TRANSCRIÇÃO MANUAL E AUTOMÁTICA DE TEXTOS HISTÓRICOS MANUSCRITOS ATRAVÉS DO SOFTWARE LAPELINC TRANSCRIPTOR

Bruno Silvério Costa
Universidade Estadual do Sudoeste da Bahia – UESB (Brasil)
Endereço Eletrônico: bsilveriocosta@gmail.com

Jorge Viana Santos
Universidade Estadual do Sudoeste da Bahia – UESB (Brasil)
Endereço Eletrônico: viana.jorge.viana@uesb.edu.com

Cristiane Namiuti
Universidade Estadual do Sudoeste da Bahia – UESB (Brasil)
Endereço Eletrônico: cristianenamiuti@uesb.edu.br

INTRODUÇÃO

As Humanidades digitais possibilitaram o desenvolvimento de novos métodos e técnicas para a pesquisa na área de Linguística de Corpus, criando sobretudo formas de aprimoramento da fidelidade e velocidade no tratamento de textos históricos digitalizados (PORTELA, 2013). No contexto de desenvolvimento desta área do conhecimento humano, está sendo compilado no âmbito do projeto "Memória conquistense: implementação de um corpus digital" o Corpus DOViC (Corpus de Documentos Oitocentistas de Vitória da Conquista), utilizando-se de técnicas e ferramentas para a transcrição e anotação de documentos históricos em meio digital (SANTOS e NAMIUTI, 2016).

Sendo o DOViC um corpus baseado nas diretrizes do CTB - Corpus Tycho Brahe (GALVES, 2019), vários conceitos teóricos e recursos tecnológicos foram herdados como diretrizes para sua compilação e anotação. Apesar do DOViC utilizar o software e-Dictor (PAIXÃO DE SOUSA, KEPLER e FARIA, 2010) para o processo de transcrição de fac-símiles de documentos históricos, existem limitações na expressão de elementos do documento original, uma vez que o referido software apresenta apenas recursos primários de auxílio à transcrição, direcionando as suas funcionalidades principalmente ao processo de edição dos textos digitais. Objetivando resolver estes e outros problemas no processo de transcrição paleográfica do corpus DOViC, foi desenvolvido um software para a assistência ao processo de transcrição paleográfica: o LAPELINC TRANSCRIPTOR.

O LAPELINC TRANSCRIPTOR, a partir de uma interface que permite a visualização e manipulação adequada dos fac-símiles, fornece funcionalidades como a ampliação/redução de áreas selecionadas, o fatiamento da imagem em linhas de transcrição, o alinhamento entre texto transcrito e imagem, e a construção de um abecedário que permite a consulta da forma de escrita dos grafemas dos escritores e copistas. O software permite ainda a representação das ocorrências do texto original de forma fidedigna, relacionando imagem e transcrição para texto em subscrito, sobrescrito, símbolos especiais (MARTÍNEZ, 1988), abreviaturas (BORGES NUNES, 1981) e nomes, além de recursos de visualização com tratamento da imagem, auxiliando ao paleógrafo durante a leitura para a transcrição do documento histórico. Todas essas funcionalidades caracterizam os recursos disponíveis no software e constituem o processo de Transcrição Manual Assistida por Computador (TMAC).

Os resultados da TMAC são utilizados para o treinamento da Inteligência Artificial (IA) interna do software, possibilitando a utilização do LAPELINC TRANSCRIPTOR para o reconhecimento automático de textos manuscritos. As rotinas para o reconhecimento automático de textos manuscritos (Handwritten Text Recognition-HTR) utilizam a base de dados de transcrições manuais para o treinamento do sistema de IA-HTR. Uma vez treinada, a IA do LAPELINC TRANSCRIPTOR é capaz de reconhecer textos manuscritos a uma mão e grafias similares. Sua interface permite ainda a correção de erros de transcrição da IA possibilitando retroalimentar o sistema para aprimoramento de seu processamento e aumento da precisão dos modelos internos de transcrição automática.

METODOLOGIA

Para a obtenção dos resultados relatados neste trabalho, foi utilizado o método LAPELINC como arcabouço teórico e pragmático no estabelecimento do fluxo de trabalho para a implementação das atividades de descrição topográfica dos documentos trabalhados (SANTOS; NAMIUTI, 2017; SANTOS; NAMIUTI, 2016). A construção do software foi conduzida mediante parâmetros apresentados no LAPELINC FRAMEWORK, conforme apresentado em Costa (2019); Costa, Santos e Namiuti (2022) e Costa, Santos, *et al.* (2022).

As rotinas de software que implementam o LAPELINC TRANSCRIPTOR foram escritas em Linguagem de Programação Java™, utilizando os *frameworks* Java Server Pages (JSP) e Java Persistense API (JPA), além das linguagens *client-side*

JavaScript e HTML5. Para ajuste de layout foi utilizada a linguagem de estilo CSS (DEITEL&DEITEL, 2005). A Inteligência Artificial foi implementada utilizando a biblioteca TensorFlow (TensorFlow, 2022) e rotinas em Linguagem Python (PYTHON SOFTWARE FOUNDATION, 2022), seguindo diretrizes apresentadas no trabalho de Harald Scheidl para a construção de HTRs baseados em algoritmos de aprendizado de máquina (SCHEIDL, 2018).

RESULTADOS E DISCUSSÃO

Após o carregamento das imagens associados ao facsímile do documento, o LAPELINC TRANSCRIPTOR possibilita a transcrição de um documento seguindo um fluxo formado por três etapas: zonear, trascrever e treinar. O processo de Transcrição Assistida (TMAC) é composto apenas pelas duas primeiras etapas (zonear e transcrever), possibilitando o delineamento de zonas de texto da imagem e associação com linhas de texto transcrito, realizando o processo de alinhamento texto-imagem. A

Figura 1 apresenta a tela do software demonstrando a etapa de zonear:

Figura 1 – Etapa 1 - Zonear do LAPELINC TRANSCRIPTOR

Fonte: Banco de imagens dos autores

A etapa de transcrição segue à de zoneamento, possibilitando informar manualmente o texto correspondente à leitura paleográfica para a zona da imagem selecionada. Caso a pesquisa demande apenas a transcrição manual, o processo de transcrição finaliza nesta etapa. No entanto, se o objetivo da transcrição manual for o treinamento para a transcrição automática, será necessário seguir para a próxima etapa:

treinar. A etapa treinar possibilita a segmentação da imagem e da transcrição em unidades denominadas *tokens*, permitindo o aprendizado orientado a unidades do texto delimitadas por espaços em branco. A Figura 2 apresenta o estágio inicial da segmentação com um *token* marcado:

Figura 2 - Tela da atividade de Treinar

Fonte: Banco de imagens dos autores

A Figura 3 permite a visualização do detalhe de uma linha completa segmentada:

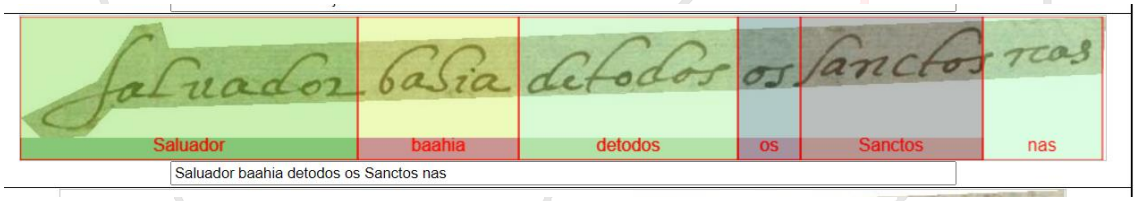

Figura 3 - Detalhe da segmentação de uma linha completa na etapa Treinar
Fonte: Banco de imagens dos autores

CONCLUSÕES

O processo de transcrição paleográfica de documentos manuscritos é uma tarefa complexa, demandando conhecimento especializado das áreas de paleografia e filologia, além de negável experiência. Se apropriar a expertise humana ao transcrever um texto histórico, fornecendo a algoritmos de aprendizado de máquina transcrições feitas por humanos, permite a criação de uma base de dados de treinamento para automatizar esse processo.

Treinar um computador através de mecanismos de aprendizado baseados em Inteligência Artificial, permite a construção de uma ferramenta capaz de reconhecer e decifrar textos escritos manualmente, possibilitando a implementação de corpora digitais de textos históricos de forma mais rápida e com menores custos humano e financeiro.

O LAPELINC TRANSCRIPTOR, como ferramenta de transcrição assistida e automática, agrega valor à área de Linguística de Corpus, auxiliando na construção de corpora de maneira mais simples e menos custosa, auxiliando ao pesquisador formador de corpora na construção de bases de dados textuais amplas e com maior regularidade na observância aos critérios de transcrição estabelecidos.

PALAVRAS-CHAVE: Transcrição paleográfica. Handwritten Text Recognition. Transcrição Manual Assistida por Computador.

REFERÊNCIAS

BORGES NUNES, E. **Abreviaturas Paelográficas Portuguesas**. Lisboa: [s.n.], 1981.

COSTA, B. S. Um Framework integrado para a criação, o gerenciamento e a disponibilização de corpora digitais em língua portuguesa. Projeto de doutoramento desenvolvido sob a supervisão de Jorge Viana Santos e Cristiane Namiuti, UESB (**Programa de Pós-Graduação em Linguística)**, Vitória da Conquista, 2019.

COSTA, B. S. et al. The Systematic Construction of Multiple Types of Corpora Through the Lapelinc Framework. **Computational Processing of the Portuguese Language**, Switzerland, p. 401-406, 2022. Disponivel em: <https://link.springer.com/chapter/10.1007/978-3-030-98305-5_37>.

COSTA, B. S.; SANTOS, J. V.; NAMIUTI, C. Uma proposta metodológica para a construção de corpora através de estruturas de trabalho: o LAPELINC Framework. **(No prelo) Revista Brasileira de Humanidades Digitais.**, Rio de Janeiro, n. 1, 2022.

GALVES, C. O Corpus Tycho Brahe: Um corpus sintaticamente anotado do português histórico. **RBBA**, Vitória da Conquista, v. 8, p. 181-204, julho 2019. ISSN 23161205. Disponivel em: <https://periodicos2.uesb.br/index.php/rbba/issue/view/339>.

MARTÍNEZ, T. M. **Paleografia y Diplomática**. Madrid: Universidad Nacional de Educacion a Distancia, 1988.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. P. F. **E-Dictor:** Novas perspectivas na codificação e edição de corpora de textos históricos In: Tania Shepherd; Tony Berber Sardinha; Marcia Veirano Pinto. (Org.). Caminhos da linguística de corpus. Campinas: Mercado das Letras, 2010.

PORTELA, M. Humanidades digitais: as humanidades na era da Web 2.0. Revista **Rua Larga (Revista da Reitoria da Universidade De Coimbra)**, Coimbra, v. 38, 2013.

PYTHON SOFTWARE FOUNDATION. Welcome to Python.org. **Python Language - Getting Started**, 2022. Disponivel em: <https://www.python.org/>. Acesso em: maio 2022.

SANTOS, Jorge Viana; BRITO, Giovane Santos. A Transposição de documentos manuscritos históricos jurídicos para o meio Digital através da Fotografia Digital: O Método Lapelinc. "E-Book do Congresso de Humanidades Digitais em Portugal: Construir pontes e quebrar barreiras na era digital – 2015". Lisboa: Universidade Nova de Lisboa, 2017.

SANTOS, J. V.; NAMIUTI, C. **Documentos Oitocentistas de Vitória da Conquista. Memória Conquistense. UESB/LAPELINC, Vitória da Conquista-Bahia/Brasil**, 2016. Disponivel em: <http://memoriaconquistense.uesb.br/websinc.>. Acesso em: 16 agosto 2019.

SANTOS, J. V.; NAMIUTI, C. De manuscritos históricos a corpora anotados: do Documento Fisico (DF) ao Documento Digital Imagem (DDI). Revista **A Cor das Letras**, Feira de Santana, 2016.

SCHEIDL, H. **Handwritten Text Recognition in Historical Documents**. Tese de doutoramento (Faculty of Informatics at the TU Wien), Vienna, 2018.

TENSORFLOW. Uma plataforma completa de código aberto para machine learning: TensorFlow, maio 2022. Disponivel em: <https://www.tensorflow.org/>.