DOI: 10.14393/DL51-v16n3a2022-10

Transcrição e anotação de dados linguísticos usando as ferramentas ELAN e LancsBox

Transcription and linguistic data annotation using ELAN and LancsBox tools

Marta Deysiane Alves Faria SOUSA[*]
Victor Renê Andrade SOUZA[**]

RESUMO: Objetiva-se com este trabalho demonstrar como as ferramentas de transcrição de dados ELAN 5.9 (2020) e de análise de *corpora* LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020) têm contribuído para a transcrição de entrevistas sociolinguísticas realizadas no escopo do Grupo de Estudos em Linguagem, Interação e Sociedade (GELINS) da Universidade Federal de Sergipe, bem como para extração automatizada de fenômenos linguísticos variáveis. Para tanto, apresenta-se as normas pelas quais as entrevistas são transcritas, formas de utilizar o ELAN 5.9 (2020) para transcrição, e por fim, a maneira de fazer a etiquetagem morfológica dos dados e buscas por fenômenos variáveis nos dados de fala utilizando-se a ferramenta LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020). As duas ferramentas têm se mostrado eficientes para uma transcrição alinhada com áudio, para anotação morfológica e buscas automáticas em grandes volumes de

ABSTRACT: This study aims at demonstrating how the transcription tool ELAN 5.9 (2020) and the corpus analysis LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020) have been contributing to transcribe sociolinguistics interviews made at Grupo de Pesquisa em Linguagem, Interação e Sociedade (GELINS) as well as to the automatically extract variable linguistic phenomena. In order to do that, norms through which the interviews are transcribed, ways of using ELAN 5.9 (2020) for transcription, and the way to morphologically tag data and to search speech data for linguistic variable phenomena using LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020) are presented. The two tools have been useful to an aligned transcription, a morphological annotation, and an automated search on large amounts of speech data. This text contributes to the exploration of tools that enables a faster and more accurate transcription of speech data as well as more automated searches on large

[*] Doutoranda em Letras, Universidade Federal de Sergipe. ORCID: https://orcid.org/0000-0002-0480-0422. professoramarta2018@outlook.com
[**] Mestre em Letras, Universidade Federal de Sergipe. ORCID: https://orcid.org/0000-0003-0392-2839. victor.andrade573@gmail.com

textos orais. Este texto contribui para exploração de ferramentas que permitam uma transcrição mais rápida e acurada de dados orais bem como buscas mais automatizadas de grandes volumes de dados.

amounts of data.

PALAVRAS-CHAVE: Dados orais. Sociolinguística. ELAN. LancsBox. Bancos de dados linguísticos.

KEYWORDS: Speech data. Sociolinguistics. ELAN. LancsBox. Linguistic Databases.

1 Introdução[1]

Atualmente, no Brasil, bancos de dados linguísticos[2], tanto de fala quanto de textos escritos, têm fornecido subsídios a análises descritivas de variedades do português a partir de múltiplas perspectivas de análise linguística (FREITAG *et al.*, 2021) e têm servido também como recurso para a produção de material didático a programas de ensino de língua materna (FREITAG, 2013).

No entanto, a constituição de banco de dados sociolinguísticos é tarefa dispendiosa, requer tempo e recursos financeiros (GONÇALVES; TENANI, 2008;

[1] Este texto é resultado de ações conjuntas do Grupo de Estudos em Linguagem, Interação e Sociedade (GELINS) para o desenvolvimento e aplicação de normas de transcrição e anotação morfológica. Como ações, o GELINS desenvolve uma série de cursos extensionistas voltados à capacitação de recursos humanos para realizar transcrições conforme o sistema de normas adotado, como por exemplo, o curso de extensão "Ferramentas computacionais para a documentação linguística". Ademais, o grupo tem oferecido cursos sobre etiquetagem e buscas automáticas em dados de fala, como o "Ciclo de estudos em etiquetagem de dados linguísticos" e o "Corpus linguístico: do som para letra", este último em parceria com o grupo Corpus Infantil Longitudinal (CIL) da Universidade Federal dos Vales do Jequitinhonha e Mucuri.

[2] Conforme Cianconi (1987, p. 54, grifo nosso), "[na] literatura na área de Informação, pode-se conceituar **Base de Dados** como um conjunto de dados interrelacionados, organizados de forma a permitir recuperação de informações. **Banco de Dados**, embora frequentemente encontrado como sinônimo de base de dados, pode ser visto como um conjunto de bases de dados". Acreditamos que os bancos de dados sociolinguísticos brasileiros se adequam à terminologia de Cianconi (1987) por se tratarem de conjuntos de amostras de fala que se relacionam e são organizados possibilitando a recuperação de informações, por exemplo, tipo de comunidade, tipo de amostra, tipo de coleta (interação/ entrevista), além de ser um termo empregado amplamente pelos pesquisadores da área de Sociolinguística para designar seus conjuntos de dados, a exemplo do Banco de Dados Iboruna, do Varsul, do Projeto SP2010 e do Banco de Dados Falares Sergipanos.

FREITAG, 2013). A compilação de amostras de fala para a construção de bancos de dados sociolinguísticos envolve os seguintes procedimentos: a coleta dos dados propriamente, com a seleção de informantes e gravação de material linguístico; a transcrição, anotação e o armazenamento dos dados; e a disponibilização para a comunidade científica.

Tendo em vista os objetivos dos bancos de dados sociolinguísticos, a padronização dos procedimentos metodológicos é um elemento fulcral, por permitir uma análise mais acurada e automatizada e por possibilitar a comparabilidade com outros bancos já constituídos, pois, como defendem Freitag, Martins e Tavares (2012, p. 918), "[...] só a padronização dos procedimentos metodológicos permitirá a realização de estudos contrastivos entre as variedades [...]", e, por consequência, uma descrição mais ampla do Português Brasileiro.

Para atender a esses desafios, diversas ações têm sido desenvolvidas no âmbito nacional. O simpósio *Descrição linguística: gestão de dados linguísticos no Abralin Ao Vivo*[3] discutiu questões latentes relacionadas ao gerenciamento de dados linguísticos (FREITAG, *et al*., 2021; CARDOSO, 2020). Dentre os desafios apresentados, está a busca por ferramentas mais adequadas para a vitalidade dos conjuntos de dados linguísticos.

Duas das etapas da constituição de banco de dados linguísticos que requerem particular atenção são a de transcrição e anotação dos dados. A partir dessa demanda específica, neste artigo, de natureza procedural, apresentamos duas ferramentas computacionais que auxiliam no processo de transcrição e anotação de dados linguísticos: o ELAN 5.9 (2020) e o LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020). Já existem tutoriais sobre essas ferramentas, mas a maioria é escrita em inglês (ROSENFELDER, 2011; NAGY; MEYERHOFF, 2015; TACCHETTI, 2017; STARTING, 2017). Nosso objetivo é, portanto, apresentá-las em português com a

[3] Disponível em: https://aovivo.abralin.org/lives/gestao-de-dados-linguisticos/.

finalidade de popularizar e incentivar o uso dessas ferramentas na constituição dos bancos de dados sociolinguísticos brasileiros.

Para tanto, em um primeiro momento, descrevemos as normas de transcrição ortográfica adotadas pelo Banco de Dados Falares Sergipanos (FREITAG, 2013), chanceladas pelo Grupo de Estudos em Linguagem, Interação e Sociedade (GELINS), na constituição da amostra *Deslocamentos*[4] e instrumentalizamos a transcrição de entrevistas sociolinguísticas no *software* ELAN 5.9 (2020), seguindo o protocolo de transcrição do referido banco. Na sequência, apresentamos os procedimentos de anotação morfológica dos dados que constituem a referida amostra. Para tanto, faremos uma breve introdução do que seja a anotação morfológica e, em seguida, indicaremos como utilizar a ferramenta LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020) para fazer tal anotação e como realizar buscas por fenômenos linguísticos de forma automática.

2 Transcrição de dados linguísticos

Transcrever significa "transpor o discurso falado, da forma mais fiel possível, para registros gráficos mais permanentes" (PAIVA, 2003, p. 135) de modo padronizado, a fim de automatizar análises (OUSHIRO, 2014). Em Sociolinguística, o procedimento de anotação dos *corpora* tem contribuído com o processo de compartilhamento de dados, automatização das análises e, sobretudo, com o

[4] A amostra *Deslocamentos* faz parte do projeto *Como fala, lê e escreve o universitário?* (FREITAG, 2018), coordenado pela Profª Drª Raquel Meister Ko. Freitag, e integra o Banco de Dados Falares Sergipanos (FREITAG, 2013), sendo composta por 60 entrevistas sociolinguísticas realizadas com estudantes da Universidade Federal de Sergipe (UFS) *campus* Prof. José Aloísio de Campos, localizado no município de São Cristóvão, Sergipe, Brasil. A amostra é estratificada conforme os parâmetros do Banco de Dados Falares Sergipanos (FREITAG, 2013) no que diz respeito à mobilidade acadêmica, quanto a sexo/gênero, tempo de curso e deslocamento geográfico dos estudantes.

reaproveitamento por outros pesquisadores – medidas que estão em concomitância com as diretrizes do movimento Ciência Aberta[5].

O procedimento de transcrição requer recursos tecnológicos adequados, normas de transcrição estabelecidas e uma equipe de anotadores capacitada. Cada banco de dados adota normas, mais ou menos fixas[6], para transcrição, de modo a alcançar uma padronização (mesmo relativa) que atenda aos interesses do grupo de pesquisa e/ou pesquisador – e também à comunidade científica – e que permita uma busca automatizada e mais acurada dos fenômenos a serem investigados. Na próxima seção, descrevemos as normas adotadas pelo GELINS para a transcrição dos dados de fala coletados, seja por meio de entrevistas sociolinguísticas, como no caso da amostra *Deslocamentos*, foco deste estudo, seja por outros métodos, como interações conduzidas.

2.1 Normas de transcrição do Banco de Dados Falares Sergipanos

As normas de transcrição adotadas para a amostra *Deslocamentos* do Banco de Dados Falares Sergipanos (FREITAG, 2013) foram construídas com base nos sistemas de transcrição de outros bancos de dados linguísticos já existentes. A transcrição das gravações é realizada no nível da ortografia. Freitag (2013, p. 162-163) esclarece o porquê da adoção de um sistema de transcrição ortográfico:

> O processo de transcrição é baseado na audição impressionística do áudio, mas, diferentemente de outros bancos de dados que seguem o protocolo de coleta da entrevista sociolinguística, que adotam uma transcrição de base fonética adaptada, no banco de dados Falares Sergipanos seguimos um

[5] Exemplos de ações que incorporam as diretrizes da Ciência Aberta (*Open Science*) podem ser encontradas no site da Associação Brasileira de Linguística (ABRALIN): https://www.abralin.org/site/ciencia-aberta/; e são discutidas por Freitag *et al.* (2021).

[6] O não estabelecimento de regras pode acarretar modelos de transcrição heterogêneos e dificultar a busca automatizada de fenômenos.

> modelo de transcrição tomando como referência os princípios ortográficos da escrita do português, o que nos permite fazer uso das ferramentas computacionais da Linguística de Corpus – como os softwares concordanceadores – e tornar o trabalho de levantamento de dados mais otimizado. [...] É possível buscar aproximações entre as áreas, com a padronização ortográfica de corpora sociolinguísticos e o desenvolvimento de etiquetas XML para codificar a informação social de modo a ser processável por softwares concordanceadores. Para possibilitar a manipulação dos dados, a transcrição ortográfica padrão é mais eficiente; marcações fonológicas podem ser feitas após a seleção dos contextos pela transcrição ortográfica, com tratamentos acústicos em softwares específicos, garantindo resultados mais acurados.

Observa-se, então, que a transcrição ortográfica se torna importante devido ao fato de ela permitir que ferramentas computacionais, como o LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020), ou o AntConc (ANTONY, 2020), desenvolvidos para a Linguística de *Corpus* para o processamento de grandes volumes de dados, sejam também utilizadas na Sociolinguística.

As normas de transcrição adotadas pelo Banco de Dados Falares Sergipanos seguem detalhadas abaixo. No quadro 1, são apresentadas as ocorrências, os sinais empregados em cada uma delas e uma exemplificação da notação.

Quadro 1 – Normas de transcrição ortográfica adotadas pelo Banco de Dados Falares Sergipanos.

Ocorrência	**Sinal**	**Exemplo**
Qualquer tipo de pausa, substituindo todos os sinais específicos da língua escrita que desempenham tal função: ponto e vírgula, ponto final, dois pontos e vírgula	...	Não é o que era antigamente...onde a gente não...sabia de nada
Interrogação	?	Sabe o que é?
Comentário do transcritor sobre o que está acontecendo no ambiente	(( ))	((RISOS)) ((PIGARRO))
Estímulo do locutor	(est)	Olhe aqui a marquinha (est) olhe ela aqui
Hesitação do locutor	(hes)	Foi (hes) uma brincadeira bem interessante

Truncamento de palavra	-	Come-começou
Nomes próprios, profissões, nomes de cursos, filmes	Iniciais maiúsculas	...fui a Petrópolis uma vez...
Palavras não dicionarizadas	<< >>	<<bora>> <<afugiado>>
Discurso direto	" "	Eu saio pra apresentar trabalho fora eles têm orgulho "ah ela saiu pra outro estado tá apresentando trabalho da universidade" então de certa forma isso é um apoio...
Números	Por extenso	Eu tenho vinte e oito anos
Incompreensão do que ouviu	( )	
Hipótese do que ouviu	(hipótese)	Ter que estudar lá no campus de São Cristóvão ía re- ía reque- requerer da minha (como a associação) que eu teria que pagar todos os meses
Onomatopeias e siglas	Caixa alta	A questão do incentivo de participação de eventos porque assim de eventos por exemplo o OCMEA ela é incentivado por todos os professores

Fonte: Elaboração própria com base nas normas adotadas pelo Banco de Dados Falares Sergipanos (FREITAG, 2013).

Sobre essas normas, é relevante fazer alguns esclarecimentos. Nas transcrições, tem-se como objetivo representar a fala o mais fidedignamente possível. No caso do Banco de Dados Falares Sergipanos, isso é feito de acordo com as normas do acordo ortográfico vigente. Por isso, em realizações de verbos no modo infinitivo, por exemplo, ainda que o /r/ não seja percebido de oitiva, grafamo-lo na transcrição. As variações fonológicas não são consideradas, ou seja, caso ocorra, por exemplo, monotongação de ditongo decrescente oral em palavra como "cadeira", não consideramos o apagamento da semivogal, transcrevendo a palavra conforme a grafia dicionarizada.

As variações morfossintáticas, por outro lado, são transcritas. Isso significa que se um falante disser "num", "pra", "tava", "tou", "ocê", "cê" e demais reduções, a transcrição é realizada respeitando a pronúncia do informante. É válido destacar que as representações ortográficas de casos de variação são transcritas desde que elas não estejam no nível fonológico.

Além dessas marcas relacionadas à grafia das palavras, elementos suprassegmentais também são considerados e transcritos. Assim, é norma do Banco de Dados Falares Sergipanos que, por exemplo, pausas preenchidas[7] sejam transcritas como "ah", "eh", "uh".

Na próxima seção, trataremos da transcrição dos dados utilizando o *software* ELAN 5.9 (2020).

2.2 Transcrevendo dados orais com o ELAN 5.9 (2020)

Explicitadas as normas de transcrição adotadas pelo Banco de Dados Falares Sergipanos, nesta seção, apresentamos o procedimento de transcrição de entrevistas sociolinguísticas utilizando o *software* ELAN 5.9, seguindo os critérios estabelecidos pelo referido banco.

2.2.1 O ELAN 5.9 (2020)

Para a sincronização do áudio e da transcrição, adotamos o *software* ELAN 5.9 (2020). Esse *software* é utilizado para transcrever/anotar gravações de áudio e vídeo e foi desenvolvido na Holanda pelo Instituto de Psicolinguística Max Planck. É um transcritor gratuito (https://archive.mpi.nl/tla/elan/download), compatível com Windows, Mac IOS ou Linux. A instalação é autoexplicativa e o *software* apresenta interface intuitiva, podendo ser utilizado em língua portuguesa. Com o ELAN 5.9

[7]A noção de pausa preenchida adotada é discutida por Freitag, Pinheiro e Silva (2017).

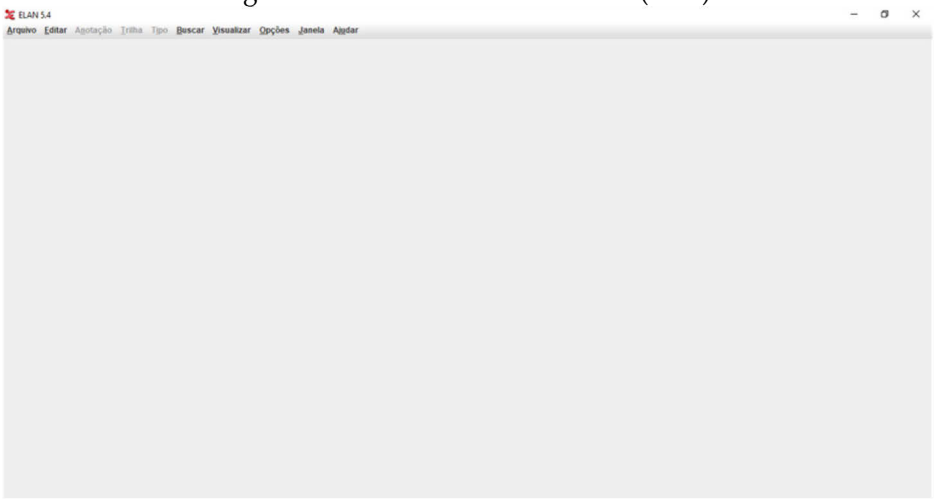
(2020) é possível realizar a sincronização entre o arquivo de mídia e a anotação, a criação de múltiplas trilhas, e a exportação dos arquivos para múltiplos formatos (txt; TextGrid; HTML, entre outros) (OUSHIRO, 2014; NAGY; MEYERHOFF, 2015), o que viabiliza a compatibilidade com outros softwares; e, principalmente, automatiza a busca de fenômenos dentro de um *corpus*.

Recomendamos a leitura do tutorial de Oushiro (2014), sobre transcrição de entrevistas sociolinguísticas com o ELAN 5.9 (2020), que compõe o livro "Metodologia de coleta e manipulação de dados em Sociolinguística" (FREITAG, 2014). Oushiro (2014) ensina como fazer a instalação e apresenta as principais funcionalidades do *software*. Para o escopo desse texto, no entanto, interessa-nos demonstrar/instruir a transcrição de entrevistas sociolinguísticas seguindo as normas adotadas pelo Banco de Dados Falares Sergipanos (FREITAG, 2013).

2.2.2 Fluxo de trabalho no ELAN 5.9 (2020) seguindo as normas de transcrição do Banco de Dados Falares Sergipanos

Devidamente coletados e organizados os arquivos de áudio, o primeiro passo da transcrição é abrir o *software*. A tela inicial do ELAN 5.9 (2020) é a janela em branco a seguir (Figura 1).

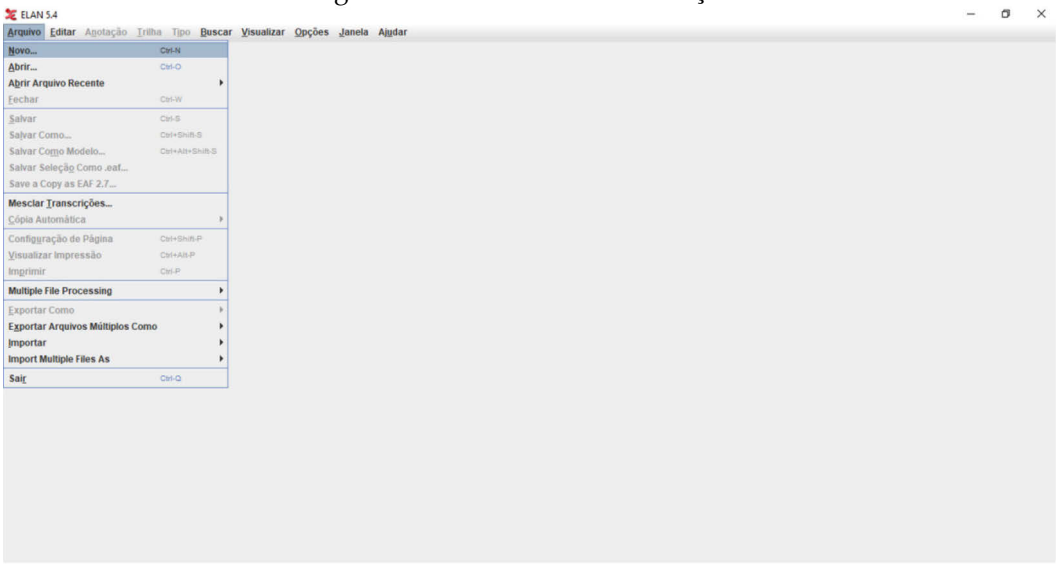
Figura 1 – Tela inicial do ELAN 5.9 (2020).



Fonte: extraída do software ELAN 5.9 (2020).

Na sequência, devemos criar uma nova transcrição. Para isso, clicamos em *Arquivo* > *Novo*, como ilustra a Figura 2, a seguir.

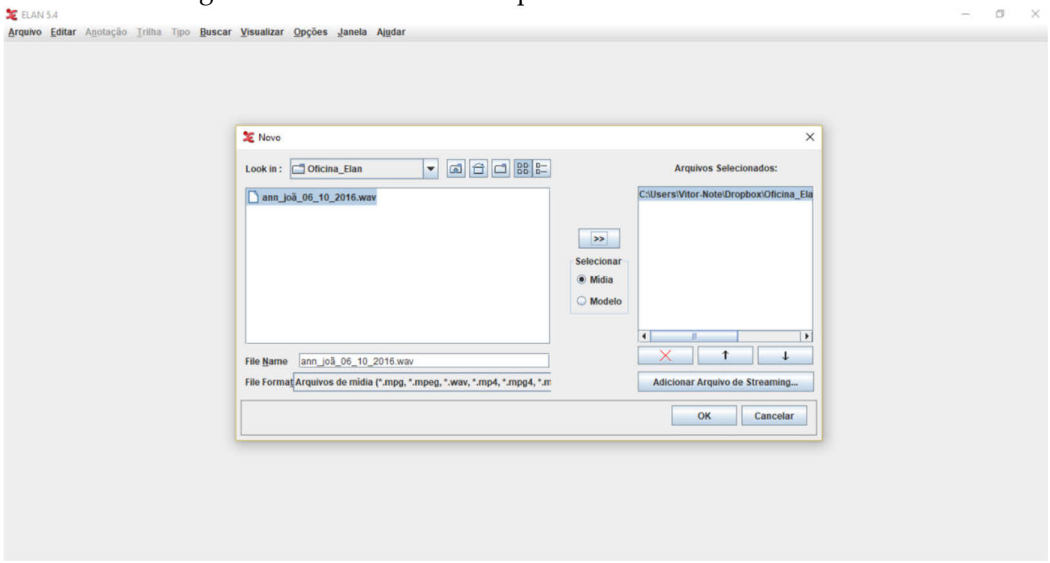
Figura 2 – Criando nova transcrição.



Fonte: extraída do software ELAN 5.9 (2020).

Uma nova janela se abrirá (Figura 3) para selecionar o arquivo de áudio a ser anotado. Do lado esquerdo, selecionamos o arquivo de áudio a ser transcrito e clicamos em >> . O arquivo aparecerá em *Arquivos Selecionados*. Feito isso, clicamos em *OK*.

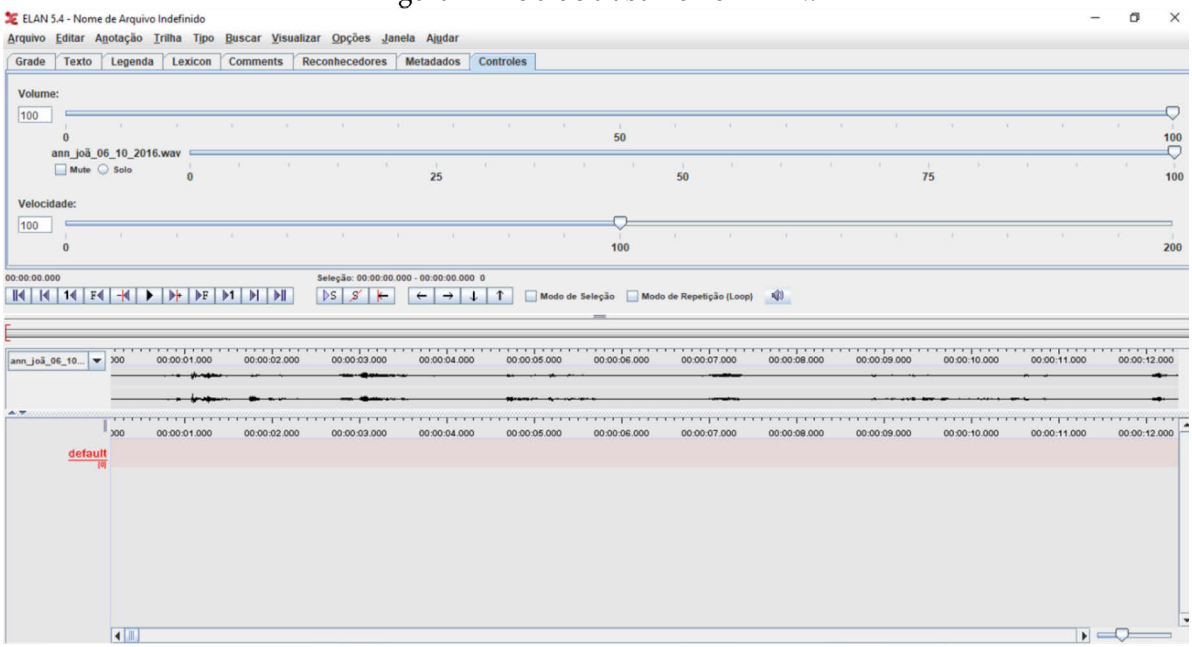
Figura 3 – Selecionando o arquivo de áudio a ser transcrito.



Fonte: extraída do software ELAN 5.9 (2020).

Após seleção, o arquivo de áudio será aberto na janela principal do ELAN 5.9 (2020), como ilustra a Figura 4, abaixo. Para conhecer detalhadamente as principais funcionalidades da janela principal do ELAN 5.9 (2020), bem como seus atalhos, remetemos novamente a Oushiro (2014).

Figura 4 – Tela de trabalho no ELAN.

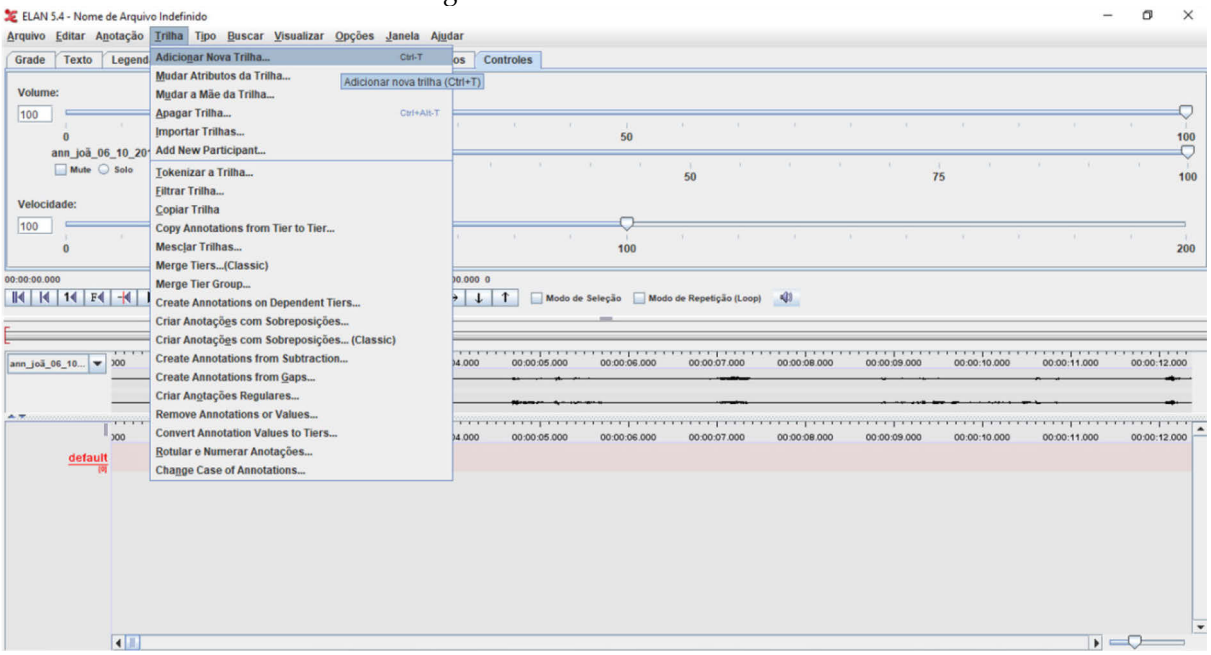


Fonte: extraída do software ELAN 5.9 (2020).

Com áudio a ser transcrito aberto, devemos criar as trilhas de anotação, ou seja, o espaço onde ficará o texto transcrito com as respectivas marcações de tempo. No caso de documentação sociolinguística, é recomendado uma trilha por participante da interação. Assim, em uma entrevista sociolinguística, por exemplo, teremos uma trilha para o documentador e uma trilha para o informante; em interações conduzidas, por sua vez, teremos uma trilha para cada interactante.

Para adicionar trilhas, clicamos em *Trilha* > *Adicionar nova trilha*, conforme Figura 5.

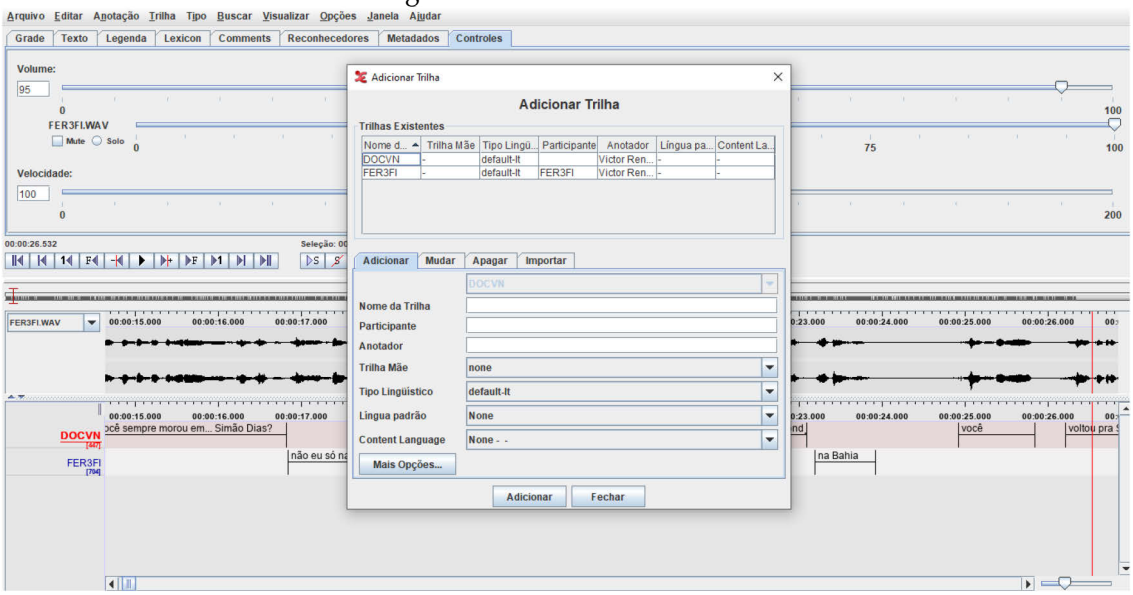
Figura 5 – Adicionando trilhas.



Fonte: extraída do software ELAN 5.9 (2020).

Uma nova janela se abrirá (Figura 6), com uma lista das trilhas na parte superior e abas e campos na parte inferior, onde a aba *Adicionar* estará em destaque.

Figura 6 – Adicionando trilhas.



Fonte: extraída do software ELAN 5.9 (2020).

No Banco de Dados Falares Sergipanos, as trilhas são identificadas de modo padronizado. Na janela da Figura 6, inserimos as seguintes informações para cada trilha:

Nome da trilha: intitulamos a trilha do informante com as três primeiras letras do nome do falante, o número do deslocamento, o sexo/gênero, e o período (início ou fim) do curso, no caso de estudante universitário, que foi a condição dos informantes da amostra *Deslocamentos*. Se o informante se chamar José, e ele pertencer ao deslocamento 4, for do sexo/gênero masculino e estiver no final do curso, a trilha chama-se JOS4MF. A trilha do documentador é intitulada de DOC + Inicial do primeiro e segundo nome do entrevistador. Assim, a título de exemplificação, se a documentadora se chamar Maria Santos, intitulamos a trilha da seguinte maneira: DOCMS.

Participante: no campo participante, adicionamos o código relativo ao informante ou ao documentador, a depender da trilha;

Anotador: nome completo do transcritor/anotador.

Criadas as trilhas necessárias, clicamos sobre a trilha *default* do ELAN 5.9 (2020) e a apagamos, clicando em *Apagar*.

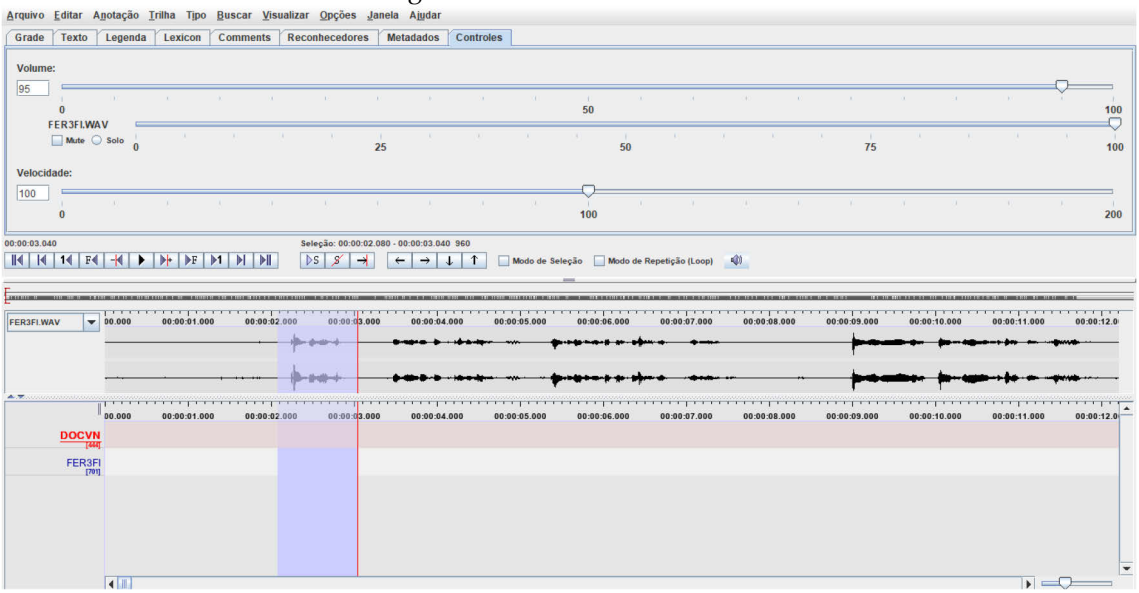
Em seguida, o arquivo está pronto para iniciar o processo de transcrição. Adotamos – e sugerimos – o fluxo de trabalho proposto por Oushiro (2014, p. 128):

> 1) ligue o Modo de Seleção (e o Modo de Repetição, se quiser) clicando sobre as respectivas caixas de seleção acima da onda sonora;
> 2) o cursor estará no início do arquivo. Pressione [Shift] + [Space] para começar a tocar. À medida que o cursor se mover, ele selecionará o intervalo tocado;
> 3) deixe o cursor tocar até antes da primeira sentença. Pressione [Shift] + [Space] para pausar a gravação;
> 4) use os controles de navegação de mídia para mover o cursor exatamente para o ponto em que você quer começar um nova anotação.

As teclas [Ctrl] + [←] e [Ctrl] + [→] voltam/adiantam 1 segundo e as teclas [Shift] + [←] e [Shift] + [→] voltam/adiantam um frame (provavelmente esses últimos serão mais utilizados);
5) desfaça a seleção atual com [Esc].
6) inicie o playback novamente com [Shift] + [Space]. O cursor agora vai começar a selecionar o trecho de fala.
7) pause o playback logo após o final da fala com [Shift] + [Space]. Se necessário, use os controles de navegação de mídia novamente para mover o cursor exatamente para o ponto final da nova anotação. Você pode ouvir o trecho selecionado com os comandos [Ctrl] + [Space];
8) ative a trilha do falante correspondente onde você quer a nova anotação usando [Ctrl] + [↑]/[↓];
9) se você ligou o Loop Mode, pressione [Ctrl] + [Space] para começar a tocar a seleção;
10) pressione [Shift] + [Enter] para criar uma nova anotação na trilha ativa;
11) após transcrever a fala, pressione [Enter] para salvar a transcrição;
12) pressione [Ctrl] + [Space] para parar o loop da seleção atual;
13) pressione [Shift] + [Space] para recomeçar o playback da atual posição do
cursor;
14) repita os passos (3) a (13).

Seguindo esse procedimento, primeiro reproduzimos o áudio e ajustamos o cursor até os limites da produção linguística, como ilustra a Figura 7, a seguir.

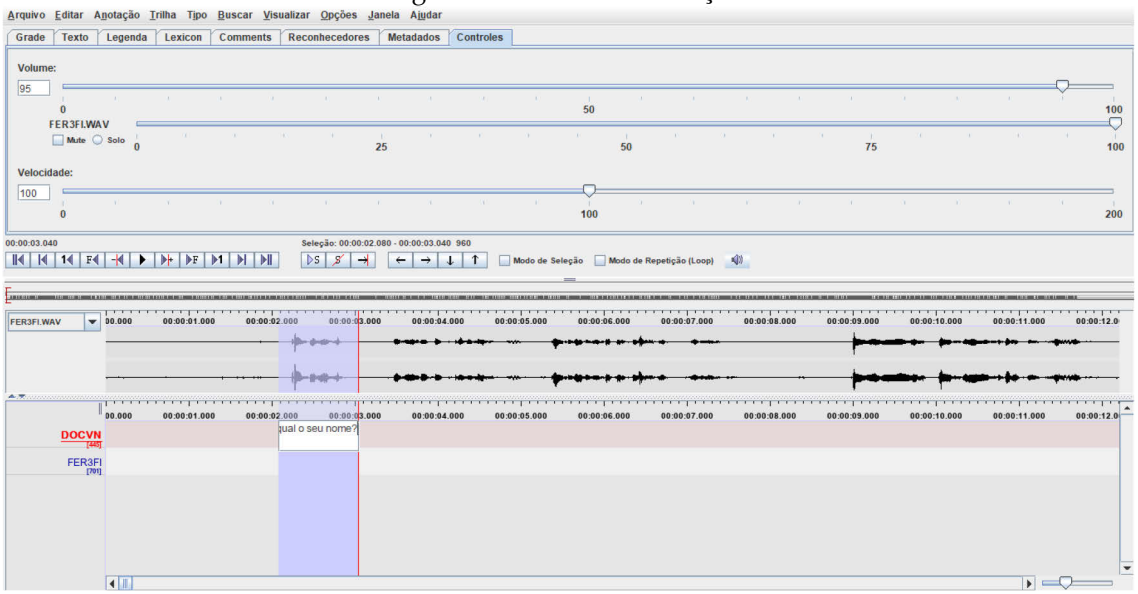
Figura 7 – Selecionando trecho.



Fonte: extraída do software ELAN 5.9 (2020).

Selecionado o trecho a ser anotado, damos um duplo clique sobre a linha vermelha na altura da trilha correspondente ou pressionamos [Shift] + [Enter] para criar uma nova anotação na trilha ativa. Na figura abaixo, por exemplo, no trecho selecionado, o documentador pergunta "qual o seu nome?" e a fala é transcrita seguindo as normas pré-determinadas, sem iniciais maiúsculas e com o sinal de interrogação indicando pergunta.

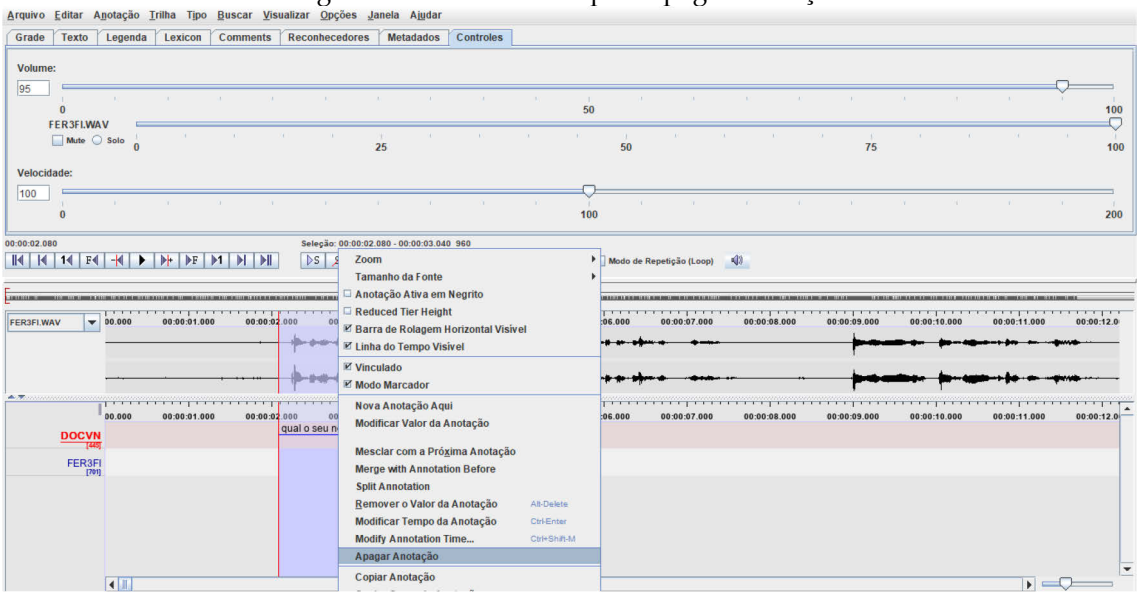
Figura 8 – Criando a anotação.



Fonte: extraída do software ELAN 5.9 (2020).

Para apagar uma anotação, clicamos sobre ela com o botão direito do mouse e selecionamos *Apagar anotação*, como podemos observar na Figura 9.

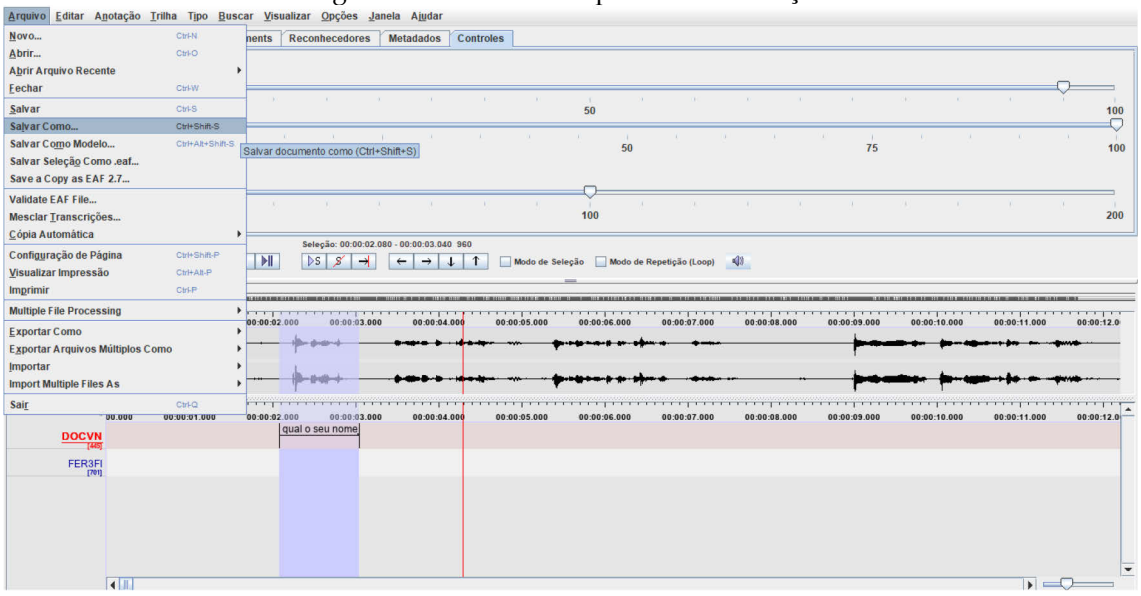
Figura 9 – Procedimento para apagar anotação.



Fonte: extraída do software ELAN 5.9 (2020).

Antes de fechar o arquivo, devemos salvar sempre as alterações. Para isso, clicamos em *Arquivo*, na barra superior, e, na sequência, selecionamos *Salvar como*.

Figura 10 – Salvando arquivo de transcrição.



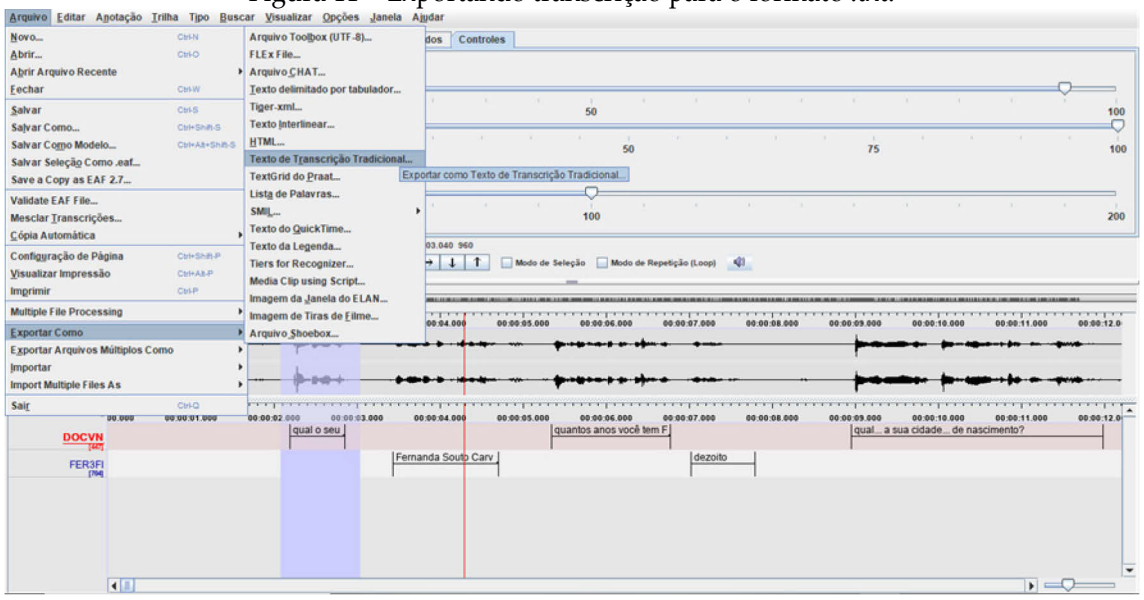
Fonte: extraída do software ELAN 5.9 (2020).

Esse é o fluxo de trabalho repetido durante toda a transcrição. É um trabalho árduo, mas automatiza os passos seguintes da constituição e anotação do banco de dados.

Após procedimento de transcrição, ocorre o de validação das transcrições. Este é o momento em que os arquivos de transcrição são revisados por pares devidamente treinados, normalmente integrantes do grupo de pesquisa, para garantir a homogeneidade das normas adotadas.

No caso da amostra *Deslocamentos*, utilizamos arquivos exportados do ELAN 5.9 (2020) em formato .txt, por ser um formato aceito por diferentes *softwares* tanto de análise linguística quanto de análise estatística, mas há a possibilidade de exportá-lo para múltiplos formatos (txt; TextGrid; HTML, entre outros) a depender dos interesses do pesquisador. Para realizar a exportação, clicamos em *Arquivo* > *Exportar Como* > *Texto de Transcrição Tradicional* (Figura 11).

Figura 11 – Exportando transcrição para o formato .txt.

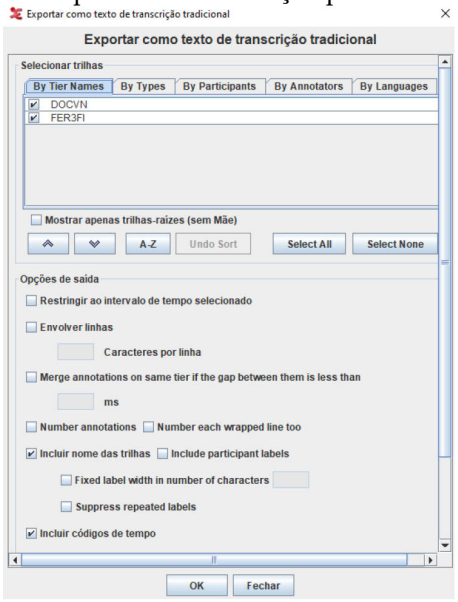


Fonte: extraída do software ELAN 5.9 (2020).

Na sequência, a aba representada na Figura 12, abaixo, se abrirá. Marcamos as caixas correspondentes às trilhas do documentador e do informante na parte superior da aba, em seguida, marcamos a caixa responsável por incluir códigos de tempo, na

parte inferior da aba, para que possamos saber onde no áudio se encontra o fenômeno a ser pesquisado.

Figura 12 – Exportando transcrição para o formato .txt.



Fonte: extraída do software ELAN 5.9 (2020).

Ao final do procedimento de exportação para o formato .txt, temos um arquivo como na Figura 13:

Figura 13 – Transcrição em formato .txt.

```
TIPO DE AMOSTRA: ENT
 DOCUMENTADORA: VIVIANE
 COMUNIDADE: UFS
 ANO: 2019
 DESLOCAMENTO: 3
 TEMPO DE CURSO: INICIO
 INF: FER3FI
 sexo: FEM
 escolaridade: S
 Idade: 18
 Cidade Natal: SIMÃO DIAS, SE

DOCVN    qual o seu nome?
         00:00:02.190 - 00:00:02.850

FER3FI   Fernanda Souto Carvalho
         00:00:03.430 - 00:00:04.700

DOCVN    quantos anos você tem Fernanda?
         00:00:05.340 - 00:00:06.770

FER3FI   dezoito
         00:00:07.020 - 00:00:07.790

DOCVN    qual... a sua cidade... de nascimento?
         00:00:08.950 - 00:00:11.980

FER3FI   Simão Dias
         00:00:12.450 - 00:00:13.140

DOCVN    ah você sempre morou em... Simão Dias?
         00:00:14.350 - 00:00:16.820

FER3FI   não eu só nasci lá sempre morei em Paripiranga e depois que eu vim morar por aqui
         00:00:16.840 - 00:00:21.480
```

Fonte: entrevista transcrita da amostra Deslocamentos.

Após a transcrição dos dados, o ELAN 5.9 (2020) possibilita a busca por palavras ou sequências de caracteres (OUSHIRO, 2014, p. 129) em múltiplos arquivos de transcrição através da ferramenta *Buscar*, o que automatiza o processo de localização de traços variáveis e, por conseguinte, facilita o trabalho do pesquisador. No entanto, para alguns fenômenos variáveis, como a variação no preenchimento da posição de determinante antes de possessivos pré-nominais, não é possível fazer uma única busca para coletar apenas os determinantes (o, a, os, as) e os pronomes possessivos e suas formas flexionadas (meu(s) minha(s), seu(s) sua(s), teu(s)/tua(s), seu(s), sua(s), nosso(s), nossa(s), vosso (s), vossa(s)).

Na próxima seção, então, explicaremos o estágio que segue a transcrição. Serão descritos os procedimentos de anotação morfológica realizados utilizando a ferramenta LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020), *software* utilizado também para automatizar as buscas por fenômenos linguísticos nas amostras do Banco de Dados Falares Sergipanos.

3 Anotação e etiquetagem de dados linguísticos em entrevistas sociolinguísticas

Conforme Berber Sardinha (2004), a etiquetagem de dados linguísticos se faz por meio de marcações morfológicas, sintáticas, semânticas ou discursivas em diferentes *corpora*. Em linhas gerais, essas marcações colaboram grandemente no processamento de grandes volumes de texto e buscas automatizadas. Além disso, por meio da anotação morfológica, a identificação de padrões lexicais e gramaticais, bem como a desambiguação lexical são favorecidas (SARDINHA, 2004).

3.1 A ferramenta LancsBox 5.1.2 (2020)

Com o avanço da linguística computacional e também pelo interesse comercial que o processamento de linguagem natural desperta, encontram-se hoje disponíveis diversas ferramentas que fazem a anotação morfossintática automática de dados

linguísticos (TreeTagger (SCHMID, 1994), Aelius 0.9.7 (ALENCAR, 2013), Sketch Engine (KILGARRIFF, 2014), entre outras). No entanto, grande parte delas não é gratuita, tem licenças de uso restritas, ou precisam de conhecimentos de linguagem de programação (Python, por exemplo) para que possam ser operacionalizadas no computador.

Adotamos a ferramenta LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020) para realizar a anotação morfológica dos dados que constituem a amostra *Deslocamentos* por ser gratuita, possuir uma interface de fácil manuseio, ter suporte com vídeos[8] e manuais disponíveis em *website* próprio e não necessitar de conhecimentos aprofundados de linguagem de programação para realizar buscas no *corpus* de trabalho.

Criado por pesquisadores da Universidade de Lancaster, o LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020) possui várias outras funcionalidades além da etiquetagem como visualização gráfica de coligados e colocações, listas de frequências, cálculo de frequência, *n-grams*, entre outras. No entanto, os objetivos desta seção se limitam a guiar o leitor a entender como utilizamos essa ferramenta para fazer a anotação morfológica da amostra *Deslocamentos* e como realizamos as buscas por fenômenos pesquisados no âmbito do Grupo de Estudos em Linguagem, Interação e Sociedade (GELINS), citando como exemplo o fenômeno estudado por Siqueira (2021).

3.1.1 Onde baixar a ferramenta? E como instalar?

O LancsBox5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020) encontra-se disponível no seguinte endereço: http://corpora.lancs.ac.uk/lancsbox/index.php, na aba *downloads*. Nessa aba, a própria página identifica para o usuário a versão a ser

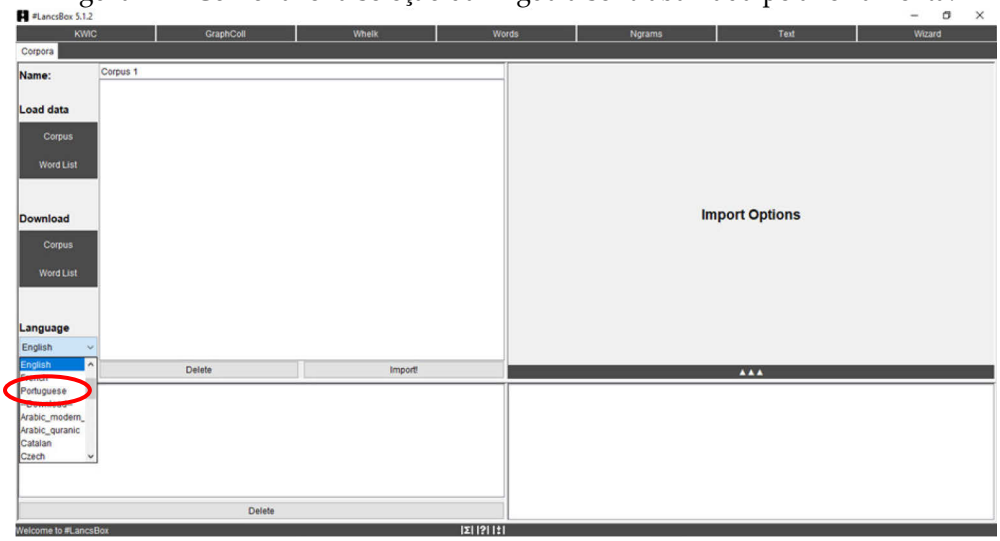
[8] STARTING with #LancsBox v. 3.0. 2017. 1 vídeo (6min 54s). Publicado pelo canal de Vaclav Brezina. Disponível em: https://www.youtube.com/watch?v=7SFJMFUP83Y. Acesso em: 20 jul. 2021.

adotada em conformidade com o sistema operacional do computador em que a ferramenta está sendo baixada. Em seguida, executamos o instalador da ferramenta habilitando as permissões correspondentes ao sistema operacional do computador e ela já estará pronta para uso.

3.1.2 Carregando o *corpus* para fazer a anotação morfológica

Ao abrir a ferramenta, antes de fazer o carregamento do *corpus*[9], é importante selecionar a língua. Como a Língua Portuguesa não possui suporte total do LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020), devemos baixá-la no próprio *software* antes de utilizá-lo, clicando em *Languages* e selecionando *Portuguese*. Em seguida, realizamos o carregamento do *corpus* em *Load Data* escolhendo-se a opção *Corpus*. Aparecerá, então, uma janela para que possamos escolher os arquivos a serem processados, lembrando que, no nosso caso, optamos por selecionar todas as entrevistas para realizar a etiquetagem. Nas figuras 14 e 15, estes procedimentos estão demonstrados.

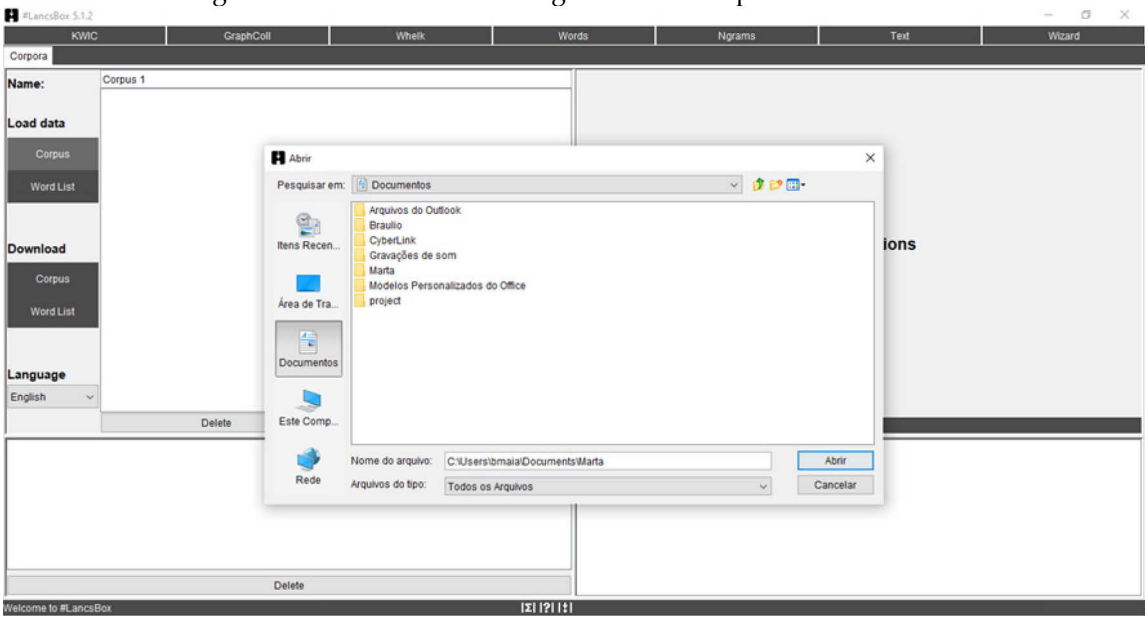
Figura 14 – Como fazer a seleção da língua a ser trabalhada pela ferramenta.



Fonte: extraída do software LancsBox5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020).

[9] No caso da amostra *Deslocamentos*, utilizamos arquivos exportados do ELAN em formato .txt. Salientamos, contudo, que o LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020) processa diferentes formatos de texto (.pdf; .docx, entre outros).

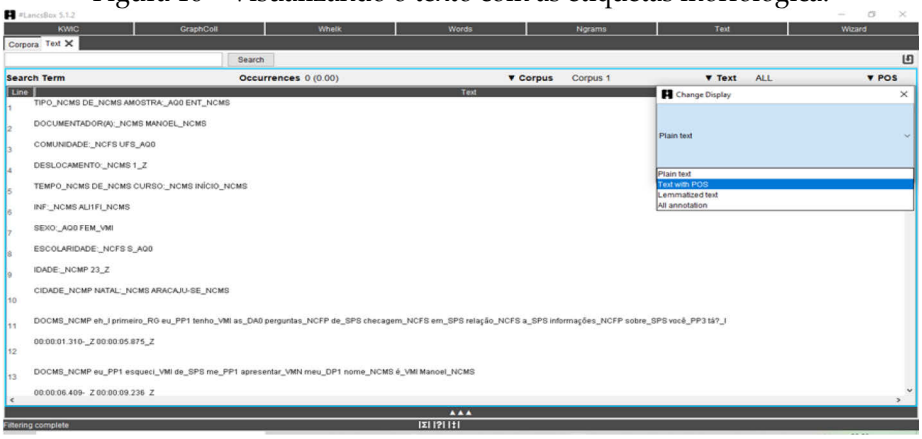
Figura 15 – Como fazer o carregamento do corpus na ferramenta.



Fonte: extraída do software LancsBox5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020).

Feito o carregamento do *corpus*, clicamos em *Import* para iniciar a importação dos dados, sendo possível ver no rodapé do programa se o processo de etiquetagem está sendo realizado por meio da frase *Tagging*. Após essa etapa, podemos visualizar o texto etiquetado por meio da aba *Text*, clicando, em seguida, na aba *Display*, selecionando a opção *Text with POS* (texto com a classe de palavra), e, posteriormente, clicando em *Apply*, conforme figura 16, abaixo.

Figura 16 – Visualizando o texto com as etiquetas morfológica.



Fonte: extraída do software LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020).

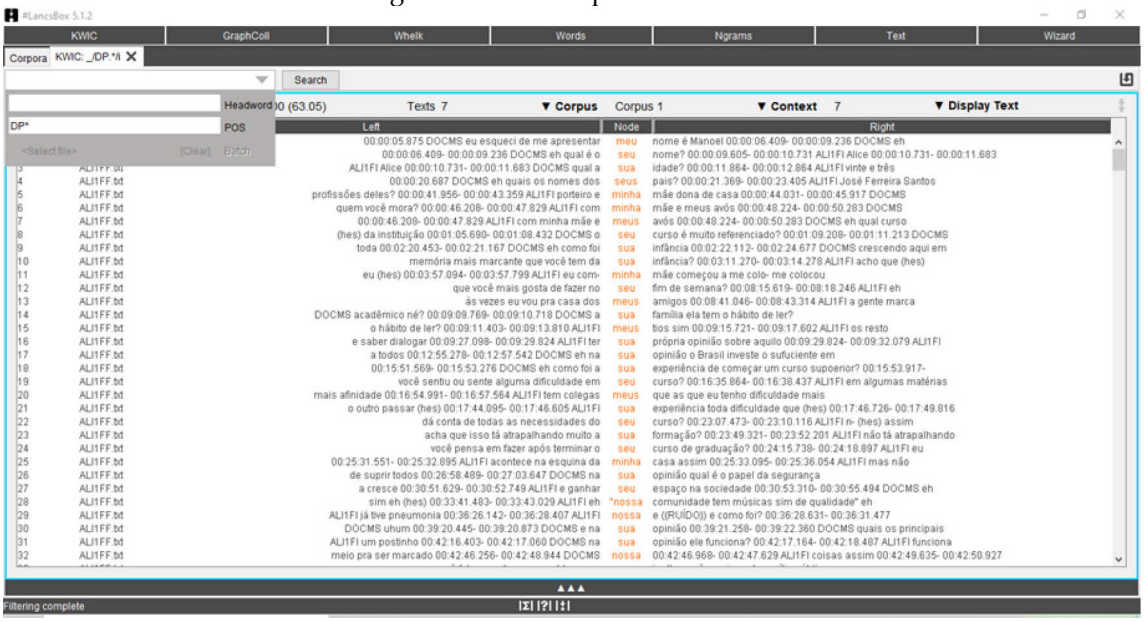
O LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020) utiliza um conjunto de etiquetas que também foram feitas para o *TreeTagger* (SCHMID, 1994). No caso da Língua Portuguesa, por exemplo, "V" é a etiqueta para "verbos" e "P" para "pronomes". Para identificação de cada etiqueta para a Língua Portuguesa utilizada na ferramenta, remetemos ao seguinte endereço que direciona para as etiquetas utilizadas: encurtador.com.br/hkoF8.

3.2 Como realizar buscas por fenômenos variáveis?

Nesta seção, explicamos o mecanismo de buscas KWIC (*Key Word In Context*), por ser a função do LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020) que permite a visualização do fenômeno linguístico pesquisado em seu contexto. Ela também pode ser usada para encontrar a frequência de uma palavra, classe de palavras, no *corpus*, bem como encontrar estruturas mais complexas no caso de línguas que possuem suporte completo, entre outras funcionalidades. Após importar dados, clicamos em KWIC para iniciar as buscas (Figura 17).

A título de exemplo, citamos o estudo de Siqueira (2021), no qual o foco é a variação no preenchimento da posição de determinante antes de possessivos pré-nominais. Iniciamos nossa busca clicando na seta em *search* e colocando em POS a etiqueta DP (para determinante possessivo) seguida de ponto (.) e asterisco (*), conforme Figura 18. Ressaltamos que para realizar buscas por classe de palavras, preenchemos o campo POS com a etiqueta disponível no site das etiquetas do *TreeTager* de acordo com a língua escolhida.

Figura 17 – Buscas pela aba KWIC.



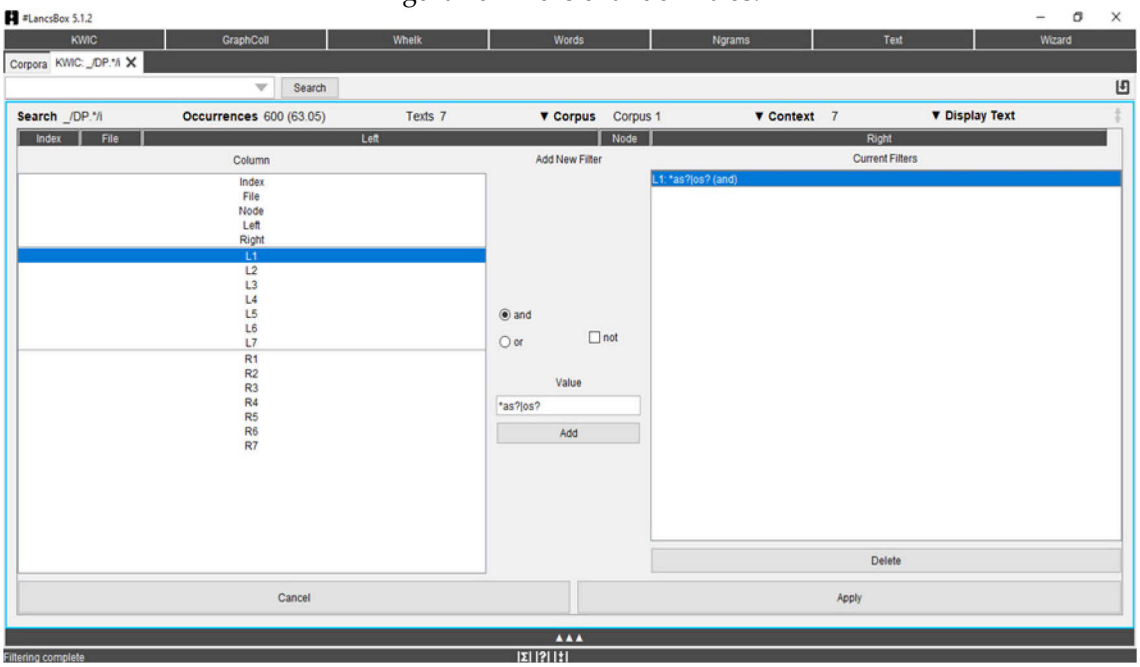
Fonte: extraída do software LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020).

Na Figura 17, vemos como o LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020) apresenta as ocorrências. Na parte esquerda da ferramenta, está o contexto precedente à ocorrência desejada, e na parte direita o contexto seguinte. É importante destacar que, na primeira coluna à esquerda está a ordem da ocorrência, na segunda, o nome do arquivo onde se encontra a ocorrência, e, a coluna em vermelho, chamada *Node*, apresenta a ocorrência. É possível refinar a quantidade de palavras antes e depois da ocorrência, clicando em *Context* e selecionando o número de palavras desejado.

Devido à grande quantidade de ocorrências, é possível refinar a pesquisa para localizar apenas os contextos em que a última palavra termine em *o*, *a* e suas variações no singular e plural, o que possibilita recuperar o preenchimento em que artigo e preposições formam uma palavra só como em "na" (em+a) e "do" (de+o). Para tanto, utilizamos a função *Filter*, que aparece ao clicarmos com o botão direito do mouse em cima de *Left* ou *Right*. Como queremos filtrar a primeira palavra que antecede o determinante, devemos selecionar na janela que aparecerá L1 e *and* porque queremos L1 "e" os resultados da expressão regular, e, em *value* digitamos a expressão regular

as?|os? . Em seguida, a expressão aparecerá em *Current Filters* onde a selecionamos e, em seguida, clicamos em *Apply*, o que pode ser visualizado na Figura 18.

Figura 18 – Adicionando Filtros.



Fonte: extraída do software LancsBox 5.1.2(BREZINA; WEILL-TESSIER; MCENERY, 2020) .

Embora o LancsBox seja uma ferramenta de fácil manuseio, acreditamos ser importante saber mais sobre expressões regulares (REGEX), uma vez que elas viabilizam uma busca mais refinada, diminuindo o tempo de procura pelos fenômenos a serem pesquisados. Portanto, para maiores informações sobre o que são e como se usam expressões regulares, recomendamos os tutoriais disponíveis em: http://corpus.futurelingua.com/.

4 Considerações finais

As duas ferramentas descritas neste texto vêm sendo amplamente utilizada nas pesquisas realizadas no escopo do nosso grupo de trabalho, o GELINS. Relatamos aqui o uso delas para a transcrição da amostra *Deslocamentos* e também para a anotação morfológica e busca automática de fenômenos variáveis.

No caso das transcrições, observamos que a ferramenta ELAN 5.9 (2020), ao permitir o alinhamento do áudio com a transcrição, contribui qualitativamente para dados mais acurados e também para encontrar fenômenos em interface com o nível fonológico, facilitando a escuta. Além disso, destacamos a potencialidade do *software* em exportar os arquivos de transcrição para múltiplos formatos, compatíveis com outros programas computacionais utilizados para análise linguística.

Já a ferramenta LancsBox5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020) foi relevante na automatização do processo de levantamento das ocorrências de diferentes fenômenos variáveis, que, tradicionalmente tem sido feito de maneira manual, sendo um processo altamente dispendioso (OTHERO; AYRES, 2019). Destacamos no texto, o caso da análise do uso variável de determinantes em sintagmas nominais possessivisados (SIQUEIRA, 2021), mas as buscas pelas ocorrências de outros fenômenos dentro do escopo do GELINS também foram automatizadas com o uso dessa ferramenta, como na busca de verbos para se analisar a variação na terceira pessoa do plural (NOVAIS, 2021) e na busca de preposições para análise da variação na regência de complementos locativos verbos de movimento (RODRIGUES, 2021).

A anotação morfológica permitiu, então, o processamento de mais de 60 textos com aproximadamente 500.000 palavras de uma única vez. Além disso, por meio da busca automática foi possível localizar as ocorrências dos fenômenos com seus contextos antecedentes e seguintes. Foi possível, também, identificar o informante, a qual deslocamento ele pertence, seu sexo/gênero, sua escolaridade e sua idade, uma vez que o nome do arquivo já possui esses dados, otimizando a tabulação dos dados.

Este texto, de natureza procedural, traz duas implicações para a pesquisa com dados de entrevistas sociolinguísticas. Primeiro, ao apresentar as normas de transcrição adotadas na constituição do Banco de Dados Falares Sergipanos e propor um roteiro de procedimentos para a transcrição de dados linguísticos orais no *software* ELAN 5.9 (2020), buscamos contribuir para tornar o processo de transcrição mais

acurado e também apontar diretrizes para a padronização de transcrições na área de sociolinguística. Por fim, ao apresentar como a anotação morfológica tem sido feita, bem como a buscas por um fenômeno linguístico variável dentro do escopo da amostra *Deslocamentos*, utilizando a ferramenta LancsBox 5.1.2 (BREZINA; WEILL-TESSIER; MCENERY, 2020), buscamos contribuir para tornar o trabalho do pesquisador menos dispendioso, tanto em termos de tempo, quanto de recursos financeiros.

Referências

ALENCAR, L. F. **Aelius 0.9.7 User's Manual**. 2013. Disponível em: http://aelius.sourceforge.net/manual.html. Acesso em: 25 fev. 2020.

ANTHONY, L. **AntConc** v. 3.5.9 [Computer Software]. Tokyo, Japan: Waseda University. Disponível em: https://www.laurenceanthony.net/software. Acesso em: 20 ago 2020.

BERBER SARDINHA, T. **Linguística de Corpus**. Barueri: Manolo, 2004.

BREZINA, V.; WEILL-TESSIER, P.; MCENERY, A. (2020). **#LancsBox** v. 5.1.2 [software]. Disponível em: http://corpora.lancs.ac.uk/lancsbox. Acesso em: 20 ago. 2020.

CARDOSO, P. B. O paradoxo entre a transparência dos dados e a privacidade dos informantes na gestão de dados linguísticos. **Revista da ABRALIN**, v. 19, n. 2, p. 1-9, 24 ago. 2020. Disponível em: https://revista.abralin.org/index.php/abralin/article/view/1631. Acesso em: 05 jul. 2021. DOI https://doi.org/10.25189/rabralin.v19i2.1631

CIANCONI, R. B. Banco de dados de acesso público. **Ciência Da Informação**, v. 16, n 1, p. 53-59, 1987. Disponível em: https://revista.ibict.br/ciinf/article/view/271. Acesso em: 08 jul. 2022. DOI

ELAN (Version 5.9) [Computer software]. (2020). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Disponível em: https://archive.mpi.nl/tla/ELAN.

FREITAG, R. M. Ko; MARTINS, M. A.; TAVARES, M. A. Bancos de dados sociolinguísticos do português brasileiro e os estudos de terceira onda: potencialidades

e limitações. **Alfa**: Revista de Linguística, v. 56, n. 3, 2012. Disponível em: https://www.scielo.br/j/alfa/a/J6ZcH9z3RPYz5ZGxnQkZJkr/abstract/?lang=pt. Acesso em: 12 jan. 2021. DOI https://doi.org/10.1590/S1981-57942012000300009

FREITAG, R. M. Ko. Banco de dados falares sergipanos. **Working Papers em Linguística**, v. 14, n. 2, p. 156-164, 2013. Disponível em: https://periodicos.ufsc.br/index.php/workingpapers/article/view/1984-8420.2013v14n2p156. Acesso em: 05 jul. 2021. DOI https://doi.org/10.5007/1984-8420.2013v14n2p156

FREITAG, R. M. Ko (org.). **Metodologia de Coleta e Manipulação de Dados em Sociolinguística**. São Paulo: Blucher, 2014. DOI https://doi.org/10.5151/BlucherOA-MCMDS

FREITAG, R. M. Ko.; PINHEIRO, B. F. M.; SILVA, L. S. Análise variacionista de pausas preenchidas em fronteiras de constituintes. *In*: FREITAG, R. M. KO.; LUCENTE, L. **Prosódia da fala**: pesquisa e ensino. São Paulo: Blucher, 2017. DOI https://doi.org/10.5151/9788580392593-07

FREITAG, R. M. K. **Projeto de pesquisa:** A língua do universitário: fala, leitura e escrita para o letramento acadêmico. 2018. Disponível em: https://url.gratis/5V6QBR. Acesso em: 20 abr. 2020.

FREITAG, R. M. K.; MARTINS, M. A. R.; ARAÚJO, A.; BATTISTI, E.; COELHO, I. M. W. DA S.; SOUSA, M. D. A. F.; SILVA, R. G. DA; LIMA-LOPES, R. E. DE. Desafios da gestão de dados linguísticos e a Ciência Aberta. **Cadernos de Linguística**, v. 2, n. 1, p. 01-19, abr. 2021. Disponível em: https://cadernos.abralin.org/index.php/cadernos/article/view/307. Acesso em: 05 jul. 2021. DOI https://doi.org/10.25189/2675-4916.2021.v2.n1.id307

GONÇALVES, S. C. L.; TENANI, L. E. Problemas teórico-metodológicos na elaboração de um sistema de transcrição de dados interacionais: o caso do projeto ALIP (Amostra Lingüística do Interior Paulista). **Gragoatá**, n. 25, p. 165-183, 2008. Disponível em: https://periodicos.uff.br/gragoata/article/view/33148. Acesso em: 05 jul. 2021.

KILGARRIFF, A.; BAISA, V.; BUŠTA, J.; JAKUBÍČEK, M.; KOVÁŘ, V.; MICHELFEIT, J.; RYCHLÝ, P.; SUCHOMEL, V. The Sketch Engine: ten years on. **Lexicography**, v.1, p. 7-36, 2014. DOI https://doi.org/10.1007/s40607-014-0009-9

NAGY, N.; MEYERHOFF, M. Extending ELAN into variationist sociolinguistics. **Linguistics Vanguard**, v. 1, n. 1, 2015, p. 271-281. Disponível

em: https://doi.org/10.1515/lingvan-2015-0012. Acesso em: 20 jul. 2021. DOI https://doi.org/10.1515/lingvan-2015-0012

NOVAIS, V. S. **Variação na concordância verbal de terceira pessoa do plural na fala de universitários sergipanos**. 2021. Dissertação (Mestrado em Letras) – Universidade Federal de Sergipe, São Cristóvão, 2021.

OTHERO, G. A.; AYRES, M. R. Anotação morfológica automática de corpus de língua falada: desafios ao Aelius. **Texto Livre**: linguagem e tecnologia, v. 7, n. 2, p.44-60, 2014. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/textolivre/article/view/6123/5959. Acesso em: 01 jul. 2019. DOI https://doi.org/10.17851/1983-3652.7.2.44-60

OUSHIRO, L. Transcrição de entrevistas sociolinguísticas com o ELAN. *In*: FREITAG, Raquel Meister Ko (org.). **Metodologia de Coleta e Manipulação de Dados em Sociolinguística**. São Paulo: Blucher, 2014. DOI https://doi.org/10.5151/BlucherOA-MCMDS-9cap

PAIVA, M. C. Transcrição de dados lingüísticos. *In*: MOLLICA, M. C.; BRAGA, M. L. (org.). **Introdução à Sociolingüística**: o tratamento da variação. São Paulo: Contexto, 2003. p. 135-146.

RODRIGUES, F. G. C. **Variação na regência de complementos locativos verbos de movimento na fala de universitários da UFS**. 2021. Dissertação (Mestrado em Letras) – Universidade Federal de Sergipe, São Cristóvão, 2019.

ROSENFELDER, I. **A short introduction to transcribing with elan**. University of Pennsylvania, 2011. Disponível em: https://www.ling.upenn.edu/~wlabov/L560/ELAN_introduction.pdf. Acesso em: 20 jul. 2021.

SCHMID, H. Improvements in Part-of-Speech Tagging with an Application to German. *In*: Proceedings of the ACL SIGDAT-Workshop. 1995, Dublin. **Proceedings** […]. Dublin, 1994.

SIQUEIRA, M. Análise contrastiva da estrutura do sintagma nominal possessivizado no português brasileiro. **Matraga**, v. 28, n. 52, 2021. Disponível em: https://www.e-publicacoes.uerj.br/index.php/matraga/article/view/53146. Acesso em: 02 jul. 2021. DOI https://doi.org/10.12957/matraga.2021.53146

STARTING with #LancsBox v. 3.0. 2017. 1 vídeo (6min 54s). Publicado pelo canal de Vaclav Brezina. Disponível em: https://www.youtube.com/watch?v=7SFJMFUP83Y. Acesso em: 20 jul. 2021.

TACCHETTI, M. **User's Guide for ELAN Linguistic Annotator**. 2017. Disponível em: https://www.mpi.nl/corpus/manuals/manual-elan_ug.pdf. Acesso em: 20 jul. 2021.

#LancsBox 5.1 manual. Lancaster University. Disponível em: http://corpora.lancs.ac.uk/lancsbox/docs/pdf/LancsBox_5.1_manual.pdf. Acesso em: 20 jul. 2021.

Artigo recebido em: 20.07.2021 Artigo aprovado em: 02.03.2022