamente em: catalogo-ahu-sul.streamlit.app/. Por limitação de memória do servidor gratuito, essa versão pública usa o modelo semântico mais leve descrito na Seção 2; a versão completa, com o modelo E5, está sendo preparada para um servidor próprio.
- **GitHub:** O histórico de versões e a documentação técnica do código-fonte estão disponíveis em: github.com/saulorrrp/catalogo-ahu-cone-sul. Ali também estão os dados do Catálogo AHU-Brasil completo, tanto o corpus organizado por coleções e camadas quanto os arquivos usados pela aplicação, além dos scripts da coleta, do controle de qualidade e da transcrição automática descritos na Seção 1.1, publicados como rodaram, e o passo a passo da ampliação do corpus (docs/METODO_COLETA.md).

Quem utilizar estes recursos deve citar formalmente o 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 um tratamento mais sofisticado dos dados, que oferece um novo paradigma para a exploração de arquivos de grande extensão. O banco de dados resultante e a sua interface não encerram a investigação historiográfica ou linguística; antes, oferecem 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, que permite acessar os documentos coloniais do Brasil no AHU de um modo até então inatingível exclusivamente pelos métodos filológicos tradicionais.

7. Referências

ARQUIVO CIENTÍFICO TROPICAL DIGITAL (ACTD). Conselho Ultramarino: catálogos dos documentos avulsos. Lisboa: Universidade de Lisboa, s.d. Disponível em: <https://actd.iict.pt/>. Acesso em: 22 out. 2026.

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

- BENDER, Emily M.; GEBRU, Timnit; McMILLAN-MAJOR, Angelina; SHMITCHELL, Shmargaret. On the dangers of stochastic parrots: can language models be too big? In: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT '21), 2021, [evento virtual]. Proceedings [...]. New York: ACM, 2021. p. 610-623. DOI: 10.1145/3442188.3445922.
- 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. Tradução de Luciana Ferraz. 1. ed. Rio de Janeiro: Alta Books, 2019.
- CORMACK, Gordon V.; CLARKE, Charles L. A.; BÜTTCHER, Stefan. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Boston: ACM, 2009. p. 758-759.
- KABATEK, Johannes. Tradições discursivas e mudança lingüística. In: LOBO, Tânia; RIBEIRO, Ilza; CARNEIRO, Zenaide; ALMEIDA, Norma (org.). *Para a história do português brasileiro: volume VI: novos dados, novas análises, tomo II*. Salvador: EDUFBA, 2006. p. 505-527.
- LOSE, Alicia 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, 2009. Edição online.
- 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 SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL*, 17., 1994, Dublin. Proceedings [...]. Dublin: ACM, 1994. p. 232-241.
- SERPA, Élio Cantalício; FLORES, Maria Bernardete Ramos (org.). *Catálogo de documentos avulsos manuscritos referentes à Capitania de Santa Catarina (1717-1827)*. Florianópolis: Editora da UFSC, 2000.
- STREAMLIT. Streamlit: a faster way to build and share data apps. Versão 1.55.0. [S.l.]: Snowflake, 2026. Disponível em: <https://pypi.org/project/streamlit/1.55.0/>. Acesso em: 28 out. 2026.
- TULKENS, Stephan; VAN DONGEN, Thomas. Model2Vec: fast state-of-the-art static embeddings. [S.l.]: MinishLab, 2024. Disponível em: <https://github.com/MinishLab/model2vec>. Acesso em: 28 out. 2026.
- 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 Financiamento

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, no repositório GitHub e no depósito do Zenodo descritos na Seção 5; os arquivos do Zenodo estão sob embargo até 25 de fevereiro de 2027, e até lá o código e os dados estão abertos no GitHub.

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 exhaustivamente detalhado e auditado no manuscrito. Na ampliação do corpus (Seção 1.1), o modelo Gemini Flash foi usado para a transcrição automática de catálogos escaneados, que serviu apenas como conferência, e o modelo sonar-pro, da Perplexity, foi usado na busca de dados biográficos dos remetentes da coleção AHU-Sul, com a confiança de cada resposta calculada por regras; nenhum resumo do catálogo foi escrito ou reescrito por IA. 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, a reflexão teórica ou a 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.

12. Contribuição do Autor (CRediT)

Saulo Rogério Pacheco Rocha: conceituação, curadoria de dados, análise formal, investigação, metodologia, administração do projeto, software, validação, visualização, escrita do rascunho original, revisão e edição.

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.