LINGUÍSTICA COMPUTACIONAL E TRANSCRIÇÃO DE MANUSCRITOS: TRATAMENTO DE RASCUNHOS DE CARTAS PARA INSERÇÃO EM *SOFTWARES*[1]

Stephanne da Cruz Santiago (PPGEL/UEFS)[2]
stephannesantiago@gmail.com

Liliane Lemos Santana Barreiros (PPGEL/UEFS)[3]
lilianebarreiros@uefs.br

RESUMO

O presente trabalho propõe a discussão de possíveis critérios de adaptação para as transcrições dos rascunhos de carta do caderno *Farmácia São José*, do escritor baiano Eulálio Motta, visando a leitura em *softwares*. Sabe-se que a filologia, além de desenvolver pesquisas próprias, fornece *corpora* para várias áreas do conhecimento, como literatura, história, linguística. Os estudos lexicais, especialmente em uma perspectiva histórica, utilizam edições filológicas por fornecerem dados linguísticos confiáveis que, futuramente, servirão de estudo. Além disso, a linguística computacional vem, progressivamente, desenvolvendo e disponibilizando *softwares* que promovem o estudo mais automatizado da língua, beneficiando, entre outros, os estudos lexicais. Contudo, nem sempre os *softwares* conseguem realizar a leitura das transcrições dos textos, principalmente as de rascunhos. O *corpus* desta pesquisa trata-se da transcrição do rascunho de carta 'Meu caro Eudaldo: Saudações', presente no caderno *Farmácia São José*, de Eulálio Motta, que foi adaptada para ser analisada no *software AntConc*, de maneira que contemple a natureza do texto em questão. A discussão referente à linguística de corpus será pautada nas pesquisas de (BERBER SARDINHA, 2004; BARREIROS, 2017; OTHERO, 2006) e a discussão filológica terá por base (BARREIROS, 2013; 2015; CAMBRAIA, 2005).

Palavras-chave: Filologia. Linguística computacional. Estudos lexicais. Rascunho de carta. Eulálio Motta.

---

[1] Este trabalho faz parte da dissertação em andamento intitulada *"Meu caro Eudaldo": edição e estudo do vocabulário religioso dos rascunhos de cartas do caderno Farmácia São José, de Eulálio Motta*.

[2] Bolsista CAPES/CNPq. Mestranda do Programa de Pós Graduação em Estudos Linguísticos (PPGEL) pela Universidade Estadual de Feira de Santana. E-mail: stephannesantiago@gmail.com

3 Doutora em Língua e Cultura pela Universidade Federal da Bahia. Professora do Departamento de Letras e do Programa de Pós Graduação em Estudos Linguísticos (PPGEL) da Universidade Estadual de Feira de Santana. E-mail: lilianebarreiros@uefs.br

ABSTRACT

The present work proposes a discussion on possible adaptation criteria for transcriptions of drafts of the notebook *Farmácia São José*, by the Baiano writer Eulálio Motta, in order to read them in software. It is known that Textual Studies, in addition to developing applied research, provides corpora for various areas of knowledge, such as literature, history, linguistics. Lexical studies, especially from a historical perspective, use editions to provide linguistic data that, in the future, will serve for study. In addition, computational linguistics is progressively developing and making available software that promote a study more automated of the language, benefiting, among others, lexical studies. However, software are not always able to read text transcriptions, especially drafts'. The corpus of this research is the transcription of the draft letter 'Meu caro Eudaldo: Saudações', present in the notebook *Farmácia São José*, by Eulálio Motta, which was adapted to be analyzed in the *AntConc* software, in a way that it contemplates the nature of this type of text. The discussion regarding the corpus linguistics will be guided by the research of (BERBER SARDINHA, 2004; BARREIROS, 2017; OTHERO, 2006) and a Textual Studies discussion will be based on (BARREIROS, 2013; 2015; CAMBRAIA, 2005).

Keywords: Textual Studies. Computational linguistics. Lexical studies. Draft letter. Eulálio Motta.

1. Introdução

A edição de textos que representam a micro-história de uma comunidade é de grande interesse dos estudos históricos, antropológicos, literários, sociais e linguísticos. Numa perspectiva literária, esses tipos de textos promovem a valorização de escritores locais que não seriam amplamente conhecidos, caso não fossem editados e publicados. É importante, em muitas instâncias, a valorização de textos escritos em situações de 'pouco prestígio' pelas tradições clássicas, tendo em vista que o interesse de se explorar uma história não contada, uma literatura não canônica ou um grupo social pouco estudado é valorizado por pesquisadores que buscam uma constante renovação em suas áreas de pesquisa.

Considerando essa importância, são realizados no Núcleo de Estudos Interdisciplinares em Humanidades Digitais, no âmbito dos projetos de pesquisa *Edição das obras inéditas de Eulálio Motta* (UEFS/CONSEPE, Resolução Nº 128/2008 e Nº 070/2016) e *Estudos lexicais no acervo de Eulálio Motta* (UEFS/CONSEPE, Resolução Nº

137/2017), estudos com o acervo do escritor Eulálio de Miranda Motta, nascido no município de Mundo Novo, que se localiza na região do semiárido baiano.

O escritor arquivou seus manuscritos e outros documentos, compondo um respeitável acervo pessoal. Grande parte da documentação desse acervo corresponde a textos literários inéditos em forma de rascunhos, projetos e anotações para a composição dos textos. O acervo também conta com diários, rascunhos de cartas, fotografias, postais, coleções de jornais etc. Além desses documentos, também há 15 cadernos manuscritos, compostos em sua maioria por rascunhos, sendo documentos que revelam o processo de escrita do autor, tanto de textos literários, quanto não literários. Dentre eles, encontra-se o caderno *Farmácia São José*, composto majoritariamente de uma escrita cursiva, com exceção de uma colagem de um poema tiposcrito.

O *corpus* deste trabalho, o rascunho de carta 'Meu caro Eudaldo: Saudações', foi retirado do caderno *Farmácia São José*, que também conta com outros rascunhos de cartas, de textos literários, além de anotações pessoais, financeiras e cotidianas. Neste sentido, o presente artigo busca propor uma adaptação transcrição do rascunho de carta citado, para que ele possa ser processado no programa computacional de análise e estatística léxica *AntConc*, além de discutir brevemente a área da linguística computacional, bem como sua relação com a paleografia e a filologia, áreas essenciais para a realização de transcrições de textos.

2. Diálogos entre Linguística Computacional, Filologia e Paleografia

A linguística de corpus, de acordo com Berber Sardinha (2004), se ocupa da coleta e exploração de *corpora*, ou seja, de conjuntos de dados linguísticos textuais, e esses *corpora* têm como propósito servir para investigação de uma língua ou de variedades linguísticas, dedicando-se a exploração da linguagem por meio de evidências extraídas por computador. A linguística de corpus de base eletrônica, utilizando texto escrito, tem como marco inicial a composição do *Brown*, o primeiro *corpus* linguístico eletrônico, em 1964, contando com todas as adversidades que os tempos primórdios da informática poderiam oferecer. É claro que já havia constituição de *corpora* antes do computador e a mudança fundamental que veio junto ao surgimento da

era digital é que, anterior a ela, o tratamento dado aos *corpora* era outro, o levantamento e análise de dados eram feitos manualmente e agora pode-se contar com computadores para realizar ou otimizar essa tarefa.

Berber Sardinha (2004) também enfatiza que os *corpora* atuais foram moldados conforme *corpora* não-computadorizados, como o *Survey of English Usage* (*SEU*) que serviu de base para a constituição do *Brown* e de muitos outros que vieram em seguida. Havia problemáticas ao se tentar processar manualmente *corpora* linguísticos muito grandes, como é o caso do *SEU*, considerando a falibilidade humana que tornaria esses *corpora* muito passíveis de equívocos, portanto, pouco confiáveis. O uso de computadores calhou perfeitamente para otimizar e reduzir as margens de erros que seriam cometidos por mãos humanas ao manipularem *corpora* grandes e essa tecnologia passou a figurar fortemente nos ambientes de pesquisas linguísticas. Berber Sardinha (2004) afirma que, fora da esfera europeia, a linguística de corpus não se desenvolveu tanto e que, no Brasil, se encontra em estágio inicial, tendo a lexicografia como uma das áreas que mais realiza pesquisa em *corpus*, salientando o primeiro trabalho em estudos lexicais em perspectiva computacional no Brasil *Análise Computacional de Fernando Pessoa (Ensaio de Estatística Léxica)* (1969), de Maria Tereza Biderman, que incentivou diversos outros trabalhos nesse âmbito.

É importante pontuar que a linguística de corpus integra uma área maior, a linguística computacional. De acordo com Othero (2006, p. 342), a linguística computacional "[...] pode ser didaticamente dividida em duas subáreas: a Lingüística de Corpus e o Processamento de Linguagem Natural (PLN)". Considerando essa informação, cabe estabelecer a diferença entre a linguística de corpus e o PLN. Para Othero (2006), a linguística de corpus se preocupa com o trabalho a partir de *corpora* eletrônicos formados com base em amostras de linguagem natural e podem ser de diversas fontes, como por exemplo, língua falada, escrita, textos literários, jornalísticos, entre outros. Othero (2006) também salienta que nem sempre os trabalhos em linguística de corpus objetivam a produção de *softwares* ou aplicativos e focam nos estudos de fenômenos linguísticos e suas ocorrências em grandes amostras de dada língua ou de uma variedade, modalidade ou dialeto dela. Já o PLN se volta para o estudo da linguagem em busca do desenvolvimento de *softwares*, aplicativos, sistemas computacionais específicos, como tradutores e *parsers* (analisadores).

Cada vez mais buscam-se maneiras proveitosas para otimizar a atividade do pesquisador, como, por exemplo, o desenvolvimento de programas computacionais para levantamento, análises linguísticas (*parsers*) e até para a elaboração de obras lexicográficas. A linguística computacional tem se desenvolvido brilhantemente nesse aspecto, com a criação e constante atualização de programas que auxiliam os estudos linguísticos em vários níveis, como o E-Dictor, a nível morfossintático, e a nível lexical temos o *Antconc*, *FieldWorks Language Explorer* (*FLEx)*, *Unitex/GramLab*, *WordSmith Tools*, entre outros.

Além de fornecer programas para estudos linguísticos, a linguística computacional também vem buscando desenvolver programas que sirvam à filologia, realizando transcrições de textos. Editar um texto é uma tarefa que desafia qualquer pesquisador, pois ainda que experiente, ele precisará articular uma série de habilidades para executar a decodificação de um documento realizando sua transcrição, e assim, partir para a sua edição. A leitura de um texto envolve um conhecimento aprofundado da língua utilizada, considerando a ambientação temporal de sua escrita, além de mobilizar outras noções de paleografia, pois é preciso compreender a materialidade do texto para decifrá-lo. Outra questão importante é a sócio-história do texto, sendo necessário levar em conta aspectos que circundam o documento, como a sua produção, difusão e recepção, em busca de conhecer o máximo possível sobre o documento editado. Uma vez que o pesquisador tem consciência desses fatores, a transcrição e a edição do documento serão feitas com mais precisão.

Os aspectos positivos da linguística computacional na filologia são inúmeros, como indica Cambraia (2005):

> Na *elaboração* do texto, a informática possibilitou uma grande flexibilidade, pois programas de edição de textos oferecem ao usuário uma constelação de recursos *ad libitum* para elaborar os textos: escreve-se, apaga-se, substitui-se, muda-se a ordem, altera-se a formatação (mancha, fonte, cor, etc.) – tudo com simples toques sobre um teclado ou sobre um *mouse* (CAMBRAIA, 2005, p. 176).

Esta descrição feita por Cambraia (2005) nos remete a programas como o *Word*, com o qual vários filólogos realizam transcrições/edições dos documentos. Esse programa se enquadra na

categoria dos não-automáticos, em que o processo depende totalmente do editor, mas já existem programas que fazem transcrições de forma automática ou semi-automática. Nesses casos, o editor atua estabelecendo critérios para um plano de revisão dessas transcrições e futuras edições. É importante dizer que nem sempre os programas possuem uma interface amigável, ou seja, são de fácil manipulação aos menos experientes com o meio digital, o que acaba requerendo do editor um certo conhecimento de linguagem computacional.

Nesse cenário, alguns pesquisadores podem vir a questionar o papel do editor e sua importância, ou até a necessidade, em alguns casos mais extremos, nas transcrições de textos, visto que já existem programas que realizam essa tarefa. Biderman (2001) diz que "[a]tualmente a manipulação de textos via computador no domínio da Lingüística e das Humanidades transformou o editor de textos em uma banalidade". Apesar de serem muito úteis e facilitarem as transcrições esses programas estão longe de tornar banal o ofício do editor, uma vez que são muito passíveis de erros e não se aplicam a todos os tipos de escrita, além de ser uma visão reducionista do que é ser um filólogo.

Podemos tomar como exemplo as ferramentas de *Optical Character Recognition* (OCR), que existem em vários tipos de mecanismos, como apresenta Mendonça (2008), para tratamento de imagens, com caracteres padronizados (padrão ANSI), reconhecimento de números e até de escrita manuscrita. Mendonça (2008) propôs uma arquitetura de um sistema para tratamento e reconhecimento automático de documentos paleográficos por meio de OCR e realizou testes em dois mecanismos de OCR, o *Pytesser* e o *ABBYY Fine*, para o reconhecimento de diferentes escritas: padronizada (impressa), manuscrita moderna e paleográfica. Os sistemas de OCR cometeram um total de 11 erros para 11 caracteres, se apresentando como não satisfatórios para o uso. Nos outros tipos de escrita, eles se desempenharam melhor, o *Pytesser* obtendo 9 acertos de 11 em escrita manuscrita e 10 acertos de 11 em escrita padronizada; e o *ABBYY Fine* obteve 11 de 11 acertos nas escritas padronizada e manuscrita. Vale ressaltar que o *Pytesser* é um OCR gratuito e o *ABBYY Fine* é pago.

É importante salientar que Mendonça (2008) diz que os mecanismos de OCR para escrita manuscrita cursiva têm sido utilizados em situações em que as frases são curtas, de conteúdo controlado, em

que haja a possibilidade de validar o resultado com outra informação do mesmo documento. Em textos com períodos extensos e rasuras, características próprias de rascunhos, que apresentam processos de escrita, como é o caso do *corpus* desse trabalho, essa dificuldade é aumentada, sendo inviável propor uma transcrição por OCR, revalidando, novamente, a natureza essencial de um editor.

A relação entre a filologia, a paleografia, a linguística geral e a linguística computacional é de grande importância, uma vez que textos editados com rigor filológico e paleográfico fornecem dados linguísticos confiáveis e a linguística de corpus se vale desses dados para realizar suas análises. Além disso, o PLN utiliza esses dados linguísticos para expandir o desenvolvimento de seus *softwares* e *parsers*, visto que ele se vale da linguística de corpus para elaborar esses programas. O PNL também pode contar com a avaliação de filólogos e linguistas para o *upgrade* de seus programas, pois, de certa forma, é o usuário quem consegue visualizar melhor as limitações. Por fim, todos esses dados linguísticos que a linguística computacional consegue levantar, analisar e armazenar irão contribuir para os estudos em linguística teórica e aplicada, e estes irão fornecer conhecimento de língua, seja ele histórico, social, intrassistêmico, que auxiliarão o filólogo e o paleógrafo na prática de edição de textos e estudo da escrita, tudo isso resultando em um ciclo de relações simbióticas entre as áreas.

3. O software Antconc *e o rascunho de carta 'Meu caro Eudaldo: saudações'*

O *AntConc* é um *software* de levantamento e análise de *corpus*, desenvolvido por Anthony Lawrence, pesquisador da Faculdade de Ciências e Engenharia da Universidade de Waseda, Japão. Lawrence também desenvolveu outros programas que buscam o processamento da linguagem natural. O *AntConc* é um *software* gratuito executável em *Windows*, *Macintocsh* e *Linux*. Não se trata de um *parser* online, basta realizar o download uma única vez e ele será executado pelo computador, podendo inclusive ser executado por pendrive em outros computadores.

O *AntConc* está em sua versão 3.5.8 [4] e na própria página inicial do programa se encontra seu manual em diversas línguas. A versão do manual em português foi feita por Julia S Borba Gonçalves, associada ao *Laboratory of New Technologies in International Relations* - LANTRI (Laboratório de Novas Tecnologias e Relações Internacionais). Além disso, Lawrence mantém um diálogo aberto com pesquisadores que utilizam seus *softwares* e enviam sugestões para melhorias. Desde 2019, já estão na página inicial do *AntConc* algumas melhorias agendadas, como a aceitação de outros formatos de texto pelo programa, como PDF, por exemplo; redesenhar a arquitetura do banco de dados para lidar com corpora massivo, entre outras.

Geralmente, para se inserir um texto em programas computacionais é preciso utilizar o formato .txt, por não carregar a formatação original do texto, apenas os caracteres, além de que esse formato é facilmente aberto e lido por qualquer programa que realiza leitura de textos, sendo considerado um formato universal. Por ora, como o programa ainda não aceita o formato PDF ou DOC, é preciso utilizar um *converter* para converter os formatos dos textos. Neste sentido, utilizou-se para essa pesquisa o *AntFileConverter*, do mesmo criador do *AntConc*, que se encontra disponível no mesmo site.

O programa foi utilizado no processo de elaboração do *Vocabulário de Eulália Motta* por Barreiros (2017), em sua tese de doutorado, e este modelo de vocabulário foi adotado pelo projeto de pesquisa *Estudos lexicais no acervo de Eulália Motta* (UEFS/CONSEPE, Resolução Nº 137/2017). Dentre as diversas vantagens oferecidas pelo *AntConc*, pode-se destacar algumas ferramentas, como apresenta Barreiros (2017):

> Sua praticidade de uso possibilita a extração de listas de palavras (*Word List*), listas de concordâncias (*Concordance*) e de palavras-chaves (*KeyWord*), além de gerar gráficos com os dados analisados. Estas ferramentas são de grande relevância para o linguista, em especial, para o lexicógrafo, pois fornece o conjunto das combinações e das colocações que a palavra pode ter em um determinado corpus (BARREIROS, 2017, p. 220).

[4] Cf. Anthony (2014).

O *corpus* utilizado para realizar este trabalho faz parte do caderno *Farmácia São José*, que é composto por 149 folhas (reto e verso), porém a mancha escrita ocupa apenas 146 folhas. Possui capa dura azul com uma colagem de papel personalizada da Farmácia São José, em que consta o nome do autor e a data de 1º de outubro de 1940, localizada no centro da capa. Os textos do caderno estão em escrita cursiva, em sua maioria com tinta preta e a lápis grafite, contudo há passagens, geralmente correções, marcações e acréscimo de palavras, que se encontram feitas com lápis de cor azul e vermelha; além de um endereço escrito na parte superior do fólio 2v. com tinta azul. Mede 166mm de largura, 237mm de comprimento e 20mm de espessura. O caderno encontra-se bastante conservado e os textos não sofreram apagamentos com a ação do tempo. Foi feito com capa dura e encadernação artesanal brochura e as folhas estão em bom estado de conservação, pouco oxidadas.

Por se tratar de rascunhos manuscritos, a maioria deles possui rasuras, borrões, emendas, cancelamentos, e, por conta disto, a leitura do caderno foi dificultada, sendo em determinados contextos inviável. Por meio das marcas presentes no texto, podemos observar o processo de criação e desenvolvimento da escrita do autor. Existem diversos tipos de edições/transcrições filológicas, algumas com maior interferência do editor do que outras. Neste trabalho, segue-se os critérios estabelecidos por Barreiros (2013; 2015), que é o modelo utilizado no projeto de pesquisa *Edição das Obras Inéditas de Eulálio Motta* (UEFS/CONSEPE, Resolução Nº 128/2008 e Nº 070/2016):

(1) Indica-se o fólio;
(2) Enumera-se as linhas de 5 em 5 à margem esquerda da folha;
(3) Transcreve-se o texto como se encontra no original, interferindo apenas com marcadores genéticos estabelecidos;
(4) Corresponde a uma transcrição linearizada acomodando as rasuras, substituições, correções e acréscimos na sequência lógica do texto, não obedecendo a topografia do original;
(5) São mantidas as interpolações, os lapsos do autor, a ortografia, a acentuação, o uso de maiúsculas, a pontuação e registraram - se todas as correções, emendas, rasuras e acréscimos, através da utilização de símbolos;
(6) A rubrica do autor indica-se entre colchetes;

Quanto aos operadores genéticos utilizados nas transcrições:

(1) { } seguimento riscado, cancelado;

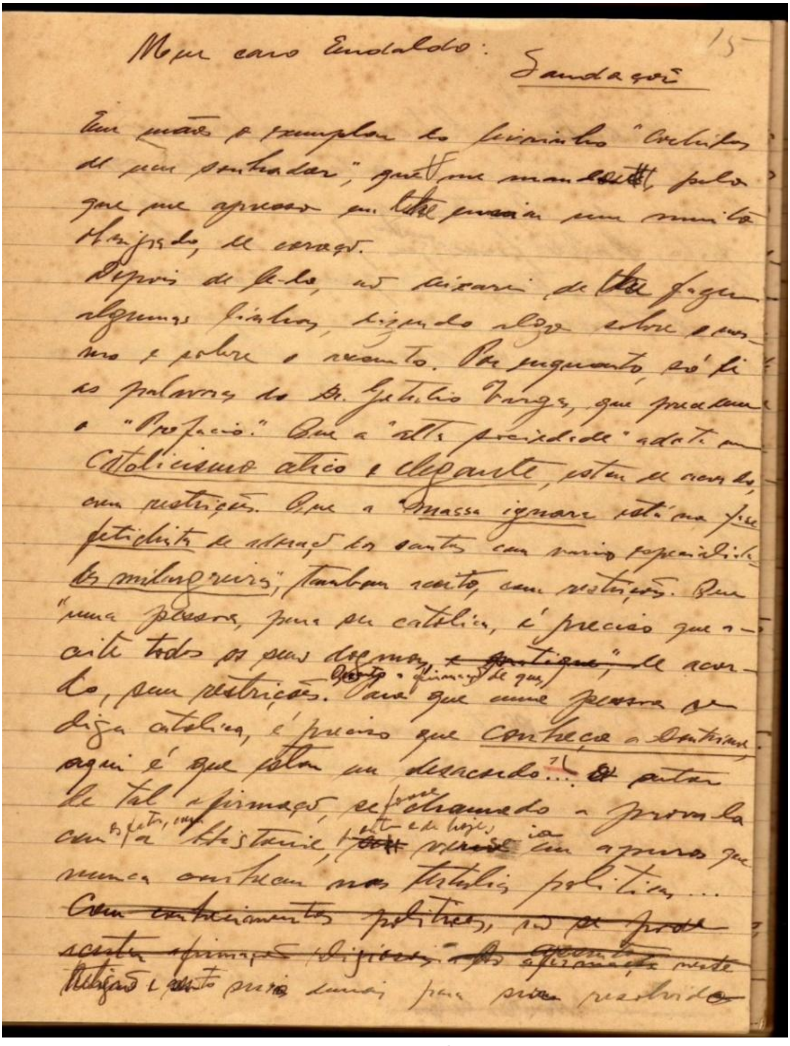
(2) {†} seguimento ilegível;
(3) {†} / \ segmento ilegível substituído por outro legível na relação {ilegível} /legível\;
(4) { } / \ substituição por sobreposição, na relação {substituído} /substituto\;
(5) { } [↑] riscado e substituído por outro na entrelinha superior;
(6) { } [→] riscado e substituído por outro na margem direita;
(7) { } [←] riscado e substituído por outro na margem esquerda;
(8) [ ] acréscimo no curso da linha;
(9) [↑] acréscimo na entrelinha superior;
(10) [↓] acréscimo na entrelinha inferior;
(11) [→] acréscimo na margem direita;
(12) [←] acréscimo na margem esquerda;
(13) [↑{ }] acréscimo na entrelinha superior riscado;
(14) [↑{†}] acréscimo na entrelinha superior ilegível;
(15) [↑{ }/ \] acréscimo na entrelinha superior riscado e substituído por outro na sequência;
(16) [↑{†} / \] acréscimo na entrelinha superior ilegível e substituído por outro na sequência;
(17) [↓{ }] acréscimo na entrelinha inferior riscado;
(18) [↓{†}] acréscimo na entrelinha inferior ilegível;
(19) [↓{ }/ \] acréscimo na entrelinha inferior riscado e substituído por outro na sequência;
(20) [↓{†} / \] acréscimo na entrelinha inferior ilegível e substituído por outro na sequência;
(21) [*↑] parte do texto localizada à margem superior indicada pelo autor através de seta, linha ou números remissivos;
(22) [*↓] parte do texto localizada à margem inferior indicada pelo autor através de seta, linha ou números remissivos;
(23) [*→] parte do texto localizada à margem direita indicada pelo autor através de seta, linha ou números remissivos;
(24) [*←] parte do texto localizada à margem esquerda indicada pelo autor através de seta, linha ou números remissivos;
(25) [*(f. ou p.)] parte do texto localizada em outro fólio ou página indicada pelo autor a partir de números e letras remissivos ou anotações. Nesses casos, o número do fólio ou da página aparece entre parênteses;
(26) / * / leitura conjecturada;
( ) intervenção do editor (acréscimos e informações).

Devido à grande quantidade de operadores que são utilizados nas transcrições, ela se enquadra no tipo de alto grau de interferência do editor, por buscar marcar o processo da gênese da escrita dentro do corpo do texto. Considerando que o *corpus* deste trabalho é um rascunho de carta que conta com várias rasuras e apresenta grandes marcas de processo de escrita, a transcrição mais adequada a ser aplicada a ele foi a

genética, pois ela busca marcar, por meio dos operadores genéticos já apresentados, os processos de escrita feitos pelo autor.

As alterações intravocabulares feitas pelo autor foram marcadas na transcrição, pois elas revelam o processo de escrita e escolhas linguísticas feitas por Eulálio Motta. Por conta disso, o processamento das transcrições no programa *AntConc* foi altamente comprometido, uma vez que ele não reconhece as unidades lexicais se estas estiverem com operadores genéticos dentro de sua estrutura, sendo necessária a adaptação dessas transcrições para que o processamento se dê de forma que contabilize todas as lexias. Pode-se conferir na figura 1 o fac-símile seguido da transcrição da primeira folha do rascunho de carta ‘Meu caro Eudaldo: Saudações’ (f. 9v):

Figura 1 - Fac-símile do rascunho de carta ‘Meu caro Eudaldo: Saudações’ (f. 9r)



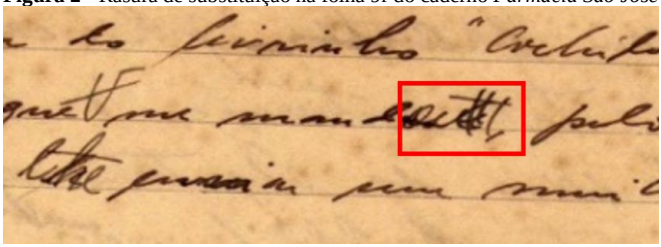
Fonte: Acervo de Eulálio Motta.

Meu caro Eudaldo:
Saudações
Em mãos o exemplar do livrinho “cochilos
de um sonhador”, que {V} me mand{astes}/ou\, pelo
que me apresso em {lhe}/te\ enviar um muito
obrigado, de coração.
Depois de le-lo, não deixarei de {lhe}/te\ fazer
algumas linhas, dizendo algo sobre o mês-
mo e sobre o assunto. Por enquanto, só li
as palavras do Dr. Getulio Vargas, que precedem
o “Prefacio”. Que a “alta sociedade” adota um
Catolicismo cetico e elegante, estou de acordo,
com restrições. Que a massa ignara está na fase
fetichista de adoração dos santos com varias especialida-
des milagreiras”, tambem aceito, com restrições. Que
“uma pessoa, para ser catolica, é preciso que a-
ceite todos os seus dogmas, {e pratique”,} de acor-
do, sem restrições. [↑Quanto a afirmação de que,] Para que uma pessoa se
diga catolica, é preciso que conheça a Doutrina[↓,]
aqui é que estou em desacordo...[*(14)] [↑1() {†}/O\ autor
de tal afirmação, se [↑fosse] chamado a prova-la
com [↑os fatos, com] a Historia, [↑de ontem e de hoje,] {ver} ver{ia}/-se-[↑ia]\ em
apuros que
nunca conheceu nas tertulias politicas...
{Com conhecimentos politicos, não se pode}
{acertar afirmação religiosas: {As}/Os\{afirmações}/[↑apresento]\ nes{s}/t\e}
{T{†}}/Religião\ e {†}/assunto\ seri{†}/o\ demais para ser{em} resolvid{os}/o\

(A página possui enumeração 15, feita a lápis, na parte superior direita. Há, na linha 20, um traço feito em lápis de cor vermelho, abaixo no número remissivo 1)

Como já foi dito, a transcrição genética é de alto grau de interferência do editor, por conta das marcações da gênese do texto de maneira integrada ao corpo do texto na transcrição. Uma vez que o texto é transcrito nesse modelo, é necessário fazer adaptações na transcrição para que o texto seja processado de maneira adequada por *softwares* e que contemple todos os dados linguísticos encontrados nele. Um exemplo de como essa leitura seria comprometida é o caso das unidades lexicais “mandaste” e “mandou”, que se encontram na linha 4 da folha 9r. No texto, ela apresenta uma rasura de substituição sobrescrita e na transcrição essa rasura é marcada dentro da estrutura da unidade, visto que ela não foi totalmente substituída, apenas a parte final de sua estrutura foi rasurada, resultando nesta forma editada: mand{astes}/ou\.

Figura 2 - Rasura de substituição na folha 9r do caderno *Farmácia São José*



Fonte: Acervo de Eulálio Motta.

Considerando que ambas as formas do verbo foram utilizadas pelo escrevente e ambas configuram como dado linguístico válido, é preciso que as duas formas sejam lidas e reconhecidas pelo programa de levantamento e análise lexical. Pensando nisso, foi feito um teste para observar como o *AntConc* realizaria a leitura dessa unidade. Num primeiro momento, foram feitas as limpezas já realizadas de praxe pela linguística de corpus, como apagamento da enumeração da folha, das linhas e das notas do editor, além de unir as separações silábicas feitas por quebra de linha. Então, converteu-se o texto para o formato .txt via *AntFileConverter* e inseriu-se no *AntConc*. Ao se realizar a busca pela forma "mandou" do verbo, o programa não reconheceu a unidade, pois estava com interferência dos operadores genéticos em sua estrutura, reconhecendo apenas a forma "mand", que é até onde havia a preservação da estrutura da unidade. Quando isso ocorre, compromete a estatística léxica, além de comprometer o levantamento das lexias.

Portanto, num segundo momento, foi feita a adaptação dessa transcrição para realizar o processo novamente. Para essa adaptação foram tomados os seguintes critérios:

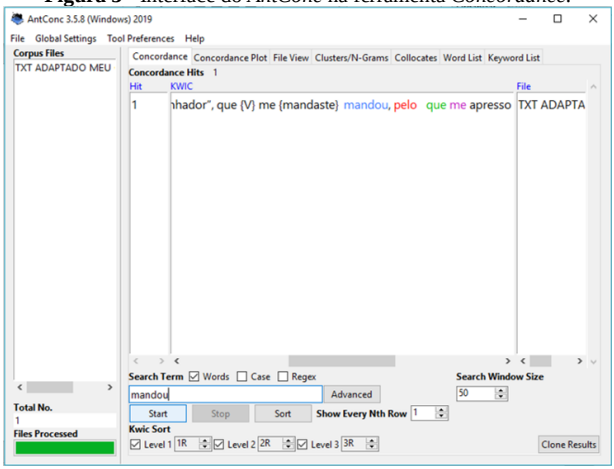
1) Foram retirados todos os sinais críticos do texto para facilitar a visualização dentro do programa;
2) Adotou-se o sinal { } para as formas excluídas pelo autor no texto e para as versões anteriores no processo da escrita, tanto para unidades lexicais, quanto para sentenças inteiras, limitando-se ao espaço da linha;
3) Transcreveu-se a forma da unidade lexical na íntegra dentro das chaves { } para que a lexia seja lida e contabilizada pelo programa;

Para ilustrar, podemos ver a versão adaptada da folha 9r da carta 'Meu caro Eudaldo: Saudações':

Meu caro Eudaldo:
Saudações
Em mãos o exemplar do livrinho “cochilos
de um sonhador”, que {V} me {mandaste} mandou, pelo
que me apresso em {lhe} te enviar um muito
obrigado, de coração.
Depois de le-lo, não deixarei de {lhe} te fazer
algumas linhas, dizendo algo sobre o mesmo
e sobre o assunto. Por enquanto, só li
as palavras do Dr. Getulio Vargas, que precedem
o “Prefacio”. Que a “alta sociedade” adota um
Catolicismo cetico e elegante, estou de acordo,
com restrições. Que a massa ignara está na fase
fetichista de adoração dos santos com varias especialidades
milagreiras”, tambem aceito, com restrições. Que
“uma pessoa, para ser catolica, é preciso que aceite
todos os seus dogmas, {e pratique”,} de acordo,
sem restrições. Quanto a afirmação de que, Para que uma pessoa se
diga catolica, é preciso que conheça a Doutrina,
aqui é que estou em desacordo... 1 {†} O autor
de tal afirmação, se fosse chamado a prova-la
com os fatos, com a Historia, de ontem e de hoje, {ver} {veria} ver-se-ia em apuros que
nunca conheceu nas tertulias politicas...
{Com conhecimentos politicos, não se pode}
{acertar afirmação religiosas: {As} Os {afirmações} apresento {nesse} neste}
{T{†}} Religião e {†} assunto serio demais para {serem} ser {resolvidos} resolvido

Desta maneira, ao inserir o texto no *AntConc*, o programa conseguiu realizar a leitura de todas as unidades. Tomando por base o exemplo anterior, ao buscar pela unidade “mandou” e “mandaste”, ambas foram indicadas pelo programa, não desprezando nenhum dado linguístico. Além disso, por meio do sinal crítico { } ainda é possível compreender a cronologia do texto ao fazer a busca pelo contexto da unidade com a ferramenta *Concordance*:

Figura 3 - Interface do *AntConc* na ferramenta *Concordance*.



Fonte: Elaborada pelo pesquisador.

É importante considerar que estas necessidades referem-se ao tipo de *corpus* utilizado para a pesquisa, uma vez que o rascunho apresenta processos de escrita. Caso se estabelecesse primeiro uma edição crítica do *corpus*, seriam perdidos dados linguísticos, uma vez que as unidades eleitas para comporem a versão final da edição seriam apenas a pertencentes a última vontade do autor, excluindo as outras unidades que ocorreram anteriores a elas. Como por exemplo, na versão final de uma edição crítica desse *corpus*, a unidade representativa seria "mandou", enquanto "mandaste" teria sua ocorrência desprezada.

Em busca de contemplar a natureza do tipo de texto que foi editado, o rascunho, e contemplar também todas as unidades lexicais presentes nele, estabeleceu-se a versão adaptada para que a leitura no programa fosse realizada satisfatoriamente e a lógica textual, sua cronologia, fosse preservada. Esse modelo de adaptação será empregado nos demais textos do caderno *Farmácia São José* que apresentarem as mesmas necessidades, em virtude da elaboração de um vocabulário religioso dos rascunhos de carta destinados a Eudaldo Lima, amigo de infância do escritor. Caso seja necessário, ao manipular as outras cartas e surgirem novas questões, pode haver a expansão ou modificação dos critérios utilizados para a adaptação apresentada neste trabalho.

4. Considerações finais

A pesquisa com o acervo de Eulálio Motta é promissora, principalmente no sentido linguístico, sociológico, cultural, literário e histórico, pois compreende-se a importância de estudar um escritor que represente, até certo ponto, o semi-árido baiano. Do ponto de vista filológico e paleográfico, o estudo da escrita de Eulálio Motta no caderno *Farmácia São José* possibilitou sua transcrição e o plano de revisão das transcrições, figurando como uma área essencial para a produção de transcrições e de edições, sendo um dos pontos de partida do editor, além de que a transcrição de rascunhos é uma tarefa árdua e cheia de peculiaridades, cabendo ao editor tomar diversas decisões importantes durante o processo. O modelo de transcrição que foi empregado ao *corpus* é singular, pois integrar ao corpo do texto toda a gênese do processo de escrita, considerando que, às vezes, os rascunhos se apresentam com inúmeras rasuras, tornando-se uma proposta bastante desafiadora para o editor.

Graças as contribuições da linguística computacional, mesmo que com suas limitações, trabalhar com análises linguísticas e constituição de *corpora* vem sendo mais viável e, cada vez mais, conta com o aprimoramento de programas para realizar os estudos. O diálogo entre as áreas é essencial para seus desenvolvimentos e a otimização de pesquisas. Por fim, a adaptação de transcrições de rascunhos para a seu processamento em programas computacionais ainda é uma discussão em andamento, visto que cada rascunho apresenta suas características peculiares que devem ser respeitadas pelo editor e pelo linguista.

REFERÊNCIAS BIBLIOGRÁFICAS

ANTHONY, Laurence. AntConc (Versão 3.5.8) [*Software* de Computador]. Tóquio, Japão: Universidade de Waseda. 2014. Disponível em: <http://www.laurenceanthony.net/>. Acesso em: 10 abr. 2020.

BARREIROS, Liliane L. S. *Vocabulário de Eulálio Motta*. 360f. Tese (Doutorado – Programa de Pós-Graduação em Língua e Cultura). Universidade Federal da Bahia, Instituto de Letras, Salvador, 2017.

BARREIROS, Patrício N. *O pasquineiro da roça:* a hiperedição dos panfletos de Eulálio Motta. Feira de Santana-BA: UEFS Editora, 2015.

BARREIROS, Patrício N. *O pasquineiro da roça:* edição dos panfletos de Eulálio Motta. 325f. Tese (doutorado em Letras e Linguística) – Instituto de Letras, Universidade Federal da Bahia, Salvador, 2013.

BERBER SARDINHA, Tony. *Linguística de Corpus*. Barueri, SP: Manole, 2004.

BIDERMAN, M. T. C. *Análise Computacional de Fernando Pessoa* (Ensaio de Estatística Léxica). Tese (Doutorado em Filologia e Língua Portuguesa). Faculdade de Filosofia, Letras e Ciências Humanas/USP, São Paulo, 1969.

BIDERMAN, Maria Tereza C. *Teoria linguística*: teoria lexical e linguística computacional. 2 ed. São Paulo: Martins Fontes, 2001.

CAMBRAIA, César Nardelli. *Introdução à crítica textual*. São Paulo: Martins Fontes, 2005.

MENDONÇA, Fábio Lúcio Lopes de. *Proposta de arquitetura de um sistema com base em OCR neuronal para resgate e indexação de escritas paleográficas do sec. XVI ao XIX*. 2008. 118 f. Dissertação (Mestrado em Engenharia Elétrica) - Universidade de Brasília, Brasília, 2008.

OTHERO, G. A. *Lingüística Computacional:* uma breve introdução. Letras de Hoje, v. 144, 2006.

UEFS/CONSEPE. Resolução CONSEPE Nº 137/2017. Aprova o Projeto de Pesquisa *Estudos lexicais no acervo de Eulálio Motta*, sob a coordenação da Profa. Dra. Liliane Lemos Santana Barreiros, do Departamento de Letras e Artes, desta Universidade. Feira de Santana-BA: D.O.E., 12 dez. 2017.

UEFS/CONSEPE. Resolução CONSEPE No 070/2016. Aprova o Projeto de Pesquisa Edição das Obras Inéditas de Eulálio de Miranda Motta (IV Etapa), sob a coordenação do Prof. Dr. Patrício Nunes Barreiros, do Departamento de Letras e Artes, desta Universidade, financiado pela FAPESB. Feira de Santana-BA: D.O.E., 2 set. 2016.

UEFS/CONSEPE. Resolução CONSEPE No 128/2008. Aprova o Projeto de Pesquisa Edição das Obras Literárias Inéditas de Eulálio de Miranda Motta, sob a coordenação do Prof. Patrício Nunes Barreiros, do Departamento de Letras e Artes, desta Universidade. Feira de Santana-BA: D.O.E., 27 ago. 2008.