UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS

MARTA DEYSIANE ALVES FARIA SOUSA

PROTOCOLO PARA ANOTAÇÃO LINGUÍSTICA E GERENCIAMENTO DE AMOSTRAS SOCIOLINGUÍSTICAS: O CASO DA AMOSTRA DESLOCAMENTOS 2019

São Cristóvão - SE
2023

MARTA DEYSIANE ALVES FARIA SOUSA

PROTOCOLO PARA ANOTAÇÃO LINGUÍSTICA E GERENCIAMENTO DE AMOSTRAS SOCIOLINGUÍSTICAS: O CASO DA AMOSTRA DESLOCAMENTOS 2019

Tese apresentada ao Programa de Pós-Graduação em Letras da Universidade Federal de Sergipe, como requisito parcial à obtenção do título de Doutora em Letras. Área de concentração: Estudos Linguísticos. Linha de Pesquisa: Descrição, Análise e Usos Linguísticos.

Orientadora: Profa. Dra. Raquel Meister Ko. Freitag

São Cristóvão -SE

2023

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

S725p

Sousa, Marta Deysiane Alves Faria.
Protocolo para anotação linguística e gerenciamento de amostras sociolinguísticas : o caso da amostra deslocamentos 2019 / Marta Deysiane Alves Faria Sousa ; orientadora Raquel Meister Ko. Freitag. – São Cristóvão, SE, 2023.
148 f. ; il.

Tese (doutorado em Letras) – Universidade Federal de Sergipe, 2023.

1. Sociolinguística. 2. Ciência. 3. Amostragem (Estatística). 4. Ferramentas. I. Freitag, Raquel Meister Ko., orient. II. Título.

CDU 81'272

MARTA DEYSIANE ALVES FARIA SOUSA

PROTOCOLO PARA ANOTAÇÃO LINGUÍSTICA E GERENCIAMENTO DE AMOSTRAS SOCIOLINGUÍSTICAS: O CASO DA AMOSTRA DESLOCAMENTOS 2019

Tese apresentada ao Programa de Pós-Graduação em Letras da Universidade Federal de Sergipe, como requisito parcial à obtenção do título de Doutora em Letras. Área de concentração: Estudos Linguísticos. Linha de Pesquisa: Descrição, Análise e Usos Linguísticos.

Aprovada em: 27/02/2023.

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
RICARDO JOSEH LIMA
Data: 11/07/2023 14:55:03-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Ricado Joseh Lima
Universidade do Estado do Rio de Janeiro
Examinador – Externo à instituição

gov.br Documento assinado digitalmente
ALISSON HUDSON VERAS LIMA
Data: 11/07/2023 15:44:34-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Álisson Hudson Veras Lima
Instituto Federal de Alagoas – IFAL
Examinador – Externo à instituição

gov.br Documento assinado digitalmente
JULIANA PEREIRA SOUTO BARRETO
Data: 12/07/2023 14:10:14-0300
Verifique em https://validar.iti.gov.br

Prof.ª Dr.ª Juliana Pereira Souto Barreto
Universidade Federal de Sergipe
Examinadora - Externa ao Programa

gov.br Documento assinado digitalmente
ROANA RODRIGUES
Data: 12/07/2023 15:12:07-0300
Verifique em https://validar.iti.gov.br

Prof.ª Dr.ª Roana Rodrigues
Universidade Federal de Sergipe
Examinadora – Interna ao programa

gov.br Documento assinado digitalmente
RAQUEL MEISTER KO FREITAG
Data: 14/07/2023 06:50:23-0300
Verifique em https://validar.iti.gov.br

Prof.ª Dr.ª Raquel Meister Ko. Freitag
Universidade Federal de Sergipe
Presidente – Orientadora

À minha filha, Júlia, por ser minha fonte de alegria e força, dedico.

AGRADECIMENTOS

Quando comecei a frequentar aulas de Sociolinguística em 2017, eu ainda não tinha ouvido falar em “rede de apoio”. No entanto, ao engravidar, “rede de apoio” aparecia em todas as conversas, nas redes sociais, em todo lugar. Hoje, relembrando a trajetória de trabalho para tese, a expressão “rede de apoio” faz todo o sentido para tudo que vivenciei e vivencio fazendo parte do GELINS.

É muito bonito ver as reuniões do grupo, nas quais realmente trocamos saberes e afetos. Por isso, agradeço imensamente à minha orientadora, a Professora Dra. Raquel Freitag, por ter me acolhido. Mesmo sabendo de todos os riscos envolvidos em ter uma orientanda se tornando mãe, em meio a uma pandemia, ela apostou em mim e me deu a oportunidade de me desafiar navegando pelo mundo da programação e da Sociolinguística. Escutei e ainda escuto atenta a todos os seus conselhos, acadêmicos e de vida, aprendi muito nessa jornada.

Agradeço também ao GELINS e aos DEVS do GELINS. Em especial a Paloma e Manoel pelas parcerias, leituras atentas, comentários e discussões, a Túlio (meu “filho” acadêmico), que esteve mais que presente na reta final, quebrando a cabeça comigo, realizando parcerias e me ensinando muito sobre o pensamento de programador. Agradeço também a Thaís, Lucas e Verônica, que sempre foram muito solícitos e acolhedores comigo.

Aos professores da banca de qualificação e de defesa, as Professoras Doutoras Juliana Barreto e Roana Rodrigues e os Professores Doutores Ricardo e Álisson, agradeço por possibilitarem que minha rede de apoio fosse para além de Sergipe e também da Sociolinguística. A leitura cuidadosa e aprofundada de vocês foi fundamental para que eu pudesse entregar minha tese.

Por fim, agradeço a três pessoas que se esforçaram ao máximo para que eu chegasse até aqui, mesmo que de maneira inconsciente: Braulio, meu esposo, minha filha, Júlia, e minha sogra, Cirlei. Obrigada por respeitarem meu espaço, por entenderem as minhas ausências, e por alegrarem a minha existência.

RESUMO

Bancos de dados linguísticos são ferramentas que propiciam aos pesquisadores acesso ágil a amostras de língua (textos orais ou escritos), cruzamento entre dados de diferentes regiões e um acervo linguístico de um determinado período e localidade, servindo não só a propósitos científicos, mas também didáticos (FREITAG; MARTINS; TAVARES, 2012; GONÇALVES, 2019; SILVA, 2015). Tanto no contexto brasileiro quanto internacional, a preocupação com a documentação e arquivamento de amostras sociolinguísticas pode ser explicada pela importância desses dados para o avanço das pesquisas na área (KENDALL, 2013), pelas demandas da Ciência Aberta quanto ao compartilhamento dos dados e pelos avanços tecnológicos em termos de armazenamento e anotação linguística (VANN, 2021). No entanto, assim como no exterior, no Brasil, empreendimentos nesse sentido têm-se dado no nível individual, sem padronização na metodologia, codificação e disponibilização de dados, o que dificulta a replicabilidade e consequente cotejamento de fenômenos variáveis entre diferentes bancos de dados. Ademais, não há amostras sociolinguísticas linguisticamente anotadas entre aquelas que já se encontram disponíveis online, assim como protocolos de gerenciamento de dados e códigos para realização de análise estatística. Objetivamos com este trabalho, elaborar um protocolo de sistematização e divulgação Amostra Deslocamentos 2019 (FREITAG, 2018) seguindo os preceitos da Ciência Aberta. Nossa tese é a de que é possível utilizar recursos abertos e gratuitos para anotação linguística e sistematização de amostras sociolinguísticas seguindo o paradigma da Ciência Aberta. Para defendermos essa tese, delimitamos os seguintes objetivos específicos: i) testar duas ferramentas computacionais (*LancsBox* 6.0 e *spaCy* 3.5) gratuitas na etiquetagem da Amostra Deslocamentos 2019; ii) avaliar a etiquetagem empreendida pelas ferramentas; iii) comparar o desempenho das ferramentas em relação à buscas e funcionalidades para uma pré-análise do fenômeno da variação do preenchimento de determinante antes de possessivo pré-nominal; iv) descrever ações para a divulgação e o compartilhamento dos dados da amostra Deslocamentos 2019; v) sistematizar as ações desenvolvidas em forma de protocolo. Os resultados gerais confirmam nossa tese de que é possível sistematizar e anotar linguisticamente, no nível gramatical (*part-of-speech*) amostras sociolinguísticas usando apenas recursos gratuitos disponíveis para a língua portuguesa. As ferramentas também contribuem para buscas mais acuradas e com resultados com maior número de ocorrências do que uma busca manual. Por outro lado, em relação à divulgação dos dados, o armazenamento e a hospedagem do site é ainda uma limitação a respeito do uso de recursos abertos e gratuitos.

Palavras-chave: Sociolinguística Variacionista. Ciência Aberta. Dados de fala. Anotação Linguística. PLN.

ABSTRACT

Linguistic data bases are considered tools that provide researchers with fast access to language samples (written or oral texts), crossing among data from different regions, and a linguistic collection of a certain period of time and place, being useful not only to scientific purposes, but also to didactic ones (FREITAG; MARTINS; TAVARES, 2012; GONÇALVES, 2019; SILVA, 2015). Both in Brazilian and international scenarios, there is a concern with the documentation and archiving of sociolinguistic samples, which may be explained due to the importance of these data to the advance of the research in this field (KENDALL, 2013), to the Open Science demands in relation to the sharing of data, and also to the technological advances regarding archiving and linguistic annotation (VANN, 2021). However, as in the international scenario, in Brazil, such endeavors have been individually made, without standardization in the methodologies, codes, and data availability, which makes it difficult to replicate and, consequently, compare different variable phenomena from different databases. In addition to it, there are no sociolinguistic samples linguistically tagged among those that are already available online as well as data storage and management protocols and codes to perform statistical analysis. With this study, we aim at creating a protocol to systematize and disseminate the sample Displacements 2019 (FREITAG, 2018) from *Falares Sergipanos* database following Open Science principles. Our thesis is that it is possible to use open and free resources to linguistically tag and systematize sociolinguistic samples according to Open Science paradigm. In order to support our thesis, we set the following specific goals: i) to test two free computational tools (*LancsBox* 6.0 e *spaCy* 3.5) to linguistically annotate the sample Displacements 2019; ii) to evaluate the annotation performed by each tool; iii) to compare the performance of the two tools in relation to searches and functionalities for a pre-analysis of the phenomenon the filling of the determiner position before possessives in pre-nominal position; iv) to describe actions to disseminate and share the data of the sample Displacements 2019; v) organize the actions taken in a protocol. The general results confirm our thesis that it is possible to systematize and linguistically annotate sociolinguistic samples using only free resources available for the Portuguese language. The tools tested also contributed to searchers that are more accurate and with a greater number of occurrences of the phenomenon in comparison to a manual search. On the other hand, it is still a limitation to host and store a web site with a high number of data using free resources.

Keywords: Variationist Sociolinguistics. Open Science. Speech Data. Linguistic annotation. PLN.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

A Sociolinguística Variacionista, subárea que concatena aspectos linguísticos com o comportamento social do indivíduo, objetiva descrever a variação linguística, fenômeno no qual há diferentes formas que exprimem o mesmo valor representacional (LABOV, 1978), e os fenômenos linguísticos em processo de mudança das línguas (LABOV, 2006 [1966]). No Brasil, os estudos desenvolvidos por Labov têm sido a maior fonte de referência para as pesquisas na área (FREITAG, 2016), influenciando o trabalho de descrição, documentação e também a criação de bancos de dados do português brasileiro, os quais podem possibilitar aos pesquisadores a realização de estudos com vistas a uma descrição mais abrangente da variedade brasileira do português (FREITAG, 2013).

"Bancos de dados", conforme Cianconi (1987) são um agrupamento de bases de dados, que, por sua vez, são descritas como: "um conjunto de dados interrelacionados, organizados de forma a permitir recuperação de informações." (CIANCONI, 1987, p. 54). Acreditamos que a definição proposta por Cianconi (1987) também abarca os bancos de dados sociolinguísticos, uma vez que estes bancos são constituídos por amostras de fala, e sua organização permite a recuperação de informações sobre a forma da coleta, a comunidade onde foi realizada a pesquisa entre outros. Destacamos também que, nesta tese utilizamos amostras e *corpus* como sinônimos, uma vez que as amostras sociolinguísticas podem ser concebidas como *corpora* menos prototípicos (GRIES; BEREZ, 2017), além de ser um termo que vem sendo amplamente utilizado na linguística de maneira geral (ALUÍSIO; ALMEIDA, 2006).

A constituição de bancos de dados tem sido assunto recorrente (SILVA, 2015), não somente por sua relevância enquanto ferramenta de armazenamento, mas também pela discussão acerca da metodologia de coleta empregada e dos aspectos éticos concernentes à divulgação dos dados (FREITAG; MARTINS; TAVARES, 2012), e pela possibilidade de se atestar teorias linguísticas baseadas no uso por meio dados empíricos (COLLISCHONN; MONARETTO, 2012).

Dentre as várias finalidades dos bancos de dados linguísticos, Gonçalves (2019) destaca:

> (i) tornar disponível, em meio eletrônico, amostras de fala e/ou de escrita em uma base de dados que agilize pesquisas diversas e torne possível a verificação de hipóteses e postulados teóricos acerca dos efeitos do uso sobre a gramática da língua; (ii) oferecer a pesquisadores interessados bases de dados operacionalizáveis por meio de recursos computacionais; (iii) verificar a produtividade de expressões linguísticas na língua; (iv) possibilitar estudos baseados em extensas amostras de dados efetivamente atestados, de modo a se obter subsídio para a elaboração de gramáticas, dicionários, material para o ensino de língua etc. (GONÇALVES, 2019, p. 278)

Os bancos de dados linguísticos se apresentam, então, como ferramenta que propicia aos pesquisadores acesso ágil a amostras da língua (textos orais ou escritos), cruzamento entre dados de diferentes regiões e um acervo linguístico de um determinado período e localidade. Ademais, esses bancos servem não só a propósitos científicos, mas também didáticos, como exemplo, mencionamos a Gramática do Português Falado, que se serve dos dados do projeto Norma Linguística Urbana Culta (NURC) para as análises empreendidas nos diversos volumes dessa gramática (CASTILHO, 2021).

A relevância dos bancos de dados linguísticos não se encerra dentro da área de estudos de descrição e análise linguística. De acordo com Hirschberg e Manning (2015), o Processamento de Linguagem Natural (doravante PLN), área que pesquisa formas de ensinar máquinas a realizar ações que envolvem a compreensão e produção de linguagem humana (JURAFSKY; MARTIN, 2009), necessita de dados linguísticos para poder criar tecnologias como sistemas de diálogo falado (*Spoken Dialogue Systems*). Tais sistemas, conforme Hirschberg e Manning (2015), auxiliam o usuário a ter acesso a informações (como fazem a *Cortana* da *Microsoft* ou a *Siri* da *Apple*), a fazer pequenas tarefas, como auxiliar uma pessoa a se localizar em um prédio e acompanhá-la até o seu destino e a tomar decisões financeiras, por exemplo. Outros exemplos de aplicações do PLN incluem máquinas de tradução automática (*machine translation*) e pesquisas na web (*web based question answering*), como o pesquisador do *Google* (JURAFSKY; MARTIN, 2009).

Além da importância dos bancos de dados para o fortalecimento das pesquisas na área de descrição e análise linguística, bem como de PLN, outros fatores podem contribuir para influenciar os pesquisadores a sistematizar os dados coletados na área de Sociolinguística Variacionista, sendo dois deles as demandas do paradigma da Ciência Aberta e o avanço das tecnologias de armazenamento, transcrição e sistematização de dados (VANN, 2021). De acordo com Silva e Silveira (2019), o movimento da Ciência Aberta é uma iniciativa para se romper com as práticas de pesquisa tradicionais nas quais os resultados dos estudos circulavam apenas na academia. Esse paradigma incentiva a transparência, desde a concepção da pesquisa, com a publicação aberta do projeto de pesquisa, por exemplo, até seu produto, como o acesso a dissertações e teses, bem como publicação dos resultados em periódicos científicos de acesso gratuito. Busca-se também, por meio da Ciência Aberta, um maior detalhamento de metodologias e gerenciamento de dados de forma que eles possam ser facilmente acessados.

Diferentes periódicos na área de linguística, como a Revista da Abralin, revista nacional, o *International Journal of Applied Linguistics*, o *Open Linguistics,* o *Language*

Communication, entre outros têm aderido às demandas da Ciência Aberta, adicionando nas suas políticas editoriais práticas de acesso aos dados originais das pesquisas para que os estudos possam ser reproduzidos, replicados e terem sua confiabilidade atestada (LYON, 2016). Assim, a sistematização de dados linguísticos facilita o seu compartilhamento quando da publicação dos resultados das pesquisas em periódicos da área.

Apesar desse crescente interesse em sistematizar dados linguísticos em bancos, essa prática na Sociolinguística Variacionista brasileira, assim como no exterior (KENDALL, 2013), tem se dado de maneira individual e pouco padronizada. Diferentes estudos fizeram levantamentos dessas iniciativas individuais de se criarem bancos de dados sociolinguísticos e disponibilizá-los na internet, como Salomão (2011), Silva (2015), e mais recentemente, o Grupo de Trabalho de Sociolinguística (GT de Sociolinguística) da Associação Nacional de Pós-Graduação e Pesquisa em Letras e Linguística (ANPOLL), este em andamento.

Até o momento, observamos que há diversidade nas formas de coleta dos dados como estratificação por cotas fixas a exemplo do VARSUL (Variação Linguística na Região Sul do Brasil) e do SP2010, coletas diacrônicas e sincrônicas como o VARPORT (Análise Contrastiva de Variedades do Português) e coletas com variedades do Português Europeu e Brasileiro, caso do CORPORAPORT (Corpora de Variedades do Português em Análise). Há também amostras com pequenos trechos disponíveis para ilustração de cada comunidade, caso do VARSUL[1] (o restante da amostra continua restrito, ou seja, seu acesso se faz por pedido aos pesquisadores responsáveis).

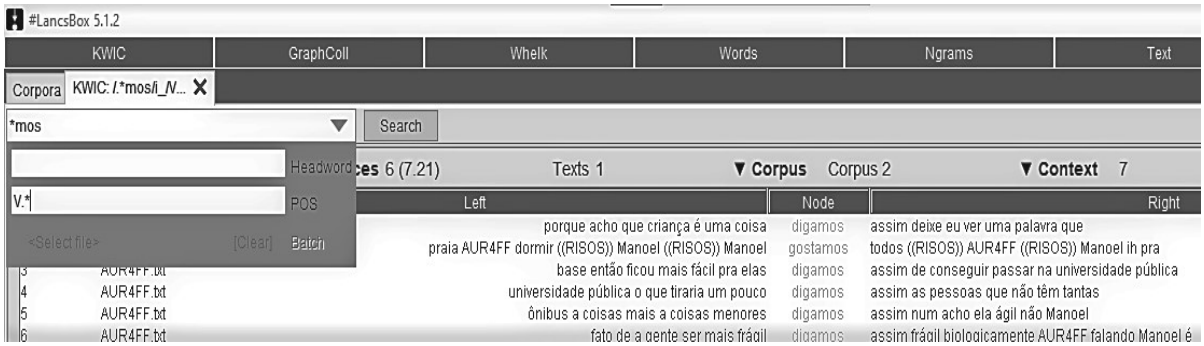
Com exceção da Amostra do NURC de Recife, nenhuma das amostras sociolinguísticas disponíveis online referenciadas nos estudos de Salomão (2011), Silva (2015) e do levantamento GT de Sociolinguística possui etiquetagem gramatical, morfológica ou sintática, sendo as buscas nos respectivos sites feitas principalmente pelas características sociodemográficas das amostras, como comunidade, faixa-etária e escolaridade. Assim, caso o pesquisador queira fazer uma comparação de fenômenos linguísticos variáveis em relação a fatores linguísticos, ele terá que baixar todos os arquivos das amostras e realizar as buscas manualmente. Com amostras etiquetadas, contudo, é possível criar uma interface de buscas considerando aspectos linguísticos.

[1] O projeto VARSUL conta com dois conteúdos: O banco de dados Varsul e a Amostra Digital Varsul. Não há, para o banco de dados informações mais específicas sobre a quantidade de tempo de cada áudio disponibilizado, consta apenas que há trechos. Já a Amostra Digital Varsul compõe-se de áudios entre cinco e 15 minutos de fala, sendo 24 das capitais do Sul do Brasil (8 de cada capital), oito do Banco Monguilhott e oito do Banco Brescancini e Valle.

A etiquetagem ou anotação linguística se refere à adição de informações linguísticas a dados, sejam de fala ou escrita, atribuindo a eles uma interpretação sobre a linguagem utilizada naquele material (LEECH, 2013; LEECH, 2004; GRIES; BEREZ, 2017). Uma anotação morfológica faz marcações concernentes ao gênero e número de um substantivo, já uma anotação gramatical ou de classe de palavras (*Part of Speech*) atribui à palavra uma etiqueta referente à sua classe gramatical, se ela é um verbo ou um substantivo, por exemplo. Como argumentam Othero e Ayres (2014), é possível que os pesquisadores que trabalham sob a perspectiva da Sociolinguística Variacionista desconheçam a potencialidade que dados anotados podem oferecer, principalmente no que tange à automatização das buscas. Por meio de linguagem de programação ou de ferramentas de busca em *corpora*, é possível fazer um levantamento por fenômenos linguísticos variáveis no *corpus* de trabalho utilizando apenas as etiquetas.

Na figura 1, apresentamos um exemplo de busca automática feita por um *software* de análises linguísticas.

Figura 1 - Busca automática na ferramenta *LancsBox* 6.0.



Fonte: Elaboração própria a partir do uso de uma entrevista da Amostra Deslocamentos 2020 no software *LancsBox* 5.1.2

Na figura 1, exemplificamos uma busca por todos os verbos flexionados na primeira pessoa do plural, por meio da expressão *mos (que nos retorna palavras que terminam em "mos") e com a etiqueta V (que indica verbos) seguida de .* (ponto e asterisco) para delimitar ocorrências apenas de verbos. Esse exemplo serve como ilustração da potencialidade de uma busca automatizada para analisar a variação da primeira pessoa do plural, utilizando atributos que retornem ocorrências em que o sujeito é nulo.

É importante ressaltar que anotar dados linguísticos é uma tarefa que demanda muito tempo, recursos financeiros e humanos, porém, uma vez anotados, diferentes pesquisadores podem reutilizá-los para análise de diferentes fenômenos linguísticos (LEECH, 2004; LEECH, 2013; HOVY; LAVID, 2010; SARDINHA, 2000). Além disso, a partir da anotação, a

reprodução dos estudos já feitos se torna mais rápida, facilitando a verificação dos resultados e a observação de possíveis inconsistências ou ratificação das análises (BEREZ-KROEKER, *et al.*, 2018).

Visando atender à lacuna de bancos de dados sociolinguísticos anotados, considerando a escassez de recursos financeiro para a área da linguística, o que consequentemente limita o quantitativo de recursos humanos e a busca por otimizar o uso de ferramentas gratuitas de anotação, nossa tese é a de que é possível utilizar recursos abertos e gratuitos para etiquetagem e sistematização de amostras sociolinguísticas seguindo o paradigma da Ciência Aberta. Essa tese, então, parte da seguinte pergunta de pesquisa: "Como utilizar recursos abertos para etiquetar e sistematizar dados de amostras sociolinguísticas seguindo o paradigma da Ciência Aberta?. Para defendermos a nossa tese, objetivamos criar um protocolo de sistematização e divulgação da amostra Deslocamentos 2019 que integra o banco de dados Falares Sergipanos (edital 02/2015 SENACON/MJ; edital CAPES/FAPITEC/SE 10/2016 PROMOB), aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal de Sergipe no processo CAAE: 0386.0.107.000-11. Esta tese também foi contemplada pelo edital nº 10/2022 POSGRAP/CINTTEC/UFS, Minha Patente UFS, que tinha como escopo o desenvolvimento tecnológico e proteção de propriedade intelectual.

Para atingir o objetivo geral, delimitamos os seguintes objetivos específicos: i) testar duas ferramentas computacionais (*LancsBox* 6.0 e *spaCy* 3.5) gratuitas na etiquetagem da Amostra Deslocamentos 2019; ii) avaliar a etiquetagem empreendida pelas ferramentas; iii) comparar o desempenho das ferramentas em relação à buscas e funcionalidades para uma pré-análise do fenômeno da variação do preenchimento de determinante antes de possessivo pré-nominal; iv) descrever ações para a divulgação e o compartilhamento dos dados da amostra Deslocamentos 2019; v) sistematizar as ações desenvolvidas em forma de protocolo.

Para realizar a produção do protocolo, escolhemos a amostra Deslocamentos 2019 do banco de dados Falares Sergipanos porque ela já se encontrava com a coleta das entrevistas realizadas e as transcrições feitas, além de já ter subsidiado o desenvolvimento de duas dissertações de mestrado (CORREA, 2019; RIBEIRO, 2019). Escolhemos como fenômeno linguístico a variação no preenchimento de determinantes antes de possessivo pré-nominal para a testagem das ferramentas em termos de etiquetagem e funcionalidades, por ser um fenômeno no nível morfossintático, descrito em diferentes cidades do Brasil (GUEDES, 2019; SEDRINS; PEREIRA; SILVA, 2019), e pesquisado também com amostras do banco de dados Falares Sergipanos (SILVA, 2020; SIQUEIRA, 2020). Mais detalhes acerca da amostra, assim como

do fenômeno que serviu como teste para o protocolo serão oferecidos em seções específicas da tese.

A sistematização da amostra Deslocamentos 2019 do banco de dados "Falares Sergipanos", visando principalmente a etiquetagem POS, e sua disponibilização em livre acesso para outros acadêmicos permitirá o cotejamento com outros bancos, otimizando a descrição do português brasileiro. Além disso, como assinala Silva (2015), a constituição de bancos de dados linguísticos deve ser divulgada, visto que, internacionalmente, os conhecimentos das pesquisas sobre a língua portuguesa ainda são incipientes e não possuem "divulgação adequada entre os estudiosos do português" (SILVA, 2015, p. 169). Ainda de acordo com a autora, os bancos de dados servem como registro e contribuem para traçar um perfil sociolinguístico das variedades da língua portuguesa, facilitando a identificação das particularidades e semelhanças entre elas. Por fim, ao criar um protocolo, rotinas procedurais são estabelecidas, tornando o trabalho de mapeamento de variáveis linguísticas de uma determinada comunidade menos dispendioso, subsidiando a replicabilidade, e comparabilidade com outras comunidades, gerando análises mais robustas.

Este texto está dividido em quatro capítulos. Nesta introdução, apresentamos nosso objeto de estudo e delimitamos os objetivos a serem alcançados com esta tese. No capítulo 1, "Ciência Aberta e bancos de dados linguísticos: acessibilidade e transparência" inserimos nosso estudo no escopo da Ciência Aberta, apontando para a necessidade de metodologias mais transparentes e a adoção de práticas de acessibilidade para o intercâmbio de dados subsidiando a reprodução e replicação de pesquisas. No capítulo 2, "Bancos de dados sociolinguísticos etiquetados: possibilidades", destacamos a importância dos bancos de dados para diferentes áreas, apresentando o banco de dados Falares Sergipanos como fonte de formação de recursos humanos. Estabelecemos também como os procedimentos de etiquetagem e disponibilização de dados de fala advindos de uma metodologia de pesquisa inserida no escopo da Sociolinguística Variacionista pode ser importante para avanços tanto nessa área como na área de PLN. No terceiro capítulo, apresentamos os processos e os resultados para defender a nossa tese. No quarto e último capítulo, tecemos nossas considerações finais.

1 CIÊNCIA ABERTA E BANCOS DE DADOS LINGUÍSTICOS: ACESSIBILIDADE E APLICAÇÕES

Objetivamos, com este capítulo, inserir nosso protocolo de sistematização e divulgação de uma amostra do banco de dados Falares Sergipanos dentro do paradigma da Ciência Aberta. Dividimos o capítulo em cinco seções principais: na primeira, discutimos a necessidade de se tornar a acessibilidade como prática a ser adotada pelos pesquisadores, apresentando brevemente fatores que levaram à crise na ciência e como o movimento Ciência Aberta tenta contribuir para minimizar esses efeitos.

Na segunda seção, descrevemos como tem sido feita a sistematização de amostras sociolinguísticas no contexto brasileiro. Já na terceira, fazemos um breve panorama dos bancos de dados de fala sociolinguísticos brasileiros e no exterior no que tange a transparência no acesso aos dados e na metodologia, evidenciando os pontos positivos da abertura dos dados e da metodologia para área de Sociolinguística Variacionista. Na quarta seção, apresentamos as aplicações dos bancos de dados linguísticos, de maneira geral, para diferentes propósitos, reforçando a necessidade de maior acesso a dados linguísticos. Por fim, na quinta seção, apresentamos o banco de dados Falares Sergipanos, foco de nossa tese, para mostrar sua relevância para a descrição da variedade sergipana do português brasileiro e também para a formação de recursos humanos.

1.1 Crise na ciência e Ciência Aberta

Terraplanismo, negacionismo, descrença. A ciência nunca foi tão questionada quanto nos últimos anos. Como reporta Andrade (2019), em pesquisa realizada pelo Instituto Gallup com um grupo de 140 mil pessoas em diferentes países, foi observado que o nível de desconfiança de brasileiros, franceses e japoneses em relação à ciência está acima de 70%, evidenciando que o problema não reside apenas em países em desenvolvimento, como o Brasil, mas também em países desenvolvidos como Japão e França.

Em uma perspectiva antropológica centrada na explicação digital para a pós-verdade, Cesarino (2021) discute como o populismo digital e o neoliberalismo têm favorecido uma desestabilização do sistema de peritos, fazendo surgir o que é chamado de regime de pós-verdade. Nesse sentido, novas formas de conhecimento emergentes criadas a partir de experiências idiossincráticas, acentuadas pela velocidade de propagação desse conhecimento,

pelo populismo digital e o neoliberalismo, vêm ocupando o espaço da ciência produzida na academia. Nesse sentido, há um esvaziamento dos sentidos produzidos pela academia e uma validação de informações falsas e daquelas produzidas pela experiência, ambas rapidamente disseminadas em mídias sociais.

Para Fidler e Wilcox (2021), dentro da dinâmica dos modos de se fazer ciência, entre os fatores que têm ocasionado o questionamento das ciências, encontra-se o fato de que projetos de reprodutibilidade em diferentes áreas tiveram resultados pouco satisfatórios. A partir de um levantamento bibliográfico, os autores verificaram que existe uma crise de reprodutibilidade e replicabilidade nas ciências que está relacionada às seguintes questões: a) ausência de replicação de estudos em várias áreas; b) falha na reprodução de estudos; c) enviesamento das publicações; d) práticas questionáveis na pesquisa; e e) falta de transparência e completude seja nos métodos, dados e/ou análises nas publicações.

O movimento da Ciência Aberta, conforme Spellman, Gilbert e Corker (2017) é uma proposta para contribuir na solução dessa crise, trazendo, então, discussões em torno das questões reportadas no parágrafo anterior e procurando estabelecer critérios para tornar a ciência mais transparente e acessível, resultando em uma ciência mais replicável e robusta. Em outras palavras, há um maior detalhamento de metodologias e de gerenciamento dos dados coletados (SILVA; SILVEIRA, 2019), facilitando o acesso a esses dados e, por consequência, a reprodução e a replicação dos estudos.

De acordo com Wilkinson *et al*. (2016), os princípios norteadores de gerenciamento de dados, dentro do paradigma da Ciência Aberta, podem ser resumidos pelo acrônimo FAIR, segundo o qual F significa *findable* (ser localizável), A significa *accessible* (ser acessível), I, *interoperable* (ser interoperável) e R, *reusable* (ser reutilizável). No quadro 1 abaixo, reproduzimos os descritores de cada um dos princípios FAIR.

Quadro 1 – Descrição dos princípios FAIR.

Os Princípios Norteadores FAIR
Ser localizável: F1 Atribuir um identificador exclusivo e persistente aos dados e metadados. F2 Descrever dados com metadados ricos (definidos pelo R1 abaixo). F3 Incluir o identificador dos dados descritos nos metadados de maneira clara e explícita. F4 Indexar e registrar (meta) dados em um repositório pesquisável **Ser acessível:** A1. Usar protocolos padronizados para recuperar dados e metadados pelo seu identificador. A1.1 Os protocolos devem ser abertos, gratuitos e universalmente aplicáveis. A1.2 Os protocolos devem permitir um procedimento de autenticação e autorização, se necessário. A2. Garantir a acessibilidade aos metadados, mesmo que os dados não estejam mais disponíveis. **Ser interoperável**: I1 Usar linguagens formais (acessíveis, compartilhadas e padronizadas) para representar o conhecimento nos (meta)dados. I2 Usar um vocabulário que também segue os princípios do FAIR nos meta(dados). I3 Incluir referências a outros (meta)dados. **Ser reutilizável**: R1 Descrever os (meta)dados com uma pluralidade de atributos precisos e relevantes. R1.1. Publicar (meta)dados com uma licença de uso clara e acessível. R1.2. Usar detalhes de proveniência na descrição dos (meta)dados R1.3. Seguir padrões comuns da área de conhecimento a que os (meta)dados se referem.

Fonte: Adaptado de Wilkinson *et al.* (2016) para a língua portuguesa.

Os princípios acima descrevem em linhas gerais o que um pesquisador deve fazer para garantir que seu conjunto de dados e metadados estejam organizados e gerenciados para que eles possam ser compartilhados de maneira segura e padronizada na comunidade científica a que o pesquisador pertence. Os descritores apontam, por exemplo, que o depósito dos (meta)dados seja feito em repositórios pesquisáveis, para facilitar a sua localização por outros pesquisadores. Apontam também que, para um conjunto de dados ser considerado acessível, seus (meta)dados devem ser recuperáveis por meio de um identificador padronizado por um protocolo de comunicação aberto, gratuito, que permita autenticação e autorização, e também que, mesmo em caso de os dados não estarem mais disponíveis, seus metadados sejam ainda acessíveis. Em outras palavras, espera-se que qualquer usuário consiga o acesso ao conjunto sem a necessidade de ferramentas ou conhecimentos especializados, e que mesmo na indisponibilidade dos dados por algum motivo, seus metadados ainda sejam recuperáveis.

Para que um conjunto de dados seja considerado reutilizável, por sua vez, os meta(dados) devem ser liberados por meio de uma licença de uso clara e acessível, ter sua procedência detalhada e estar de acordo com domínios relevantes para a comunidade. Nesse caso, para que o usuário de uma determinada comunidade acadêmica verifique se os dados são úteis a ele, é necessário que os metadados estabeleçam uma descrição pormenorizada dos dados em diferentes camadas. Na Sociolinguística Variacionista, isso se daria, por exemplo, no detalhamento das informações sociodemográficas da comunidade e dos informantes pesquisados, bem como dos critérios de seleção desses informantes, dos modelos estatísticos empregados, das ferramentas computacionais empregadas. Percebemos pelos descritores dos princípios, então, que a padronização nos procedimentos de organização e gerenciamento de dados possibilita uma melhor comunicação entre aqueles interessados em divulgar e compartilhar seus dados, assim como reutilizar dados de outros pesquisadores.

Lyon (2016) argumenta que a maior transparência na condução e publicação das pesquisas acarreta não só maior fortalecimento do rigor metodológico e experimental, mas também "o fortalecimento de percepções públicas sobre a qualidade e integridade das pesquisas e confiança nos resultados, afirmações, conclusões e asserções derivadas das atividades de pesquisas"[2] (LYON, 2016, p. 161).

A ideia da Ciência Aberta, por meio desses princípios, é incentivar a transparência. Ao trabalhar nesse paradigma, podemos contribuir para a diminuição da crise da reprodutibilidade nas pesquisas ocasionada principalmente pelo não detalhamento da metodologia e da disponibilização dos dados utilizados em estudos científicos, no nosso caso, amostras de fala, e aumentar a confiança pública nos resultados alcançados.

1.2 A sistematização de amostras sociolinguísticas e acessibilidade

No Brasil, apesar de já existirem repositórios para gerenciamento de dados e de publicações em áreas como ciências agrárias[3] e biológicas[4] aderentes ao paradigma da Ciência

[2] Tradução nossa para o original: "strengthen public perceptions of research quality, integrity and trust in the results, claims, conclusions and assertions derived from research activities.".

[3] No repositório ALICE (Acesso Livre à Informação da Embrapa), os pesquisadores da Embrapa (Empresa Brasileira de Agropecuária) depositam dissertações, teses, notas técnicas, entre outros documentos como forma de disseminar o conhecimento produzido pelos pesquisadores da empresa. Esse repositório encontra-se disponível em: https://www.alice.cnptia.embrapa.br/. Acesso em: 12 jul. 2021.

[4] O SisGen (Sistema Nacional de Gestão do Patrimônio Genético e do Conhecimento Tradicional Associado) é um repositório que permite ao usuário depositar e acessar dados do patrimônio genético nacional, por exemplo, amostras de plantas, dados genéticos entre outros. O SisGen está disponível em: https://sisgen.gov.br/paginas/InstallSolution.aspx. Acesso em: 12 jul. 2021.

Aberta, a pesquisa na área de Sociolinguística carece desse tipo de gerenciamento (FREITAG *et al.*, 2021). Freitag (2017a) constatou que o acesso a bancos de dados sociolinguísticos ainda é restrito, sendo que grande parte deles se encontra em dispositivos de armazenamento móveis do próprio pesquisador. Isso, muitas vezes, dificulta a reprodução de estudos e análises contrastivas. Além desses fatores, situações prosaicas como problemas com HDs, queima de computadores e mofo são fatos que tornam desejável a adoção de um plano de preservação de dados para que eles não sejam perdidos (TAGLIAMONTE, 2006).

Avanços tecnológicos, contudo, têm favorecido o armazenamento, sistemas de anotação e digitalização de amostras sociolinguísticas, e também a disponibilização dessas amostras. Vann (2021) argumenta que devido a esse avanço, ferramentas que possibilitam a gravação e a disseminação de áudios levam a uma nova concepção de registro linguístico, tornando a transcrição impressa obsoleta. Nesse sentido, *corpora* orais de terceira geração são considerados aqueles possuem transcrição e áudios alinhados como aponta Mello (2021). Esse tipo de alinhamento entre áudio e transcrição tem se consolidado como rotina procedural no âmbito do banco de dados Falares Sergipanos.

Adicionalmente, devido ao movimento da Ciência Aberta, os periódicos têm incentivado o acesso aos dados, evidenciando a necessidade de ter um plano de gerenciamento para torná-los totalmente digitais e sistematizados. Como exemplo de incentivo à disponibilização de dados, o periódico *Open Linguistics*[5] encoraja os autores a fornecer acesso irrestrito a tudo aquilo que foi necessário para gerar os resultados das pesquisas a serem publicadas pelo periódico, incluindo os dados crus, códigos, tabulações entre outros. Conforme a política editorial desse periódico, a acessibilidade aos dados favorece a replicação, a robustez e a solidez do trabalho, aumentando a qualidade das pesquisas da área e, consequentemente, o aumento quantitativo do fator de impacto e citações, métricas importantes para a concessão de fundos para pesquisa. Seguindo a mesma política editorial em consonância com o movimento Ciência Aberta, a Revista da Abralin[6] também incentiva o compartilhamento dos dados, *scripts* para análises estatísticas e quaisquer outros materiais dessa natureza. Esses exemplos de ações demonstram o crescente interesse em sistematizar os dados advindos de pesquisas

[5] Na seção "*Suplementary Materials*" (materiais suplementares) do site deste periódico, existe um documento para download chamado "Data Sharing Policy" (política de compartilhamento dos dados) no qual constam as diretrizes para o compartilhamento dos dados. Disponível em: https://www.degruyter.com/journal/key/opli/html. Acesso em: 10 out. 2021.

[6] No site da revista, na aba "Sobre", no tópico "Ciência Aberta", estão contidas as medidas que o periódico toma em consonância com os preceitos de Ciência Aberta, entre elas, o incentivo ao compartilhamento dos dados e materiais usados em pesquisa. Disponível em: https://revista.abralin.org/index.php/abralin/about. Acesso em: 10 out. 2021.

sociolinguísticas em bancos de dados alocados, preferencialmente, em sites ou repositórios online.

Em 2020, diferentes simpósios realizados no evento ABRALIN Ao Vivo, notadamente "Grandes Projetos da Linguística em rede: "Tycho Brahe, PROHPOR e PHPB"[7], "*Archiving and Language Documentation*"[8] e "Descrição linguística: gestão de dados linguísticos"[9], discutiram ações sobre o gerenciamento de dados linguísticos frente às demandas da Ciência Aberta. O painel temático "Futuros Possíveis para Dados Sociolinguísticos"[10], no Festival do Conhecimento da Universidade Federal do Rio de Janeiro em 2021, também teve como foco a gestão de dados sociolinguísticos evidenciando o interesse por projetos de sistematização, gerenciamento e preservação de dados linguísticos.

Kendall (2013) relata que, nos Estados Unidos, vários pesquisadores já vinham apontando para a necessidade de ações conjuntas para o desenvolvimento de *corpora* sociolinguísticos, no qual *corpus* é considerado em um sentido menos prototípico. No sentido menos prototípico, agrupam-se conjuntos de dados linguísticos que deixam de atender a um ou mais critérios que caracterizam um *corpus* tradicionalmente descrito na Linguística de *Corpus*, como representatividade, balanceamento, coleta naturalística, e possibilidade de ser lido por máquinas (GRIES; BEREZ, 2017). Kendall (2013) afirma que mesmo com o interesse de ações conjuntas, a sistematização dos dados linguísticos nos Estados Unidos tem se dado no nível individual. O mesmo tem acontecido no Brasil, em que alguns pesquisadores já disponibilizam algumas de suas amostras online de forma individual, porém, diversas publicações sobre o tema (FREITAG; MARTINS; TAVARES, 2012; FREITAG, 2014; 2016; 2017a; 2017b) consideram necessária a padronização tanto em termos de coleta quanto de disponibilização.

Em 2019, o GT de Sociolinguística decidiu de maneira conjunta prospectar e começar a realizar ações para que um repositório consorciado da área seja criado. A partir de 2021, o projeto tomou fôlego e se concretizou em uma proposta intitulada "Plataforma Digital da Diversidade Linguística Brasileira", a qual foi apresentada em setembro no Congresso Internacional da Abralin[11] (MACHADO VIEIRA *et. al*, 2021) e posteriormente publicada.

Um dos aspectos de implementação da proposta da plataforma é o mapeamento de amostras sociolinguísticas compiladas no Brasil (tabela 1).

[7] Disponível em: https://www.youtube.com/watch?v=gJgrqArfDIw. Acesso em 10 nov. 2020.
[8] Disponível em: https://www.youtube.com/watch?v=uQY4dZnQKds&t=269s. Acesso em 15 ago. 2020.
[9] Disponível em: https://www.youtube.com/watch?v=S7YS57i7ogs&t=7348s. Acesso em: 15 ago. 2020.
[10] Disponível em: https://www.youtube.com/watch?v=ZrZxsd5QQns. Acesso em: 12 jul. 2021
[11] Disponível em: https://www.youtube.com/watch?v=BsCvqcTo-qc&t=10s. Acesso em 24 set. 2021

Tabela 1 – Bancos de dados com site.

Nome do Projeto	Sigla	Página na internet	Indica como citar	Local	Amostras no site
Variação Linguística no Português Alagoano	PORTAL	https://www.portuguesalagoano.com.br	Não.	AL/BR	Sim.
Corpora de Variedades do Português em Análise	COPORAPORT	https://corporaport.letras.ufrj.br/	Sim.	RJ/BR, PT e MOÇ	Sim.
Análise Contrastiva de Variedades do Português	VARPORT	https://varport.letras.ufrj.br/	Não.	BR e PT	Sim.[1]
Vertentes do Português Popular do Estado da Bahia	VERTENTES	http://www.vertentes.ufba.br/	Não.	BA/BR	Não.
Variação Linguística na Região Sul do Brasil	VARSUL	http://www.varsul.org.br/	Não.	RS, SC e PR/BR	Sim.
Amostra Linguística no Interior Paulista	ALIP	https://www.alip.ibilce.unesp.br/	Sim.	SP/BR	Sim.
Núcleo de Estudos sobre Interlínguas	NEIS	https://corpusneis.wixsite.com/home	Não.	RJ/BR	Não.
Corpus Histórico da Língua Portuguesa	HISTLING	https://histling.letras.ufrj.br/index.php	Não.	Variado	Sim.
Projeto Variação Linguística no Estado da Paraíba	VALPB	http://projetovalpb.com.br/index.html	Não.	PB/BR	Não.[2]
Programa de Estudos Sobre o Uso da Língua	PEUL	https://peul.letras.ufrj.br/	Não.	RJ/BR	Sim
Grupo de Estudos Variacionistas	GEVAR	https://sites.google.com/site/uftmgevar/	Não.	MG/BR	Não.
Núcleo de Estudos do Português em Uso	PORUS	http://porus.sites.uff.br/	Não.	RJ e MG/BR	Sim.
Não possui nome por extenso.	LínguaPOA	https://www.ufrgs.br/linguapoa/	Sim.	RS/BR	Não.

1. O acesso aos dados se faz por meio de *links*, mas alguns não funcionam. 2. Os *links* de acesso aos dados direcionam para pastas vazias, ou informam que os *links* não estão disponíveis. **Fonte:** Elaboração própria a partir do levantamento do GT de Sociolinguística em 2021 e apresentado em Machado Vieira *et. al* (2021)

Até setembro de 2021, o GT já havia mapeado 24 bancos de dados, em termos de 24 respondentes ao formulário de mapeamento. De acordo com Machado Vieira *et. al* (2021), a maioria dessas amostras está vinculada a programas de pós-graduação e localizada na faixa litorânea do Brasil, o que pode estar relacionado à densidade populacional nessas áreas, conforme a pesquisadora.

Pode ser observado pela tabela 1 que: sete, dos 13 projetos que possuem site na internet, disponibilizam amostras em seus sites; três possuem amostras coletadas na região nordeste do Brasil (PORTAL, VERTENTES e VALPB); dois, na região sul (VARSUL e LÍNGUAPOA); cinco, na região sudeste (ALIP, NEIS, PEUL, GEVAR, PORUS); dois (CORPORAPORT e VARPORT), no Brasil (Rio de Janeiro) e em outros países; um com coleta de cartas e documentos produzidos por informantes com origem variável. Sobre referência aos dados, apenas três indicam as formas de citá-los (CORPORAPORT, ALIP e LÍNGUAPOA), evidenciando a necessidade de se adotar medidas mais claras em relação aos colaboradores.

As ações do GT de Sociolinguística juntamente com os projetos que já possuem site, independentemente de já disponibilizarem suas amostras ou não, ou de indicarem formas de citar os dados, demonstram a preocupação na sistematização dos dados bem como a sua salvaguarda. Além disso, essas ações corroboram a afirmação de Kendall (2013) de que nas diferentes áreas da linguística tem havido a busca por tornar os dados compartilháveis, mas que, na Sociolinguística Variacionista, esse tipo de ação tem se concretizado ainda no nível individual (atrelados a grupos de pesquisas isolados), o que pode estar impedindo avanços na área. De fato, dados coletados no âmbito da Sociolinguística Variacionista têm maior variabilidade, caso já estivessem há mais tempo disponíveis, seria possível termos ferramentas de processamento linguístico mais sensíveis à variação, por exemplo.

O número maior de sites com dados compartilháveis também confirma as experiências de Kendall (2008) e Tagliamonte (2006) sobre ter amostras sistematizadas digitalmente. Os autores perceberam que com a sistematização das amostras em um repositório, houve maior integração entre os dados, melhora nas análises devido ao fácil acesso aos dados, maior facilidade em colaborar com pesquisas e compartilhamento de dados e resultados.

Contudo, no Brasil, conforme Freitag (2021b), ainda existem alguns desafios na constituição desses repositórios que precisam ser superados em relação aos seguintes aspectos: depreciação, autoria, compartilhamento e financiamento. A depreciação e o financiamento são aspectos que estão interligados. Nesse sentido, toda a constituição de uma estrutura para se produzir uma pesquisa com dados linguísticos bem como a sua salvaguarda demanda

financiamento. A formação de pesquisadores em diferentes níveis (graduação e pós-graduação), a digitalização de amostras antigas, procedimentos de *backup* para amostras já digitalizadas são limitadas pela escassez de recursos para tal fim.

Sobre a autoria, como visto acima, dos bancos de dados que possuem site, apenas três indicam a forma de como citar os dados. O CORPORAPORT e o ALIP indicam os coordenadores dos projetos que deram origem aos bancos de dados como seus respectivos autores, já o LINGUAPOA indica a citação ao site, seguindo as recomendações Tromsø [12] para citações de dados em Linguística**.** Essa diferença na forma de citação aos bancos de dados confirma a necessidade de se refletir sobre questões relacionadas à autoria, como aponta Freitag (2021b). A pesquisadora recomenda a descrição dos papéis desempenhados por cada pesquisador na compilação e gerenciamento dos dados utilizando a taxonomia CRediT[13], o que assegura que o trabalho de cada um dos envolvidos na pesquisa seja reconhecido, e também recomenda o uso de *copyright*, uma vez que a constituição das amostras é um produto intelectual. Nesta tese, defendemos que a autoria seja reconhecida e seja dado o devido crédito a todos que colaboraram para a criação de amostras e de bancos de dados sociolinguísticos, principalmente para fortalecer as redes de colaboração dentro da Sociolinguística.

No que diz respeito ao compartilhamento, Freitag (2021b) afirma que essa é uma das ações importantes na constituição de repositórios, visto que o acesso aos dados vem sendo motivado por periódicos, como descrito acima, e também para conferir maior transparência e confiabilidade às pesquisas. A autora salienta que como pesquisadores, instituições e controladores envolvidos na pesquisa assumem a responsabilidade ética e legal pelos dados coletados, estes não podem ser disponibilizados à revelia apenas por se tratarem de produtos gerados a partir de recursos públicos. Além disso, dados linguísticos são importantes para o PLN, sendo considerados de grande valor comercial de forma que empresas privadas não devem obter acesso a eles sem a garantia de uma contrapartida (cf. FREITAG, 2021b; GARELLEK et al., 2020). Assim como a pesquisadora, argumentamos a favor da utilização de licenças de uso assegurando níveis de acesso aos dados.

Nesta seção, apresentamos sistematização e divulgação de amostras sociolinguísticas, evidenciando não somente a importância dessas ações para desenvolvimentos no âmbito das

[12] Essas recomendações seguem princípios de ciência aberta e encontram-se disponíveis no site: https://www.rd-alliance.org/group/linguistics-data-ig/outcomes/troms%C3%B8-recommendations-citation-research-data-linguistics. Acesso em 13 dez. 2021.

[13] A taxonomia CrediT retrata os papéis desempenhados pelos colaboradores no desenvolvimento da pesquisa desde a conceptualização até a escrita final. Para maiores detalhes sobre como usar e a descrição de cada papel veja o website: https://credit.niso.org/

pesquisas, mas também em outras áreas. Ao final, apontamos alguns desafios para se atingir os objetivos de sistematizar e divulgar as amostras. Dito isso, baseando-nos nos princípios norteadores de gerenciamento do *Linguistic Data Consortium* (LDC), acreditamos que o conhecimento sobre gerenciamento de dados (acesso, uso e arquivamento, entre outros) é um passo importante para assegurar a efetividade das pesquisas baseadas em dados.

1.3 Ciência Aberta e bancos de dados sociolinguísticos no exterior e no Brasil: metodologia e transparência

Na Sociolinguística Variacionista, devido à preocupação em se entender como a mudança ocorre, a língua é estudada dentro da comunidade, e, por isso, consegue-se explicar a variação em situações de contato, a coexistência de diferentes formas linguísticas para indicar mesmo valor referencial dentro de uma mesma comunidade, e como os usos da língua são sistemáticos e estruturados conforme a sociedade (WEINREICH; LABOV; HERZOG, 1968; LABOV, 2006 [1966], 2008 [1972]). Para se fazer análise linguística dentro dessa perspectiva, dados são coletados, principalmente por meio de entrevistas sociolinguísticas. As análises recorrem a modelos estatísticos para explicar quais fatores (de natureza linguística, social ou cognitiva) interferem no uso de determinada variante em relação a outra (LABOV, 1969; TAGLIAMONTE, 2012; WEINREICH; LABOV; HERZOG, 1968).

No modelo tradicional de coleta de dados, parte-se da premissa de que a fala dos indivíduos representa as normas da comunidade de fala (TAGLIAMONTE, 2012), logo, para se captar os padrões linguísticos de uma dada comunidade, é preciso ter uma amostra que reflita esses padrões. A escolha por entrevistas, então, não é feita ao acaso. Apesar dos efeitos do paradoxo do observador, conforme Labov (1981), a entrevista sociolinguística representa o melhor meio para se capturar o vernáculo em grandes volumes (entre uma e duas horas de duração) e com melhor qualidade em termos de gravação.

Tradicionalmente a escolha dos indivíduos para compor a amostra a ser estudada pode ser feita de duas maneiras: amostra aleatória simples e amostra aleatória estratificada (SILVA, 2004). De acordo com Silva (2004), na amostra aleatória simples busca-se pela quantidade de informantes proporcionalmente à sua representação na população. Já a amostra aleatória estratificada é dividida em estratos (ou células sociais), as quais são compostas por indivíduos aleatoriamente selecionados, mas que compartilham das mesmas características sociais. No estudo empreendido por Labov em Nova Iorque, por exemplo, foi empregada a amostragem

aleatória estratificada, pois esta pode oferecer uma representação adequada do grupo-alvo a ser estudado, no caso, falantes de inglês considerados nativos de Nova Iorque (LABOV, 2006[1966]).

Para uma amostragem ser considerada aleatória, a seleção dos informantes deve ser feita ao acaso, ou seja, qualquer indivíduo na população teria chances iguais de ser selecionado. Esse procedimento pode ser feito utilizando informações obtidas por agentes de saúde do Programa Saúde da Família, como feito na constituição da amostra do Povoado Açuzinho em Sergipe (FREITAG; SANTANA; ANDRADE, 2014), ou por meio de dados do censo como levantado pelo projeto NORPORFOR (Norma Popular Oral de Fortaleza) (ARAÚJO, 2011).

Em relação à quantidade de informantes, Labov (2006 [1966]) argumenta que, embora importante, apenas volume não indica a eficácia do método, sendo o detalhamento da metodologia de coleta (seleção da área, detalhes geográficos, composição da *survey*, detalhamento dos critérios de amostragem, quantidade de respondentes e fontes de erro) também significativo. Ademais, fazendo um retrospecto acerca da composição da amostragem empregada no estudo de Nova Iorque, Labov (2001, p.80) argumenta que:

> embora haja uma variação individual considerável dentro de cada grupo, ela não é normalmente grande o suficiente para perturbar a regularidade do padrão quando entre 5 e 10 falantes são incluídos em cada grupo. Indivíduos cujo desvio da média é suficientemente grande para perturbar o padrão são marcados por histórias sociais irregulares.

Dessa forma, para preencher as células sociais adequadamente, elas devem ter o mínimo de 5 informantes para que não haja discrepâncias no padrão linguístico apresentado. Feagin (2013) argumenta que, na prática, a Sociolinguística Variacionista faz sua análise baseada na quantidade de *tokens*[14] por informante, ou seja, a quantidade de dados por informante torna-se mais relevante do que a quantidade de informantes por si. Freitag (2018a), por outro lado, salienta a importância do fenômeno para a amostragem, argumentando, com base em Meyerhoff, Schleef e Mackenzie (2015), que resultados com boa confiabilidade e acurácia são atingidos quando cada fator apresenta 30 ocorrências por célula, de forma que em fenômenos em níveis de análise mais altos e raros sejam necessários mais informantes e horas de gravação. Por outro lado, fenômenos fonológicos podem ser encontrados em um dimensionamento menor da amostra.

[14] *Tokens* nesta tese são definidos como unidades linguísticas por exemplo, palavras, sons, frases; já *types* seriam os padrões linguísticos dessas unidades, ou seja, combinações morfológicas, sintáticas, semânticas entre elas, sendo a relação entre frequência de *type* e de *token* calculada para aferir a produtividade de uma variante (BYBEE, 2007).

De acordo com Freitag (2017b; 2018a), a amostragem, na maioria das pesquisas brasileiras, tem sido feita: em termos de cotas, ou seja, são determinadas as quantidades dos participantes por células sociais a priori; conveniência, os documentadores buscam aqueles informantes que se disponibilizam; ou julgamento, os informantes são selecionados por sua adequação à geração de dados para pesquisa. Silva (2004), por exemplo, aponta que existem amostras no Brasil com células (estratos sociais) de até dois informantes (como no caso de Araújo e Almeida (2014)).

Araújo e Almeida (2014, p. 44) ao descreverem a constituição da amostra de fase 3 do projeto "A Língua Portuguesa no Semiárido Baiano" relatam algumas razões para justificar um tamanho diferente daquele considerado ideal: "falta de financiamentos, dificuldades em se conseguir informantes com certos perfis, perda de entrevistas por problemas técnicos e tempo para conseguir fazer a coleta". Esses mesmos desafios também foram relatados na constituição do Banco de Dados Fala-Natal (TAVARES; MARTINS, 2014), do Projeto SP2010 (MENDES, 2011) e do Banco RECI (SILVA; COELHO, 2020).

Labov (2001) argumenta que estudos não aleatórios têm produzido resultados que oferecem informações relevantes sobre variáveis sociolinguísticas influenciadas por "graus de distância social, e iluminam os mecanismos sociais que levam à conformidade e diversidade linguística"[15] (p. 39), ou seja, por meio desses estudos, é possível observar a atuação das variáveis sociais na diversidade linguística. Contudo, ainda segundo o mesmo autor, os resultados advindos desses estudos não possuem um caráter explanatório que permita generalizações acerca do comportamento linguístico da comunidade de fala.

Não estamos fazendo uma crítica aos trabalhos acima. Ao contrário, estamos relatando como o levantamento de dados no escopo da sociolinguística no Brasil é condicionado por questões contextuais que são discrepantes em relação aos estudos clássicos empreendidos por Labov na década de 60 nos Estados Unidos. Além disso, dificuldades em se replicar a metodologia de Nova Iorque também são encontradas por pesquisadores em outras partes do mundo, como em Toronto (TAGLIAMONTE, 2012), e outros estudos apontados pelo próprio Labov (2001)[16], como o de Haeri (1996) no Cairo e Milroy e Milroy (1978) em Belfast.

O trabalho de constituição de amostras de fala, como visto acima, não é somente caro, mas também dispendioso em termos do tempo gasto para se realizá-lo, e esse fato, como

[15] Tradução nossa para o original: "degrees of social distance, and illuminate the social mechanisms that lead to linguistic conformity and diversity".

[16] Uma lista com mais estudos que não seguiram estritamente a metodologia utilizada em Nova Iorque encontra-se em Labov (2001, p. 39). Esses mesmos estudos foram também descritos no capítulo 15 de Labov (2006).

apontam Freitag, Martins e Tavares (2012), não consegue ser refletido na metodologia das publicações na área de sociolinguística. Os autores relatam que existe uma padronização na escrita, sempre em voz passiva e com construções como "como *corpus* foram selecionados X informantes do banco de dados Y, estratificados em Z células sociais" (FREITAG; MARTINS, TAVARES, 2012, p. 917), de forma que o delineamento da pesquisa, os desafios em se fazer a amostragem, a entrada em campo, detalhes de transcrição, tempo dispendido, scripts e softwares utilizados, ficam em segundo plano, como também reportado por Freitag (2017b).

Salientamos, conforme Freitag e Rost-Snichelotto (2015) e Freitag (2018a), que a forma como é feita a amostragem traz implicações metodológicas, pois ao se manter cotas fixas, a comparabilidade com outras amostras torna-se possível, mas uma amostragem proporcional à estratificação consegue representar melhor a população em estudo, sendo que em qualquer uma das metodologias empregadas, para se manter um rigor estatístico, grandes quantidades de dados devem ser geradas. Devido a isso, ressaltamos a necessidade apontada em diferentes estudos (COLLISCHONN; MONARETTO, 2012; FREITAG, 2016, 2018a; FREITAG; MARTINS; TAVARES, 2012; LABOV, 2006 [1966]) de que os critérios de estabelecimento das amostras estejam evidentes nas publicações da área para que a confiabilidade dos resultados possa ser atestada.

Os fatos mencionados na seção anterior acerca do armazenamento e disponibilização de bancos de dados sociolinguísticos contrastam com o paradigma da Ciência Aberta. A falta de acessibilidade aos dados e ao delineamento metodológico que fundamenta as pesquisas na área de sociolinguística vai de encontro aos princípios *Accessible* e *Reusable*, isto é, aos princípios de os dados serem acessíveis e reutilizáveis. Gilmore, Kennedy e Adolph (2018) argumentam por mais transparência e abertura das pesquisas. Segundo os autores, o compartilhamento da pesquisa como um todo facilita a reprodução dos resultados originais e também a identificação de fatores que podem levar a resultados discrepantes. Além disso, segundo esses pesquisadores, o compartilhamento leva ao reuso dos dados, o que pode favorecer novas descobertas científicas.

No contexto internacional, por outro lado, a adoção do alinhamento entre a constituição de bancos de dados sociolinguísticos às práticas da Ciência Aberta, considerando-se o que está preconizado pelos princípios FAIR tem sido uma prática recorrente (CALAMAI; FRONTINI, 2018). Pesquisadores têm buscado publicar as formas como eles transcrevem e etiquetam os dados e também disponibilizar esses dados em plataforma com níveis de acesso, principalmente

em repositórios consorciados, como o *Linguistic Data Consortium*[17], *The Language Archive*[18], *Talk Bank*[19] entre outros.

Oez (2018), por exemplo, descreve como foi feita a documentação do dialeto Beth Qustan, que pertence a uma língua neo-aramaica chamada Turyo falada na região central da província Mardin no sudeste da Turquia, reportando tanto informações de procedimentos de amostragem e preceitos éticos como também a transcrição e o nível de anotação do *corpus*. O autor menciona também as ferramentas ELAN (utilizada para transcrição e segmentação) e FleX para a tradução e anotação morfológica. Além disso, o pesquisador alocou seus dados em um repositório consorciado (ELAR – *Endengered Language Archive*[20]) com acesso livre. Embora se trate da descrição de uma variedade ameaçada, uma vez disponibilizado pelo autor, o protocolo pode ser replicado em outras situações.

Outros bancos de dados sociolinguísticos disponíveis e de livre acesso no contexto internacional são: *Corpus del español en el sur de Arizona* (CESA) (CARVALHO, 2012) que contém um protocolo de coleta e transcrição dos dados, mas não possui anotação morfológica; *The Miami corpus* que possui protocolo com detalhamento da transcrição e, também, marcações morfológicas; e o *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América* (PRESEEA, 2008) que possui protocolo de transcrição e etiquetagem de aspectos contextuais e interacionais, mas não possui marcação morfológica.

Kendall e French (2006) também reportam a constituição do banco de dados *North Carolina Sociolinguistic Archive and Analysis Project*, apresentando com ênfase como foram feitas as transcrições e anotações prosódicas no *corpus*. Contudo, embora disponíveis online, o acesso aos dados é restrito.

No contexto internacional, os bancos de dados sociolinguísticos acima citados evidenciam exemplos de práticas que se alinham com o princípio da acessibilidade e reutilização, ao disponibilizarem os dados na internet seja com níveis de acesso ou totalmente abertos e ao reportarem os processos de transcrição e etiquetagem em diferentes níveis como o morfológico, fonológico e discursivo. Essas práticas promovem também maior transparência na pesquisa o que, conforme visto na seção anterior, contribui para a replicabilidade e reprodutibilidade dos estudos, favorecendo o fortalecimento da ciência (LYON, 2016).

[17] https://www.ldc.upenn.edu/
[18] https://archive.mpi.nl/tla/
[19] https://talkbank.org/
[20] https://www.elararchive.org/

Práticas de disponibilização de protocolos de constituição de amostras, como as citadas nos parágrafos anteriores acerca dos bancos de dados sociolinguísticos no exterior, são desejáveis para que sejam garantidas a confiabilidade e intersubjetividade das análises, ou seja, que ao seguir a mesma metodologia, a reprodução dos estudos deve encontrar os mesmos resultados em relação ao mesmo fenômeno linguístico (BAILEY; TILLERY, 2004). O contexto nacional, por outro lado, ainda necessita de estudos e de bancos de dados que sigam as práticas abertas, de forma a tentar atingir um padrão transparente na metodologia das pesquisas em Sociolinguística.

A exemplo dos bancos de dados de fala bem como da divulgação dos protocolos dos respectivos bancos no contexto internacional, defendemos uma publicação mais detalhada dos procedimentos metodológicos de coleta e sistematização dos dados de amostras sociolinguísticas. Apresentamos a metodologia de coleta advinda da pesquisa em Sociolinguística Variacionista em termos de coleta dando relevo à importância de se deixar os critérios de amostragem explícitos para a confiabilidade dos resultados possa ser atestada. Além disso, dada a natureza da amostragem a ser feita, observamos que, na sociolinguística, a quantidade de dados gerados para se analisar um fenômeno deve ser relevante para subsidiar as análises, explicações e resultados mais acurados (FREITAG, 2018a). Por fim, discutimos práticas de acessibilidade que possibilitam a reutilização de dados no contexto internacional, que ainda são incipientes no contexto brasileiro.

Neste capítulo, abordamos questões relacionadas à Ciência Aberta e principalmente no que tange à acessibilidade aos dados para posterior reuso. Apresentamos como os bancos de dados têm sido sistematizados no Brasil e no exterior, apontando para maior necessidade de aderência dos bancos nacionais aos preceitos de Ciência Aberta. Por fim, relatamos os critérios metodológicos de amostragem. No próximo capítulo, discutimos questões concernentes à etiquetagem de dados linguísticos, sua importância no contexto da Ciência Aberta, e apresentamos nossa proposta de sistematização da amostra Deslocamentos 2019 (FREITAG, 2018b) que compõe o banco de dados Falares Sergipanos.

2 BANCOS DE DADOS SOCIOLINGUÍSTICOS ETIQUETADOS: POSSIBILIDADES

Neste capítulo, temos por objetivo estabelecer uma proposta de disponibilização de dados de fala advindos de uma metodologia de pesquisa inserida no escopo da Sociolinguística Variacionista. Para consecução desse objetivo maior, dividimos nosso capítulo em três seções. Na primeira, apresentamos as funções dos bancos de dados linguísticos, evidenciando suas contribuições não somente para a descrição linguística, mas também para a esfera educacional, mostrando o interesse crescente pela disponibilização desses dados para a comunidade científica. Na segunda, apresentamos um breve panorama do banco de dados Falares Sergipanos para ilustrar a relevância dos bancos de dados linguísticos para descrição linguística e formação de recursos humanos.

Na terceira seção, abordamos questões relativas às possibilidades de disponibilização de dados de fala anotados, evidenciando as vantagens de dados anotados tanto para a Sociolinguística Variacionista quanto para o PLN. Embora o foco do nosso trabalho seja a anotação automática no nível gramatical, descrevemos, na seção final do capítulo, diversos níveis de anotação linguística.

2.1 Aplicações dos bancos de dados linguísticos

Na Sociolinguística Variacionista, a compilação de bancos de dados tem sido fonte de referência para a descrição das diferentes variedades do português brasileiro. Além de fonte para a descrição da língua, a compilação de dados sociolinguísticos pode oferecer subsídios para outras funções sociais. No entanto, como Labov (2020) aponta em sua palestra, "Justice as a Linguistic Matter", proferida no evento Abralin Ao Vivo, nem sempre ficam evidentes as contribuições da Sociolinguística Variacionista como ferramenta na promoção de justiça social, por isso, nessa palestra, ele deu relevo às contribuições dos seus estudos para a sociedade, o que também nos propomos a fazer nesta seção. Conforme Labov (2020), a realização do trabalho no Harlem observando a estrutura do inglês vernacular afro-americano (*African American Vernacular English* - *AAVE*) culminou na criação de um material didático, o "*The Reading Road*". Esse material, baseado na realidade dos estudantes de comunidades afro-americanas de periferia, tem como objetivo disponibilizar aos alunos do ensino fundamental conhecimentos sobre leitura e práticas de letramento. Além disso, conforme o pesquisador, com a coleta e análise dos dados do "*Atlas of North-American English*", foi possível desenvolver

softwares de análise acústica identificando diversos fenômenos fonético-fonológicos variáveis que caracterizavam as diferentes variedades do inglês norte-americano. O conhecimento adquirido nesse empreendimento foi decisivo para inocentar um homem injustamente acusado de ameaçar uma companhia aérea com bombas.[21]

A exemplo das contribuições da Sociolinguística Variacionista citadas por Labov (2020), no Brasil, essa área da linguística tem subsidiado a reformulação de políticas educacionais. Conforme Freitag (2016), a variação e o reconhecimento da diversidade linguística são questões abordadas em documentos oficiais como nos Parâmetros Curriculares Nacionais (BRASIL, 1998), no Programa Nacional do Livro Didático (BRASIL, 2021), na Base Nacional Comum Curricular (BRASIL, 2017) e no Programa de Mestrado Profissional em Letras[22] (notadamente nas ementas das disciplinas "Fonologia, variação e ensino" e "Gramática, variação e ensino").

Além disso, a compilação de dados de fala em bancos tem oferecido recursos linguísticos que podem ser operados por computadores e também ser utilizados como fonte para construção de materiais didáticos, gramáticas, dicionários, desenvolvimento de teorias linguísticas e desenvolvimento de recursos humanos (GONÇALVES, 2019; FREITAG, 2021). Como exemplos de desenvolvimento de recursos educacionais a partir de bancos e dados de fala, mencionamos: a "Gramática do Português Culto Falado no Brasil", uma coleção em 7 volumes produzida com dados de fala do projeto Norma Linguística Urbana Culta (NURC) com análise linguística em diferentes níveis e sob diferentes perspectivas teóricas (CASTILHO, 2021); o Banco de Dados RECI, que forneceu dados para a construção de uma proposta de ensino dos modos indicativo e subjuntivo da Língua Portuguesa levando em consideração contextos mais formais e informais de comunicação (SILVA; COELHO, 2020).

A relevância dos bancos de dados linguísticos não compreende somente as áreas de descrição e análise linguística ou educacional. Para Burke e Chelliah (2021), bancos de dados linguísticos funcionam como uma fonte de dados de comunidades de fala, bem como de suas histórias e costumes, sendo, portanto, um recurso de pesquisa para diferentes áreas do

[21] Em 1984, Paul Prinzivalli, havia sido detido sob acusação de ameaças de bomba à Pan American, uma companhia aérea americana. No entanto, ao se comparar a voz que fazia as ameaças e a de Prinzivalli, foi percebido que a pessoa dos áudios possuía traços fonéticos característicos do sudeste da Nova Inglaterra, enquanto Prinzivalli era um falante com traços fonéticos de Nova Iorque. Labov foi até o tribunal para apontar essas diferenças e devido a essa compilação do "*Atlas of North American English"* e o desenvolvimento de técnicas de análise acústica, o juiz se convenceu de que o réu era inocente (LABOV, 2020).

[22] Esse programa é realizado em âmbito nacional em rede, ou seja, em qualquer uma de suas unidades o regime didático e a matriz curricular é a mesma. Mais informações sobre esse programa estão disponíveis em: https://profletras.ufrn.br/. Acesso em: 20 nov 2021

conhecimento, como antropologia, agricultura e história da arte. Ademais, os bancos de dados linguísticos oferecem subsídios para o desenvolvimento de ferramentas para o PLN. Especificamente para o processamento da língua portuguesa, citamos algumas ferramentas de análise de *corpora* que foram feitas com base em bancos de dados disponíveis: o parser *PALAVRAS* (BICK, 2000), um analisador sintático automático que se beneficiou da disponibilidade de *corpora* do NURC[23] (CASTILHO, 2021), Tycho Brahe[24] (GALVES, ANDRADE, FARIA, 2017), NILC[25] (KUHN; ABARCA; NUNES, 2000) entre outros, para treinamento; o *AELIUS*, outro analisador sintático automático, que também utilizou o *corpus* Tycho Brahe (GALVES; ANDRADE; FARIA, 2017) e o Mac-Morpho[26] (ALUÍSIO *et al.*, 2003) para treinamento, e, no caso do Tycho Brahe, utilizou também o conjunto de etiquetas; o *spaCy* 3.5 que usa o modelo Universal Dependencies do *corpus* Bosque[27] (RADEMAKER *et al.*, 2017); e o etiquetador *TreeTagger*, que também criou seu modelo baseado no Bosque e no CETEMPúblico[28] (SANTOS; ROCHA, 2001).

É possível observar, nesta seção, a importância da compilação de dados linguísticos em formato de bancos de dados. Podemos verificar também que a quantidade de *corpora* disponíveis que contribuíram para o desenvolvimento de ferramentas computacionais na área de PLN da Língua Portuguesa é, em sua grande maioria, oriunda de pesquisas na área da Linguística de *Corpus*, primordialmente *corpora* de textos escritos. Existe, então, uma demanda por dados de fala, com os quais a Sociolinguística Variacionista pode contribuir. Além disso, os bancos citados como fonte para o desenvolvimento dessas ferramentas encontram-se disponíveis online e estão sistematizados em conformidade com as intenções de pesquisa que culminaram em sua produção. Essa sistematização e disponibilização ainda é incipiente no Brasil na área da Sociolinguística Variacionista, com poucos bancos de dados online com amostras disponíveis. Na próxima seção, apresentamos um panorama da constituição do banco de dados Falares Sergipanos, bem como suas contribuições.

[23] Por se tratar de um projeto interinstitucional, nem todos os dados do NURC encontram-se digitalizados. Neste site: https://fale.ufal.br/projeto/nurcdigital/, encontram-se os dados referentes à coleta realizada em Recife. Acesso em: 10 jan. 2023.
[24] Disponível em: https://www.tycho.iel.unicamp.br/corpus/. Acesso em 10 jan. 2023.
[25] Disponível em: http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm. Acesso em 10 jan. 2023.
[26] Disponível em: http://nilc.icmc.usp.br/macmorpho/. Acesso em: 10 jan. 2023.
[27] Disponível em: https://github.com/UniversalDependencies/UD_Portuguese-Bosque. Acesso em 10 jan. 2023.
[28] Disponível em: https://www.linguateca.pt/CETEMPublico/. Acesso em 10 jan. 2023.

2.2 Banco de dados Falares Sergipanos: produção e formação

De acordo com Freitag (2013), a constituição do banco de dados Falares Sergipanos se deu pelos seguintes objetivos: pela variedade sergipana do português brasileiro ser pouco explorada e oferecer amostras linguísticas empiricamente coletadas como fonte de recursos para descrição dessa variedade e, por consequência, do português brasileiro, bem como subsidiar recursos para aplicações educacionais. Ainda segundo a pesquisadora, o banco tem seguido duas linhas de coleta: estratificação homogeneizada (indivíduos divididos em células/estratos sociais) e de comunidades de práticas (indivíduos que compartilham práticas em comum).

Atualmente, o projeto conta com 18 amostras coletadas por meio de diferentes metodologias em conformidade com os diferentes fenômenos a serem investigados. A amostra Grupo de Pesquisa em Educação Física, por exemplo, seguindo a metodologia de comunidades de práticas, foi coletada por meio da gravação de entrevistas sociolinguísticas e de interação entre os participantes, a amostra Deslocamentos 2020 possui entrevistas sociolinguísticas com gravações em áudio e vídeo para capturar elementos paralinguísticos (expressões faciais, intensidade de voz) no momento da fala, com estudantes dos mais variados cursos de graduação da Universidade Federal de Sergipe, a amostra Como Fala o Universitário foi coletada por meio da gravação de leitura de textos feita por universitários.

Até o último levantamento feito em julho de 2021, o banco de dados Falares Sergipanos contava com um acervo de dados composto por: 133 arquivos de vídeo em formato mp4; 459 arquivos de áudio (135 em formato mp3 e 324 em formato wav); 297 arquivos no formato eaf (ELAN); 140 arquivos formato trs (Transcriber); 380 arquivos de texto (303 em formato txt e 77 em formato docX); 36 arquivos em formato TextGrid (Praat); e 265 arquivos de imagem em formato jpg. Para gerar essa quantidade de dados, em termos de desenvolvimento de recursos humanos, foram formados mais de 20 alunos em nível de iniciação científica, mais de 15 em nível de mestrado. Atualmente, há 6 estudantes em processo de doutoramento e 1 atualmente no mestrado.

Adicionalmente, foram pesquisados diversos fenômenos linguísticos variáveis visando à descrição da variedade Sergipana do Português Brasileiro, desde o nível fonológico, como a palatalização de /t/ e /d/ (CORRÊA, 2019; SILVA, 2021) ao discursivo, como as funções de "tipo" (SANTANA, 2019) e variação nos usos de "(Eu) acho que" (CARDOSO, 2021).

Podemos observar que a compilação de amostras sociolinguísticas deve servir a diferentes frentes de pesquisa. Conforme Freitag (2017c), a documentação feita no escopo da

Sociolinguística pode subsidiar aplicações na descrição de usos linguísticos, na preservação do patrimônio linguístico, e no ensino de línguas, entre outras. O banco de dados Falares Sergipanos, mesmo ainda recente no cenário da Sociolinguística brasileira, tem atendido à descrição dos usos linguísticos e à preservação do patrimônio linguístico, mapeando a variedade sergipana do português, como pudemos ver nesta seção. Contudo, os dados precisam de sistematização no que tange ao arquivamento, padronização de transcrições mais antigas e anotação linguística. É isso que estamos fazendo com a criação deste protocolo, a partir da testagem e avaliação de ferramentas gratuitas estabelecer parâmetros para o arquivamento dos dados, etiquetagem automática e divulgação das amostras.

2.3 Etiquetagem e variação: caminhos possíveis

Nesta seção, mostramos a possiblidade de se etiquetar gramaticalmente, morfologicamente e sintaticamente uma amostra de dados colhida por meio de entrevistas sociolinguísticas e sistematizá-la para sua futura disponibilização. Dessa forma, esta seção se divide em duas sub-seções. Na primeira, discutimos a relação entre a etiquetagem e a Sociolinguística Variacionista e Ciência Aberta, em seguida discutimos a interseção entre etiquetagem PLN e Sociolinguística. Finalizamos com os tipos de etiquetadores e de anotação linguística que podem ser feitas em *corpora*.

2.3.1 Aplicações da etiquetagem para a Sociolinguística Variacionista na Ciência Aberta

Devido à natureza da pesquisa sociolinguística de vertente variacionista, dados são gerados para se buscar captar o vernáculo dos informantes, subsidiando informações sobre os padrões de uso linguístico de uma determinada co,\munidade, seja ela de fala, conjunto de indivíduos que compartilham o mesmo significado social dos usos linguísticos, ou de práticas, conjunto de indivíduos que compartilham valores e conhecimentos em comum (FREITAG; MARTINS; TAVARES, 2012). Além de contribuir para a automatização da análise do volume de dados gerados, a etiquetagem de dados linguísticos pode contribuir em diferentes frentes para a pesquisa em Sociolinguística Variacionista. Dentre as vantagens de se anotar *corpora*, na Linguística de Corpus, Hovy e Lavid (2010, p. 29) apontam a "investigação teórica, criação

de dicionários ou de bases de dados lexicais, compilação de *corpus*"[29], que também são vantagens da criação de bancos de dados linguísticos (GONÇALVES, 2019). Para Hovy e Lavid (2010), as vantagens podem ser resumidas em três aspectos: reutilização, estabilidade e reprodutibilidade, preceitos que se coadunam com o paradigma da Ciência Aberta.

Ao lidar com grandes volumes de dados, muitas vezes o pesquisador na área de Sociolinguística Variacionista, sem saber da existência de ferramentas que possam contribuir para uma análise mais automatizada dos dados dispendem de um tempo substancialmente grande para levantar os fenômenos sob investigação (OTHERO; AYRES, 2014). A etiquetagem de dados linguísticos, por outro lado, é prática comum na Linguística de *Corpus*, vista como uma forma de enriquecer os dados com informações linguísticas que podem subsidiar novas pesquisas e desenvolvimentos (LEECH, 2013).

Hovy e Lavid (2010) definem a anotação de *corpus*, ou etiquetagem, como a adição de informações linguísticas ou de outra ordem, inseridas por humanos ou máquinas, de forma a atender propósitos teóricos ou práticos. Por outro lado, para Leech (2013, p. 2), a anotação em *corpus*, ou etiquetagem de *corpus*, pode ser descrita como "a adição de informações linguísticas e interpretativas a um *corpus* eletrônico de dados de fala ou escrita"[30]. Em comum, os autores compartilham o fato de que há uma adição ao texto de uma interpretação linguística, ou seja, a etiquetagem traz informações sobre a linguagem empregada no texto, não sobre o que ele discorre.

Existem diferentes tipos de etiquetagem de *corpus* e eles dependem dos diferentes propósitos de pesquisa e níveis de análise linguística, sendo as mais comuns delas a morfológica e a de classe de palavras, também chamada de classe gramatical (*part of speech tagging* ou POS) (LEECH, 2013). Como exemplo, o *TreeTagger* (SCHIMID, 1995) faz as seguintes marcações logo após cada palavra, a frase "seu nome" fica marcada da seguinte forma: *DET.Masc.Sing* nome *NOUN.Masc.Sing*", em que *DET* marca a classe determinante, *Masc.* marca o gênero masculino, *Sing.* o número do substantivo, no caso, singular, e *Noun.* marca que a palavra é um substantivo. Leech (2013) considera que a etiquetagem morfológica e de classe de palavra são os princípios base para fornecer informações linguísticas para começar o desenvolvimento de ferramentas que operam em níveis mais altos, como o sintático e o semântico.

[29] Tradução nossa para o original: "theoretical investigations, the creation of dictionaries or lexical databases, further *corpus* compilation".

[30] Tradução nossa para o original: "adding interpretive, linguistic information to an electronic *corpus* of spoken and/or written data".

Conforme Leech (2004), uma das grandes vantagens em se ter um *corpus* anotado reside no fato de ele permitir o processamento e a análise automáticos. O autor cita como exemplo a criação de listas de frequência em termos de léxico e de categorias gramaticais e também a análise sintática automática. A adição de informação em um *corpus*, ou seja, etiquetagem, é um passo anterior à extração de informação dele. No caso de uma transcrição ortográfica, por exemplo, é importante se ter a informação da classe de palavra para que se possa fazer a correta extração por categorias, principalmente em relação a palavras homógrafas (LEECH, 2004). Além disso, caso o pesquisador queira extrair todas as ocorrências de verbos no plural para verificar o fenômeno de variação da concordância em terceira pessoa do singular, basta buscar pela etiqueta referente a verbos no plural e no singular para quantificar as ocorrências, no caso do *TreeTagger*, as etiquetas seriam "VERB.Fin.Plur" e "VERB.Fin.Sing", respectivamente.

Para diferentes autores, como Sardinha (2000), Leech (2004, 2013), e Hovy e Lavid (2010), a etiquetagem é um processo em que se gasta muito tempo e recursos financeiros. Assim, como afirmam esses autores, outra vantagem de se já possuir um *corpus* anotado é a sua reutilização para busca de outros fenômenos e diversos tipos de análise.

Como vimos na seção sobre o paradigma da Ciência Aberta, a reutilização dos dados é um de seus princípios, de forma que os dados devam estar padronizados atendendo aos interesses da comunidade. Ressaltamos aqui o fato de que Leech (2013) afirma que não existem anotações que possam ser consideradas como padrão ouro. Portanto, para aderir aos preceitos de Ciência Aberta, os esquemas de anotação devem ser detalhados para que a comunidade acadêmica faça seu escrutínio e crítica (HOVY; LAVID, 2010). Em outras palavras, a anotação é disponibilizada para que outros possam reutilizá-la, e, quando julgarem necessário, construir seus modelos a partir dela, não tendo que começar o processo do zero. Embora razões existam para que haja esquemas diferentes de anotação, como o propósito inicial da anotação e também o tipo de dados, a busca pela padronização é importante para que o intercâmbio entre pesquisas, em termos de recursos, *softwares*, seja facilitado (LEECH, 2013).

Leech (2004) cita como exemplo de reutilização o fato de a anotação dos *corpora LOB Corpus* e o *BNC Sampler Corpus* ter servido como base para milhares de outros estudos. No contexto brasileiro, *corpora* de fala anotados morfossintaticamente, como o *Corpus* Brasileiro[31] e o C-ORAL[32], disponíveis para consulta, também foram reutilizados em diferentes pesquisas, como Costa Jr. (2018), que estudou discurso direto, e Silva (2017), que faz uma análise

[31] Disponível em: https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS
[32] Disponível em: https://www.c-oral-brasil.org/

construcional do conector "a hora que", ambos realizados com o C-ORAL[33]. O acesso a esses *corpora* se faz mediante requisição para usos acadêmicos e não-comerciais, pois, na interface disponível online para consulta, os resultados são limitados a uma pequena amostragem das ocorrências.

No que tange à estabilidade, Hovy e Lavid (2010) argumentam que um *corpus* anotado serve como referência e base estáveis para análise linguística, de forma que estudos posteriores utilizem essa mesma base para comparação e crítica. Os autores ainda ressaltam que se torna possível verificar hipóteses acerca do funcionamento da língua de forma empírica, pois, ao se debruçar sobre os dados, os pesquisadores podem se deparar com novas questões que, talvez, um único ponto de vista teórico possa não ser suficiente para explicar o fenômeno, sendo necessária a ampliação ou redefinição teórica.

Por fim, de acordo com Berez-Kroeker *et al.* (2018), o objetivo da reprodutibilidade é o de facilitar o acesso aos dados gerados em uma pesquisa por outros estudiosos de forma a garantir sua confiabilidade. Conforme os mesmos autores, replicar estudos nas ciências humanas é uma tarefa pouco tangível devido às inúmeras razões contextuais que são difíceis de serem controladas. No caso da pesquisa em Sociolinguística Variacionista, como visto na seção anterior, o fato de as amostras não serem aleatórias dificulta ainda mais que as condições metodológicas sejam replicadas.

Berez-Kroeker *et al.* (2018) defendem que na área de linguística a reprodutibilidade seja uma métrica mais tangível para garantir a confiabilidade dos estudos. Assim, para atingir esse objetivo, os autores defendem a transparência em duas frentes: nos métodos de coleta e análise de dados e na fonte dos dados (como eles podem ser acessados). A etiquetagem de dados linguísticos oferece esse caminho, pois, como já visto anteriormente, a etiquetagem é uma adição de informação linguística aos textos, podendo ser inspecionada por outros estudiosos (HOVY; LAVID, 2010), e, também, como aponta Leech (2013, p. 4), "*corpora* só são úteis se pudermos extrair conhecimento ou informações deles"[34]. Em outras palavras, ao adicionar informações linguísticas a um texto transcrito ortograficamente, a análise e a teoria já estão dadas para que possam ser reproduzidas, confirmando sua confiabilidade ou apontando caminhos para um melhor modelo de anotação.

[33] Uma lista com diferentes publicações que usaram o C-ORAL está sistematizada e disponível em: https://www.c-oral-brasil.org/. Não há uma lista como esta disponível para o *Corpus* Brasileiro.

[34] Tradução nossa para o original: "Corpora are only useful if we can extract knowledge or information from them".

Apesar de termos discorrido sobre as vantagens em se disponibilizar amostras de fala anotadas, esse ponto não é consenso entre pesquisadores na área da Linguística de *Corpus*. Sinclair (2004a; 2004b) defendem a disponibilização de *corpora* anotados separados e dos textos originais sem a adição dessas etiquetas. Já Beck *et al.* (2020) apontam para as falhas inerentes a qualquer modelo de anotação, mas reconhecem o valor que ela possui para os estudos linguísticos e para o PLN.

Para Sinclair (2004a; 2004b), a anotação codifica informações que são extras ao texto, isto é, que não são possíveis de serem recuperadas pelo texto original. Existe para o autor uma perda da integridade do original quando são adicionadas etiquetas ao texto original sem que haja uma cópia dele antes dessa adição, principalmente, quando se tenta removê-las.

O autor argumenta também que existem diferentes tipos de etiquetas para diferentes propósitos de pesquisa, ou seja, não há universalidade no esquema de anotação. De fato, como já apontado por Leech (2013) não existe uma padronização das etiquetas até mesmo para o mesmo nível de análise, embora ela seja desejável. No entanto, é importante ressaltar que assim como existem diferentes esquemas de anotação, a constituição de *corpora* seja na área de Linguística de *Corpus* ou na Sociolinguística Variacionista, também se faz levando em consideração os tipos de fenômenos a serem pesquisados (FREITAG, 2018a; GRIES; BEREZ, 2017).

Outra limitação imposta pela etiquetagem, como aponta Sinclair (2004a), é o fato de a medida de acurácia em etiquetadores não ser baseada em evidência linguística. Leech (2004) confirma esse fato, mostrando que a medida da acurácia tem referência naquilo que o esquema de anotação propicia e quais modelos estatísticos são utilizados para se determinar essa medida. Sinclair (2004a) ainda aponta que a sistematicidade do erro dos computadores leva a duas questões: o computador estaria fornecendo novas informações ao invés de cometer erros ou as etiquetas erroneamente atribuídas estão naquelas ocorrências que o pesquisador gostaria de confiar.

Além disso, para Sinclair (2004b), os computadores, ao lerem um texto com anotação, ignoram a informação do texto, trabalhando somente com as etiquetas, de forma que pode haver questões negligenciadas, como os casos de ambiguidade, quando as etiquetas não são sensíveis aos diferentes sentidos de uma palavra. Essa limitação também é apontada por Beck *et al.* (2020) no caso de ambiguidades em que o etiquetador ora atribui uma etiqueta com maior probabilidade de acerto ora ignora completamente o item, gerando incerteza em relação à

etiquetagem uma vez que esse tipo de situação não é frequentemente explicitado ao usuário final.

Apesar de não ser contra a disponibilização de *corpora* anotados, Beck *et al*. (2020) apontam outras três limitações do processo de etiquetagem: “variação”, “erro” e “viés”. “Variação”, em uma etiquetagem, não se refere à sua classificação ou identificação, mas ao fato de ligar uma ou mais ocorrências a uma mesma variável. Pode haver a existência de um estágio intermediário de univerbação, em que múltiplas palavras que formavam uma expressão foram reanalisadas como uma só, como no caso de "embora", que surgiu da expressão "em boa hora", e as duas instâncias podem ser encontradas em amostras de fala históricas e marcadas de maneira diferente.

Beck *et. al* (2020) argumentam também que “erros” de anotação são comuns mesmo havendo inspeção por outros dois anotadores. Conforme Leech (2004), etiquetadores automáticos têm alcançado uma média de 98% de acurácia para a língua inglesa, porém, essa medida não é realista, pois, ao ser submetida à inspeção por humanos, ambiguidades presentes na etiquetagem podem ser revistas e desfeitas, diminuindo esse índice. Em relação ao “viés”, assim como já mencionado, existem várias ferramentas que fazem a etiquetagem linguística, e cada uma delas vai atribuir certas etiquetas e, consequentemente, diferentes serão os resultados, ou seja, haverá diferentes vieses, de acordo com o esquema de anotação, o que é inevitável (BECK *et al*., 2020; BEREZ; GRIES, 2017; LEECH, 2004).

Todas as razões expostas nesta seção acerca das vantagens de se disponibilizar *corpora* já anotados se coadunam com os preceitos de Ciência Aberta apontados na seção “1.1 Crise na ciência: acessibilidade como prática”, na qual argumentamos para uma necessidade de se adotar preceitos de acessibilidade e reutilização para aumentar a confiança nos resultados de pesquisas científicas. Em outras palavras, acreditamos que a disponibilização de amostras sociolinguísticas anotadas oferece subsídios para a reprodução de estudos, reutilização dos dados para pesquisas com diferentes fenômenos e também como referência para a constituição de outras amostras. Conforme Gries e Berez (2017) apontam, a anotação é uma prática que está consolidada e tem rendido bons frutos tanto na área da Linguística de *Corpus* quanto de PLN, e esperamos também explorar suas potencialidades na área de Sociolinguística Variacionista. No entanto, consideramos importante abordar as limitações do método para que o próprio leitor possa avaliar a adequação da etiquetagem para seus propósitos de pesquisa.

2.3.2 Contribuições da etiquetagem para o Processamento de Linguagem Natural e a interação com a Sociolinguística Variacionista

O PLN é compreendido como "o uso de técnicas computacionais para produzir, entender e aprender conteúdos da linguagem humana" (HIRSCHBERG; MANNING, 2015, p. 261). O que distingue o PLN de outros tipos de processamento, para Jurafsky e Martin (2009), é a demanda pelo conhecimento da língua. Os autores citam diferentes usos do conhecimento linguístico para o PLN: a) para computar o número de palavras, o processador precisa saber o que significa ser uma palavra; b) para reconhecer a fala e fazer síntese de fala, o processador precisa de conhecimentos fonético-fonológicos; c) para reconhecer e produzir palavras no singular e no plural é necessário conhecimento morfológico. Resumindo, a complexidade das tarefas na área de PLN demandam conhecimento linguístico e, para todas essas tarefas, *corpora* anotados são primordiais.

Como aponta Leech (2013), a utilidade dos *corpora* advém da possibilidade de se obter informações ou conhecimento por meio deles e é nesse ponto que o PLN se beneficia de dados linguísticos que já possuam algum tipo de etiquetagem, principalmente nos níveis morfológico e sintático. Palmer e Xue (2010) afirmam que para avanços substanciais na área de PLN, com foco em aprendizagem de máquinas, são necessários *corpora* com etiquetas linguísticas para que essas máquinas sejam treinadas, de forma que modelos supervisionados continuam e continuarão sendo as melhores fontes para que tal objetivo seja alcançado. Modelos de aprendizagem de máquinas supervisionados são aqueles em que o treinamento e o teste de um sistema são feitos em um texto que já possui uma análise presente (categorias morfossintáticas, entidades semânticas definidas entre outras), logo, é primordial a disponibilidade de *corpora* anotados para esse tipo de desenvolvimento tecnológico.

De acordo com Candito e Liberman (2019), a anotação de *corpora* aliada à pesquisa estatística em mensagens ampliou o escopo das tarefas realizadas por computador, que antes faziam principalmente a codificação e transmissão mensagens, e hoje realizam o reconhecimento de fala, a tradução por máquinas, a anotação morfológica, o *parsing*, e outras formas de trabalho com a língua. Hirschberg e Manning (2015) também argumentam que a área da linguística computacional tem se desenvolvido muito nos últimos anos, não só em termos de pesquisa científica como também em desenvolvimentos de produtos do nosso dia-a-dia. Ademais, o PLN pode ser considerada uma área estratégica, pois, a partir de seus avanços, por meio de raspagem/coleta de dados (*webscrapping*) e usando modelos estatísticos em termos de

linguagem, como exemplo, é possível recolher informações demográficas, verificar opiniões e crenças sobre diferentes assuntos, reconhecer notícias falsas e identificar nichos sociais das pessoas que interagem juntas online (cf. HIRSCHBERG; MANNING, 2015).

Esse desenvolvimento tecnológico não se deu ao acaso, Hirschberg e Manning (2015, p.261) elencam quatro fatores que foram preponderantes:

> (i) um vasto crescimento no poder da computação, (ii) a disponibilidade de grandes quantidades de dados linguísticos, (iii) o desenvolvimento de métodos de aprendizagem de máquinas bem-sucedido; (iv) uma compreensão maior da estrutura da língua e seu emprego em contextos sociais. [35]

É no escopo desses fatores que nossa tese se insere, visto que é a partir da disponibilização de *corpora* de dados espontâneos de fala que estaremos contribuindo diretamente para estudos descritivos e, por consequência, uma maior compreensão da língua e seus usos, o que por fim favorece o poder da linguística computacional no desenvolvimento e aprimoramento de ferramentas para análise linguística e também em outras aplicações, como texto-para-fala, síntese de fala, mineração de dados, entre outras.

Ide (2017) afirma que dados anotados são a base para avaliação de tecnologias com linguagem humana, além de oferecerem subsídios para o desenvolvimento de modelos estatísticos para essas tecnologias, de forma que o aumento na disponibilidade de dados linguísticos anotados leva a um aumento no desenvolvimento das ferramentas tanto de armazenamento quanto da criação de *corpora* etiquetados. A anotação, então, pode ser considerada como um ponto de intersecção entre cientistas da computação que trabalham com PLN e com linguistas pois permite a ambos:

> a) coletar dados para investigar um fenômeno de interesse e ao final modelá-los ou explicá-los; b) produzir teorias e modelos linguísticos suficientemente claros e sistematizados para serem aplicados a dados reais; e c) testar, validar e/ou melhorar teorias e modelos. Ao mesmo tempo, *corpora* consistentemente anotados são meios para os cientistas da computação desenvolverem, treinarem e testarem seus sistemas. (ALUÍSIO; PARDO; DURAN, 2014, p.307).[36]

[35] Tradução nossa para o original: "i) a vast increase in computing power, (ii) the availability of very large amounts of linguistic data, (iii) the development of highly successful machine learning (ML) methods, and (iv) a much richer understanding of the structure of human language and its deployment in social contexts.".

[36] Tradução nossa para o original: "In addition, a consensus has been reached that annotating a *corpus* is one of the main meeting points for linguists and computer scientists working with NLP and *corpus* linguistics as it allows linguists and related professionals (a) to harvest data in order to investigate phenomena of interest and eventually to model or explain them; (b) to make linguistic theories and models clear and systematized enough to be applied to actual data; and (c) to test, validate and/or improve theories and models. At the same time, reliably annotated *corpora* are a means for computer scientists to develop, train and test their systems.".

Como Aluísio, Pardo e Duran (2014) argumentam, existe um sistema de retroalimentação, em que linguistas oferecem dados empiricamente atestados e etiquetados, e os pesquisadores da PLN desenvolvem a tecnologia para a automatizar a análise desses dados.

Embora tenhamos no contexto do português brasileiro inúmeros trabalhos descritivos no âmbito da Sociolinguística Variacionista e disponibilidade de *corpora*, estes principalmente advindos da Linguística de *Corpus*, no âmbito do PLN, para as variedades do português são consideradas ainda como de poucos recursos (cf. AGUIAR DE LIMA; COSTA-ABREU, 2020), quais sejam etiquetadores, *parsers*, e *softwares* que fazem análise de sentimento, entre outros. Esse fato dificulta trabalhos na área de Sociolinguística sobre o português que possuem grandes volumes de dados, no sentido de que *softwares* que fazem anotação morfológica, sintática, ou semântica automaticamente são pagos ou possuem uma interface pouco amigável para pesquisadores que não conhecem linguagem de programação. Portanto, disponibilizar dados de fala coletados em situações espontâneas, como no caso dos dados gerados na Sociolinguística Variacionista, pode favorecer o maior desenvolvimento desses recursos, aumentando sua disponibilidade.

Pelas razões expostas acima, a etiquetagem é um recurso que tem grande valia para o desenvolvimento de tecnologias que dependem da linguagem humana. Por meio de *corpora* anotados, *chatbots* (como o Bahianinho das Casas Bahia e Aura da Vivo), sistemas de conversação (como a Alexa e a Siri) são criados, e por meio do PLN, são criados etiquetadores automáticos de classes gramaticais (*Part of speech taggers*), de classes sintáticas, semânticas, pragmáticas e discursivas, facilitando a criação de *corpus* etiquetados, formando um sistema de intersecção entre a Sociolinguística e o PLN[37].

2.3.3 Etiquetagem e níveis de análise linguística

Como visto no capítulo um, a anotação de dados linguísticos compreende a adição de informações descritivas e analíticas aos dados (GRIES; BEREZ, 2017; IDE, 2017; LEECH, 2013). Dados linguísticos anotados têm sido a base para o desenvolvimento de tecnologias com linguagem humana, além de oferecer subsídios para o desenvolvimento de modelos estatísticos para essas tecnologias, de forma que o aumento na disponibilidade de dados linguísticos

[37] Embora nosso foco seja na disponibilização de corpora para a construção de recursos de PLN, existem outras aplicações desta área para os estudos da linguagem. Por exemplo, algoritmos de aprendizagem de máquina podem validar a classificação realizada por humanos acerca de fenômenos linguísticos como feito por Freitag *et al.* (2021), na classificação de diminutivos e dos sentidos de "(eu) acho que" e por Rodrigues, Souza e Santos (2022) nas classificações de verbos locativos do espanhol.

anotados tem levado a um aumento nas ferramentas tanto de armazenamento quanto de anotação de *corpora* (IDE, 2017). Atualmente, etiquetadores automáticos conseguem fazer a marcação de diferentes níveis de análise linguística: a anotação de lema, a gramatical, a morfológica, a sintática, a semântica e a discursiva (KLÜBER; ZINSMEISTER, 2015; LEECH, 2004; LEECH, 2013).

Para Klüber e Zinsmeister (2015, p.45), lema pode ser definido como "a forma base de uma palavra, como ela é representada em um léxico"[38]. Para o português, por exemplo, que é uma língua que possui gênero, convencionalizou-se que a forma base para adjetivos e substantivos é a forma masculina, como em "amigo", já para verbos, a forma base é o infinitivo. As pesquisadoras afirmam que a anotação de lema é importante uma vez que ela permite a busca por todas as ocorrências de uma palavra. Em outras palavras, ao fazermos uma busca pelo lema "correr", teremos como possíveis resultados em um *corpus*: "corri", "correm", "correram" entre outras flexões desse verbo.

As autoras relatam que a lematização pode ser feita de maneira automática e que, como ela é feita baseada em regras, de maneira geral, pode ser considerada segura, porém, ressaltam que a performance varia em conformidade com a cobertura lexical do lematizador. No caso de uma palavra não existir no léxico, o lematizador pode deixar a palavra como ela aparece no texto ou "adivinhar" o lema por meio dos sufixos. No seguinte exemplo, temos uma representação da etiquetagem feita pelo *TreeTagger*, um etiquetador automático e gratuito, que faz tanto lematização quanto a marcação morfológica e gramatical, em um trecho de uma entrevista da Amostra Deslocamentos 2019.

Figura 2: Anotação morfológica e gramatical.

```
elas  PRON.Fem.Plur    ele
não   ADV   não
gostam      VERB.Fin.Plur    gostar
de    ADP   de
muita DET.Fem.Sing     muito
de    ADP   de
muita DET.Fem.Sing     muito
intimidade NOUN.Fem.Sing    intimidade
```

Fonte: Elaboração própria a partir do uso do *TreeTagger* na amostra Deslocamentos 2019.

Na figura 2, a sentença “elas não gostam de muita de muita intimidade” está dividida em *tokens* (separação por palavras) na primeira coluna, na segunda, encontram-se a marcação

[38] Tradução nossa para o original: “the basic form of a word, as it is represented in a lexicon.”

gramatical e morfológica, na terceira, os lemas. Como pode ser visualizado, o lema da palavra "elas" é "ele" e de "gostam", gostar, e de muita, muito.

Kübler e Zinsmeister (2015) descrevem a anotação morfológica, por sua vez, como aquela que oferece informações inflexionais como número, gênero, caso e definitude para substantivos; grau, gênero, número e caso para adjetivos; e número, pessoa, tempo, modo para verbos, ressaltando que essas informações variam em conformidade com o conjunto de etiquetas de cada etiquetador disponível para cada língua. As autoras afirmam que existem etiquetadores que oferecem informações derivacionais. A diferença entre o primeiro e o segundo tipo de anotação morfológica é que o segundo pode alterar a marcação de classe de palavra, e o primeiro não altera nem o significado nem a classe de palavra.

Na figura 2 acima, também há um exemplo de anotação morfológica, pois, como pode ser observado, a classe gramatical (NOUN, substantivo) é seguida pelas seguintes informações inflexionais "Fem", que indica gênero feminino e "Plur" que indica o número plural.

A anotação gramatical de classe de palavras (POS), considerada como a base para outros tipos de anotação (GRIES, 2017; LEECH, 2013), está intrinsecamente relacionada à morfológica, sendo que esta ocorre posteriormente àquela, conforme Kübler e Zinsmeister (2015). É possível encontrar etiquetadores que façam as duas análises integradas, como o *TreeTagger* e o *LancsBox 6.0*, porém, isso dependerá do aparato feito para a língua e do conjunto de etiquetas. Na figura 2 acima, a anotação POS mostra que "elas" é um pronome (PRON), "não", "um" advérbio (ADV), "gostam", um verbo finito (VERB Fin), "de", uma adposição[39], "muita", um determinante (DET), e "intimidade", um substantivo (NOUN).

Kübler e Zinsmeister (2015) descrevem como anotadores sintáticos aqueles que marcam as relações entre os constituintes de uma sentença, ou seja, as relações estruturais entre os elementos de uma frase. Segundo as autoras, a etiquetagem sintática tende a obedecer a dois tipos de formalismos: os baseados em estrutura dos constituintes e os baseados em estrutura de dependência.

Na estrutura de constituintes, "as palavras são agrupadas em sintagmas" (KÜBLER; ZINMEISTER, 2015, p. 58), como sintagmas adverbiais, adjetivais, nominais entre outros. Em cada um desses sintagmas existe uma palavra, nomeada de "cabeça", que caracteriza a categoria sintática do sintagma. De acordo com Kübler e Zinsmeister (2015), esses sintagmas são

[39] Adposição se refere a um conjunto de palavras que podem preceder ou suceder um complemento formado por um sintagma nominal de forma a evidenciar as relações gramaticais e semânticas dentro de uma oração de acordo com a descrição das etiquetas do Universal Dependencies. Disponível em: https://universaldependencies.org/u/pos/ADP.html. Acesso em 10 ago 2021. No caso específico acima, trata-se de uma preposição que relacionam gramaticalmente o verbo "gostar" e seu complemento "muita intimidade"

hierarquizados de maneira decrescente, isto é, dos sintagmas maiores até as orações. Para ilustrar, as autoras utilizam o seguinte exemplo:

Figura 3: Anotação Sintática de Constituintes.

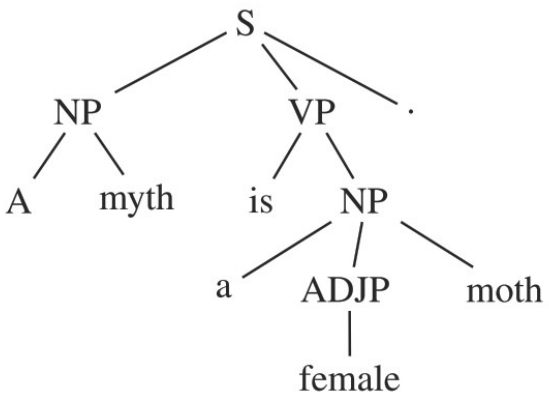
(1) $[_{S}\ [_{NP}$ A myth$]\ [_{VP}$ is $[_{NP}$ a $[_{ADJP}$ female$]$ moth$]]]$.

Fonte: Kübler e Zinsmeister (2015, p. 57).

Na sentença da figura 3 acima, os constituintes são separados por meio de colchetes e os tipos de sintagma são marcados em sobrescrito. Nessa sentença, que pode ser traduzida como "Um mito é uma mariposa fêmea", tanto "*A myth* (um mito)" quanto "*a female moth* (uma mariposa fêmea)" são considerados sintagmas nominais (NP). O sintagma nominal "*a female moth*" tem um sintagma adjetival (ADJP) que juntamente com o verbo "*is* (é)" forma um sintagma verbal (VP). Todos esses constituintes juntos formam a oração (S).

Essa mesma estrutura de colchetes pode ser visualizada em forma de árvore, como os autores ilustram:

Figura 4 - Anotação em forma de árvore.



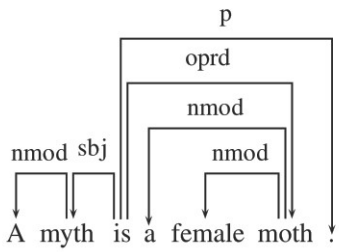
Fonte: Kübler e Zinsmeister (2015, p. 58).

Na representação presente na figura 4, conforme Kübler e Zinsmeister (2015), o nó "S" é considerado como o nó raiz da árvore. Os autores apontam que é ainda possível anotar por funções gramaticais como sujeito e predicado, em que na representação acima o primeiro sintagma nominal (NP) seria o sujeito. Embora importante para o português, que tem uma disposição livre dos constituintes na oração, a marcação das funções gramaticais dos

constituintes não é normalmente feita para o inglês, uma vez que este idioma possui disposição fixa de ordem dos constituintes.

A estrutura de dependência, por sua vez, é marcada pelas relações entre pares de palavras, sendo dependência a relação direta entre a cabeça e o seu dependente (KÜBLER; ZINMEISTER, 2015). Na figura 5 abaixo, a mesma sentença foi utilizada para representar uma relação de dependência.

Figura 5 - Anotação sintática de dependência.



Fonte: Kübler e Zinsmeister (2015, p. 58).

De acordo com Kübler e Zinsmeister (2015), na figura, "*a* (um)" é dependente de "*myth* (mito)" em uma dependência do tipo modificador nominal. A palavra "*myth* (mito)" é um dependente do tipo sujeito do verbo "*is* (é)", este por sua vez não possui relação de dependência por ser a cabeça de toda a sentença. P indica uma relação de dependência do tipo predicado, OPRD mostra que o sintagma "*a female moth* (uma mariposa fêmea)" está em uma relação de dependência com "*is* (é)" do tipo objeto predicado.

Kübler e Zinsmeister (2015) reportam que, embora pareça que os dois tipos de etiquetagem descritos possuem apenas uma teoria subjacente, na realidade para cada um dos tipos de representação há diferentes teorias que os embasam. Como exemplo, as pesquisadoras citam as gramáticas de constituintes em que pode haver diferentes representações no que tange às relações de longa distância, de forma que, em alguns bancos de dados, são assumidos elementos nulos (como no caso do *Penn TreeBank*), ou utilizadas etiquetas específicas.

Figura 6 - Representação dos constituintes com elemento nulo.

Referential null subjects are represented as (NP-SBJ *pro*).

```
( (IP-MAT (NP-SBJ *pro*)
          (VB-D Declamei)
          (PP (P contra)
              (NP (D-F a) (N vaidade)))
          (. ,))
  (TYCHO BRAHE; ID A_001_PSD,03.3))
```

Fonte: Manual do Tycho Brahe (2019).

No exemplo acima, extraído do manual de anotação do *corpus* Tycho Brahe, que segue os parâmetros de anotação sintática do *Penn Treebank*, observamos uma etiquetagem mais plana e, por isso, menos tradicional que a estrutura de árvore canônica, portanto, como seus criadores argumentam, a representação dos constituintes não se conforma com os requisitos nem estrutura de ramificação binária ou da teoria X-barra, priorizando a simplicidade da anotação do que conformidade com a teoria linguística. Ademais, como visto acima, há uma representação em que são assumidos elementos nulos, como ilustrado pela etiqueta NP-SBJ *pro*, que assinala a existência de um sintagma nominal no qual a presença do sujeito (SBJ) é nula, o que é marcado pelos asteriscos indicando a ausência de um pronome (pro).

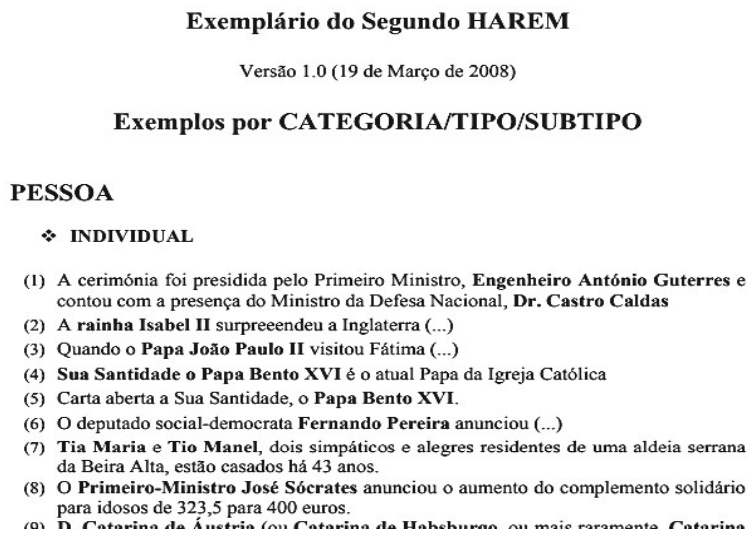
De acordo com Kübler e Zinsmeister (2015), a área da semântica, por sua vez, lida com o significado literal da linguagem, podendo ser dividida em dois tipos: semântica lexical, que lida com o significado das palavras, e a semântica composicional, que compreende os significados das frases e sentenças, no que tange a veracidade das informações dadas a partir das formas pelas quais as palavras se combinam, bem como pelos seus significados individuais.

As autoras reportam cinco tipos de anotação semântica: i) "named entity" ou "entidades nomeadas", é um tipo de etiquetagem que considera classes semânticas maiores, designando uma etiqueta a partir de um conjunto de classes pré-definido; ii) "*word sense*" ou "anotação de sentidos das palavras" fornece informações específicas de cada palavra, do tipo que distingue, por exemplo, o "banco de praça" do "banco agência bancária"; iii) "*semantic role*", "anotação de papel semântico", é desenvolvida principalmente pelo predicado, ou seja, quais papéis o verbo atribui para o sujeito e o objeto dentro de um quadro (*frame*); iv) "*temporal information*", "anotação temporal", oferece informações em relação à dêixis, a sequência em que os eventos acontecem e são hierarquizados; v) a "anotação de semântica formal", "*formal semantics*", vai

além do oferecimento de informações no nível frasal, oferecendo informações como pressuposição, tempo, e escopo.

A título de ilustração, abaixo seguem dois exemplos de anotação semântica, um de entidades nomeadas (Figura 7) e outro de sentidos de palavras.

Figura 7 - Anotação de entidades nomeadas.



Exemplário do Segundo HAREM

Versão 1.0 (19 de Março de 2008)

Exemplos por CATEGORIA/TIPO/SUBTIPO

PESSOA

❖ **INDIVIDUAL**

(1) A cerimónia foi presidida pelo Primeiro Ministro, **Engenheiro António Guterres** e contou com a presença do Ministro da Defesa Nacional, **Dr. Castro Caldas**
(2) A **rainha Isabel II** surpreeendeu a Inglaterra (...)
(3) Quando o **Papa João Paulo II** visitou Fátima (...)
(4) **Sua Santidade o Papa Bento XVI** é o atual Papa da Igreja Católica
(5) Carta aberta a Sua Santidade, o **Papa Bento XVI**.
(6) O deputado social-democrata **Fernando Pereira** anunciou (...)
(7) **Tia Maria** e **Tio Manel**, dois simpáticos e alegres residentes de uma aldeia serrana da Beira Alta, estão casados há 43 anos.
(8) O **Primeiro-Ministro José Sócrates** anunciou o aumento do complemento solidário para idosos de 323,5 para 400 euros.
(9) **D. Catarina de Áustria** (ou **Catarina de Habsburgo**, ou mais raramente, **Catarina**

Fonte: Exemplário do Segundo HAREM, versão 1.0.[40]

A figura 7 apresenta um modelo de anotação de entidade nomeada, com a categoria "Indivíduo" do tipo "Pessoa". As palavras em negrito são os exemplos que pertencem a essa categoria, como Dr. Castro Caldas e Papa João Paulo II.

Figura 8 - Anotação de sentido da palavra.

<ß>
Ninguém [ninguém] <*> **SPEC** M S @SUBJ> §EXP #1->2
gosta [gostar] <fmc> **V** PR 3S IND VFIN @FS-STA §PRED #2->0
de [de] **PRP** @<PIV #3->2
chuva [chuva] <wea-rain> **N** F S @P< §TH #4->3
. [.] PU @PU #5->0 </ß> </ß>

Fonte: Analisador automático PALAVRAS (*Visual Interactive Syntax Learning*)[41]

Na figura 8 acima, colocamos uma frase aleatória no analisador PALAVRAS. No resultado da análise, existe a anotação de lema em colchetes, a anotação morfológica e POS em

[40] Disponível em: https://www.linguateca.pt/HAREM/. Acesso em: 11 out. 2022. Esse *corpus* está presente na Linguateca, um centro de recursos para o processamento da língua portuguesa. Nela estão contidos *corpora*, dicionários, léxico e ferramentas computacionais.

[41] Disponível em: https://visl.sdu.dk/visl/pt/parsing/automatic/parse.php. Acesso em: 11 out 2021.

cor azul, a anotação sintática, seguida pelo símbolo @ (arroba), e a semântica entre < > (colchetes angulares). “Ninguém” não foi uma palavra etiquetada em termos semânticos pelo analisador <*>, o verbo “gostar” foi etiquetado como “*finite main clause heading verb* (verbo de cabeça de oração principal finito)”, “de” por ser integrante de uma classe fechada de palavras, não possui anotação semântica <PIV#3->, e “chuva” foi marcada como “*rain and other precipitation” <wea-rain>*, chuva e outras precipitações, que abarcam outras palavras como chuvisco, tromba d'água, granizo.

Por fim, a anotação discursiva compreende a integração de informações não apenas em uma sentença, mas entre diferentes sentenças para formar um todo que seja representado na mente do usuário. Por lidar com um nível que vai além do textual, os conceitos são mais difíceis de serem explicitados e muitos deles não são fixos (KÜBLER; ZINSMEISTER, 2015; LEECH, 2013). No português brasileiro, existe uma ferramenta que faz anotação discursiva baseada na Rhetorical Structure Theory (RST, Teoria da Estruturação Retórica): o DiZer. Trata-se um programa brasileiro para segmentação e anotação discursiva que resolve ambiguidades discursivas, por exemplo, “quando nenhuma relação retórica pode ser encontrada entre dois seguimentos, o DiZer assume um método padrão: ele adota uma relação de ELABORAÇÃO (que é a relação mais genérica) com o segmento que aprece primeiro no texto sendo seu núcleo” (PARDO; NUNES, 2004, p.5)[42]. Nas figuras abaixo, há um exemplo desse tipo de relação de ELABORAÇÃO.

Figura 9 - Texto segmentado para anotação discursiva.

[*Desde a sua abertura comercial, em 1993, a Internet tornou-se um meio de comunicação poderoso,*]$_1$ [*ao permitir a um usuário entrar em contato com quaisquer outros, espalhados pelo mundo todo.*]$_2$

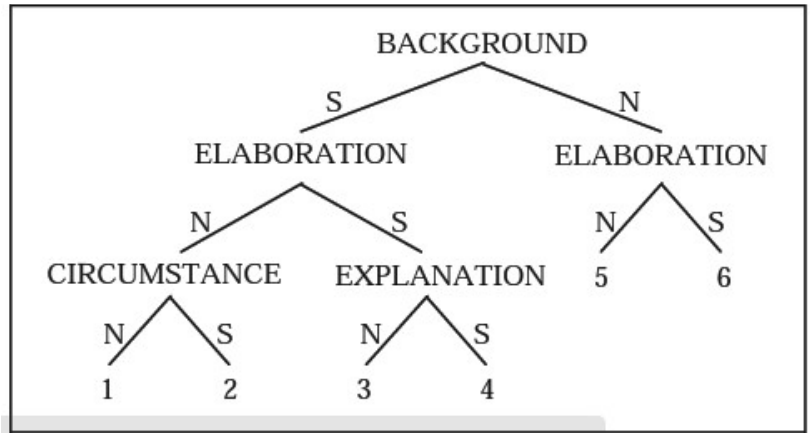
[*O comércio eletrônico é um dos novos nichos de exploração comercial da rede mundial de computadores,*]$_3$ [*pois ela torna possível realizar transações comerciais de forma global, com custo de manutenção inferior ao empregado em uma rede de comércio tradicional.*]$_4$

[*O objetivo deste trabalho é apresentar uma proposta para o projeto e implementação de um serviço de comércio eletrônico na plataforma JAMP.*]$_5$ [*Esta plataforma constitui-se em um middleware implementado em Java/RMI para desenvolvimento de aplicações multimídia distribuídas, e em particular, aplicações para World Wide Web (WWW), através de frameworks de serviços para suporte ao desenvolvimento destas aplicações.*]$_6$

Fonte: PARDO; NUNES (2004, p.6)

[42] Tradução nossa para: “when no rhetorical relation can be found between two segments, DiZer assumes a default heuristic: it adopts an ELABORATION relation (which is the most generic relation), with the segment that appears first in the text being its nucleus.

Figura 10 – Anotação Discursiva.



Fonte: PARDO; NUNES (2006, p.6)

A figura 9 apresenta um texto segmentado em seis unidades. Na figura 10, a representação das relações retóricas entre as unidades está marcada pelas palavras em maiúsculas. "S" marca a oração satélite (que fornece informações complementares) e "N", a oração núcleo (que fornece informações mais importantes), os números se referem aos números dos segmentos do texto da figura 9. "BACKGROUND" seria relação retórica de contextualização, "ELABORATION", de elaboração, "CIRCUMSTANCE", de circunstância e "EXPLANATION" de explicação. Os segmentos 1 e 2 possuem uma relação de circunstância na qual o segmento 1 é o núcleo e 2, o satélite. Os segmentos 3 e 4 estão em uma relação de explicação, na qual o segmento 3 é o núcleo e 4 o satélite. Os segmentos 1, 2, 3 e 4 estão em uma relação de elaboração, na qual 1 e 2 são o núcleo e 3 e 4, os satélites. E todos os segmentos estão em uma relação contextual, na qual os segmentos 1, 2, 3 e 4 são satélites e os segmentos 5 e 6 são o núcleo.

Neste capítulo, reportamos diferentes formas de como os dados linguísticos subsidiam não só avanços acadêmicos, mas também avanços na área de educação, bem como para o PLN. Apresentamos as possíveis contribuições de se etiquetar dados tanto para a Sociolinguística Variacionista quanto para o PLN, e como uma área pode retroalimentar a outra. Ao final, ilustramos diversas formas de se etiquetar dados para que o nosso leitor saiba o que já é possível de ser feito em termos de anotação linguística, ressaltando que, nesta tese, o nível linguístico analisado é o POS. No próximo capítulo descrevemos como conduzimos o nosso estudo, apresentando e discutindo nossos resultados.

3 FIZ A COLETA, TRANSCREVI E AGORA? A SISTEMATIZAÇÃO DA AMOSTRA

Dentro da perspectiva da Sociolinguística Variacionista, a língua é tomada como um sistema heterogêneo, regido por regras categóricas e variáveis, de forma que a variação não compromete o entendimento mútuo entre os falantes (WEINREICH; LABOV; HERZOG, 1968). É tarefa do pesquisador, nessa área, então, descrever como essas regras regem os usos da língua feitos pelos falantes e os significados sociais inerentes a esses usos. Para realizar esse empreendimento, é necessário, como apontado anteriormente, coletar amostras de dados empíricos, sendo a entrevista sociolinguística escolhida por oferecer melhores condições de gravação e também de exposição ao vernáculo do informante.

Após essa coleta e a transcrição, o pesquisador precisa sistematizar seus dados para realizar as buscas pelo fenômeno e após as análises tomar medidas para o arquivamento e o gerenciamento desses dados. Nosso protocolo, então, se inicia, após a etapa de transcrição e termina com a etapa pós-análise dos dados.

Por esta tese tratar da criação de um protocolo para sistematização da amostra Deslocamentos 2019 do banco de dados Falares Sergipanos, neste capítulo, descrevemos as ações e os resultados para atingir nosso objetivo. Primeiramente, apresentamos a amostra que escolhemos para servir de testagem e detalhamos os procedimentos de inspeção e preparação para fazermos a anotação linguística POS e a busca automática por fenômenos variáveis. Após a preparação da amostra, descrevemos o processo de seleção das ferramentas gratuitas que fazem etiquetagem linguística POS automática, em seguida, apresentamos a avaliação da etiquetagem, bem como a avaliação do modelo que utilizamos para avaliá-las. Em seguida, discorremos sobre as funcionalidades das ferramentas e seu desempenho para buscas automáticas. Por fim, traçamos as diretrizes para balizar questões éticas e as licenças de uso para divulgação da amostra em conformidade com os preceitos defendidos pelo paradigma da Ciência Aberta.

3.1 A amostra Deslocamentos 2019

3.1.1 Inspeção da amostra

A amostra Deslocamentos 2019 foi concebida no âmbito do Projeto “A língua do universitário: fala, leitura e escrita para o letramento acadêmico” (FREITAG, 2018b) e coletada

pelas pesquisadoras Thaís Regina de Andrade Corrêa e Cristiane Conceição de Santana Ribeiro no ano de 2018. Essa amostra contém 64 entrevistas sociolinguísticas realizadas com estudantes universitários pertencentes à comunidade de prática da Universidade Federal de Sergipe. Esses informantes foram estratificados por tipo de deslocamento, tempo de ingresso no curso e sexo/gênero. Deslocamento diz respeito ao acesso do estudante ao campus em termos de mobilidade:

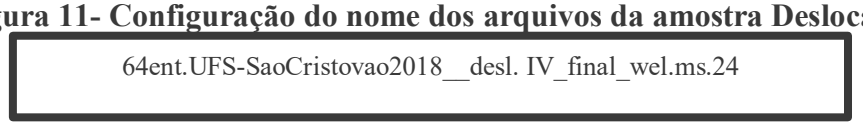
> "a) Deslocamento I: constituído por estudantes da região metropolitana de Aracaju; b) Deslocamento II: constituído por estudantes que residem no interior do estado de Sergipe que vão e voltam todos os dias para o município (zona urbana e zona rural) onde residem; c) Deslocamento III: constituído por estudantes oriundos do interior do estado de Sergipe (residentes tanto da zona rural, quanto da zona urbana) que passaram a morar na região metropolitana de Aracaju para estudar na UFS; d) Deslocamento IV: constituído por estudantes oriundos de outros estados do Brasil (de diferentes regiões) que vieram para o estado de Sergipe, mais especificamente para região metropolitana de Aracaju, para estudar na UFS." (CORRÊA, 2019, p. 72)

Já o tempo de ingresso no curso é divido em dois níveis: início, isto é, aqueles que estão cursando do primeiro ao quarto período do curso, e, final, a partir do quinto período até o último período do curso. A variável sexo/gênero foi definida como masculino ou feminino.

Inicialmente, para codificação dos arquivos das entrevistas, as pesquisadoras haviam decidido manter as três letras iniciais do primeiro nome do participante, seguidas pelo número do deslocamento no qual o participante se encaixava, o sexo/gênero (M para masculino, F para feminino), e se o estudante pertencia ao início do curso (I para 1º ao 4º período) ou ao final do curso (F para 5º ao 10º período). Exemplo: ADE4MI refere-se ao participante com iniciais ADE, pertencente ao deslocamento 4, do sexo/gênero masculino, em início de curso.

No entanto, essa codificação não traz a informação sobre o tipo de amostra (entrevista ou interação) e a comunidade onde a amostra foi coletada, o que, na criação de uma ferramenta de busca de dados para o site, dificultaria a busca por comunidade e por tipo de amostra. Adotamos, então, o padrão das amostras coletadas em Itabaiana, no qual o nome da amostra é codificado pelo número do arquivo, tipo de amostra, comunidade do informante, ano da coleta, tipo do deslocamento, tempo no curso, três letras iniciais do nome, sexo/gênero, se é estudante universitário e idade. Como exemplo de nomeação temos:

Figura 11- Configuração do nome dos arquivos da amostra Deslocamentos 2019



Fonte: Banco de dados Falares Sergipanos, Amostra Deslocamentos 2019

Na figura 11, "64" indica o número do arquivo na amostra, "ent" indica que o tipo de coleta foi entrevista, "UFS-SãoCristovao" se refere à comunidade do informante, "2018", ao ano de

coleta, “desl.IV;’, ao tipo de deslocamento, “final” indica que o participante está no final do curso, “wel” são as três primeiras letras do primeiro nome do informante, “m”, indica que o informante é do sexo/gênero masculino, “s”, que é estudante universitário, e “24;” a idade do informante.

Após a modificação do nome dos arquivos, que foi realizada de maneira manual, realizamos uma inspeção para quantificar a amostra em relação aos seus arquivos. Abaixo segue um quadro com todos os tipos de arquivos disponíveis da amostra, a duração total dos arquivos de áudio e também a quantidade de palavras total das entrevistas.

Quadro 2 – Dados Gerais da Amostra Deslocamentos 2019.

Quantidade de entrevistas	Duração total dos arquivos de áudio	Formatos de arquivos disponíveis	Quantidade de palavras total
64	49h01min20s	.wav; .eaf; .txt	460.235

Fonte: elaboração própria.

No quadro 2, estão mencionados os três tipos de formatos de arquivo das 64 entrevistas. Os arquivos em áudio já se encontravam salvos em formato .wav (Waveform Audio File Format). Salientamos que, embora seja um formato com tamanho maior de armazenamento do que um áudio em formato mp3 (MPEG-1 Audio Layer 3), por exemplo, é o padrão que tem sido utilizado como boa prática por organizações internacionais como o *The Language Archive*, *Linguistic Data Consortium*, e também recomendado pela *International Association of Sound and Audiovisual Archives*, devido à alta qualidade do som.

As entrevistas em formato .eaf (*EUDICO Annotation Format*) referem-se aos arquivos de transcrições alinhadas feitas por meio do ELAN (*EUDICO Linguistic Annotator*), sendo somente possível abrir o arquivo caso o programa esteja instalado na máquina do usuário. Além disso, para que o usuário possa ouvir o áudio e acompanhá-lo juntamente com a transcrição, é necessário que ambos arquivos de áudio e de transcrição alinhada estejam guardados em uma mesma pasta.

Por fim, os arquivos em formato .txt (*Unformatted text file)* contêm somente as entrevistas transcritas conforme as normas de transcrição ortográficas que já haviam sido elaboradas para o banco de dados Falares Sergipanos. Como essas normas podem afetar os resultados da etiquetagem automática, elas estão descritas no quadro 3 abaixo.

Quadro 3 – Normas de transcrição no 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: Sousa, Souza (2022).

Como se observa pelo quadro, a transcrição considera questões que são próprias do discurso falado como marcas de hesitação, pausa preenchida, truncamentos, dentre outros fatores que são relevantes para análises linguísticas que consideram, por exemplo, questões referentes à assimetria de relações de poder, processamento linguístico, entre outras. Assim,

essas informações devem constar dos arquivos originais. Por outro lado, essas mesmas marcas podem favorecer ou desfavorecer a etiquetagem automática, algo que será discutido mais adiante. Em relação ao formato das transcrições, a extensão .txt permite que o arquivo seja processado por outros tipos de softwares, por exemplo, etiquetadores como o *TreeTagger*, *softwares* para análise de texto e de *corpora* como o *LancsBox 6.0* e o *AntConc*, e para algoritmos de análise estatística criados por meio de linguagens de programação como o R ou o *Python*.

3.1.2 Preparação da amostra para etiquetagem

Após uma primeira inspeção para verificar a formatação dos nomes, a quantidade de arquivos, seus formatos, o número de palavras, procedemos à preparação da amostra. Inicialmente, esta fase contou com a ajuda de dois bolsistas de iniciação científica, Mohamed Malam Dabo e Calyne Porto de Oliveira. Foram feitos dois treinamentos com os estudantes para ensiná-los a baixar e instalar duas ferramentas de etiquetagem, *LancsBox* 6.0 e o *TreeTagger*. Ambos os treinamentos foram gravados e resultaram em tutoriais que foram disponibilizados para os estudantes. Após o treinamento, os arquivos em .txt foram organizados como ilustrado na figura 12.

Figura 12 - Fluxo de preparação da amostra inicial.

Fonte: elaboração própria.

Os procedimentos compreenderam i) a seleção de 14 entrevistas (20% do total), contendo 100.437 palavras (21% do total) para servir de treinamento; ii) a exclusão dos cabeçalhos das entrevistas, das marcações de tempo, das marcações de risos, barulhos, intervenções, traços de oralidade e contextuais da situação de comunicação de forma manual; iii) arquivamento das alterações em novos documentos, uma vez que a documentação linguística deve servir a diferentes propósitos, por isso uma cópia de cada uma das transcrições originais está arquivada (cf. LEECH, 2004; LEECH, 2013; OLIVEIRA JR., 2016; SINCLAIR, 2004b). Apesar de ser uma prática comum para a etiquetagem de textos, não houve a necessidade de fazer a

tokenização (divisão do texto em segmentos menores, como palavras e sinais de pontuação) das transcrições porque os etiquetadores testados já fazem esse processo.

Contudo, percebemos, durante o processo, que a extração manual das informações contidas na segunda etapa de preparação foi falha, pois ainda restaram marcas que deveriam ter sido retiradas. Além disso, como tínhamos o objetivo de testar a performance dos etiquetadores para automatizar as tarefas de extração automática de ocorrências de fenômenos linguísticos, optamos por fazer o processamento de todas as entrevistas com a transcrição original e também com as transcrições limpas excluindo da análise o *TreeTagger* por ele não oferecer um recurso de busca automática por fenômenos linguísticos. Desta vez, a limpeza das transcrições foi feita de forma automática por meio de um código criado em linguagem *Python* especificamente para isso. Esse código, construído em parceria com o estudante de iniciação científica Túlio de Sousa Góis, que é estudante de Engenharia da Computação UFS-São Cristóvão, permite também a contagem automática de palavras presentes em todos os arquivos e está disponível no repositório do projeto desta tese (SOUSA *et al.*, 2022) na plataforma *Open Science Framework* (*OSF*), ressaltando que este é um dos passos para fazer pesquisa dentro do paradigma da Ciência Aberta. Na figura 13 abaixo, apresentamos a função de limpeza:

Figura 13 – Função de limpeza.

```
import re
import string

def limpa_str(x):
    stoplist = ['eh', 'hes', 'RISOS', 'risos', 'BARULHO', 'INTERVENIENTE', 'PIGARRO', 'pigarro', 'CHORO', 'est', 'ah', 'uh']
    pontuacao = "!#$%&'*+,-./;<=>()@[]^_`{|}~"

    x = x.translate(str.maketrans('', '', pontuacao))                    #remove pontuações
    x = ' '.join([word for word in x.split() if word not in stoplist])   #remove as palavras que estejam na stoplist

    return x
```

Fonte: (SOUSA *et al.*, 2022).

Conforme observamos pela figura 13, a função de limpeza foi criada a partir das bibliotecas *re* e *string* do *Python*, que permitem o trabalho com expressões regulares[43] e *strings* (sequências ordenadas de caracteres). Podemos observar pelo código que palavras (no caso do código, *strings*) como "RISOS", "BARULHO", que oferecem informações sobre o contexto e a fala do entrevistador e do entrevistado, estão sendo selecionadas para exclusão em uma *stoplist*, isto é, em uma lista de palavras que não possuem valor para certos tipos de análise,

[43] Expressões regulares, conforme Dias (2021) são formas utilizadas para se descrever um conjunto de *strings* (sequências) que determinam um padrão. Disponível em: https://ic.unicamp.br/~mc102/aulas/aula15.pdfe. Acesso em: 19 out., 2021.

como frequência e contagem de palavras. Já na figura 14, abaixo, estão contidos a implementação da função de limpeza e o código para contagem de palavras.

Figura 14 - Implementação da função de limpeza e código para contagem de palavras.

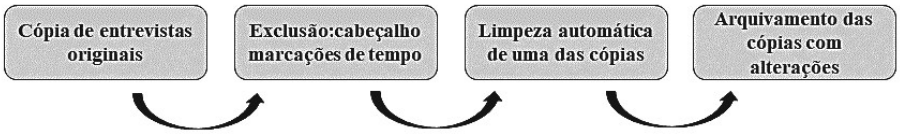
```
    #Abre o arquivo e limpa
    conteudo_arquivo = limpa_str(open(arquivo, encoding="utf8").read())
    #Cria um doc (no spacy) para o arquivo em execução
    doc = nlp(conteudo_arquivo)
    |
    #Realiza a contagem de palavras de todos os arquivos
    tokens = [token.orth_ for token in doc]
    ntokens += len(tokens)
print(ntokens)
```

Fonte: (SOUSA *et al.*, 2022).

Como nossa intenção é sermos transparentes e acessíveis, todos os códigos que criamos vêm com explicação do que cada linha representa, após os *hashtags* (#), para que as explicações não modifiquem o código. Assim como na figura 13, na figura 14 temos o uso da função "limpa_str", que vai realizar a limpeza do arquivo que está sendo aberto. Já a contagem de palavras é realizada por meio da função "token.orth_", que nos retorna apenas os tokens de palavras e da função "len", que nos retorna a quantidade de palavras. A função "token.orth_" está contida na biblioteca do *spaCy* 3.5.

O fluxo da preparação final da amostra se modificou para apenas duas etapas: i. fazer duas cópias dos arquivos originais. Uma delas não terá a anotação POS, como o apresentado na figura 15.

Figura 15 – Fluxo final de preparação da amostra.



Fonte: elaboração própria.

Como optamos por comparar os resultados da etiquetagem em entrevistas com e sem limpeza, o fluxo de trabalho final compreendeu quatro estágios: i) fazer cópias das entrevistas originais; ii) excluir os cabeçalhos e as marcações de tempo, algo que já tínhamos feito no fluxo anterior; iii) submeter uma das cópias à limpeza de marcações contextuais e de oralidade; iv) arquivamento das cópias com alterações.

3.2 O processo de anotação automática da amostra

3.2.1 A escolha das ferramentas para o estudo

Existem dois tipos de métodos para o desenvolvimento de ferramentas para o PLN (KUMAR; JOSAN, 2010; PALMER; XUE, 2010): o supervisionado com *corpora* previamente etiquetados, o que facilita ao sistema a desambiguação ou a aprendizagem de regras de etiquetagem; o não-supervisionado, que não usa *corpora* etiquetado, mas técnicas computacionais para etiquetar, ou seja, tanto calcula a informação probabilística por meio de etiquetadores estocásticos ou por meio de sistemas baseados em regras ou sistemas baseados em transformação. No entanto, conforme Palmer e Xue (2010), para se atingir melhor desempenho, os modelos supervisionados são os melhores recursos para que tal objetivo seja alcançado, o que confirma a importância da disponibilização de *corpora* anotados para o PLN.

Embora de extrema importância para o PLN, as especificações técnicas discutidas acima não foram consideradas ao apresentar as ferramentas disponíveis e gratuitas para a língua portuguesa, pois isso fugiria ao escopo dessa tese, mas acreditamos ser importante pontuar esses aspectos para evidenciar como os dados etiquetados são relevantes para a construção de etiquetadores. As ferramentas gratuitas para a língua portuguesa, então, estão descritas em relação às funções que desempenham, à necessidade de conhecimentos de linguagem de programação para operá-las, ao formato de saída dos documentos anotados, à necessidade de instalação na máquina do usuário, e à existência de interface de busca automática por padrões linguísticos na ferramenta.

O *e-Dictor* (PAIXÃO DE SOUSA; KEPPLER, 2007) foi criado com a intenção primeira de ser um editor de textos históricos de língua portuguesa para análise linguística automática. A interface do *e-Dictor* é amigável, possui manual em língua portuguesa e uma das suas funcionalidades é permitir a anotação de classe de palavras (POS). Contudo, o programa não realiza buscas automáticas em *corpora* e a inserção das etiquetas no programa é feita manualmente. O formato de saída dos arquivos pode ser tanto em .txt ou .html.

A ferramenta *Aelius* (ALENCAR, 2013) foi desenvolvida em linguagem *Python*, utilizando a biblioteca NLTK 2.0.1rc1. Essa ferramenta faz a etiquetagem POS e sintática para o português e pode também utilizar interfaces com os etiquetadores *HunPos* e *Stanford Tagger*. O *Aelius* também já faz a tokenização dos textos para posterior etiquetagem. Os arquivos de saída podem ser .txt que possuem uma anotação plana, ou XML. Conforme salienta o manual

do usuário, não são necessários conhecimentos profundos de linguagem de programação para se realizar a etiquetagem com o *Aelius*, de forma que, como afirma o autor, um curso de curta duração de 15 dias é suficiente. É possível por meio de conhecimentos de linguagem *Python* realizar buscas no corpus etiquetado, mas esta não é uma função nativa da ferramenta.

Citius Tools (GAMALLO *et al*., 2014) é uma suíte para o PLN escrita em linguagem Perl, que faz a quebra de sentenças, tokenização e a etiquetagem POS e sintática para a língua portuguesa. É preciso conhecimentos da linguagem Perl para se conseguir utilizar a suíte, porém, como afirma Sardinha (2004), conhecimentos acerca dessa linguagem são de grande valia para quem trabalha com processamento de textos, principalmente na área de Linguística de *Corpus*. O formato de saída do processamento é .txt, com representação do texto em CoNLL-X (cada linha uma entrada e cada coluna uma categoria de análise). Esta ferramenta não possui interface para buscas automáticas em *corpora*.

LX-Suite (BRANCO; SILVA, 2004) é uma suíte gratuita que faz o processamento da língua portuguesa. Entre as suas diferentes funcionalidades estão a quebra de sentença, a tokenização e a etiquetagem POS. Não é necessário fazer o download da ferramenta, sendo possível fazer o processamento online de textos. Está arquitetada em língua portuguesa e inglesa, sua interface é amigável e não necessita de conhecimentos de linguagem de programação para manuseá-la. O formato de saída dos arquivos são .txt, porém não possui interface para buscas por itens linguísticos.

O *LancsBox* 6.0 (BREZINA; WEILL-TESSIER; MCENERY, 2022) também é uma suíte que faz análise de *corpora* de diferentes línguas, sendo a língua inglesa aquela que possui maior funcionalidade. Essa ferramenta possui uma interface de fácil manuseio, não necessitando de conhecimentos de linguagem de programação para operá-la. Dentre suas funcionalidades encontram-se a lematização, a tokenização, a anotação morfológica, a visualização gráfica de coligados e colocações, listas de frequências, buscas automáticas em *corpora* do próprio usuário por meio de etiquetas e também de expressões regulares para linguagem Perl. Como formato de saída, os arquivos processados podem ser exportados como .txt.

O *TreeTagger* (SCHIMID, 1995) é uma ferramenta específica para lematização, anotação morfológica e POS. Possui uma interface em língua inglesa de fácil manuseio e instalação, sendo que conhecimentos de linguagem de programação para operá-la não são necessários. O formato de saída dos arquivos processados é o .txt.

spaCy 3.5 (HONIIBAL *et al*., 2020) é uma biblioteca aberta de PLN da linguagem de programação *Python*. Dentre as funcionalidades do *spaCy* 3.5 destacamos o fato de fazer lematização, tokenização, anotação POS e sintática, além de possuir recursos para buscas automáticas como o Matcher (busca de tokens por meio do lema, etiquetas morfológicas e POS) e o Dependency Matcher (que realiza buscas sintáticas combinadas com os outros tipos de etiquetas). É importante para o bom funcionamento da ferramenta que o usuário conheça a linguagem de programação *Python*. O arquivo de saída da anotação sintática e de dependências é exportado como .svg, ou html. É possível, também, por meio de conhecimentos de programação em linguagem *Python*, salvar os resultados das buscas bem como de outros formatos de etiquetagem (lema, morfológica, POS e sintática) em diferentes tipos de arquivo como .csv, .xlsx e .txt.

Na tabela 2 abaixo, fazemos um resumo das características das ferramentas que fazem etiquetagem para a língua portuguesa.

Tabela 2 – Resumo das características das ferramentas para etiquetagem.

Etiquetador	Tipo de anotação	Conhecimento de Programação	Arquivo de saída	Instalação na máquina	Interface para buscas
e-Dictor	POS	Não necessário	.html/ .txt	Necessário	Não
Aelius	POS e sintática	Necessário	.html/ .txt	Necessário	Sim
Citius Tools	POS e sintática	Necessário	CoNLL-X	Necessário	Não
LX-Suíte	POS	Não necessário	.txt	Não necessário	Não
***LancsBox* 6.0**	POS	Não necessário	.txt	Necessário	Sim
TreeTagger	POS	Não necessário	.txt	Necessário	Não
***spaCy* 3.5**	POS e sintática	Necessário	.html/.svg (relações de dependência) e diversos formatos (outros tipos de anotação)	Não necessário.[44]	Sim

Fonte: elaboração própria.

[44] Embora tivéssemos utilizado o *Google Colaboratory* para trabalhar com spaCy 3.5, consideramos relevante pontuar para o leitor que a utilização deste ambiente, por ser online, passa por alguns percalços como servidor instável e queda de internet. Por isso, recomendamos também que o usuário possua uma IDE off-line, como o *PyCharm* ou *Jupyter*, para trabalhar com essa biblioteca.

Todos os etiquetadores acima realizam a anotação POS, mas apenas o *spaCy* 3.5, o *Citius Tools* e o *Aelius* fazem a anotação sintática. Esse fato demonstra a necessidade de maior investimento em criação de ferramentas automáticas que façam etiquetagem no nível sintático e que sejam tanto de arquitetura aberta, para se propor modificações, quanto de acesso aberto, para que, diante de uma realidade de poucos investimentos em pesquisa, outros pesquisadores possam compilar e analisar *corpora* para fornecer recursos para o treinamento e aperfeiçoamento dessas ferramentas.

Três etiquetadores requerem conhecimentos de programação para que sejam manuseados de forma eficiente, o *Aelius*, o CititusTools e o *spaCy* 3.5. Todas as ferramentas oferecem formato de saída .txt. Já os formatos .html são oferecidos pelo o *Aelius*, o e-Dictor e o *spaCy* 3.5, sendo que este último também oferece saída em formato .svg. Dentre as ferramentas, o LX-Suíte e o *spaCy* 3.5 não requerem instalação na máquina do usuário. Além disso, apenas três, *Aelius*, *LancsBox* 6.0 e *spaCy* 3.5 possuem interface para buscas. Dessas três, apenas duas são atualizadas, o *LancsBox* 6.0 e o *spaCy* 3.5. Por funcionar apenas em linguagem *Python* 2.x e não ter recebido atualizações desde 2013, excluímos o *Aelius* de nossa análise, optamos, então, por comparar o *LancsBox* 6.0 e o *spaCy* 3.5 uma vez que além de realizar a etiquetagem e receberem atualizações periodicamente, oferecerem recursos para buscas automáticas por padrões linguísticos.

3.2.2 Descrição das ferramentas

LancsBox 6.0[45] é um *software* gratuito que faz análises linguísticas em grandes volumes de texto, sendo possível processar mais de 1000.000.000 de palavras por vez. É uma ferramenta construída para se fazer análises linguísticas em *corpora*, ou seja, é possível identificar padrões linguísticos por meio das frequências das ocorrências daquilo que se pretende estudar. É um *software* desenvolvido em diferentes linguagens de programação, como *Perl*, *Java*, *Groove*, entre outras.

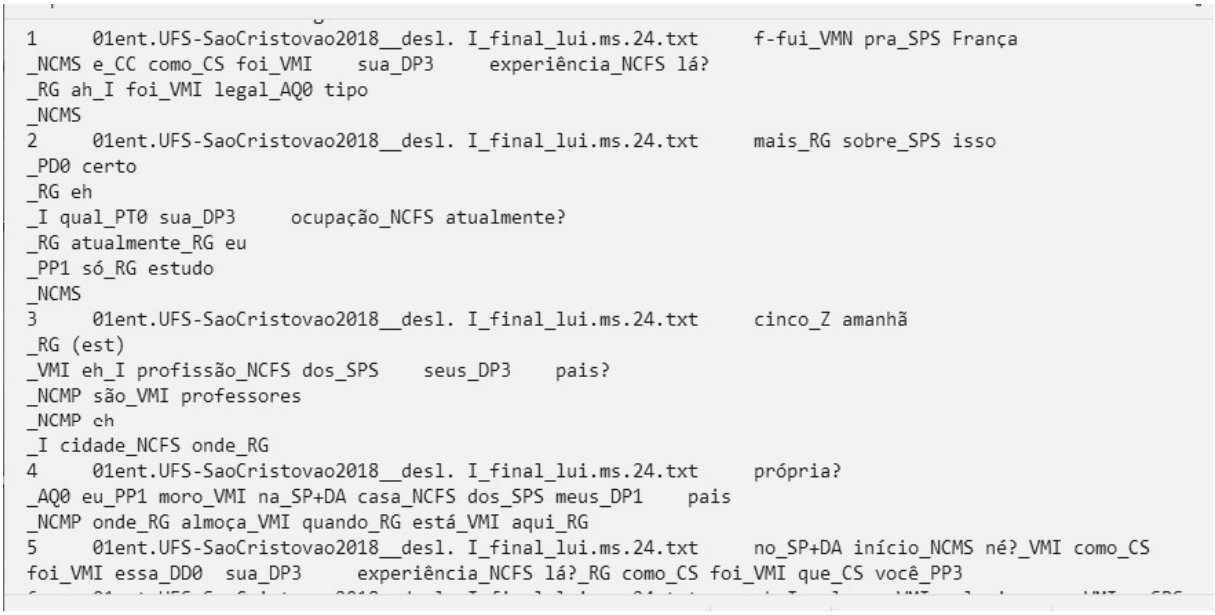
Essa ferramenta, originalmente concebida para a língua inglesa, possui funcionamento para diferentes línguas, sendo o tipo de suporte que ela oferece divido em três níveis: i) sem anotação linguística; ii) com anotação POS com interface para os parâmetros do *TreeTagger*; iii) línguas com suporte para anotação POS e de lema, pesquisas inteligentes, reconhecimento

[45]Todas as informações acerca deste software foram extraídas do site http://corpora.lancs.ac.uk/*LancsBox*/. Acesso em: 11 nov. 2022.

de abreviações e clíticos. Entre as 20 línguas que compõem o nível 2, está a língua portuguesa, o que já oferece um bom suporte para buscas automáticas por meio de atributos linguísticos, como buscas por verbos no modo indicativo, substantivos comuns e próprios, adjetivos, determinantes, entre outros. Apesar de não serem necessários conhecimentos sobre linguagem de programação para operar o *LancsBox* 6.0, ter conhecimento sobre REGEX em linguagem *Perl*, que é a interface para utilização dessas expressões, favorecem as buscas por parâmetros mais específicos, como por exemplo, palavras terminadas em "o" e "os" antes de adjetivos.

Os resultados das buscas automáticas realizadas na ferramenta podem ser diretamente salvos em arquivos com extensão .txt, com o número da ocorrência, o nome do arquivo, o contexto que antecede a ocorrência, a ocorrência e o contexto imediatamente após a ocorrência. É possível também salvar o arquivo com as marcações da anotação POS. Na figura abaixo segue um exemplo do arquivo de saída.

Figura 16 – Representação do arquivo de saída do *LancsBox* 6.0 em .txt.



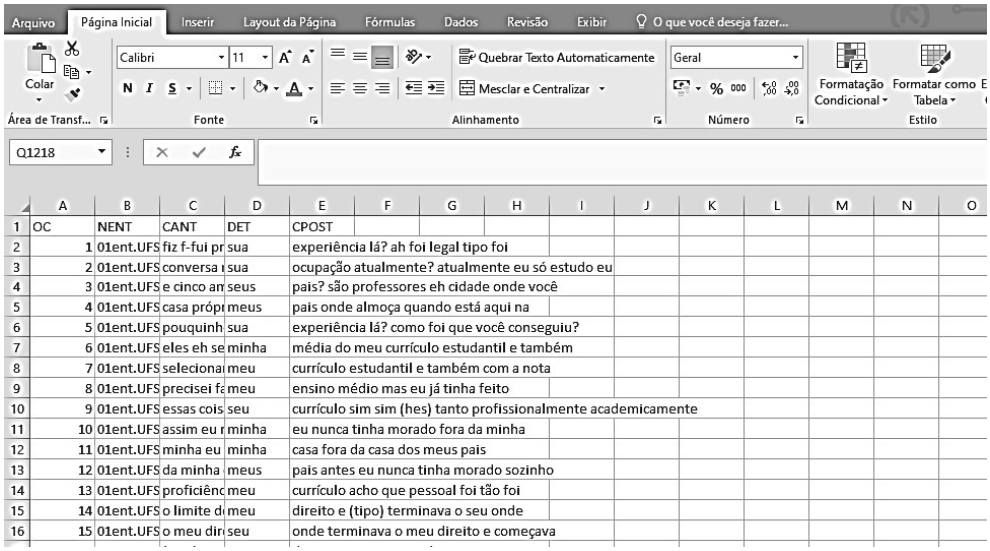
```
1     01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24.txt     f-fui_VMN pra_SPS França
_NCMS e_CC como_CS foi_VMI    sua_DP3     experiência_NCFS lá?
_RG ah_I foi_VMI legal_AQ0 tipo
_NCMS
2     01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24.txt     mais_RG sobre_SPS isso
_PD0 certo
_RG eh
_I qual_PT0 sua_DP3     ocupação_NCFS atualmente?
_RG atualmente_RG eu
_PP1 só_RG estudo
_NCMS
3     01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24.txt     cinco_Z amanhã
_RG (est)
_VMI eh_I profissão_NCFS dos_SPS    seus_DP3    pais?
_NCMP são_VMI professores
_NCMP eh
_I cidade_NCFS onde_RG
4     01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24.txt     própria?
_AQ0 eu_PP1 moro_VMI na_SP+DA casa_NCFS dos_SPS meus_DP1    pais
_NCMP onde_RG almoça_VMI quando_RG está_VMI aqui_RG
5     01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24.txt     no_SP+DA início_NCMS né?_VMI como_CS
foi_VMI essa_DD0  sua_DP3     experiência_NCFS lá?_RG como_CS foi_VMI que_CS você_PP3
```

Fonte: elaboração própria com arquivo de saída do *LancsBox* 6.0.

Na figura temos o arquivo de saída da busca por determinantes antes de possessivo. Os números no canto esquerdo indicam o número da ocorrência no *corpus*, em seguida, o nome do arquivo (no caso deste trabalho, o nome do arquivo da entrevista), o contexto anterior à ocorrência, o possessivo, o contexto posterior ao possessivo. Esse arquivo contém a anotação linguística POS, que pode ser visualizada sempre à direita de cada palavra precedida por um traço de sublinhado. É possível copiar e colar em um novo arquivo os resultados que aparecem na tela do *software* em um arquivo com extensão .csv. Em formato .csv, os dados ficam mais

organizados em colunas facilitando a visualização das ocorrências em contexto e também a manipulação em *softwares* de análise estatística, como pode ser visto na figura 17.

Figura 17 – Visualização dos dados extraídos do *LancsBox* 6.0 em planilha .csv.



	A	B	C	D	E
1	OC	NENT	CANT	DET	CPOST
2	1	01ent.UFS	fiz f-fui pr	sua	experiência lá? ah foi legal tipo foi
3	2	01ent.UFS	conversa	sua	ocupação atualmente? atualmente eu só estudo eu
4	3	01ent.UFS	e cinco an	seus	pais? são professores eh cidade onde você
5	4	01ent.UFS	casa próp	meus	pais onde almoça quando está aqui na
6	5	01ent.UFS	pouquinh	sua	experiência lá? como foi que você conseguiu?
7	6	01ent.UFS	eles eh se	minha	média do meu currículo estudantil e também
8	7	01ent.UFS	selecionar	meu	currículo estudantil e também com a nota
9	8	01ent.UFS	precisei fa	meu	ensino médio mas eu já tinha feito
10	9	01ent.UFS	essas cois	seu	currículo sim sim (hes) tanto profissionalmente academicamente
11	10	01ent.UFS	assim eu r	minha	eu nunca tinha morado fora da minha
12	11	01ent.UFS	minha eu	minha	casa fora da casa dos meus pais
13	12	01ent.UFS	da minha	meus	pais antes eu nunca tinha morado sozinho
14	13	01ent.UFS	proficiênc	meu	currículo acho que pessoal foi tão foi
15	14	01ent.UFS	o limite d	meu	direito e (tipo) terminava o seu onde
16	15	01ent.UFS	o meu dir	seu	onde terminava o meu direito e começava

Fonte: elaboração própria com resultados extraídos do *LancsBox* 6.0.

Como o *LancsBox* 6.0 utiliza interface com o *TreeTagger*, recorremos ao estudo de Gamallo e Garcia (2013) para observar o valor da acurácia para esse etiquetador. Os autores reportaram um valor de 91,3% para o português brasileiro. O conjunto de etiquetas empregado pelo *LancsBox* 6.0 é composto por um conjunto de etiquetas do padrão EAGLES[46]: adjetivo, advérbio, determinante, substantivo, verbo, pronome, conjunção, interjeição, adposição, sinais de pontuação e numerais. Esse padrão é baseado na informação que os dicionários oferecem sobre a classe gramatical de um determinado item. Ainda sob esses critérios, as 11 etiquetas especificadas anteriormente são chamadas como obrigatórias, por se tratarem de classes de palavras. Informações gramaticais como gênero (para substantivos), modo (para verbos), grau (para adjetivos) são consideradas etiquetas recomendadas, e etiquetas referentes a aspecto (para verbos), contabilidade (substantivos – contáveis ou incontáveis), por exemplo, são denominadas opcionais, por se tratarem de atributos que vão além do nível morfossintático, trazendo informações semânticas. O *LancsBox* 6.0 usa etiquetas reduzidas, ou seja, todas as

[46] EAGLES é a sigla utilizada para se referir a Expert Advisory Group on Language Engineering Standards (Grupo Consultivo de Especialistas em Engenharia de Línguas), uma iniciativa que busca estabelecer padrões para pesquisas com grandes volumes de recursos linguísticos, manipulação do conhecimento advindo dessas pesquisas e avaliação de recursos, ferramentas e produtos. O conjunto de etiquetas do padrão EAGLES usado pelo *LancsBox* 6.0 pode ser encontrado no site: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Portuguese-Tagset.html. Acesso em: 20 out. 2019.

obrigatórias e algumas recomendadas e opcionais. Assim, por exemplo, na frase: "fiz intercâmbio", o verbo "fiz" é etiquetado pelo *LancsBox* como VMI, verbo (V) principal (M) do modo indicativo (I).

O *spaCy* 3.5[47] é uma biblioteca aberta e gratuita, desenvolvida em linguagem de programação *Python*, com o objetivo de ajudar na criação de produtos que "processam e entendem grandes quantidades de texto"[48]. Essa biblioteca realiza tokenização, lematização, etiquetagem POS, etiquetagem sintática, etiquetagem de dependências, fornece subsídios para etiquetagem de entidades nomeadas, sumarização de textos, criação de *corpora*, entre outras funcionalidades. O nível de acurácia desse etiquetador para etiquetagem POS para a língua portuguesa é maior que a do *LancsBox* 6.0 com valor de 97%, já seu analisador sintático tem acurácia de 86%.

O spaCy possui um conjunto de 17 etiquetas POS no padrão *Universal Dependencies* (UD)[49]: adjetivo (ADJ), advérbio (ADV), interjeição (INTJ), substantivo (NOUN), substantivo próprio (PROPN), verbo (VERB), preposição (ADP), verbos auxiliares e cópula (AUX), conjunção coordenativa (CC), determinante (DET), numeral (NUM), prefixos e partículas negativas (não e nem) (PART), pronomes (PRON), sinais de pontuação (PUNCT), símbolos (SYM) e palavras as quais não se encaixam em nenhuma categoria POS (X).

. Embora as etiquetas morfológicas não apareçam juntas com o texto etiquetado em POS, elas podem ser visualizadas a partir de um código específico para esse tipo de etiquetagem, e, ainda, serem acessadas por meio de buscas por padrões morfológicos e gramaticais. Conforme Duran *et al.* (2022), as principais diferenças entre a anotação POS do modelo UD em relação à gramática normativa são:

> DET: a UD trabalha com o conceito de determinante e sob essa etiqueta reúne os artigos e os pronomes não nominais (demonstrativos e possessivos); NUM: apenas os numerais cardinais devem ser anotados como NUM, enquanto os numerais ordinais devem ser anotados como ADJ; PRON: apenas os pronomes nominais são anotados com essa etiqueta, e os demais pronomes, desde que estejam modificando um nominal, devem ser anotados como DET; NOUN: apenas os substantivos comuns são anotados sob essa etiqueta, pois os nomes próprios são anotados como PROPN;VERB: apenas os verbos considerados plenos e passíveis de serem classificados como predicados verbais devem ser anotados com essa etiqueta. AUX: verbos auxiliares e verbos de cópula "altamente gramaticalizados", ou seja, sem carga semântica significativa, devem ser anotados com essa etiqueta; PROPN: etiqueta destinada a anotar nomes próprios, desde que não coincidam com palavras comuns da língua. (DURAN *et al.,* 2022, p.1625).

[47] Todas as informações acerca do *spaCy* 3.5 foram retirados do seu site próprio disponível em: https://*spaCy* 3.5.io/usage/*spaCy* 3.5-101. Acesso em: 22 fev. 2022.

[48] Tradução nossa para: "process and "understand" large volumes of text". Disponível em: https://*spaCy* 3.5.io/usage/*spaCy* 3.5-101#whats-*spaCy* 3.5. Acesso em: 22 fev. 2022.

[49] *Universal Dependencies* é um quadro teórico que visa a anotação POS, morfológica e sintática. Suas etiquetas podem ser encontradas em: https://universaldependencies.org/guidelines.html. Acesso em 03 out. 2020.

Esse modelo, como poderemos ver adiante, tem desempenho melhor, uma vez que as etiquetas POS UD, por serem mais genéricas, captam contextos de determinantes em que o modelo do EAGLES não capturou.

Diferentemente do *LancsBox* 6.0, o *spaCy* 3.5 possui suporte para mais de 72 línguas, incluindo a língua portuguesa. É possível processar vários arquivos de uma só vez nessa ferramenta, sendo que para isso deve-se criar um código extra para que este processamento seja realizado, ou seja, são importantes conhecimentos de estruturas em *Python*, como estruturas de repetição, definição de funções para que a ferramenta seja manipulada de maneira eficiente. Assim como o *LancsBox* 6.0, essa biblioteca suporta buscas em *corpora* e possui também um recurso que se chama *Matcher* que permite buscas por *tokens* com base em atributos morfológicos e de POS e o *Dependency Matcher* para buscas sintáticas. Os resultados das buscas do *Matcher* podem ser visualizados em tela em linhas de concordância, como na figura 18, e em um *dataframe* a partir de usos das bibliotecas Pandas e *Wasabi* da linguagem *Python*.

Figura 18 – Visualização das saídas dos resultados do *Matcher*.

```
[dos seus, dos meus, da minha, do meu, no meu, na minha, da minha, dos meus, o meu, o seu, o meu, da
1
(est)
eh profissão dos seus pais?
são professores
eh
2 própria?
eu moro na casa dos meus pais
onde almoça quando está aqui
3 eh eles eh sele-me selecionaram a partir da minha média
```

Fonte: elaboração própria a partir do uso do *spaCy* 3.5 no Google Colaboratory.

O mesmo não ocorre para os resultados do *Dependency Matcher*, provavelmente por causa da segmentação feita pela ferramenta. Na figura abaixo temos o resultado de uma busca da relação entre uma preposição e um determinante possessivo.

Figura 19 – Visualização na tela da busca por uma relação de dependência.

```
for match in dep_matches:
  pattern_name = match[0]
  matches = match[1]
  possess, caso = matches[0], matches[1]
  print(nlp.vocab[pattern_name].text, '\t', doc[caso], doc[possess])
```

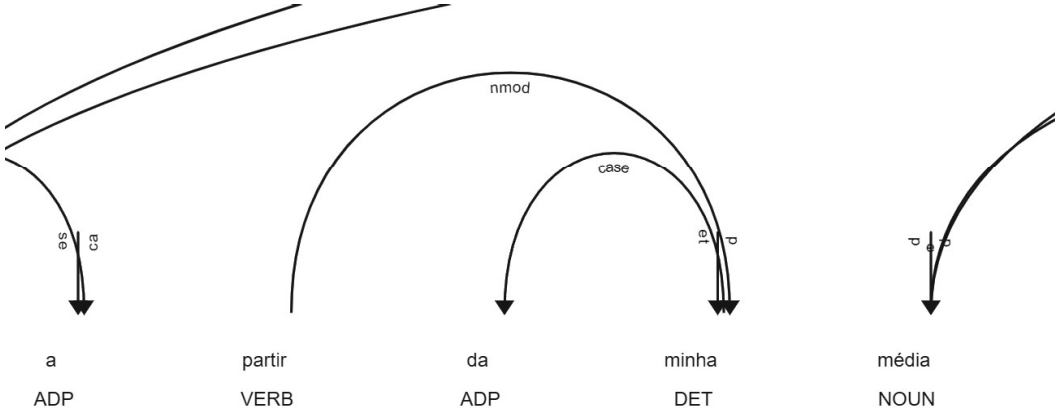
```
[(436175140319728513, [3026, 3022]), (436175140319728513, [5435, 5434])]
possess_det      de meu
possess_det      na minha
```

Fonte: elaboração própria a partir da IDE[50] Google Colaboratory.

[50] IDE (Integrated Development Environment – ambiente de desenvolvimento integrado).

Na figura 19, os números entre colchetes significam a representação das palavras encontradas pela busca, já o número maior fora dos colchetes, o nome do padrão que utilizamos para fazer as buscas. Conseguimos transformar a saída, que é uma lista de tuplas, em texto por meio da função *named tuple*. Ressaltamos que há muito mais ocorrências desse tipo de dependência, nesse mesmo documento, porém, possivelmente devido à segmentação realizada pela biblioteca e também pelo tipo de transcrição na qual não demarcamos as fronteiras sintáticas por sinais de pontuação, as relações de dependência não ficam bem marcadas o que pode gerar erro ao imprimir na tela as relações em forma de texto. Na figura 20, temos uma representação em arcos do texto que não foi captada pela busca da relação de dependência.

Figura 20 – Representação em árvore com relação de dependência de caso entre a preposição e o determinante possessivo.



Fonte: elaboração própria com o uso da biblioteca *spaCy* 3.5.

A representação acima mostra que há uma relação de caso entre “da” e “minha” na qual a cabeça é “minha”. Contudo, essa ocorrência não foi captada pela busca. Além disso, a relação entre “minha” e “média” não fica aparente, o arco está incompleto. No quadro 4, apresentamos uma comparação das duas ferramentas.

Quadro 4 – Comparação dos atributos das ferramentas.

Característica	***LancsBox* 6.0**	***spaCy* 3.5**
Tipo de suporte para Língua Portuguesa	Lema, morfológico e POS	Suporte completo
Etiquetas POS	11 etiquetas padrão EAGLES que aparecem combinadas com categorias morfológicas na superfície do texto	17 etiquetas padrão Universal Dependencies. Informações morfológicas são dadas à parte
Uso de REGEX	Sim, em linguagem Perl	Sim, em linguagem *Python*
Arquivamento dos resultados	.txt – direto da plataforma .txt, ou .csv - com recurso de copiar e colar	.txt, .csv, outros formatos – com uso de outras bibliotecas do *Python* .svg – para estruturas de dependência e para resultados do *Matcher*
Acurácia	91,3%	97%

Fonte: elaboração própria.

Em termos de suporte, como observamos no quadro 4, o *spaCy* 3.5 possui mais recursos em termos de etiquetagem do que o *LancsBox* 6.0, uma vez que este possui apenas para lema, morfologia e POS. A respeito das etiquetas, o *spaCy* possui mais categorias. Se considerarmos que a anotação linguística enriquece o texto com informações (LEECH, 2013; GRIES; BEREZ 2017; IDE, 2017), a etiquetagem do *LancsBox* 6.0 é mais simples do que a oferecida pelo *spaCy* 3.5. Porém, se analisarmos somente a etiquetagem POS, a do *spaCy* 3.5 não aglutina mais informações como a do *LancsBox* 6.0 que contém mais informações acerca das palavras em termos de gênero, pessoa, número e grau em uma única etiqueta.

Em relação ao uso de REGEX, ambas suportam. Já sobre o arquivamento, o *LancsBox* 6.0 é mais fácil por permitir o recurso de copiar e colar, o que não ocorre com o *spaCy* 3.5 que precisa da manipulação de outras bibliotecas em *Python* para realizar o arquivamento de maneira que possa ser utilizado em ferramentas/bibliotecas de análise estatística. Em termos de acurácia, o *spaCy* 3.5 possui um desempenho descrito na literatura superior ao do *LancsBox* 6.0. Os dados acima apontam que em duas características aqui descritas (acurácia e suporte) os

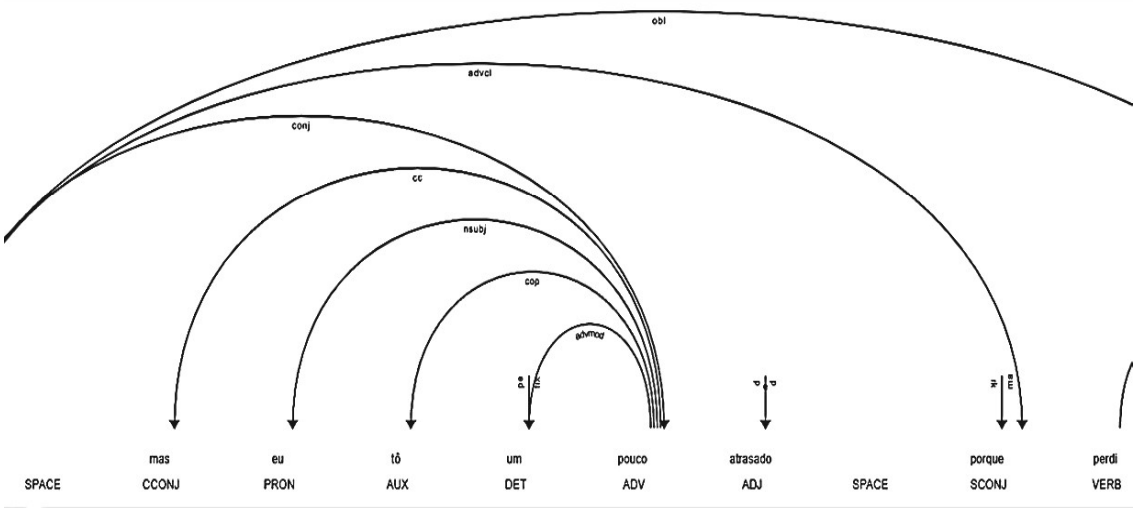
spaCy 3.5 é melhor, e o *LancsBox* 6.0., em duas também (etiquetas POS e arquivamento dos resultados).

Destacamos que o comportamento das duas ferramentas foi documentado em dados de textos escritos de jornais. Nosso objetivo é verificar os usos dessas ferramentas para automatizar as tarefas de busca e análise de dados de fala da sociolinguística, devido a isso, um escrutínio maior das ferramentas é reportado em nosso estudo.

3.2.3 O fenômeno linguístico utilizado para avaliação das ferramentas

Definimos como nível de análise o morfossintático. Essa escolha se deve: i) pelo fato de que as ferramentas produzidas para etiquetagem de textos em língua portuguesa possuem anotação automática consolidada nesse nível em relação aos demais; ii) por ser um nível que possui mais recursos automáticos para a análise de diferentes tipos de texto na língua portuguesa disponíveis gratuitamente, fato que em outros níveis de análise, os etiquetadores foram treinados em *corpora* especializados; iii) por não deixar a transcrição ilegível para outros pesquisadores, pois quando diferentes níveis são colocados no texto, recuperá-lo torna-se uma tarefa difícil (LEECH, 2004); e iv) ser a base para a formulação de outros níveis de anotação linguística (LEECH, 2013; GRIES; BEREZ, 2017). No entanto, não vamos comparar as ferramentas em relação às funções sintáticas e as relações de dependência entre os termos das orações devido ao fato de que o *LancsBox* 6.0 não faz esse tipo de anotação e também a busca pelas relações sintáticas no *spaCy* 3.5 não retorna dados em *spans* (partes do texto), não permitindo colocar os resultados das buscas em contexto, dificultando a tarefa do pesquisador em encontrar os dados no texto. Além disso, a representação gráfica do *spaCy* 3.5 das relações de dependência em textos maiores, como as entrevistas sociolinguísticas, precisa de uma segmentação maior do texto para ficar legível.

Figura 21 – Visualização das relações de dependência.



Fonte: elaboração própria com uso da biblioteca *spaCy* 3.5.

Na figura, é possível ver os arcos que representam as relações de dependência, porém, à medida que retiramos o zoom da página para vermos as relações dos arcos maiores, não é possível ver a frase e nem as relações estabelecidas nos arcos mais internos.

Em relação ao fenômeno variável, escolhemos a variação no preenchimento de determinante antes de possessivo pré-nominal. Esse fenômeno é caracterizado pelo fato de que os falantes do português brasileiro podem alternar entre preencher o determinante[51] antes de possessivo pré-nominal (1) ou não (2).

(1) "eu já fui assaltada n- na porta d**a** minha casa"

(2) "**Ø** minha irmã mais velha tem dezenove" (Excertos da entrevista 02ent.UFS-SaoCristovao2018__desl. I_inicio_lor.fs.18)

Em (1), o possessivo "minha" vem antecedido pelo determinante "a", já em (2), o mesmo possessivo "minha" não vem antecedido por determinante, ou seja, não há o preenchimento do determinante no sintagma.

Escolhemos esse fenômeno porque, além de ser um fenômeno morfossintático, ele já foi descrito em diferentes variedades do português brasileiro, como a variedade das capitais Recife, Salvador, Rio de Janeiro, São Paulo e Porto Alegre (CALLOU; SILVA, 1997); a variedade de Carnaíba em Pernambuco (SEDRINS; PEREIRA; SILVA, 2019); e a de Vitória no Espírito Santo (CAMPOS JR., 2012). Ademais, existem estudos sobre essa variação em condição de contato dialetal, na migração de Paraibanos para o estado de São Paulo (GUEDES, 2019) e na

[51] No caso do fenômeno em questão, o preenchimento de determinante o qual estamos descrevendo é o de determinante artigo.

comunidade de práticas da Universidade Federal de Sergipe em Sergipe (SILVA, 2020; SIQUEIRA, 2020). Ressaltamos que o estudo de Siqueira (2020) foi realizado com um recorte dos dados da amostra Deslocamentos 2019, já Silva (2020) utilizou dados da amostra Deslocamentos 2020, que teve parâmetros de coleta similares ao da Amostra Deslocamentos 2019, ambas pertencentes ao banco de dados Falares Sergipanos.

Os estudos acima citados, demonstram consonância no condicionamento de fatores morfossintáticos para a variação no preenchimento do determinante antes de possessivos pré-nominais: tipo de sintagma e função sintática, apesar de que Campos Jr. (2012) não controlou a função sintática. Além disso, a partir dos resultados percebemos também o caráter dialetal do fenômeno em que falantes da região sudeste tendem a preencher mais que aqueles do nordeste. Na tabela abaixo sistematizamos os resultados do fenômeno dando relevo às variáveis tipo de sintagma (TS), função sintática (FS) mais favorecedora do preenchimento, preenchimento de artigo (PA) e à comunidade de coleta (Comunidade).

Tabela 3 – Resultados de estudos sobre a variação no preenchimento de determinante antes de possessivo pré-nominal

ESTUDO	COMUNIDADE	TS	FS	PA
CALLOU E SILVA (1997)	Recife (PE); Salvador (BA); Rio de Janeiro (RJ); São Paulo (SP); Porto Alegre (RS)	Sintagma preposicionado	Sujeito	Recife (59/98) 60 % Salvador (57/87) 66 % Rio de Janeiro (280/399) 70 % São Paulo (147/209) 70 % Porto Alegre (26/33) 79 %
CAMPOS JR. (2012)	Vitória (ES)	Sintagma preposicionado	Não controlou	(331/1016) 33%
SEDRINS, PEREIRA E SILVA (2019)	Carnaíba (PE)	Sintagma preposicionado	Objeto indireto	(67/293) 23%
GUEDES (2019)	PB, PBSP, SP (falantes de João Pessoa (PB), falantes de João Pessoa (PB) em São Paulo (SP) e paulistas nativos, respectivamente).	Sintagma preposicionado	Adj. Adverbial	PB 42% PBSP 51% SP 54%
SIQUEIRA (2020)	Amostra Deslocamentos 2019 (falantes de Sergipe e de outros estados).	Sintagma preposicionado	Não controlou	(599/1.268) 47%
SILVA (2020)	Amostra Deslocamentos 2020 (falantes de Sergipe, Bahia e Alagoas).	Sintagma preposicionado	Genitivo	1.017/2.326 (44%)

Fonte: elaboração própria.

Pelos dados acima, percebemos que o sintagma preposicionado foi uma variável significativa para o preenchimento do determinante antes de possessivo pré-nominal, sendo as preposições aglutinadoras as que mais favoreceram a ocorrência do preenchimento em todos os estudos. A função sintática não foi controlada pelos estudos de Campos Jr. (2012) e Siqueira (2020). Essa variável, contudo, foi significante para todos os outros estudos apesar de não haver consenso sobre o tipo de função que foi maior favorecedora do preenchimento.

Não reportamos as variáveis extralinguísticas e semânticas descritas nos estudos, pois o nosso foco é nas características morfossintáticas do fenômeno. Salientamos, porém, que nos códigos criados, para o trabalho com as amostras, há a descrição de como organizar um *dataframe* (dados em formato de tabela) para que essas outras variáveis possam constar da análise do pesquisador.

3.2.4 Os critérios para avaliar os usos das ferramentas para tarefas sociolinguísticas

3.2.4.1 A classificação das etiquetas

Quando se avalia uma etiquetagem linguística, normalmente a métrica se faz em relação à precisão ou acurácia, ou a concordância entre anotadores, quando a validação é feita por humanos. Baker (2013) aponta que existem pontos de vista nos quais assume-se que é desnecessário um humano corrigir uma anotação automática, dentre esses pontos de vista o autor enumera o de Sinclair (1992) e de Church (1996). Para o primeiro, conforme Baker (2013), não deixar ambiguidades que são intrínsecas às línguas naturais obscurece o estado atual das gramáticas das línguas. Já o segundo acredita que sempre há cinco por cento de palavras em que não haverá concordância entre anotadores, uma vez que o autor assume que esse fato é intrinsecamente relacionado à indeterminação e desordenação das línguas humanas. Baker (2013) também evidencia uma outra hipótese que dialoga com as posições contrárias a uma revisão humana da anotação automática, na qual a revisão da anotação por um humano pode diminuir a consistência interna da etiquetagem, uma vez que o computador não foge à sua regra interna, mas o humano pode cometer erros por desatenção ou apatia, trazendo inconsistência à etiquetagem feita por meio de anotação automática.

Embora tenha apresentado argumentos de posição contrária, Baker (2013) acredita que apesar das inconsistências advindas da natureza humana em oposição à consistência dos etiquetadores automáticos, somente o humano pode atestar a qualidade da etiquetagem. Para

provar essa hipótese, o pesquisador conduziu um experimento para verificar se a edição humana pós-etiquetagem aumentaria a inconsistência da anotação. Os resultados refutaram esse argumento, mostrando que a consistência dos anotadores foi maior do que a do computador sozinho, o que também é argumentado por Leech (2004). Para este pesquisador, embora a junção da etiquetagem automática com a correção humana não alcance o ideal de 100% de acurácia, os 99,5% de acurácia já alcançados nessa junção são ainda melhores do que 96% ou 97% de acurácia automática.

Para Palmer e Xue (2010), os seguintes aspectos devem ser considerados ao avaliar uma anotação: a validade e consistência e a concordância entre anotadores. Assim como para Carletta (1996) e Leech (2004), os autores acreditam que o coeficiente *kappa*, *k*, oferece uma avaliação com maior grau de acurácia. Essa técnica, que também pode ser aplicada em outras áreas, como na sociolinguística em estudos de percepção (FREITAG, 2019), consiste em subtrair a concordância esperada (E) entre os anotadores da concordância observada (A), em seguida dividindo isso por 1 menos a concordância esperada (E), como na fórmula:

$$(1)\ k = A - E \div 1 - E$$

Outro padrão para se realizar o cálculo de acurácia de um etiquetador é feito por meio da razão entre o número de etiquetas corretamente empregado pelo número de etiquetas total empregado no corpus. Contudo, dentro desse mesmo padrão, Paroubeck (2007) faz algumas ressalvas: que haja um *corpus* de referência, que o tipo de segmentação do novo *corpus* seja o mesmo do *corpus* de referência, e, por fim, que o conjunto de etiquetas seja o mesmo empregado no *corpus* de referência. Como estamos no processo de elaborar procedimentos para a criação de um *corpus* de referência, fazer o cálculo como sugerido por Paroubek (2007) não é possível. Além dos fatores acima descritos, devemos nos atentar para o fato de que a caracterização das funções desempenhadas por uma forma linguística depende, também, da avaliação subjetiva do pesquisador, principalmente no caso dos nossos dados em que a situação comunicativa vai se criando à medida em que a entrevista acontece. Assim, para tornar o nosso protocolo replicável, implementamos uma avaliação da etiquetagem de ferramentas por meio da criação de um modelo para identificação dos erros e acertos contidos nas etiquetagens em relação ao fenômeno do preenchimento de determinante antes de possessivo pré-nominal com base nos estudos de Callou e Silva, (1997), Campos Jr. (2012), Sedrins, Pereira e Silva (2019), Silva (2020) e Siqueira, (2020).

Foram consideradas para análise todas as etiquetas empregadas às palavras imediatamente anteriores ao possessivo, bem como aos itens marcados como possessivos. Na

tabela abaixo, apresentamos como categorizamos os erros para os dois etiquetadores, porém utilizamos exemplos apenas do *LancsBox* 6.0:

Tabela 4 – Modelo para identificação de etiquetas (in)corretas.

Tipo	Explicação	Exemplo
ERR	Erro do etiquetador	"do_SPS", deveria receber etiqueta SP+DA
ETRANS	Erro ocasionado pela transcrição	"ten-_VMI", não deveria receber etiqueta, trata-se de um truncamento
NULL	Não há erro	"minha_DP1", etiqueta correta

Fonte: elaboração própria.

Na tabela, visualizamos como "ERR" o erro em que o etiquetador possui a etiqueta para designar uma determinada forma, mas não a emprega. Por exemplo, o etiquetador do *LancsBox* 6.0 possui a etiqueta "SP+DA" que indica a junção de uma preposição "SP" mais um determinante artigo "DA". Essa ferramenta foi categórica ao fazer isso para a junção da preposição "em" + artigo definido "o(a)" ou "os(as)", mas não fez o mesmo para a junção da preposição "de" + artigo definido "os(as)". Interjeições como "nossa", "meu Deus", não foram marcadas com a etiqueta "I", mas como "DP1", determinantes possessivos de primeira pessoa, no caso de "nossa" e "meu" e "N" para "Deus". Determinantes possessivos de segunda pessoa "DP2", que foram marcados como de terceira, como em "você e seu pai", em que "seu" foi marcado como "DP3", quando, na verdade, deveria ser "DP2" pelo *LancsBox* 6.0.

"ETRANS" foi utilizado para marcar erros ocasionados pela transcrição, seja por erros ortográficos, como "seu", em que o falante disse "se eu", levando o etiquetador a marcar "DP3" devido a transcrição, seja por traços característicos da fala, como no uso de marcadores discursivos "né" e "tipo", e em expressões multipalavras, como em "minha filha", quando não há relação de posse entre "minha" e "filha", sendo uma expressão multipalavra, mas os itens foram marcados pelos dois etiquetadores como possessivo e nome comum respectivamente. Outro tipo de contexto marcado como "ETRANS" é o truncamento, como em "i-minha" que foi anotado pelo etiquetador do *spaCy* 3.5 como "PROPN", ou seja, um nome próprio, quando, na verdade, é um "DET", determinante possessivo, erro que pode ter sido provocado pelo tipo da transcrição.

"NULL" é a marcação utilizada para indicar que não houve erro por parte dos etiquetadores, ou seja, houve concordância entre revisor e anotador automático. No caso de "na

minha sala", os itens "na" e "minha" foram etiquetados pelo *spaCy* 3.5 como "ADP" e "DET", respectivamente, e pelo *LancsBox* 6.0 como "SP+DA" e "DP1", respectivamente, indicando no caso do *spaCy* 3.5 uma classificação mais genérica, "preposição" e "determinante", e mais específica no caso do *LancsBox* 6.0, "preposição + artigo" e "determinante possessivo de primeira pessoa".

A análise da classificação dos dados das etiquetas em termos de frequência está descrita abaixo e foi feita utilizando o pacote data table (DAWLE *et. al*, 2022) na IDE RStudio da linguagem R. Analisamos os contextos das entrevistas com limpeza e sem limpeza para cada um dos etiquetadores em relação ao contexto anterior aos possessivos e também em relação aos possessivos.

Sobre os dados do LancsBox 6.0, ficaram fora da nossa análise os dados marcados como PX (Pronome Possessivo) na amostra, uma vez que as buscas automáticas por meio das etiquetas não os detectaram por não estarem etiquetados como DP. As entrevistas sem limpeza tiveram 1.136 dados marcados como ERR, 168 de ETRANS e 3.680 dados marcados como NULL, em um total de 4.984 etiquetas analisadas, indicando que o etiquetador acertou mais vezes que errou. A classificação também nos permitiu verificar que a etiqueta mais frequente com o possessivo para os dados sem limpeza é a preposição[52] simples ("SPS", 1.149/4.984).

Em comparação com entrevistas com limpeza, em números absolutos, houve uma queda na quantidade de ETRANS (97), aumento de ERR (1.172) e de NULL (3.704). Apesar do aumento de ERR, o etiquetador melhorou sua quantidade de acertos, mostrando que ele ainda acertou mais que errou. A etiqueta SPS também foi a etiqueta que mais acompanhou o possessivo (1.150/4.973).

Nas entrevistas sem limpeza os dados dos erros para os possessivos foram 1.896 (ERR), 7 (ETRANS) e 3.077 (NULL) e nas limpas, 1.891 (ERR), 8 (ETRANS) e 3.072 (NULL). O aumento na classificação do erro e a diminuição do acerto, em números absolutos, podem ser consequência do tamanho da amostra. As ocorrências mais frequentes nos dados dos possessivos foram de "DP1", 3.188/4.980, para entrevistas sem limpeza e 3.183/4.971, para entrevistas com limpeza. Esse dado pode ser explicado por se tratar de uma entrevista, logo, as ocorrências de possuidor de primeira pessoa tendem a ser maiores. Além disso, interjeições como "Meu Deus!", "Nossa!", "Meu!", costumeiras no discurso falado, são também marcadas como "DP1".

[52] Como nos dados não foram encontrados outros tipos de adposição, durante a análise, nos referimos a esta etiqueta somente como preposição.

Dos 5.209 dados resultantes da busca pelo fenômeno e dos ajustes [53]na amostra, os dados para a classificação dos erros das etiquetas do contexto anterior do *spaCy* 3.5 foram: 155 itens marcados como ERR, 170 como ETRANS e 4.884 como NULL. Havendo queda dos valores de ERR (117), ETRANS (102) e aumento de NULL (4.983) em 5.202 dados limpos. Nos dois tipos de dados, o etiquetador teve alto índice de etiquetas que classificamos como corretas. A etiqueta mais frequente no contexto anterior ao possessivo foi "ADP" para os dois conjuntos de dados, com valores de 1.789/5.209 e 1810/5.202 para entrevistas sem e com limpeza, respectivamente.

Como os dados dos possessivos para o *spaCy* 3.5 só possuem uma etiqueta "DET", comparamos os resultados da classificação dos erros apenas para essa etiqueta. Ter apenas essa etiqueta sem maiores informações morfológicas favoreceu o acerto do etiquetador, pois, ao não fazermos uma busca considerando a flexão de pessoa, ele não erra. Mas, ao buscar por possessivos de 2P, o etiquetador não considera "seu" e suas flexões como de 2P, retornando apenas resultados concernentes à pessoa gramatical da gramática normativa. Nas entrevistas com dados sem limpeza, 5.010 de 5.209 ocorrências de "DET" foram classificadas como NULL, já para os dados limpos 5.018 de 5.208 ocorrências, evidenciando, assim como para os dados do contexto anterior, um nível alto de etiquetas marcadas como corretas. ETRANS diminuiu seus valores assim como ERR. Os valores para ETRANS e ERR foram de 19 e 180, respectivamente em 5.209 dados sem limpeza e de 17 e 173, ETRANS e ERR, respectivamente, para dados em 5.208 dados limpos.

O quadro 5, representa a comparação das duas ferramentas em relação às classificações das etiquetas e de suas frequências.

[53]Para realizar a avaliação do modelo de classificação dos erros, tivemos que amalgamar algumas classes para evitar erros nos resultados. No entanto, para a contagem dos erros foram utilizados os dados completos.

Quadro 5 – Resumo dos resultados acerca das etiquetas do contexto anterior ao possessivo para as ferramentas *LancsBox* 6.0 e *spaCy* 3.5.

Conjunto de dados	Etiquetadores					
	LancsBox 6.0			*SpaCy* 3.5		
	Item	**Dados sujos**	**Dados limpos**	**Item**	**Dados sujos**	**Dados limpos**
Dados do contexto anterior	ERR	1.136 22,7%	1.172 23%	ERR	155 2,9%	117 2,2%
	ETRANS	168 3%	97 1,09%	ETRANS	170 3,2%	102 1,9%
	NULL	3.680 73,8%	3.704 74,4%	NULL	4.884 93,7%	4.983 95,7%
	OCOR	4.984	4.973	OCOR	5.209	5.202
	Mais frequente	SPS 1.149 23%	SPS 1.150 23%	Mais frequente	ADP 1.767 33,9%	ADP 1.810 34,7%

Fonte: elaboração própria.

Após o ajuste nos dados, no contexto anterior, o *spaCy3*.5 teve resultados melhores comparando-se com o *LancsBox* 6.0, tanto para dados sujos (93,7% em relação a 73,8% do *LancsBox* 6.0) como limpos (95,7% em relação a 74,4% do *Lancs Box* 6.0). O conjunto de dados total evidencia que a limpeza nas entrevistas diminui os erros ocasionados por marcas orais ou do contexto comunicativo da entrevista (ETRANS), trazendo implicações importantes para o modelo de transcrição que nós adotamos. Em todos os dados, a frequência de preposições com os possessivos foi maior do que outras classes de palavras, indicando um padrão relevante do comportamento dos determinantes nos dados das entrevistas da Amostra Deslocamentos 2019. É importante ressaltar, contudo, que a porcentagem maior de dados de preposições antecedendo o possessivo foi maior em proporção para o *spaCy* 3.5 devido à ADP incluir tanto preposições aglutinadas como "em" + "artigo", quanto não aglutinadas. Já no caso do *LancsBox* 6.0, o número foi relativamente menor porque SPS foi a etiqueta usada para preposições não aglutinadas e, curiosamente, para "de+artigo".

Os dados dos possessivos seguem no quadro 6.

Quadro 6 – Resumo dos resultados acerca das etiquetas dos possessivos para as ferramentas *LancsBox* 6.0 e *spaCy* 3.5.

Conjunto de dados	Etiquetadores					
	LancsBox 6.0			*SpaCy* 3.5		
	Item	**Dados sujos**	**Dados limpos**	**Item**	**Dados sujos**	**Dados limpos**
Dados dos possessivos	ERR	1.896 38%	1.891 38%	ERR	180 3,4%	173 3,3%
	ETRANS	7 0,14%	8 0,16%	ETRANS	19 0,3%	17 0,3%
	NULL	3.077 61,7%	3.072 61,7%	NULL	5.010 96%	5018 96%
	OCOR	4.980	4.971	OCOR	5.209	5.208
	Mais frequente	DP1 64%	DP1 64%	Mais frequente	-	-

Fonte: elaboração própria.

Os dados revelam que, no contexto de possessivos, a classificação do *spaCy* 3.5 também foi melhor que a do *LancsBox* 6.0, considerando a quantidade de etiquetas marcadas como NULL, ressaltando que o *spaCy* 3.5 não agrupa as características morfológicas dentro de uma mesma etiqueta. O *LancsBox* 6.0, por outro lado, agrupa características POS e morfológicas em uma mesma etiqueta, logo, grande parte em que o possessivo "seu" e suas flexões apareciam no contexto de segunda pessoa, marcamos como erro porque a etiqueta empregada para todos os contextos desse possessivo foi "DP3", ou seja, determinante possessivo de terceira pessoa.

O *LancsBox* 6.0 manteve a proporção de etiquetas marcadas como ETRANS e ERR e NULL. Comportamento semelhante foi observado para o *spaCy* 3.5. No entanto, em relação à etiqueta mais frequente, que foi "DP1" para o *LancsBox* 6.0, não pode ser comparada entre os dois etiquetadores devido ao fato de o *spaCy*3.5 qualquer possessivo, seja de primeira ou segunda pessoa, a etiqueta POS é a mesma.

Os resultados dessa análise trouxeram implicações importantes para a forma como iremos proceder em relação à transcrição dos áudios das entrevistas coletadas. Como reportado, as entrevistas com limpeza obtiveram resultados melhores para o contexto anterior do que as entrevistas sem limpeza. Embora as marcações referentes ao contexto sejam importantes para

diferentes tipos de pesquisa, adotamos o padrão de descrever essas situações em uma trilha separada no ELAN. Além disso, observamos que, quando a marca de truncamento "-"fica mais próxima da palavra cortada, os etiquetadores lidam melhor com elas, conseguindo até mesmo classificá-las corretamente. Por fim, a partir da análise, percebemos a necessidade de implementar outra revisão nas transcrições, tanto de ortografia quanto para adequação às adaptações introduzidas na norma original. Devemos nos atentar para o fato de que as ferramentas computacionais para o processamento de linguagem são feitas com base em formalismos, tais quais a morfologia inflexional, as gramáticas de dependências e constituintes dentre outros, como vimos no capítulo anterior. Já os dados linguísticos utilizados em pesquisa sociolinguística, por outro lado, são ricos em variabilidade. Nossa proposta de pesquisa, portanto, busca oferecer um caminho para superar essa limitação entre formalismos e dados variáveis, otimizando os recursos disponíveis e gratuitos para o processamento da língua portuguesa, já que é do interesse da pesquisa em Sociolinguística trabalhar com a descrição da língua que é efetivamente utilizada pelos falantes.

Para analisar se a identificação dos erros e acertos dos etiquetadores foi consistente e tornar nosso modelo replicável, utilizamos Florestas Aleatórias (WITTEN; FRANK; HALL, 2011) que é uma técnica de aprendizagem de máquinas (doravante AM) supervisionada de agrupamento (*ensemble*). As Florestas Aleatórias se baseiam em árvores de decisão individuais para prever uma classificação (como em nosso trabalho) ou uma regressão. Para implementar o algoritmo, utilizamos a linguagem de programação R, por meio da IDE RStudio e as bibliotecas *tidyverse* WICHKHAM (2022), *partykit* (HOTHORN; SEIBOLD; ZEILEIS, 2022), *data table* (DAWLE *et. al*, 2022), *caret* (KUN *et. al*, 2022) e *random forest* (BREIMAN; CUTLER; LIAW; WIENER, 2022). Fizemos esse procedimento tanto com os dados de entrevistas com limpeza quanto com dados sem limpeza. Esse procedimento está descrito na próxima seção. O conjunto de dados para a avaliação, assim como os códigos estão disponíveis no repositório desta tese (SOUSA *et. al,* 2022).

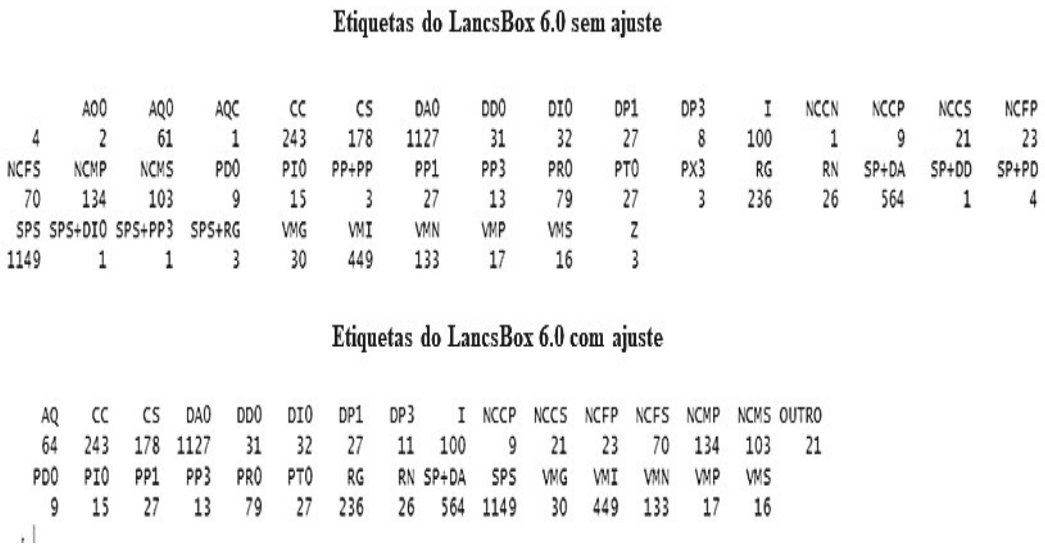
3.2.4.2 A avaliação do modelo de classificação das etiquetas para o LancsBox 6.0

A busca pelo fenômeno de determinantes antes de possessivo pré-nominais em entrevistas sem limpeza, retornou um quantitativo de 4.984 dados de determinantes possessivos e 4.980 dados de diferentes classes de palavras do contexto anterior, perfazendo um total de

9.964 dados, em entrevistas sem limpeza. As entrevistas com limpeza, por outro lado, retornaram 4.976 dados de possessivos e 4.972 de classes diferentes.

Como relatado na descrição das ferramentas, o *LancsBox* 6.0 possui um conjunto de 11 etiquetas[54]: adjetivos (A), advérbios (R), Determinantes (D), substantivos (N), verbos (V), pronomes (P), conjunções (C), interjeições (I), Preposições (S), sinais de pontuação (F), cifras e numerais (Z). Essas etiquetas se desmembram em etiquetas mais refinadas por meio das características morfológicas e gramaticais das palavras. A classe de determinantes, por exemplo, é marcada por 3 categorias: determinante, tipo e pessoa. Assim, a palavra "meu" é uma palavra etiquetada como DP1, que significa Determinante Possessivo de Primeira Pessoa. Devido a esse fato, ao montar o *dataframe* dos dados do *LancsBox* 6.0, as variáveis com as etiquetas do contexto anterior ficaram com mais de 40 níveis, sendo necessário amalgamar algumas classes do contexto anterior que tiveram poucas ocorrências e também aquelas de contexto vazio, ou seja, sem palavra que antecedesse o possessivo, para que não houvesse sobreajuste (*overfitting)* e nem aumento na complexidade no modelo (WITTEN; FRANK; HALL, 2011; FREITAG, 2021a). Para ilustrar, na figura 22, temos os dados do contexto anterior ao possessivo antes e após o ajuste.

Figura 22 – Ajuste dos dados do contexto anterior ao possessivo das etiquetas do *LancsBox*.



Fonte: elaboração própria.

[54]O conjunto dessas etiquetas está disponível em: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Portuguese-Tagset.html

Nos dados sem ajuste há a representação dos níveis vazios e níveis com frequência menor que cinco[55]. As etiquetas AO0, AQ0 e AQC, adjetivo ordinal de grau zero, adjetivo qualificador de grau 0 e adjetivo qualificador de grau diminutivo respectivamente, em AQ, conjunto de adjetivos. Os níveis DP3 e PX3, determinante possessivo de 3P e pronome possessivo respectivamente, como DP3, uma vez que entendemos que os possessivos no português brasileiro atuam como especificadores do sintagma nominal (CASTILHO, 2010). As categorias que possuíam frequência menor que 5 e não se correlacionavam em termos estruturais e de suas funções, amalgamamos na categoria OUTRO, são elas: as categorias vazias, os numerais (Z), substantivo comum de gênero comum e número invariável (NCCN), junção de preposição e determinante indefinido sem pessoa (SPS+DI0), junção preposição e pronome pessoal de terceira pessoa, preposição mais advérbio (SPS+RG), junção de preposição e preposição (PP+PP), preposição mais determinante definido (SP+DD), preposição mais pronome definido (SP+PD).

O mesmo procedimento foi realizado para as etiquetas revisadas e consideradas como mais adequadas para as palavras etiquetadas, conforme figura 23.

Figura 23 – Contexto anterior ao possessivo com etiquetas revisadas do *LancsBox* antes do ajuste e após o ajuste dos dados.

Etiquetas revisadas sem ajuste dos dados

	AQ0	AQC	CC	CONTEXT	CS	DA0	DD0
4	39	1	243	14	191	1126	30
DI0	DP1	DP2	DP3	I	MD	MWE	NCCP
16	27	10	2	35	41	5	9
NCCS	NCFP	NCFS	NCMP	NCMS	NPFP	NPFS	NPMP
24	20	66	74	93	2	6	4
NPMS	PAUSA	PD0	PI0	PIO	PP1	PP2	PP3
19	71	10	14	1	24	6	5
PR0	PT0	RG	RN	SP+DA	SP+DD	SP+PD	SPS
30	65	248	26	1332	3	5	410
SPS+DI0	SPS+PP3	SPS+RG	TRUNCAMENTO	VMG	VMI	VMM	VMN
1	1	3	26	31	411	6	131
VMP	VMS	Z					
12	8	3					

Etiquetas revisadas com ajuste dos dados

AQ	CC	CS	DA0	DD0	DI0	DISC	DP1	DP2	I	NCCP
43	242	191	1125	30	16	132	27	10	35	9
NCCS	NCFP	NCFS	NCMP	NCMS	NPFMP	NPFS	NPMS	OUTRO	PD0	PI0
24	20	66	74	93	6	6	19	22	10	15
PP1	PP2	PP3	PR0	PT0	RG	RN	SP+DA	SPS	TRUNCAMENTO	VMG
24	6	5	30	65	248	26	1332	409	26	31
VMI	VMM	VMN	VMP	VMS						
409	6	131	12	9						

Fonte: elaboração própria.

[55] Definimos esta frequência por ser aquela em que os dados já não são mais possíveis de serem pareados nos conjuntos de dados sujos e limpos.

Diferentemente do etiquetador, ao revisarmos as etiquetas, percebemos que havia informações contextuais que foram classificadas pelo *LancsBox* 6.0 de maneira inadequada, separamos então, informações contextuais, truncamentos, expressões multipalavras (MWE), pausas, o que acarretou aumento dos níveis. Logo, fizemos o mesmo procedimento de amalgamar os dados em classes maiores para diminuir o número de níveis. Na categoria DISC foram amalgamadas informações sobre o contexto (CONTEXTO), marcadores discursivos (MD), expressões multipalavras (MWE) e pausas (PAUSAS). Na categoria OUTRO, foram agrupadas as etiquetas: DP3, Z, SPS+ DI0, SP+DD, SP+PD e SPS+PP3.

Não foi necessário agrupar os níveis da variável etiquetas de determinantes possessivos em nenhuma das análises porque ela já possui poucos níveis, porém, eliminamos um único contexto em que houve erro de transcrição e o etiquetador marcou uma conjunção subordinativa como DP3, porque foi digitado "seu", quando deveria ser "se eu".

Feito o ajuste para os dados do *LancsBox* 6.0, avaliamos o modelo com base na estimativa do erro OOB (*out-of-bag,* estimativa do erro fora da reamostragem agregada), ou seja, em cada reamostragem para se criar uma árvore de decisão, alguns fatores não entram na amostragem, logo, o OOB é a previsão do erro dos dados não utilizados (BREIMAN, 1996; WITTEN; FRANK; HALL, 2011; FREITAG, 2021a). No caso dos dados do contexto anterior ao possessivo, em entrevistas sem limpeza, a previsão do erro dos dados de teste foi igual a OOB 4.56%, o que nos indica que a previsão de acurácia do modelo para os dados de teste é de 95,44%. Para os dados do contexto anterior de entrevistas limpas, o mesmo índice foi de 4,98%, uma variação baixa de 0,42%, e uma previsão de acurácia de 95,02% para os dados de teste. O algoritmo oferece ainda a taxa do erro por classe, representados na tabela 5.

Tabela 5 - Taxa de erros por classe do contexto anterior ao possessivo para o *LancsBox*.

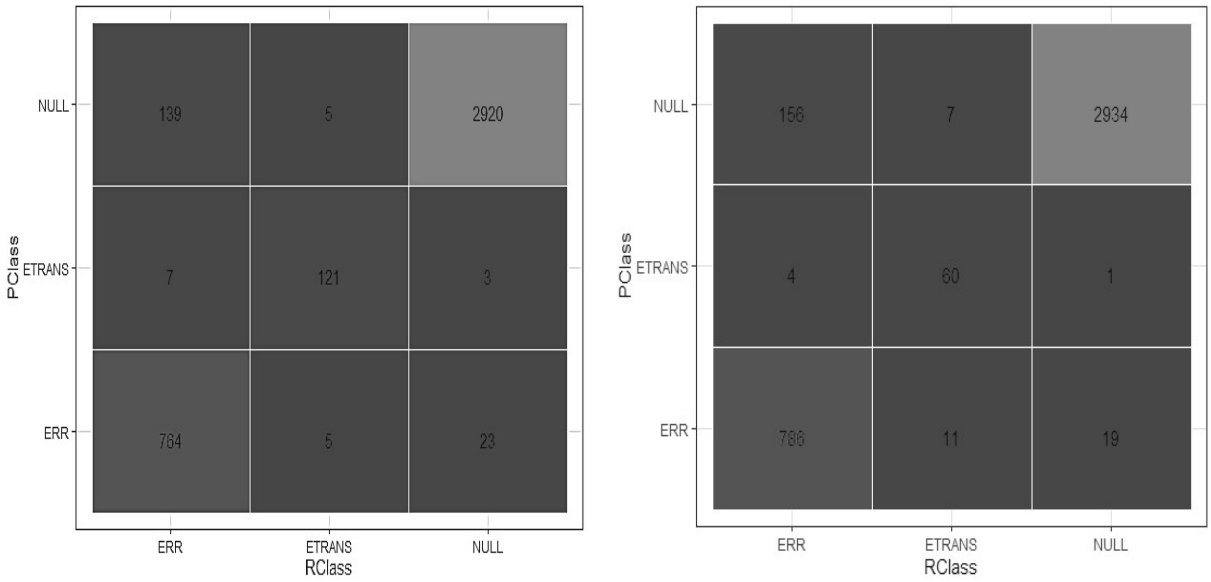
Tipos de erro	Entrevistas sem limpeza	Entrevistas com limpeza
ERR	0,160	0,156
ETRANS	0,076	0,060
NULL	0,008	0,006

Fonte: elaboração própria.

Pela tabela 5, tanto para entrevistas sem limpeza à esquerda e entrevistas com limpeza à direita, foi mais difícil classificar os erros do próprio etiquetador (ERR), com valores da média estimada do erro de classificação para 0,160 dados sem limpeza e 0,156 para dados com limpeza, em comparação com os índices dos outros tipos de erro.

Com os dados das matrizes de confusão dos contextos anteriores das entrevistas sujas e das entrevistas limpas, figura 24, podemos observar a quantidade de vezes em que cada etiqueta foi corretamente classificada. Por exemplo, o classificador classificou ERR como ERR 764 vezes, e ERR como ETRANS 5 vezes e como NULL 23 vezes, em entrevistas sem limpeza.

Figura 24 – Matrizes de confusão da classificação dos erros do contexto anterior ao possessivo em entrevistas sem e com limpeza do *LancsBox*.

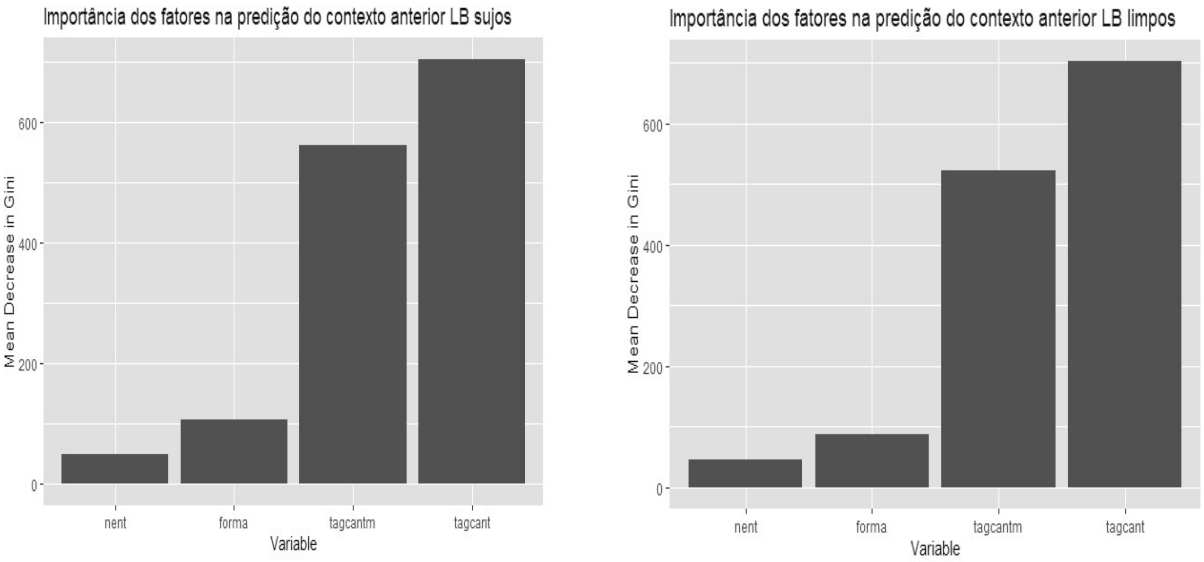


Fonte: elaboração própria.

Os dados revelam também que houve uma diminuição da quantidade de etiquetas referente ao erro ocasionado pela transcrição (ETRANS), de 131 para 85, o que pode ter acontecido pelo fato de que as entrevistas com limpeza perderam as marcas contextuais e de oralidade presentes no texto e também pelo fato de que as entrevistas com limpeza geraram menos dados do que as sem limpeza. A diminuição da amostra também pode ter influenciado o valor maior para o OOB nos dados limpos.

Além da matriz de confusão e do OOB, o algoritmo de árvores aleatórias nos permite visualizar qual fator mais influenciou na classificação, figura 25.

Figura 25 – Importância das variáveis na predição dos erros do contexto anterior ao possessivo nas entrevistas sem limpeza e entrevistas com limpeza do *LancsBox*.



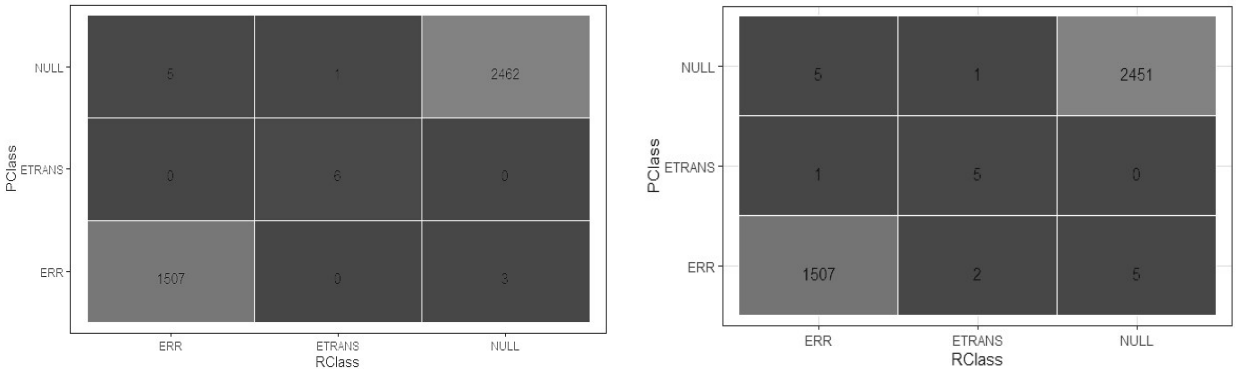
Fonte: elaboração própria.

Na representação gráfica acima, consta o índice *Mean Decrease in Gini*. Esse índice considera o valor de decrescimento do Gini (medida da diversidade de classes em um nó da árvore) a cada repartição da árvore causada por uma variável (BREIMAN, 2004; FREITAG, 2021a). Quanto maior o valor do decrescimento do desempenho da árvore em decorrência de uma variável, maior é a importância dessa variável para o modelo. Em ambos os casos do contexto anterior, seja de entrevistas com ou sem limpeza, a variável que teve maior importância na classificação foi a etiqueta empregada pelo *LancsBox* 6.0 (tagcant). A variável que teve menor importância foi a entrevista (nent), ou seja, o decrescimento do Gini foi baixo. A forma, no caso padrão (sem o preenchimento do determinante) e não-padrão (com o preenchimento), exerceu importância relativamente maior que nent, quando selecionada para fazer a repartição dos dados, sendo a variável com a segunda maior importância.

As ocorrências dos possessivos, por outro lado, tiveram melhores resultados em relação ao desempenho da classificação. No contexto de entrevistas sem limpeza, o OOB foi de 0,23%, o que demonstra uma previsão de acurácia de 99,77%, e, para entrevistas com limpeza, o índice OOB foi de 0,35%, uma previsão de acurácia de 99,65%. Observamos que a diferença entre os índices OOB do contexto anterior e dos determinantes pode ser explicada pela diminuição dos níveis da variável.

A matriz de confusão para entrevistas sem limpeza está representada à esquerda e, a com limpeza, à direita na figura 26.

Figura 26 – Matrizes de confusão dos dados da classificação dos determinantes possessivos com e sem limpeza do *LancsBox*.



Fonte: elaboração própria.

Pelas matrizes de confusão, percebemos um aumento no número de ocorrências de ETRANS, aumento nas ocorrências de ERR 1.513 e diminuição nas quantidades de NULL 2.456 nas entrevistas com limpeza (à direita). Como para o contexto anterior, a diferença na performance pode ser explicada pelo tamanho da amostra, que foi menor para as entrevistas limpas e também pelo fato de que expressões multipalavras e interjeições possuem mais ocorrências nos dados com possessivos, aumentando o erro ETRANS. Embora observemos graficamente que, nas entrevistas, a classificação foi errada em apenas uma vez para ETRANS nas entrevistas sujas e 3 vezes nas entrevistas com limpeza o erro dessa classificação foi maior, conforme tabela 6.

Tabela 6 - Taxa de erros por classe dos possessivos do *LancsBox*.

Tipos de erro	Entrevistas sem limpeza	Entrevistas com limpeza
ERR	0,003	0,003
ETRANS	0,142	0,375
NULL	0,001	0,002

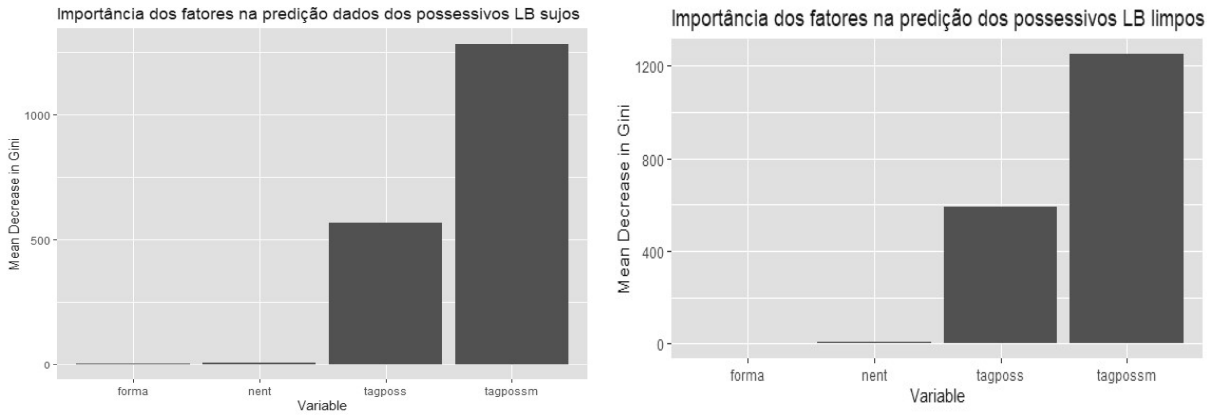
Fonte: elaboração própria.

Como observamos, as taxas dos erros de classificação para os possessivos foram menores que as taxas da classificação do contexto anterior no que se refere à classe ERR e NULL, valor de 0,003 tanto para entrevistas sem limpeza e com limpeza. Esse fato pode ter sido ocasionado pela diminuição da diversidade do número de itens classificados para os dados de determinantes, diminuindo a complexidade do modelo. Outro dado interessante é o aumento da taxa de erro para a classificação ETRANS, 0,142 e 0,375, para entrevistas sem limpeza e com limpeza respectivamente, em comparação aos dados do contexto anterior 0,076 e 0,060 para entrevistas com e sem limpeza do contexto anterior. Esse aumento na taxa pode ser

explicado devido ao desbalanceamento dos dados dos possessivos, sendo, portanto, o item mais difícil de classificar no caso do conjunto de dados de determinantes.

Em relação ao índice *Mean Decrease in Gini*, os dados dos possessivos apresentaram comportamento diferente do contexto anterior como observado pela figura 27.

Figura 27 – Importância dos fatores na predição dos erros dos determinantes possessivos do *LancsBox* 6.0.



Fonte: Elaboração própria.

A variável mais significante na predição do modelo foram as etiquetas corrigidas (tagpossm). Em segundo lugar as etiquetas do etiquetador (tagposs). Nas entrevistas sem limpeza a importância da forma e de nent foram próximas de zero. Já para entrevistas com limpeza, forma teve valor 0 e, nent, próximo de zero.

3.2.4.3 A classificação das etiquetas para o *spaCy* 3.5

No caso do *spaCy* 3.5, ao realizarmos a busca pelo fenômeno de determinantes antes de possessivo pré-nominais em entrevistas sem limpeza, obtivemos um resultado de 5.210 dados de determinantes possessivos e 5.210 dados do contexto anterior, totalizando 10.420 dados para a ferramenta. Com as entrevistas limpas, foram retornados 5.208 dados de determinantes possessivos e 5.208 dados do contexto anterior, gerando um total de 10.416 dados.

Em relação ao contexto anterior ao possessivo dos dados do *spaCy* 3.5 não amalgamamos nenhum fator por dois motivos: o conjunto de fatores já ser pequeno em comparação aos dados do *LancsBox* 6.0 e porque apenas um fator que teve frequência menor que 5 na variável etiquetas do etiquetador (tag1) teve frequência maior na variável etiquetas revisadas (tag2) e vice-versa, como ilustrado na figura 28.

Figura 28 – Etiquetas do contexto anterior ao possessivo do *spaCy*.

Etiquetas do contexto anterior

ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ
49	1789	282	196	244	1246	4	369	7	133	75	196	156
SPACE	VERB											
35	428											

Etiquetas do contexto anterior revisadas

ADJ	ADP	ADV	AUX	CCONJ	DET
37	1775	263	190	250	1236
INTJ	MD	MWE	NOUN	NUM	PAUSA
33	28	13	320	3	73
PRON	PROPN	PUNCT	SCONJ	SPACE	TRUNCAMENTO
138	30	207	194	11	44
VERB					
364					

Fonte: elaboração própria.

Como pode ser visto na figura 28, nas etiquetas do contexto anterior do etiquetador, INTJ é a etiqueta com menor frequência (4), porém, sua frequência aumenta nos dados revisados (33). O mesmo acontece nos dados revisados, nos quais a frequência de NUM é 3, e nos dados do etiquetador é 7. Os dados das entrevistas limpas seguem na figura 29.

Figura 29 – Etiquetas revisadas do contexto anterior ao possessivo do *spaCy*.

Etiquetas do contexto anterior em entrevistas limpas

ADJ	ADP	ADV	AUX	CCONJ	DET	NOUN	PRON	PROPN	PUNCT	SCONJ	VERB
56	1810	321	204	248	1251	403	151	61	160	158	379

Etiquetas do contexto anterior revisadas em entrevistas limpas

ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	MD
49	1797	302	201	247	1219	26	31
MWE	NOUN	PRON	PROPN	PUNCT	SCONJ	TRUNCAMENTO	VERB
12	360	148	33	161	200	45	371

Fonte: elaboração própria.

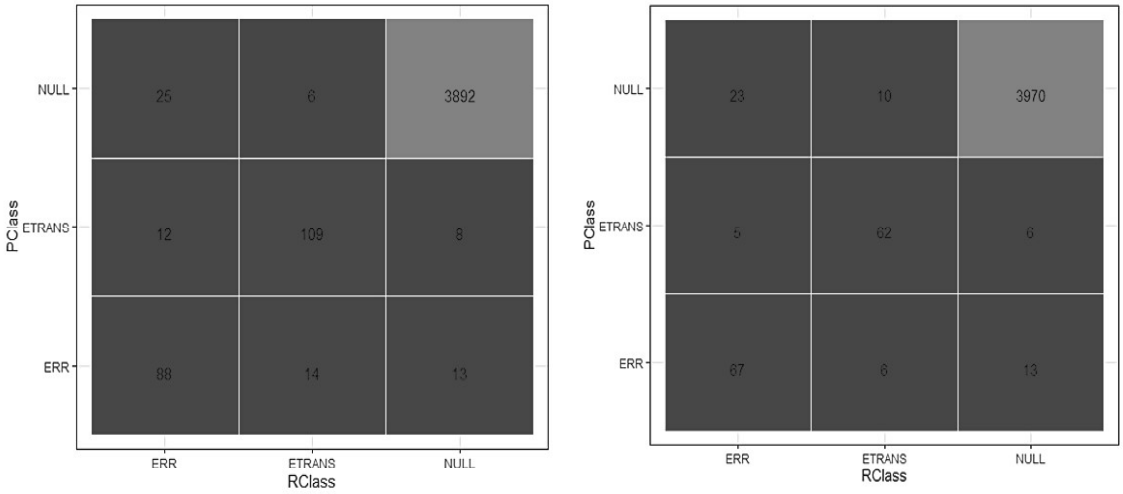
Pela figura 29, observamos que, nos dados das etiquetas do etiquetador, o fator que tem menor frequência é a etiqueta ADJ com 50 ocorrências. Já para as etiquetas revisadas, o fator que tem menor frequência é a etiqueta MWE, como 12 ocorrências, ambos fatores com frequência acima de 5. Além disso, o número de fatores é baixo.

Em relação à avaliação das etiquetas do *spaCy* 3.5, a taxa do OOB foi de 1,87% para entrevistas sujas e 1,51% para entrevistas limpas. A estimativa da acurácia para os dados de teste é, então, 98,13% e 98,49% para os dados das entrevistas sujas e limpas, respectivamente. Mostrando que o desempenho do modelo, diferentemente do resultado para o *LancsBox* 6.0, foi melhor para entrevistas com limpeza.

A figura 30 apresenta as matrizes de confusão para os dados do contexto anterior do *spaCy* 3.5 para entrevistas sem limpeza à esquerda e com limpeza à direita, facilitando a

visualização do número de vezes em que a classificação para cada variável foi correta e incorreta.

Figura 30 – Matrizes de confusão para os dados do contexto anterior ao possessivo para o *spaCy*.



Fonte: elaboração própria.

As matrizes nos mostram que houve um aumento da classificação NULL de 3.913 nas entrevistas sem limpeza para 3.989 para entrevistas com limpeza, além de uma diminuição de ERR (de 125 para 95) e ETRANS (de 109 para 78) para entrevistas com limpeza. Isso pode ser explicado devido ao fato de haver mais etiquetas corretas nas entrevistas com limpeza. As matrizes também nos possibilitam ver que a classificação de NULL como NULL foi mais fácil, enquanto a de ERR como ERR foi mais difícil para a máquina.

Em relação à taxa de erro das classes do modelo para o contexto anterior, o desempenho do modelo também foi melhor nas entrevistas com limpeza para os dados de ERR e NULL, conforme tabela 7 abaixo.

Tabela 7 - Taxa de erros por classe do contexto anterior ao possessivo para o *spaCy*.

Tipos de erro	Entrevistas sem limpeza	Entrevistas com limpeza
ERR	0,296	0,294
ETRANS	0,155	0,205
NULL	0,005	0,004

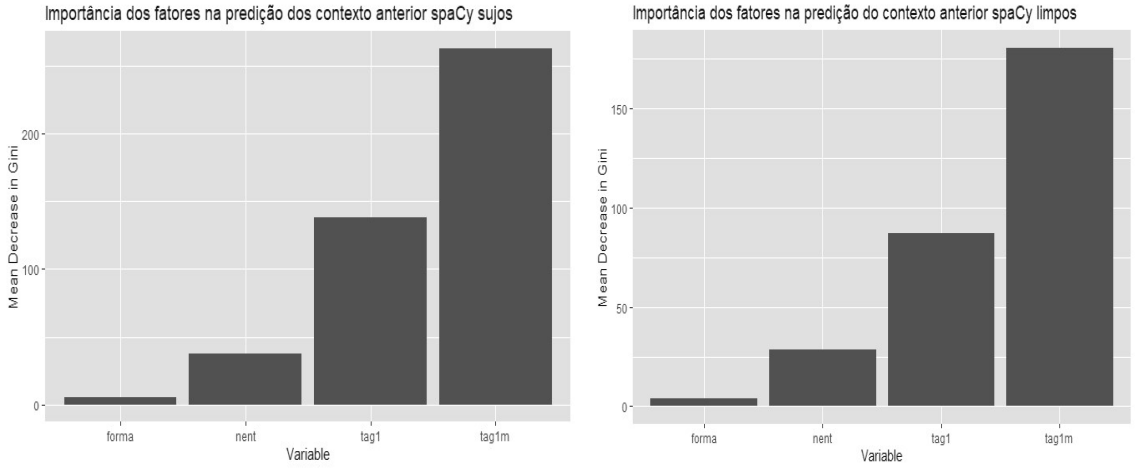
Fonte: elaboração própria.

Os dados de ERR caíram de 0,296 para 0,294, assim como os dados de NULL que foram de 0,005 para 0,004. Houve aumento na taxa de erro de ETRANS que foi de 0,155 para entrevistas

sem limpeza para 0,205 para entrevistas com limpeza. Apesar desse aumento, foi mais difícil para a máquina predizer ERR do que ETRANS, o aumento na taxa de ETRANS se explica também porque a taxa é calculada de maneira proporcional.

Diferentemente do *LancsBox* 6.0 com dados do contexto anterior, a variável com maior importância para a classificação das etiquetas do *spaCy* 3.5 foi ta1gm, ou seja, das etiquetas revisadas, tanto para entrevistas sem limpeza quanto com limpeza, conforme figura 31.

Figura 31 – Importância dos fatores na predição dos erros do contexto anterior ao possessivo para o *spaCy*.



Fonte: elaboração própria.

A ordem da importância das variáveis para a classificação também foi a mesma para os dois tipos de entrevista, sendo a variável de menor importância a forma para os dois tipos de entrevista, resultado também diferente do contexto anterior do *LancsBox* 6.0.

Embora os dados do contexto anterior para o *spaCy* 3.5 não precisassem ser ajustados, os dados dos determinantes precisavam, conforme figura 32.

Figura 32 – Etiquetas das entrevistas sem limpeza e com limpeza para os dados dos possessivos do *spaCy* antes do ajuste.

Etiquetas dos possessivos sem limpeza

```
DET
5210

      ADP   DET  INTJ   MWE  NOUN  PRON
        1  4999   188    10     2     2
TRUNCAMENTO  VERB
        7     1
```

Etiquetas dos possessivos com limpeza

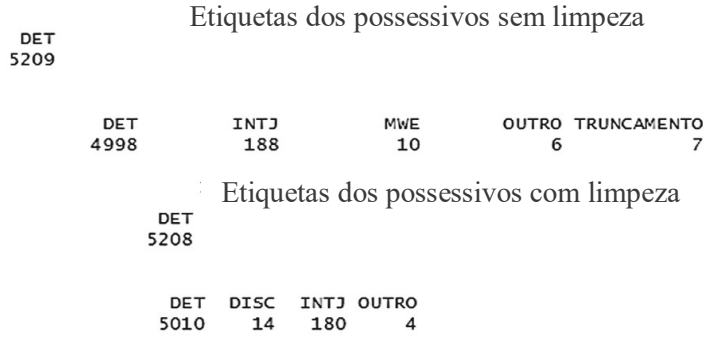
```
DET
5208

      ADJ   ADP   DET  INTJ   MWE  NOUN  PRON
        1     1  5010   180    12     1     1
TRUNCAMENTO
        2
```

Fonte: elaboração própria.

Para entrevistas sem limpeza foram amalgamados os dados de ADP, NOUN, PRON e VERB na categoria OUTRO, além de termos excluído um fator que foi incorretamente classificado como DET por erro de transcrição, quando na verdade era um "se eu" transcrito como "seu", como pode ser observado na figura 33.

Figura 33 - Etiquetas das entrevistas sem limpeza e com limpeza para os dados dos possessivos do *spaCy* após o ajuste.



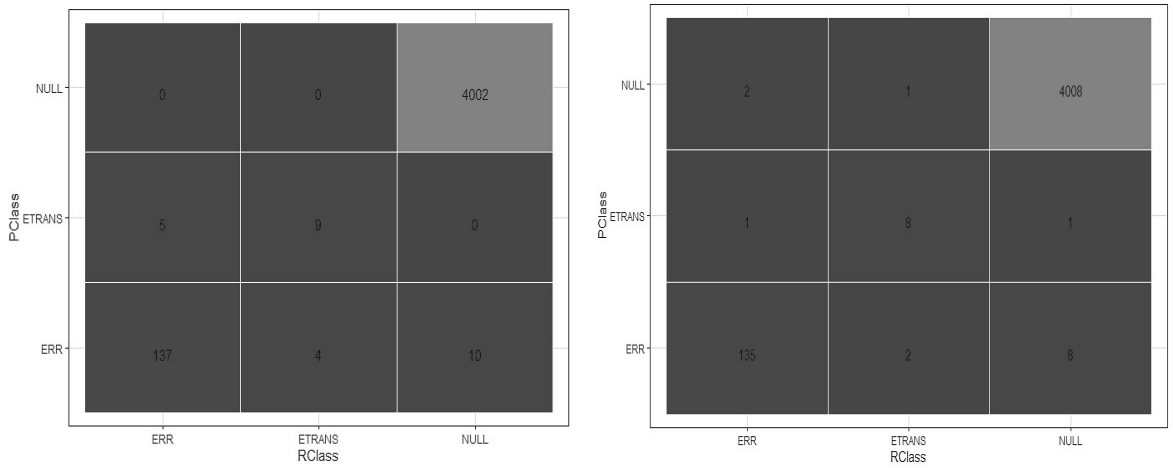
Fonte: elaboração própria.

Já para entrevistas com limpeza, as expressões multipalavras e os truncamentos foram amalgamados na categoria DISC, por se tratarem de etiquetas que também estão ligadas a traços de oralidade e, na categoria OUTRO, foram amalgamadas as etiquetas ADJ, ADP, NOUN, PRON, por possuírem frequência menor que 5.

Em relação à estimativa do erro para os dados de teste, o desempenho da classificação foi melhor para entrevistas com limpeza, OOB de 0,36%, indicando uma previsão de acurácia para os dados de teste de 99,64%. Já os dados sem limpeza, o OOB foi de 0,46%, sendo a previsão da acurácia para os dados de teste de 99,54%.

Na figura 34, as matrizes de confusão dos dados sem limpeza, à esquerda, e dos dados com limpeza, à direita, mostram que, para os dois tipos de dados, a classificação foi mais fácil para os dados de NULL, seguido de ERR, dado similar à classificação dos erros dos determinantes para o *LancsBox* 6.0.

Figura 34 – Matrizes de confusão para a classificação dos erros dos dados dos possessivos para o *spaCy*.



Fonte: elaboração própria.

Os dados também mostram um aumento na classificação de NULL nas entrevistas com limpeza, fato que é consequência também da diminuição dos níveis das variáveis tagposs e tagpossm, diminuindo a complexidade do modelo. Essa diminuição dos níveis também acarretou a diminuição dos dados de ETRANS e ERR.

Quanto aos valores dos erros das classificações por variável, houve uma diminuição da taxa para ERR e ETRANS, como pode ser visto na tabela 8.

Tabela 8 – Taxa de erros por classe dos possessivos para o spaCy.

Tipos de erro	Entrevistas sem limpeza	Entrevistas com limpeza
ERR	0,035	0,021
ETRANS	0,307	0,272
NULL	0,002	0,002

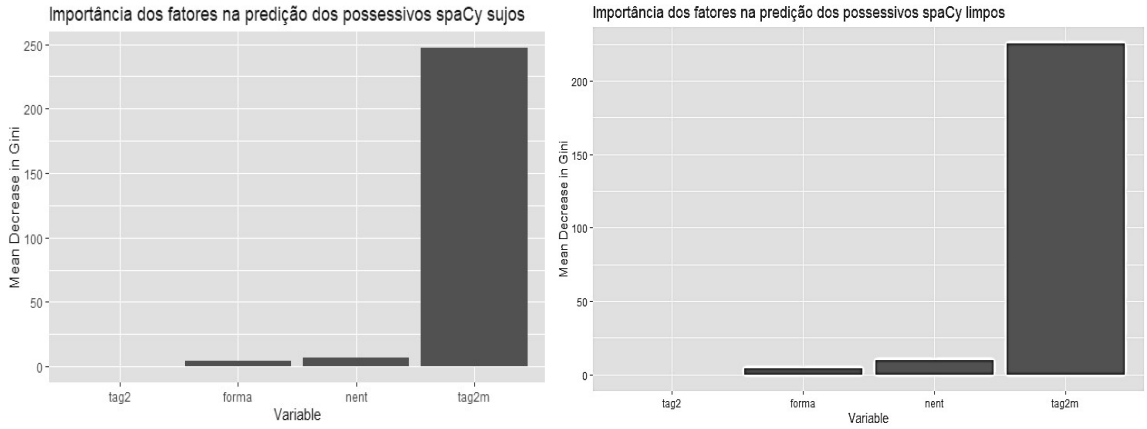
Fonte: elaboração própria.

Os valores de ERR passaram de 0,035 (entrevistas sem limpeza) para 0,021 (entrevistas com limpeza), já nos de ETRANS, a diferença foi maior de 0,307 (entrevistas sem limpeza) para 0,272 (entrevistas com limpeza). Houve manutenção da taxa para NULL, apesar do aumento quantitativo dos itens classificados como NULL. A taxa de ETRANS foi maior devido à proporção dos erros em relação aos dados, como o número de ETRANS foi muito baixo para os possessivos nos dois tipos de entrevista, houve, então, reflexo do tamanho da amostra,

comportamento diferente do contexto anterior do *spaCy* 3.5, mas similar em relação aos dados de determinantes do *LancsBox* 6.0.

Em termos de importância, figura 35, os dados apresentaram comportamento diferente de todos os outros conjuntos tanto de contexto anterior (*LancsBox* 6.0 e *spaCy* 3.5) quanto de possessivos (*LancsBox* 6.0).

Figura 35 – Importância dos fatores na predição dos erros dos possessivos para o *spaCy*.



Fonte: elaboração própria.

A variável que não exerceu influência nenhuma na repartição dos dados foi a etiqueta empregada pelo etiquetador (tag 2), sendo seus valores nulos para os dois tipos de entrevista. A forma e nent quase não foram relevantes, sendo a variável tagm a mais importante. A explicação para esse fato se deve pela quantidade de níveis da variável tag 2, que é apenas um, diferentemente dos possessivos para o *LancsBox* 6.0 que possui um conjunto de etiquetas mais detalhado e, consequentemente, maior nível dessa variável.

Nesta seção, discutimos a classificação dos erros das etiquetas, utilizando para tal o algoritmo de florestas aleatórias que é um método supervisionado de AM. No quadro 7 abaixo sintetizamos os resultados para os dois etiquetadores.

Quadro 7 – Síntese dos resultados do modelo de classificação de erros para os etiquetadores *LancsBox* 6.0 e *spaCy* 3.5.

Conjunto de dados	Etiquetadores					
	***LancsBox* 6.0**			***Spacy* 3.5**		
	Índice	**Dados sujos**	**Dados limpos**	**Índice**	**Dados sujos**	**Dados limpos**
Dados do contexto anterior	OOB	4,56%	4,98%	OOB	1,87%	1,51%
	Tx. erros	ERR 0,160	ERR 0,156	Tx. Erros	ERR 0,296	ERR 0,294
	MDG	tagcant	tagcant	MDG	tag1m	tag1m
	Índice	**Dados sujos**	**Dados limpos**	**Índice**	**Dados sujos**	**Dados limpos**
Dados dos possessivos	OOB	0,23	0,35	OOB	0,46	0,36
	Tx. erros	ETRANS 0,142	ETRANS 0,375	Tx. Erros	ETRANS 0,307	ETRANS 0,272
	MDG	tagpossm	tagpossm	MDG	tag2m	tag2m

Fonte: Elaboração própria.

Conforme quadro 7, para dados do contexto anterior, os índices de estimativa do erro das classes para os dados de teste (OOB) para dados do contexto anterior ao possessivo foram melhores para os dados anotados pelo *spaCy* 3.5, tanto para entrevistas com limpeza (1,87%) quanto sem limpeza (1,51%), em relação aos dados do *LancsBox* 6.0 (4,56% e 4,98%, para dados sem e com limpeza, respectivamente). Para os dois conjuntos de dados, a variável ERR foi a mais difícil de ser classificada, havendo diminuição da taxa de erro (Tx. Erros) para entrevistas com limpeza nos dois casos, diminuição de 0,160 para 0,156 nos dados do *LancsBox* 6.0 e de 0,296 para 0,294 para dados do *spaCy* 3.5. A respeito da importância das variáveis calculadas pelo *Mean Decrease in Gini*, as variáveis importantes para o modelo foram tagcant para os dados do *LancsBox* 6.0 e tag1m para os dados do *spaCy*, não havendo diferença entre entrevistas sem e com limpeza.

Diferentemente do contexto anterior ao possessivo, o modelo apresentou resultados melhores para dados dos possessivos do *LancsBox* 6.0, sendo o OOB 0,26 e 0,35 para entrevistas sem e com limpeza, respectivamente. Já os resultados do *spaCy* 3.5, para o índice OOB, foi de 0,46 (dados sem limpeza) e 0,36 (dados com limpeza). A taxa de erro da classificação (Tx. Erros) teve a mesma variável como a mais difícil de ser classificada ETRANS, esse fato ocorre porque, em proporção, essa variável teve ocorrências bem menores nos dois conjuntos de dados em relação às outras variáveis. Por fim, a variável de maior importância para o modelo no conjunto de dados dos possessivos, tanto nos dados do *LancsBox* 6.0 quanto para o *spaCy* 3.5, foi a etiqueta revisada, tagpossm (*LancsBox* 6.0) e tag2m (*spaCy* 3.5).

A partir da análise pelo algoritmo de florestas aleatórias, comparando os conjuntos de dados das duas ferramentas, concluímos que o modelo foi eficiente para classificar os erros das etiquetas. Além disso, os resultados da performance dos etiquetadores foram melhores para entrevistas com limpeza de forma que adotamos uma nova forma de transcrição em que as marcas contextuais ficam em uma trilha a parte, além de otimizarmos a transcrição de truncamentos, aproximando o sinal de truncamento "-" das palavras truncadas, porque, como observamos, quando o sinal de truncamento é aproximado, os etiquetadores conseguem etiquetar a palavra truncada.

3.2.4.4 A avaliação das funcionalidades das duas ferramentas

Dois dos nossos objetivos são: testar as ferramentas e comparar as funcionalidades nativas que oferecem subsídios para análise linguística. Para realizar esses objetivos, selecionamos o fenômeno da variação no preenchimento de determinantes antes de possessivos pré-nominais para realizar buscas automáticas para esse fenômeno e testar as funcionalidades das ferramentas para automatizar as análises linguísticas. Como vimos na seção 3.2.4.1, os dados das entrevistas com limpeza, comparados aos dados das entrevistas sem limpeza, ofereceram melhores resultados em relação ao contexto anterior ao possessivo para as duas ferramentas. Assim, para a análise das funcionalidades das ferramentas, usamos apenas os dados com limpeza.

Como visto na seção 3.2.3, que trata da descrição do fenômeno, a variação no preenchimento de determinante antes de possessivo pré-nominal ocorre de duas formas, com a presença do determinante, como em (4) ou com a ausência dele como em (5).

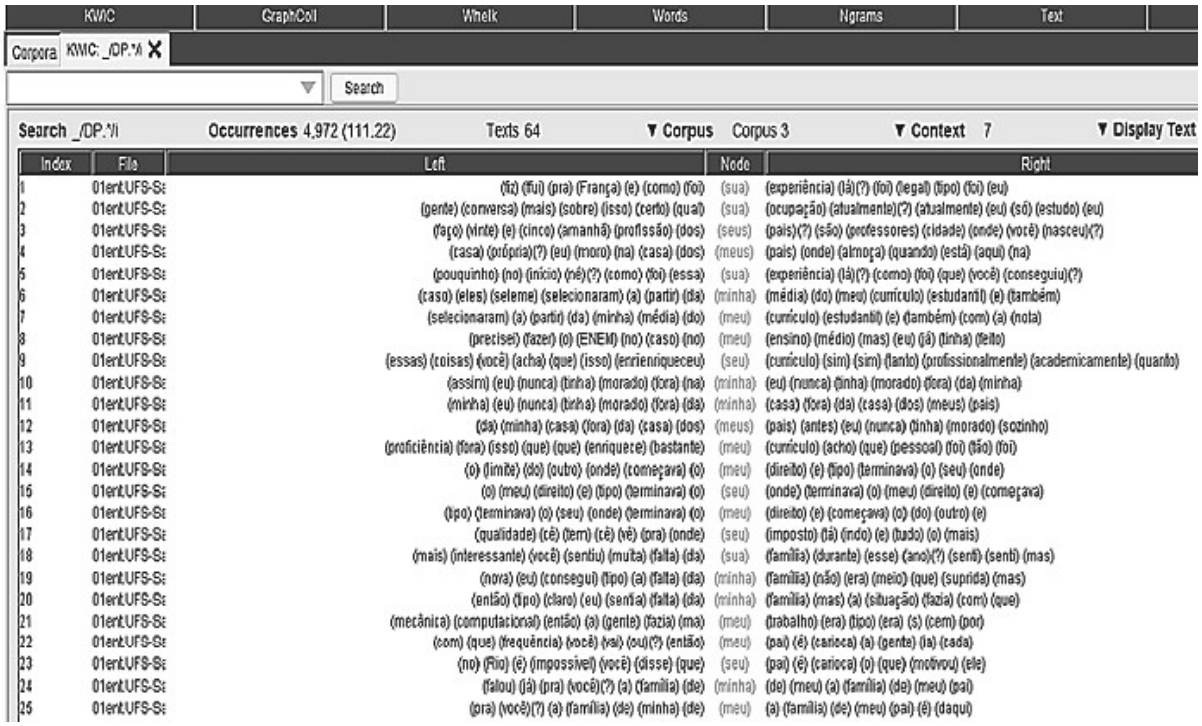
(3) cê pode informar o número **do** seu celular?

(4) então **Ø** minha consciência tá limpa (Excertos extraídos da entrevista 04ent.UFS-SaoCristovao2018__desl. I_final_cla.fs.21 da amostra Deslocamentos 2019)

Estudos realizados com dados de falantes de diferentes estados do Brasil, atestam que o preenchimento na posição de determinante é favorecido pelo tipo de sintagma, no caso, o sintagma preposicional (CALLOU; SILVA, 1997; CAMPOS JR., 2012; SEDRINS; PEREIRA; SILVA, 2019; GUEDES, 2019; SILVA, 2020; SIQUEIRA, 2020). Para realizar a busca por esse fenômeno e fazer o cômputo de sua frequência, usamos diferentes funções da ferramenta *LancsBox* 6.0.

A primeira das funções que usamos é a *Key Word in Context* (KWIC). Esta função permite que façamos a busca pelo fenômeno em seu contexto de realização, que pode ser delimitado com a quantidade de palavras precedentes e seguintes à ocorrência. Na figura 36, apresentamos a busca pelo fenômeno com o cômputo dos resultados e também o peso relativo dos dados na amostra.

Figura 36 – Resultados usando a função KWIC do *LancsBox* 6.0.



Fonte: elaboração própria a partir do *LancBox* 6.0.

Na figura 36, da esquerda para a direita, o software oferece as seguintes informações na linha acima das ocorrências: *Search*: que se refere ao valor que digitamos para fazer a busca, no caso, DP.*, que busca pelas ocorrências de todos os determinantes possessivos; *Occurrences*

oferece o valor quantidade de ocorrências e sua frequência relativa no *corpus*; *Texts*: a quantidade de textos; *Corpus*: o corpus de trabalho e *Context,* a quantidade de palavras antes e depois da ocorrência. Nas colunas, *Index* se refere ao número da ocorrência no *corpus*, *File*, ao nome do arquivo, *Left*, ao contexto anterior, *Node*, à ocorrência e *Right*, ao contexto imediatamente após a ocorrência.

A ferramenta nos retornou 4.972 ocorrências dos possessivos. A função KWIC informou também que o peso relativo das ocorrências dos determinantes nas entrevistas foi de 111,15 para cada 10.000 palavras no *corpus*. Como relatado, o sintagma preposicionado é favorecedor da ocorrência da forma preenchida, logo, podemos realizar uma busca apenas pelas preposições aglutinadoras "de", "em" e "por" com a expressão apresentada na figura 37.

Figura 37 – Busca pelos determinantes preenchidos no *LancsBox*.

```
[pos="SP.DA"] [pos="DP.*"]
```

Fonte: elaboração própria.

Na figura, pos é a etiqueta POS, sendo "SP.DA" as etiquetas de preposições mais artigos e, DP, a etiqueta para determinantes possessivos. A ferramenta retornou 564 ocorrências de preposições aglutinadas antes de possessivos pré-nominais. Contudo, como o *software* foi categórico para classificar a aglutinação de "de+artigo" como preposições simples, ele não retornou os dados desse contexto, logo, fizemos uma nova busca pelo lema "de" e "DP", mas o resultado não retornou dados com aglutinação. Optamos, então, pelo recurso de filtragem usando a expressão regular: d(as?|os?). A busca nos retornou 713 ocorrências para entrevistas, logo, a busca por preposições aglutinadoras retornou 1.277 ocorrências das 2.537 formas preenchidas. O resultado da busca aponta para uma frequência maior no preenchimento dos determinantes antes de possessivo pré-nominal 2.537/4.972 dados, o que difere dos resultados encontrados em Siqueira (2020) para falantes da mesma comunidade de práticas.

Devido a essa diferença nos resultados, optamos por replicar o estudo de Siqueira (2020) seguindo fidedignamente a mesma metodologia. Assim, utilizamos apenas dados de 32 informantes estratificados pelo tipo de deslocamento, tempo no curso e sexo/gênero, em entrevistas sem limpeza. Após a codificação dos dados, obtivemos 1.458 realizações do fenômeno, 190 realizações a mais encontradas em comparação a Siqueira (2020), que encontrou 1.268 realizações. Além dessa diferença no número total de ocorrências, a distribuição das ocorrências com determinante preenchido e sem preenchimento também foram diferentes. O

LancsBox 6.0 encontrou mais ocorrências para o preenchimento (738) do que para o não preenchimento (720), diferentemente de Siqueira que encontrou 699 ocorrências para o não preenchimento e 599 para o preenchimento.

A função *GraphColl* do *LancsBox* 6.0 nos permite analisar *collocations*, isto é, a frequência com que os itens, no caso possessivos, se agrupam com palavras à sua esquerda e à sua direita, conforme figura 38.

Figura 38 – Dados das colocações com possessivos *LancsBox* 6.0.

Freq: 4,976 - Collocates: 951 — A

Index	Status	Position	Collocate	▼ Stat	Freq (coll.)
1	o	R	que	1535.0	1535
2	o	L	a	1394.0	1394
3	o	R	eu	1267.0	1267
4	o	L	e	1263.0	1263
5	o	L	o	1153.0	1153
6	o	R	é	1144.0	1144
7	o	R	de	886.0	886
8	o	R	não	806.0	806
9	o	L	na	693.0	693
10	o	L	da	642.0	642

Freq: 4,976 - Collocates: 45 — B

Index	Status	Position	Collocate	▼ Stat	Freq (coll.)
1	o	R	VMI	7426.0	7426
2	o	R	NCFS	4883.0	4883
3	o	R	RG	4820.0	4820
4	o	L	SPS	4658.0	4658
5	o	R	NCMS	4572.0	4572
6	o	L	DA0	2398.0	2398
7	o	R	AQ0	1881.0	1881
8	o	L	CC	1876.0	1876
9	o	R	PP3	1528.0	1528
10	o	R	PP1	1467.0	1467

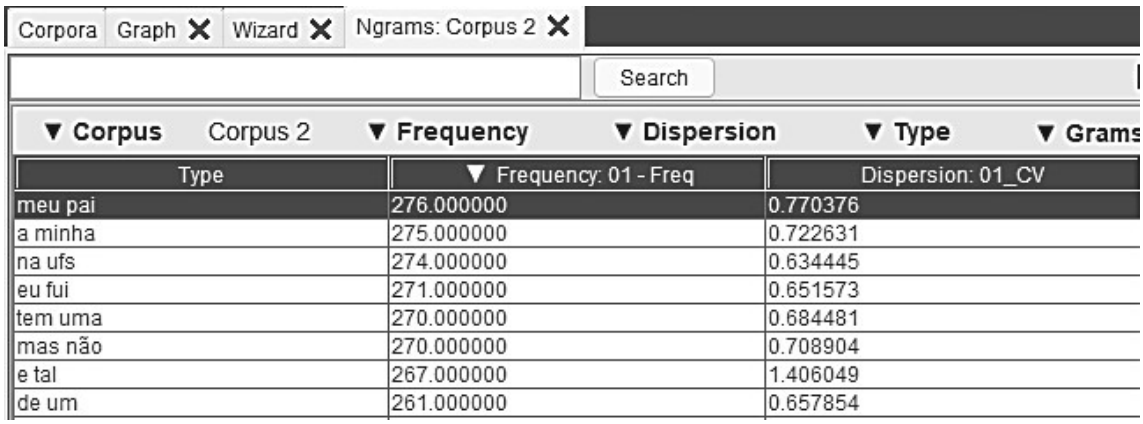
Fonte: elaboração própria a partir do *LancsBox* 6.0.

Na figura 38, em A, *Index*, indica a ordem da colocação que tem mais associação com o nó (item pesquisado). No caso, "que" é a palavra que mais se associa com os possessivos. *Status* indica se o nó está expandido ou não, de forma que as bolinhas claras representam que não está. *Position*, se o colocado está à direita (R) ou à esquerda (L), "que" está à direita, por exemplo. *Collocate* indica o colocado, *Stat* o valor de associação entre o colocado e o nó, e *Freq(coll)* (frequência) mostra a frequência da colocação no *corpus*, ou seja, "DP+que" aparece 1.535 vezes no *corpus*. Em B, temos o resultado das colocações em termos de "POS", o que confirma nossa análise das etiquetas, de que, à esquerda, a classe que mais se combina com "DP" é a "SPS" (preposições). A busca por colocações pela etiqueta POS contribuiu para que encontrássemos outro padrão na Amostra Deslocamentos 2019, a interjeição "Meu Deus" foi mais antecedida por substantivos e mais seguida por verbos. Como os dois etiquetadores foram categóricos ao analisar "Meu Deus" como possessivo e substantivo, esse padrão pode lançar uma luz para que os algoritmos dos etiquetadores considerem esse aspecto e renalisem essa ocorrência como interjeição.

O *LancsBox* 6.0 ainda oferece a função Wizard, que combina todas essas informações de frequência no *corpus* em formato de relatório, cuja primeira parte está reproduzida no anexo A. No entanto, o relatório é gerado por palavra e não por etiqueta, assim como a função *Ngrams*,

por isso não apresentamos os resultados de todos os *n-grams* com os possessivos aqui, porque teríamos que pesquisar a ocorrência de cada possessivo, como representado na figura 39.

Figura 39 – Resultados da função Ngrams do *LancsBox* 6.0.



Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
meu pai	276.000000	0.770376
a minha	275.000000	0.722631
na ufs	274.000000	0.634445
eu fui	271.000000	0.651573
tem uma	270.000000	0.684481
mas não	270.000000	0.708904
e tal	267.000000	1.406049
de um	261.000000	0.657854

Fonte: elaboração própria a partir do *LancsBox* 6.0

Como observamos pela figura 39, a função *Ngrams* retorna uma lista de n-grams (sequências de palavras) em toda a amostra, podendo ajustar o tamanho dos n-grams na aba *Grams,* o padrão da função é 2. Ao digitar "meu" na aba *Search*, a ferramenta busca pelo n-gram de maior frequência com a palavra "meu", que na amostra Deslocamentos 2019 foi "meu pai".

O *spaCy* 3.5 por outro lado, não apresenta todas essas funcionalidades que o *LancsBox* 6.0 possui de forma nativa. Como os próprios criadores da ferramenta advertem, em seu site, *spaCy* 3.5 é uma biblioteca orientada para a criação de produtos de PLN, logo, suas aplicações para o contexto acadêmico é limitada. Contudo, ressaltamos que com conhecimento de linguagem *Python* e de suas bibliotecas possibilita a criação de códigos para produzir resultados semelhantes aos das funções nativas do *LancsBox* 6.0.

A ferramenta de busca, *Matcher,* além de rápida, produz resultados mais consistentes que o *LancsBox* 6.0 para o fenômeno da variação no preenchimento de determinante antes de possessivo pré-nominal. A busca pelo fenômeno retornou 5.208 dados, possuindo taxas de erros menores proporcionalmente que o *LancsBox* 6.0, além de retornar os resultados que o *LancsBox* 6.0 classificou como "PX" (pronome possessivo).

Ao replicar o estudo de Siqueira (2020) utilizando o *spaCy* 3.5, observamos que seu desempenho também foi melhor. Por meio dessa biblioteca foram encontradas mais ocorrências do fenômeno, 1.512, em relação à 1.268 de Siqueira (2020), ou seja, uma diferença de 244 ocorrências a mais. Assim como no *LancsBox* 6.0, o *spaCy* 3.5, também teve resultados diferentes para a distribuição do fenômeno em relação ao estudo que replicamos. Essa

ferramenta localizou 754 ocorrências de não preenchimento do determinante e 758 ocorrências de preenchimento, corroborando a distribuição dos dados localizados pelo *LancsBox* 6.0. Esse fato mostra que a replicação do estudo foi consistente para as ferramentas, visto que elas tiveram resultados similares para a distribuição apesar da diferença no número total de ocorrências do fenômeno encontrado por cada uma delas.

O *spaCy* 3.5 oferece no *Matcher* uma série de atributos que possibilitam a busca pelos mais diversos tipos de padrões linguísticos. Alguns exemplos são: *MORPH* (permite a busca pelas características morfológicas), *LENGTH* (busca por extensão da palavra), *LEMMA* (busca pelo lema). Além das funções *DEP Matcher*, que possibilita buscas pelas relações sintáticas e de dependências, *Fuzzy Matcher* que permite buscas por palavras com erros ortográficos.

Sabemos que os sintagmas preposicionados são aqueles que mais favorecem o preenchimento de determinante, logo, a busca por possessivos precedidos por preposições aglutinadoras pode ser feita utilizando o atributo *ORTH*, como no exemplo extraído do nosso código representado na figura 40.

Figura 40 – Busca por determinantes preenchidos no *spaCy*.

```
matcher = Matcher(vocab=nlp.vocab)
det_poss1 = [{'TAG': 'ADP', 'ORTH': {'NOT_IN':['de','com', 'para', 'pra', 'até',
                                               'por', 'pelo', 'pela', 'ao', 'à', 'aos', 'às',
                                               'dessa','dessas', 'nessa', 'nessas']}},
             {'POS': 'DET', 'MORPH':{'IS_SUPERSET': ['PronType=Prs']}}]

det_poss2 = [{'POS': 'DET', 'MORPH': {'IS_SUPERSET': ['Definite=Def']}},
             {'POS': 'DET', 'MORPH':{'IS_SUPERSET': ['PronType=Prs']}}]
matcher.add('detposs1', [det_poss1])
matcher.add('detposs2', [det_poss2])
det_result = matcher(doc)
match = Printer()
for match_id, start, end in det_result:
  etiqueta_termo1= doc[start]
  etiqueta_termo2 = doc[end]
  ocorrencia = doc[start:end]
  contexto = doc[start-1:end+5]
  contexto_parte1 = doc[start-1:start]
  contexto_parte2 = doc[end: end+5]
  linhas.append ([numero_ent, ocorrencia,etiqueta_termo1.pos_, 'DET', 'NP', contexto])

dataframe = pd.DataFrame(linhas, columns = colunas)
```

A

```
matcher = Matcher(vocab=nlp.vocab)
detrex2 = [{'TEXT': {'NOT_IN': ['o','a', 'os','as', 'do', 'da','dos',
                                'das', 'ao', 'no', 'na', 'nos', 'nas']}},
           {'POS': 'DET', 'MORPH':{'IS_SUPERSET': ['PronType=Prs']}}]
```

B

Fonte: Sousa *et. al*, 2022.

Na figura 40, parte A, temos a busca por determinantes não-preenchidos na posição anterior ao possessivo. Pelos atributos *ORTH* e *NOT_IN*, excluímos preposições que,

inicialmente não fariam parte da análise pelo uso categórico da aglutinação em determinados contextos como “pela”, “pelo”, “nessa”. Pelo atributo *POS*, definimos na segunda linha determinantes que fossem do tipo artigo seguido de determinantes que fossem do tipo possessivo. Na figura 40, parte B, visualizamos a busca por qualquer item que anteceda os possessivos, com exceção das preposições aglutinadoras que já havíamos buscado na parte A, usando para isso os atributos *TEXT* e *NOT_IN*.

Essa busca mais refinada também é vantajosa para fenômenos em que os itens têm maior flexão, por exemplo, o fenômeno da terceira pessoa do plural. Ao usar o *LancsBox* 6.0 para fazer esse levantamento, o pesquisador deve utilizar uma expressão regular que especifique todas as flexões de verbos terminados em 3ª pessoa, uma vez que, na concordância verbal não canônica, o verbo pode não estar flexionado como 3pp (NOVAIS, 2021), como em “Eles vai cedo.”. Já no *spaCy* 3.5 a busca é mais simples, conforme figura 41.

Figura 41 – Busca pelo fenômeno de 3pp no *spaCy* 3.5.

```
matcher = Matcher(vocab=nlp.vocab)
verb3p = [{'TEXT': {'NOT_IN': ['gente', 'eu', 'cê', 'você']}},
          {"POS": "VERB", 'MORPH': {'IS_SUPERSET': ['Person=3', 'VerbForm=Fin']},
           "ORTH": {"NOT_IN": ['Têm', 'tem', 'Tem', 'têm', 'Vêm', 'Vem','vêm', 'vem', 'há']}}]

aux3p = [{'TEXT': {'NOT_IN': ['gente', 'eu', 'cê', 'você']}},
         {"POS": 'AUX', 'MORPH': {'IS_SUBSET': ['Person=3|VerbForm=Fin']},
          'ORTH': {'NOT_IN': ['tem', 'Tem', 'Têm', 'têm', 'vem', 'vêm', 'Vem', 'Vêm']}}]
```

Fonte: Sousa *et. al*, 2022.

Na figura 41, representamos dois padrões de busca, um para verbos principais e outro para verbos auxiliares. No primeiro padrão, “verb3p”, usamos o atributo *TEXT* para excluir contextos em que “gente”, “eu”, “cê” e “você” aparecem como antecedentes imediatos dos verbos, porque os verbos para esses sujeitos também possuem terminação igual à de 3p, no caso do “eu” para verbos no pretérito perfeito e no modo subjuntivo, e, nos outros casos, em todos os tempos verbais. Esse mesmo código foi utilizado no segundo padrão para exclusão de “gente”, “eu”, “cê” e “você”. Ainda no primeiro padrão, pelos atributos *POS* e *MORPH*, buscamos pela classe de verbos que sejam de terceira pessoa e estejam conjugados, o atributo *ORTH* foi usado para eliminar contextos de homófonos, ou seja, não dá para saber pela fala se a concordância foi ou não realizada. No padrão “aux3p”, fizemos o mesmo tipo de busca, mas desta vez para POS “AUX” para buscar verbos etiquetados como auxiliares.

Outra vantagem do *spaCy* 3.5 é que, por funcionar por meio de uma linguagem de programação, os dados já podem ser salvos com informações acrescentadas pelo pesquisador. Utilizando a função *data.frame* da biblioteca *pandas,* foi feita uma planilha em formato .csv já contendo as seguintes informações que ficaram onde está escrito colunas: Numero_ent (número da entrevista), Ocorrencia (o item pesquisado), TAG1 (a etiqueta da ocorrência), TAG2 (a etiqueta da palavra antecedente), Contexto (a frase onde a ocorrência está presente) e NP para formas não-padrão (formas em que houve preenchimento do determinante). Devido a isso, é possível utilizar o arquivo para outras linguagens e *softwares* de análise estatística, caso o pesquisador não tenha familiaridade com as ferramentas de análise estatística da linguagem de programação *Python*.

No quadro 8, representamos uma comparação acerca dos usos das ferramentas em termos de anotação, buscas e recursos nativos para pré-análise dos dados.

Quadro 8 – Comparação das funcionalidades das ferramentas.

Ferramenta	Anotação	Buscas	Recursos para pré-análise
LancsBox 6.0	Maior taxa de erros	Menor quantidade de resultados Menor refinamento nos atributos de busca	Maior quantidade de recursos nativos
spaCy 3.5	Menor taxa de erros	Maior quantidade de resultados Maior refinamento dos atributos de busca	Menor quantidade de recursos nativos, o que pode ser solucionado com maiores conhecimentos sobre a linguagem Python.

Fonte: Elaboração própria.

A conclusão que chegamos acerca do critério de buscas e funcionalidades das ferramentas é que o *spaCy* 3.5 foi melhor para buscas do fenômeno da variação no preenchimento de determinante antes de possessivo pré-nominal, pois, por meio da combinação dos atributos oferecidos pela ferramenta e pelo suporte de busca que permite a combinação de diferentes níveis de etiquetagem, obtivemos mais ocorrências e resultados mais precisos para a etiquetagem do contexto anterior ao possessivo, em proporção, como relatado na seção 3.4.2.1. O *spaCy* 3.5 permite também, em um mesmo ambiente, a IDE *Google Colaboratory*, por exemplo, a preparação dos dados já em planilha. Por outro lado, em termos de análise, o *LancsBox* 6.0 oferece mais recursos para além da etiquetagem, como colocações tanto em

termos de palavras quanto de etiquetas e frequências de *n-grams* sem a necessidade da criação de um código específico para essas tarefas. Indicamos, assim, o uso combinado das ferramentas para melhor aproveitamento de buscas e pré-análise dos dados.

3.3 Critérios para a disponibilização da amostra

3.3.1 Aspectos éticos da documentação da amostra

Destacamos a importância da disponibilização de dados e de uma metodologia mais detalhada frente às demandas da Ciência Aberta. Em um contexto de tamanho negacionismo em relação às ciências, práticas de pesquisa mais transparentes fortalecem o rigor metodológico dos estudos bem como a confiança pública em sua condução, além de favorecer a reprodutibilidade e replicabilidade (LYON, 2016). O compartilhamento dos dados, no entanto, não pode ser um processo feito de qualquer forma, uma vez que existe uma responsabilidade legal do pesquisador e da instituição à qual o pesquisador está vinculado sobre quaisquer danos causados àqueles que são os informantes/participantes das pesquisas (FREITAG *et al*., 2021; FREITAG, 2021b). Existe, então, um certo receio por parte dos pesquisadores em compartilhar seus dados, seja por questões éticas ou até mesmo por questões financeiras, pois dados linguísticos anotados são muito valiosos, como apontam Garelleki *et al.* (2020).

Os critérios que regem aspectos éticos em pesquisa com seres humanos nas ciências sociais no Brasil são estabelecidos pela Resolução Nº 510, de 07 de abril de 2016 do Conselho Nacional de Saúde. Em seu artigo 9º inciso V (BRASIL, 2016), a resolução diz que deve ser assegurado aos participantes o direito de "decidir se sua identidade será divulgada e quais são, dentre as informações que forneceu, as que podem ser tratadas de forma pública". Nesse inciso, observamos claramente que o informante da pesquisa deve ter seu direito de exercer o poder sobre as informações cedidas e sobre a sua identidade resguardado.

Dados de fala, especificamente, são dados passíveis de identificação por voz. Calamai e Frontini (2018) apontam como uma forma de minimizar os danos na disponibilização de dados de fala e considerando o reconhecimento pela voz, as licenças de uso e o termo de consentimento como possíveis soluções, o que também é apontado por Mello (2021) como ações importantes de compartilhamento e segurança legal no uso de dados. Em relação a esses dois pontos, quando a amostra Deslocamentos 2019 foi coletada, as pesquisadoras entregaram

e colheram a assinatura do Termo de Consentimento Livre e Esclarecido (anexo B). Nele, está previsto que o consentimento para utilização dos dados pode ser interrompido e que as entrevistas ficariam em um banco de dados e seriam utilizadas por outros pesquisadores para fins científicos, ou seja, os informantes estavam cientes dos usos dos seus dados para fins acadêmicos.

Collister (2022) aponta dois tipos de licenças para garantir os direitos do autor, a licença *GNU General Public License* (GNU GPL) e as *Creative Commons* (CC). Conforme a autora, a licença GNU GPL é menos restritiva que as CC em termos de reutilização, sendo a única restrição a de que os derivados produzidos com os dados sejam de acesso aberto. Já as CC possuem níveis de restrição de forma que, na produção do site para divulgação da amostra, decidimos utilizá-las como licenças de uso. Está no bojo das licenças CC respeitar os direitos autorais e conexos. A licença apresentada na figura 42 abaixo é um pouco restritiva, mas ideal para a proposta desta tese, no sentido de que ela protege a autoria e autoriza o desenvolvimento de novas tecnologias a partir dos dados disponibilizados.

Figura 39 - Licença *Creative Commons.*

Fonte: Site Creative Commons.

Conforme essa licença, o licenciante (o criador de conteúdo) permite ao licenciado (quem usa o conteúdo) utilizar e fazer novos trabalhos com os dados gerados, contanto que cite os licenciantes e utilize a mesma licença na criação dos derivados para propósitos não-comerciais.

Como argumenta Freitag (2021), o acesso aos dados pode ser feito em níveis diferentes. Para tanto, criamos um termo de autorização (APÊNDICE A), no qual ao acessar o site, o usuário deverá se cadastrar, preenchê-lo e assiná-lo concordando com os termos, para, a partir de então, ter acesso aos dados, caso seu cadastro seja validado. Além disso, como consta no parecer redigido pelo procurador da UFS (anexo C), o documento encontra-se em consonância com a Lei nº 13.709, de 14 de agosto de 2018, Lei Geral de Proteção de Dados Pessoais (LGPD). Esse documento foi criado nos moldes do termo de utilização do *Linguistic Data Consortium*.

3.3.2 Aspectos técnicos para tornar a amostra acessível

Inicialmente, o site estava sendo desenvolvido por meio do *Google Sites*, tarefa que vinha sendo articulada em conjunto com o aluno Kevenny de Jesus Santos, que estuda Ciências da Informação na UFS-Itabaiana e a estudante de doutorado em Letras, Vanessa Ponte. Contudo, a plataforma de sites do *Google* não permite integração entre um repositório e o site, logo, o estudante Kevenny construiu uma plataforma utilizando as linguagens Python e Dart e o *framework* Flutter, que são ferramentas gratuitas de desenvolvimento web. Foi desenvolvido também, junto com o estudante, um organizador de dados automático, que age buscando arquivos em diferentes pastas que se referem ao mesmo informante e os arquivam em uma única pasta nova. A plataforma e o organizador estão em processo de registro de *software*, sob o protocolo nº NID351-2022, cujo parecer encontra-se no anexo D.

Os dados de todas as amostras coletadas pelos pesquisadores do Grupo de Pesquisa em Linguagem Interação e Sociedade (GELINS) encontram-se alocados na nuvem *Google Drive*, que é um espaço de armazenamento pago, assim como o provedor onde o site ficará hospedado. Além disso, arquivamos os dados em um computador físico para ter maior segurança de que os dados não serão perdidos por situações do cotidiano, como aponta Tagliamonte (2006).

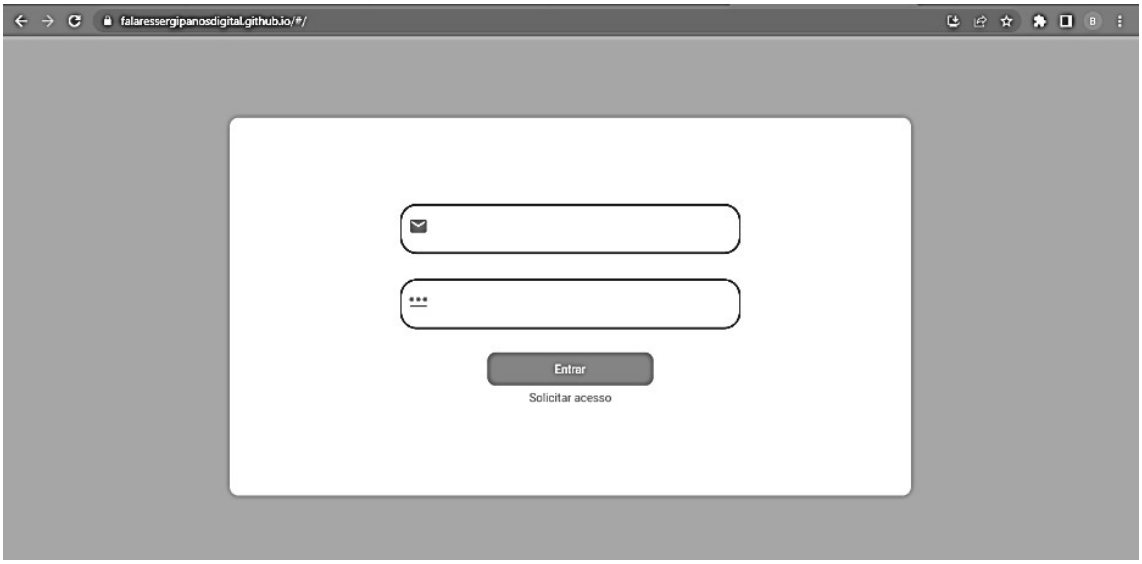
Os usuários podem navegar pelo site pesquisando por atributos, como demonstrado na figura 43.

Figura 43 – Seleção de dados por atributos.

Fonte: Banco de Dados Falares Sergipanos Digital.

Observamos, na figura 43, a seleção de um participante do sexo/gênero masculino. Na descrição do arquivo, encontram-se os metadados contextuais da entrevista, como iniciais do informante, idade, gênero, escolaridade, tempo no curso, tipo de documentação, tipo de deslocamento, ano da coleta, e a comunidade. Na figura 44, demonstramos como o usuário pode ter acesso aos dados:

Figura 44 – Aba solicitação de acesso aos dados.



Fonte: Site Falares Sergipanos Digital.

Como mencionado, aqueles interessados em nossos dados terão que fazer *login* no site, preencher o formulário de autorização de uso dos dados, para em seguida, o acesso ser liberado ou negado. O site ainda não se encontra totalmente pronto faltando ainda o design. Nossa intenção é de que ele esteja pronto para navegação a partir do segundo semestre de 2023.

O tipo de formulação para o site "Falares Sergipanos Digital" proporciona uma experiência intuitiva ao usuário e também para o administrador do site, que não precisa ter conhecimentos de linguagem de programação para adicionar novos dados, liberar acessos e atualizá-lo, evitando que o mesmo se torne obsoleto, como ocorreu com as ferramentas de gerenciamento destinadas para a criação de estímulos do projeto VIA, reportado por Fridland e Kendall (2022) que se tornaram obsoletas porque os integrantes permanentes do projeto não possuíam expertise para mantê-las a longo prazo. Nesse sentido, após a inserção dos dados novos no *Google Drive*, o link para acesso "Leitor" é criado e colado na planilha organizadora, e, de maneira automática, o dado já fica disponível no site.

O tipo de disseminação dos dados que estamos nos propondo a fazer, no momento, não é considerado como uma publicação formal por não passar por um processo de revisão dos dados, como ocorre em repositórios especializados conforme Callaghan *et al*. (2012). Ademais, a estrutura do nosso site ainda não oferece uma conexão com bases de dados científicas como o Portal Brasileiro de Publicações e Dados Científicos em Acesso Aberto (Oasisbr) e o Repositório Institucional da UFS (RI-UFS), o que conforme Champieux e Coates (2022) não permite que os dados sejam facilmente encontrados e reutilizados.

Com o objetivo de tornar nossos dados interoperáveis, encontráveis e acessíveis em conformidade com os critérios *FAIR* (WILKINSON *et al*., 2016) de gerenciamento de dados na Ciência Aberta, firmamos uma parceria com a professora Edilayne Meneses Salgueiro do Departamento de Computação da UFS para criarmos um *software* que conecte nossos dados com os repositórios citados. Para tanto, a criação dos metadados da amostra deve seguir o padrão *Dublin Core*[56]. Este padrão define uma lista de 15 elementos básicos[57] que os metadados podem ter, porém, a lista pode ser estendida ou encurtada conforme o tipo do documento do pesquisador. Representamos na figura 12 metadados de um arquivo de áudio da amostra Deslocamentos.

Figura 45 – Metadados de um arquivo da amostra Deslocamentos 2019.

```
<HEAD>
<META name="DC.title" content="01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24"/>
<META name="DC.language" content="Portuguese"
<META name="DC.description" content="Entrevista sociolinguística"/>
<META name="DC.subject" content="Assuntos pessoais seguindo moldes do roteiro de entrevista laboviano."/>
<META name="DC.contributor" content="Raquel Meister Ko. Freitag"/>
<META name="DC.contributor" content="THAÍS REGINA DE ANDRADE CORRÊA"/>
<META name="DC.contributor" content="Cristiane Conceição de Santana Ribeiro"/>
<META name="Dc.date" content="2018"/>
<META name="Dc.type" content="audio"/>
<META name="DC.format" content="wav"/>
<META name="DC.publisher" content="Condomínio de Laboratórios Multiusuários de Informática e Documentação (LAMID)"/>
<META name="DC.identifier" content="https://drive.google.com/drive/u/4/folders/1uZQcKjIvAzDxQn4YuNpbhmehYD22uFvQ"/>
<META name="DC.relation" content="01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24.txt"/>
<META name="DC.coverage" content="Brazil"/>
<META name="DC.coverage" content="Sergipe/>
<META name="DC.rights" content="Access limited to members"/>
</HEAD>
```

Fonte: elaboração própria.

Como explicitado, nem todos os 15 elementos básicos precisam estar presentes, pois, não necessariamente todos os elementos constituem o arquivo a ser descrito. No caso da amostra Deslocamentos 2019, cada arquivo deve ser descrito com 12 dos 15 elementos: título, língua do arquivo, descrição, assunto, colaboradores, data, tipo do arquivo, formato do arquivo,

[56] Site oficial do Dublin Core: https://www.dublincore.org/

[57] Os 15 elementos e sua descrição podem ser encontrados em: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#section-3

o nome da entidade que divulga o arquivo, a url da localização do documento, relação do objeto com outro item, o local coberto pelo conteúdo intelectual do arquivo e os direitos sobre o objeto. Os outros três elementos que não foram considerados são: *Creator* (criador do objeto, não usado porque em se tratando de uma entrevista sociolinguística, o nome do entrevistado deve ser preservado) e *source* (recursos que deram origem aos objetos digitais, como aqui os objetos já são digitais, não o usamos).

Os critérios estabelecidos acima oferecem subsídios para que a amostra Deslocamentos 2019 possa ser utilizada por outros estudiosos, quando da sua divulgação, resguardando o direito dos informantes, e também protegendo os pesquisadores e a instituição de ensino à qual o banco de dados Falares Sergipanos encontra-se vinculado.

3.4 Afinal, é possível? Considerações finais sobre o capítulo

Os resultados apresentados acerca da comparação entre as ferramentas *LancsBox* 6.0 e *spaCy* 3.5 confirmam a nossa tese de que é possível etiquetar e sistematizar dados sociolinguísticos nas etapas pré-análise de dados de entrevistas sociolinguísticas. A descrição dos procedimentos de arquivamento e da disponibilização dos dados por meio da criação de um site demonstrou que também é possível construir o site para divulgação das amostras com recursos abertos, porém, a hospedagem do site e dos dados em nuvem gera custos.

A avaliação da etiquetagem demonstrou que o *spaCy* 3.5, apesar de possuir uma etiquetagem POS que não integra informações morfológicas como o *LancsBox* 6.0, teve resultados mais consistentes acerca da etiquetagem tanto no contexto anterior aos possessivos quanto no contexto dos possessivos, possivelmente devido à não integração de informações morfológicas das etiquetas às etiquetas POS do *spaCy* 3.5. Ademais a análise da etiquetagem, além de fornecer informações linguísticas sobre o padrão de ocorrência dos dados do fenômeno de variação no preenchimento de determinantes antes de possessivos pré-nominais, foi significante para adotarmos uma nova revisão ortográfica nas entrevistas, bem como inserirmos as adaptações nas normas de transcrição do GELINS. Essa revisão já está acontecendo e sendo liderada por Paloma Batista Cardoso, estudante de doutorado do Programa de Pós-graduação em Letras da Universidade Federal de Sergipe.

Ainda acerca da etiquetagem, avaliamos se o nosso modelo de classificação dos erros contidos na etiquetagem foi consistente por meio de florestas aleatórias. Os resultados da

avaliação, por meio do índice OOB, permitiram-nos concluir que a nossa classificação foi consistente para os dois etiquetadores, tanto em dados com limpeza e dados sem limpeza.

No que diz respeito à testagem e avaliação das funcionalidades das ferramentas, concluímos que o *spaCy* 3.5 realiza buscas mais consistentes e com maior cobertura do que o *LancsBox* 6.0. Contudo, as funcionalidades do *LancBox* 6.0, por se tratar de um *software* específico para análise de *corpora*, possibilita a extração automática de mais informações sobre os dados, incluindo dados estatísticos, o que para o *spaCy* 3.5 demandaria mais conhecimentos da linguagem *Python* e de suas bibliotecas. Assim, sugerimos o uso combinado das ferramentas para uma pré-análise dos dados.

Questões relativas ao armazenamento e divulgação dos dados em termos de uso de ferramentas abertas evidenciaram dois pontos. O primeiro é que o uso de recursos abertos para a criação de um site com funcionalidades como liberação de acesso, ligação entre uma nuvem ao invés de um servidor, que é mais caro, é possível. O segundo é que não é possível hospedar um site que tenha uma base de dados com muita demanda de armazenamento de forma gratuita, até o momento, repositórios gratuitos, por exemplo o *Kaggle*, não suportam conjuntos de dados com mais de 20GB, e todos os dados gerados no banco de dados Falares Sergipanos já somam mais de 500GB.

Neste capítulo, tratamos da testagem e da avaliação das ferramentas *LancsBox* 6.0 e *spacy* 3.5 para tarefas pré-análise de dados. Descrevemos procedimentos também sobre a divulgação dos dados, etapa pós-análise. A seguir, tecemos as considerações finais sobre o trabalho.

4 CONSIDERAÇÕES FINAIS

Realizamos este estudo com o objetivo de estabelecer um protocolo de sistematização e etiquetagem de dados linguísticos para defender a tese de que é possível realizar essas duas tarefas com recursos abertos e gratuitos. Para criar este protocolo (apêndice B), foram realizadas os seguintes percursos i) testagem, comparação e avaliação do *LancsBox* 6.0 e do *spaCy* 3.5, ferramentas gratuitas de etiquetagem com interface nativa para buscas automáticas; ii) descrição de processos para divulgar e compartilhar a mostra dos dados do banco de dados Falares Sergipanos; iii) sistematização das ações desenvolvidas, a partir dos resultados, em forma de protocolo a ser implementado no escopo dos trabalhos de coleta e disseminação de dados do referido banco (apêndice B)

A testagem, a comparação e a avaliação da etiquetagem automática das ferramentas *LancsBox* 6.0 e *spacy* 3.5 nos levaram à conclusão de que a etiquetagem de dados obteve desempenho melhor utilizando a biblioteca *spacy* 3.5, devido ao seu conjunto de etiquetas POS que é mais simplificado. Logo, esta é a ferramenta que será utilizada para a etiquetagem de dados. As buscas automáticas também foram melhores para o *spacy* 3.5, uma vez que essa biblioteca oferece mais recursos nesse quesito. Em relação à pré-análise dos dados como interface nativa, o *LancsBox* 6.0 oferece mais recursos, por se tratar de uma ferramenta de análise de *corpora*. Contudo, ressaltamos que o conhecimento de linguagem *Python* e suas bibliotecas também permite a realização das mesmas tarefas.

Na etapa pós-análise dos dados, que trata do processo de divulgação dos dados em site, verificamos que é possível criá-lo com recursos gratuitos e de uma forma que, para gerenciá-lo, não sejam necessários conhecimentos de linguagem de programação. No entanto, o arquivamento dos dados e a hospedagem do site ainda não são possíveis em plataformas gratuitas.

Após a criação do protocolo, ele foi validado por três integrantes do GELINS, com níveis diferentes de conhecimento de recursos computacionais. O mais avançado realizou todas as etapas sem dificuldades, mas o iniciante sugeriu modificações que foram acatadas, principalmente no tocante à apresentação da interface do spaCy 3.5.

Nosso trabalho, portanto, apresenta implicações tanto para a Sociolinguística Variacionista quanto para o PLN. No primeiro caso, a criação do site vai favorecer o intercâmbio maior entre pesquisadores de outras regiões do país com a nossa, aumentando o interesse pela variedade sergipana do português brasileiro que ainda é pouco explorada, além

de também favorecer o uso das amostras do banco de dados Falares Sergipanos para profissionais da educação, podendo ser usado como recurso didático para se discutir a variação linguística.

Ainda a respeito da Sociolinguística Variacionista, a criação do organizador de dados poderá ajudar os pesquisadores a organizarem os dados que já foram coletados há mais tempo e precisaram de sistematização, como fizemos com a Amostra Deslocamentos 2019, que embora recente, tinha uma outra forma de arquivamento que não agrupavam arquivos seguindo um padrão mais detalhado. O protocolo (apêndice B) criado a partir do desenvolvimento da tese, pelo seu caráter procedural, é um produto metodológico que pode ser adaptado para outros pesquisadores na organização disseminação de dados linguísticos e buscas automáticas em suas amostras, sendo que todos os códigos criados aqui estão disponíveis no *Open Science Framework*, a partir do link: DOI 10.17605/OSF.IO/8XDZC.

Sobre a área de PLN, as implicações decorrem tanto em termos de etiquetagem quanto em termos de fornecimento de dados para melhora das ferramentas de anotação linguística automáticas para dados de fala. Sobre a etiquetagem, ressaltamos que verificamos inconsistências nas anotações devido a traços que são típicos da fala, por exemplo, as interjeições "Nossa", "Meus Deus" precisam de ajustes nos etiquetadores para conseguirem classificá-las como interjeições. Mais especificamente para o etiquetador do *LancsBox* 6.0 é preciso fazer um ajuste na etiqueta para a preposição "de" quando esta se aglutina com artigo, classificando-a como "SP+DA", assim como foi feito para as outras preposições aglutinadoras.

Apesar das implicações relatadas, este estudo também apresentou limitações. Para que generalizações maiores fossem feitas acerca dos usos das ferramentas para etapa pré-análise de dados, o ideal seríamos ter utilizado uma diversidade maior dos fenômenos. Além disso, seria importante também utilizar um conjunto maior de dados, embora tenhamos trabalhado com a análise de aproximadamente 10.000 palavras por etiquetador, é possível que o desempenho do *LancsBox* 6.0 seja melhor com mais dados.

Como sugestões de pesquisas futuras, recomendamos a replicação deste estudo para outros fenômenos, principalmente para aqueles de caráter semântico e discursivo. Além disso, com base em nossos dados, indicamos estudos que meçam a precisão e acurácia do uso de etiquetadores POS para dados de fala.

Referências

ACERVOS de dados abertos à sociedade: memória linguística e sociocultural e potencialidade de (re)uso. Mesa redonda apresentada por Maria Manuel Lopes de Figueiredo Costa Marques Borges, Ana Lígia Silva Medeiros, Marcia dos Santos Machado Vieira e Juliana Bertucci Barbosa. [s.l., s.n], 2021 1 vídeo (1h 06min 33s). Publicado pelo canal da Associação Brasileira de Linguística. Disponível em: https://www.youtube.com/watch?v=BsCvqcTo-qc&t=10s. Acesso em 21 set. 2021.

AGUIAR DE LIMA, T.; COSTA-ABREU, M. A survey on Automatic Speech Recognition systems for Portuguese language and its variations. **Computer Speech & Language**, p. 101055, 2020.

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

ALUÍSIO, S. M.; ALMEIDA, G. O que é e como se constrói um *corpus?* Lições aprendidas na compilação de vários corpora para pesquisa linguística. **Calidoscópio**, v. 4., n. 3, p.156-178, 2006. Disponível em: https://revistas.unisinos.br/index.php/calidoscopio/article/view/6002. Acesso em 16 mar. 2023.

ALUÍSIO, S. M.; PARDO, T. A. S.; DURAN, M. S. New Corpora for 'New' Challenges in Portuguese Processing. *In*: SARDINHA, T. B; FERREIRA, T. L. S. B. (Orgs.) **Working with Portuguese Corpora**. Nova Iorque: Bloomsburry Publishing, 2014. p. 303-322.

ALUÍSIO, S., PELIZZONI, J., MARCHI, A.R., DE OLIVEIRA, L., MANENTI, R., MARQUIAFÁVEL, V. 2003. An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese. In: Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language (PROPOR). **Proceedings [...]** PROPOR, 2003.

BECK, C.; BOOTH, H.; EL-ASSADY, M.; BUTT, M. Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias. In: DIPPER, S.; ZELDES, A. (Eds.) 14th Linguistic Annotation Workshop, 2020, Barcelona. **Proceedings [...]** Barcelona: Association for Computational Linguistics, 2020. p. 60-73. Disponível em: https://aclanthology.org/2020.law-1.0. Acesso em: 20 set. 2021.

ANDRADE, R. O. Resistência à Ciência. **Pesquisa Fapesp**, n. 282, p. 16-21, 2019.

ARAÚJO, A. A. o projeto norma oral do português popular de fortaleza NORPOFOR. *In*: XV CONGRESSO NACIONAL DE LINGUÍSTICA E FILOLOGIA, v. 15, n. 5, 2011, Rio de Janeiro. **Anais [...]** Rio de Janeiro: CiFEFiL, 2011. p. 835-845. Disponível em: https://url.gratis/h5Lgdm. Acesso em: 23 ago. 2019.

ARAUJO, S. S. F.; ALMEIDA, N. L. F. O Projeto A língua portuguesa no semiárido baiano – Fase 3: critérios de constituição e da amostragem do banco de dados. *In*: FREITAG, R. M. KO. (Org.). **Metodologia de Coleta e Manipulação de Dados em Sociolinguística**. São Paulo: Blücher, 2014. p. 27-48. Disponível em: https://url.gratis/fUwn0w. Acesso em: 05 ago. 2020.

BAILEY, G.; TILLERY, J. Some Sources of Divergent Data in Sociolinguistics. *In*: FOUGHT, C. (Ed.) **Sociolinguistic Variation: Critical Reflections**. Nova Iorque: Oxford University Press, 2004. p. 11-30.

BAKER, J. P. Consistency and accuracy in correcting automatically tagged data. *In:* GARSIDE, R.; LEECH, G.; McENERY, T. (Eds.) **Corpus Annotation: Linguistic information from computer text corpora**. 2ª Ed. Londres: Routledge, 2013.

BECK, C.; BOOTH, H.; EL-ASSADY, M.; BUTT, M. Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias. In: DIPPER, S.; ZELDES, A. (Eds.) 14th Linguistic Annotation Workshop, 2020, Barcelona. **Proceedings [...]** Barcelona: Association for Computational Linguistics, 2020. p. 60-73. Disponível em: https://aclanthology.org/2020.law-1.0. Acesso em: 20 set. 2021.

BEREZ-KROEKER, A. L. *et. al*. Reproducible research in linguistics: A position statement on data citationand attribution in our field. **Linguistics**, v. 56, n.1, p. 1-18. 2018. DOI https://doi.org/10.1515/ling-2017-0032. Disponível em: https://www.degruyter.com/document/doi/10.1515/ling-2017-0032/html. Acesso em: 22 set. 2021.

BICK. E. **The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. Tese (Doutorado) - Departamento de Linguística, Universidade de Aarhus, 2000.

BRANCO, A.; SILVA, J. Evaluating solutions for the rapid development of state-of-the-art pos taggers for Portuguese. In: LINO, M. T.; XAVIER, M. F.; FERREIRA, F.; COSTA, R.; SILVA R. (eds.), 4th International Conference on Language Resources and Evaluation (LREC'04), 2004, Paris. **Proceedings [...]** Paris: ELRA, 2004. p.507-510.

BRASIL. **Base Nacional Comum Curricular**. Brasília: MEC, 2017.

BRASIL. Lei nº 13.709 de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). Código Civil. **Diário Oficial da União:** seção 1, Brasília,DF, ano 139, n. 8, p. 1-74, 11 jan. 2002. PL 634/1975. Diário Oficial da União: seção 1, n. 157, p.59-64. Disponível em: encurtador.com.br/grBH4. Acesso em: 15 mai. 2020.

BRASIL. Ministério da Educação. Edital de convocação Nº 01/2021 – CGPLI. [EDITAL DE CONVOCAÇÃO PARA O PROCESSO DE INSCRIÇÃO E AVALIAÇÃO DE OBRAS DIDÁTICAS, LITERÁRIAS E PEDAGÓGICAS PARA O PROGRAMA NACIONAL DO LIVRO E DO MATERIAL DIDÁTICO - PNLD 2023. Diário Oficial da União, seção 3, Brasília, DF, n. 30, p. 47, 12 fev 2021.

BRASIL. **Parâmetros Curriculares Nacionais: Terceiro e quarto ciclos do ensino fundamental. Língua Portuguesa.** Brasília: MEC, 1998.

BRASIL. Resolução do Conselho Nacional da Saúde nº 510, de 07 de abril de 2016. **Diário Oficial [da] República Federativa do Brasil**, Poder Executivo, Brasília, DF, 24 maio 2016. Seção 1, p. 44-46. Disponível em: https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/22917581. Acesso em: 01 mar. 2020.

BREIMAN, L. CUTLER, A. **Manual--Setting Up, Using, And Understanding Random Forests** V4.0. 2004. Diponível em:https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf. Acesso em 27 jan. 2023.

BREIMAN, L. **Out-of-Bag estimation**. 1996. Disponível em: https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf. Acesso em: 24 jan. 2023

BREIMAN, L.; CUTLER, A.; LIAW, A.; WIENER, M. **Breiman and Cutler's Random Forests for Classification and Regression.** 2022. Disponível em: https://cran.r-project.org/web/packages/randomForest/index.html. Acesso em: 17 jan. 2023.

BREZINA, V.; WEILL-TESSIER, P.; MCENERY, A. 2021. **#LancsBox v. 6.0** [software]. Disponível em: http://corpora.lancs.ac.uk/lancsbox. Acesso em: 20 jun 2022.

BURKE, M.; CHELLIAH, S. L. Challenges to Representing Personal Names and Language Names in Language Archives: Examples from Northeast India. *In*: ZAVALINA, O. L.; CHELLIAH, S. L. (Eds.) International Workshop on Digital Language Archives, LanArc2021, 2021, Barcelona. **Proceedings […]** [s. l.] University of Texas, 2021. p. 44-46. Disponível em: http://hdl.handle.net/2142/111675. Acesso em: 04 out. 2021.
BYBEE, J. **Frequency of Use and the Organization of Language**. New York: Oxford University Press.

CALAMAI, S. FRONTINI, F. FAIR data principles and their application to speech and oral archives. **Journal of New Music Research**, v. 47, n. 4, 339–354, 2018. Disponível em: https://www.tandfonline.com/doi/full/10.1080/09298215.2018.1473449?scroll=top&needAccess=true. Acesso em: 30 jun. 2020.

CALLAGHAN, S. *et al*.Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres.International **Journal of Digital Curation**, v. 7, n. 1, p. 107-113. https://doi.org/10.2218/ijdc.v7i1.218. Disponível em: http://ijdc.net/index.php/ijdc/article/view/208. Acesso em: 10 mar. 2022.

CALLOU, D.; SILVA, G. M. O. O uso do artigo definido em contextos específicos. In: HORA, D. (Org.). **Diversidade Lingüística no Brasil**.João Pessoa: Idéia, 1997, p. 11-27.

CAMPOS JR., H. S. A variação morfossintática do artigo definido na capital capixaba. **PERcursos Linguísticos**, v. 2, n.5, p. 21-39. Disponível em: https://periodicos.ufes.br/percursos/article/view/3178. Acesso em: 23 dez 2022.

CANDITO, M.; LIBERMAN, M. Introduction to the special issue on annotated corpora. **Revue TAL**, ATALA, v. 2, n. 60, p. 7-17, 2019. Disponível em: https://url.gratis/qEY0TQ. Acesso em: 15 jul. 2021.

CARDOSO, P. B. **Efeitos linguísticos e paralinguísticos na inferência dos sentidos indicados por (eu) acho que em entrevistas sociolinguísticas**. 2021. Dissertação (Mestrado em Letras) – Programa de Pós-graduação em Letras, Universidade Federal de Sergipe, 2021.

CARLETTA, J. Assessing agreement on classification tasks: The Kappa Statistic. **Computational Linguistics**, v. 2, n. 22, p. 249-254, 1996. Disponível em: https://aclanthology.org/J96-2004. Acesso em: 11 out. 2021.

CARVALHO, A. M. **Corpus del Español en el Sur de Arizona (CESA)**. University of Arizona. 2012. Disponível em: cesa.arizona.edu. Acesso em: 12 set. 2020.

CASTILHO, A. T. Gramática do português brasileiro: fundamentos, perspectivas. **Cadernos de Linguística**, v. 2, n. 1, p. 01-17, 2021. Disponível em: https://cadernos.abralin.org/index.php/cadernos/article/view/252. Acesso em 08 out. 2021.

CASTILHO, A.T. **Nova gramática do português brasileiro**. São Paulo: Contexto, 2010.

CESARINO, L. Pós-verdade e a crise do sistema de peritos: uma explicação cibernética. **Ilha Revista de Antropologia**, v. 23, n. 1, p. 73-96, 2021. Disponível em: https://periodicos.ufsc.br/index.php/ilha/article/view/75630. Acesso em: 16 mar. 2023.

CHAMPIEUX, R.; COATES, H. L. Metrics for evaluating the impact of data sets. *In*: Berez-Kroeker, A. L.; McDonnell, B.; Koller, E.; Collister, L. B. (Eds.). **The Open Handbook of Linguistic Data Management**. The MIT Press, 2022. DOI: https://doi.org/10.7551/mitpress/12200.001.0001. Disponível em: https://direct.mit.edu/books/book/5244/The-Open-Handbook-of-Linguistic-Data-Management. Acesso em: 10 mar. 2022.

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.

COLLISCHONN, G.; MONARETTO, V. O. Banco de dados V ARSUL: a relevância de suas características e a abrangência de seus resultados. **ALFA: Revista de Linguística**, v.56, n.3, p. 835-853, 2012. Disponível em: https://periodicos.fclar.unesp.br/alfa/article/view/4953. Acesso em: 20 abr. 2019.

COLLISTER, L. B. Copyright and sharing linguistic data. *In*: Berez-Kroeker, A. L.; McDonnell, B.; Koller, E.; Collister, L. B. (Eds.). **The Open Handbook of Linguistic Data Management**. The MIT Press, 2022. DOI: https://doi.org/10.7551/mitpress/12200.001.0001. Disponível em: https://direct.mit.edu/books/book/5244/The-Open-Handbook-of-Linguistic-Data-Management. Acesso em: 15 mar. 2022.

CORREA, T. R. A. **A variação na realização de /t/ e /d/ na comunidade de práticas da UFS: mobilidade e integração**. Dissertação (Mestrado em Estudos Linguísticos) – Universidade Federal de Sergipe, 2019.

COSTA JÚNIOR, J. C. Compostos de Discurso Direto na fala: pesquisa no C-Oral. **Movendo Ideias**, v. 23, p. 54-64, 2018. Disponível em: https://url.gratis/jLPYX0. Acesso em: 12 set. 2021.

DAWLE, M. *et. al*. **data.table: Extension of 'data.frame'**. 2022. Disponível em: https://cran.r-project.org/web/packages/data.table/index.html. Acesso em 17 jan. 2023.

DIAS, Z. **MC102 – Aula 15 Expressões Regulares**. Campinas: Instituto de Computação, 2021. Disponível em: https://ic.unicamp.br/~mc102/aulas/aula15.pdf. Acesso em 19 out. 2021.

DUBLIN CORE. Disponível em: https://www.dublincore.org/. Acesso em: 20 ago. 2022.

DURAN *et al.* Manual de anotação como recurso de Processamento de Linguagem Natural: o modelo Universal Dependencies em língua portuguesa. **Domínios de Lingu@gem**, v. 16, n. 4, p. 1608–1643, 2022. DOI: 10.14393/DL52-v16n4a2022-13. Disponível em: https://seer.ufu.br/index.php/dominiosdelinguagem/article/view/63632. Acesso em: 24 abr. 2023.

FEAGIN, C. Entering the Community Fieldwork. *In*: CHAMBERS, J. K.; SCHILLING, N. (Eds.) **The handbook of language variation and change**. 2ª Ed. Malden: Wiley-Blackwell, 2013. p. 19-37.

FIDLER, F.; WILCOX, J. Reproducibility of Scientific Results. *In*: ZALTA, E. N. (Ed.) **The Stanford Encyclopedia of Philosophy**. Disponível em: https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/. Acesso em: 08 jul. 2019.

FREITAG, R. M. F. **Redução de variáveis e de dados**. 2021a. Disponível em: https://rkofreitag.github.io/reducao.html/. Acesso em 20 jan. 2023.

FREITAG, R. M. K., Linguistic Repositories as Asset: Challenges for Sociolinguistic Approach in Brazil. In: In: ZAVALINA, O. L.; CHELLIAH, S. L. (Eds.) International Workshop on Digital Language Archives, LanArc2021, 2021, Barcelona. **Proceedings [...]** [s. l.] University of Texas, 2021b. p. 33-35. Disponível em: http://hdl.handle.net/2142/111675. Acesso em: 04 out. 2021.

FREITAG, R. M. K. Kappa statistic for judgment agreement in Sociolinguistics. **Revista de Estudos da Linguagem**, v. 27, n.4, p.1591-1612, 2019. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/14737. Acesso em: 10 out. 2021.

FREITAG, R. M. K. (Org.). **Metodologia de coleta e manipulação de dados em sociolinguística**. São Paulo: Editora Edgard Blücher, 2014.

FREITAG, R. M. K. Amostras sociolinguísticas: probabilísticas ou por conveniência? **Revista de Estudos da Linguagem**, v. 26, n. 2, p. 667-686, 2018a. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/12412/0. Acesso em: 10 set. 2021.

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

FREITAG, R. M. K. **Documentação sociolinguística: coleta de dados e ética em pesquisa**. São Cristóvão: Editora UFS, 2017a Disponível em:

https://www.livraria.ufs.br/produto/documentacao-sociolinguistica-coleta-de-dados-e-etica-em-pesquisa/. Acesso em: 10 jun. 2020.

FREITAG, R. M.K. A dadidade (ou dadidão) do dado, **Linguística** Rio, v.3, n.1, p.1-10, 2017b

FREITAG, R.M. K. Falares sergipanos. In: ATAÍDE, C. et al. **Gelne 40 anos**. São Paulo: Blucher, 2017c.

FREITAG, R. M. K. Sociolinguística no/do Brasil. **Cadernos de Estudos Linguísticos**, v. 58, n. 3, p. 445-460, 2016.

FREITAG, R. M. K. 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: 15 mar. 2019.

FREITAG, R. M. K. *et. al*. Desafios da gestão de dados linguísticos e a Ciência Aberta. **Cadernos de Linguística**, v. 2, n. 1, p. 1-19, 2021. Disponível em: https://cadernos.abralin.org/index.php/cadernos/article/view/307. Acesso em: 10 jul. 2021

FREITAG, R. M. K.; MARTINS, M. A.; TAVARES, M. A. FREITAG, R. M. K.; MARTINS, M. A.; TAVARES, M. A. Banco de dados sociolinguísticos do português brasileiro e os estudos de terceira onda: Potencialidades e limites. **ALFA: Revista de Linguística**, v.56, n.3, p. 917-944, 2012. Disponível em: https://www.scielo.br/pdf/alfa/v56n3/a09v56n3. Acesso em: 16 set. 2020.

FREITAG, ROST-SNICHELLOTO - FREITAG, R. M. K.; ROST SNICHELOTTO, C. A. Análises contrastivas: estabilidade, variedade ou metodologia?. **Working Papers em Linguística**, v. 16, n. 1, p. 157-167, 2015. Disponível em: https://periodicos.ufsc.br/index.php/workingpapers/article/view/1984-8420.2015v16n1p157. Acesso em 01 dez 2021.

FREITAG; SANTANA; ANDRADE, 2014 - FREITAG, R. M. K.; SANTANA, C. C.; ANDRADE, T. R. C. Práticas constitutivas do povoado Açuzinho. **Revista Ambivalências**, v. 2, n. 3, p.194-217, 2014. Disponível em: https://seer.ufs.br/index.php/Ambivalencias/article/view/3129. Acesso 12 set. 2021.

FRIDLAND, V.; KENDALL, T. Managing sociophonetic data in a study of regional variation. *In*: Berez-Kroeker, A. L.; McDonnell, B.; Koller, E.; Collister, L. B. (Eds.). **The Open Handbook of Linguistic Data Management**. The MIT Press, 2022. DOI: https://doi.org/10.7551/mitpress/12200.001.0001. Disponível em: https://direct.mit.edu/books/book/5244/The-Open-Handbook-of-Linguistic-Data-Management. Acesso em: 19 mar. 2022.

GALVES, C.; ANDRADE, A. L.; FARIA, P. **Tycho Brahe Parsed Corpus of Historical Portuguese**. 2017.

GAMALLO, P. et al. Análisis morfosintáctico y clasificación de entidades nombradas en un entorno Big Data Procesamiento del Lenguaje Natural. **Procesamiento del Lenguaje Natural**, v. 53, p.17-24, 2014. Disponível em:

http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5046. Acesso em: 03 mar. 2021.

GAMALLO, P.; GARCIA, M. **FreeLing e TreeTagger:um estudo comparativo no âmbito do Português** (Relatório Técnico). 2014. Disponível em: https://gramatica.usc.es/~gamallo/artigos-web/PROLNAT_Report_01.pdf. Acesso em: 11 set. 2022.

GARELLEK, M. et al. Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls. **Journal of Speech Science**, v. 9, n. 1, p.3-16, 2020. Disponível em: https://halshs.archives-ouvertes.fr/halshs-02894375/. Acesso em: 01 jun. 2021.

GILMORE, R.; KENNEDY, J. L.; ADOLPH, K. E. Practical solutions for sharing data and materials from psychological research. **Adv Methods Pract Psychol Sci**, v. 1, n. 1, p. 121-130, 2018. Disponível em: https://journals.sagepub.com/doi/full/10.1177/2515245917746500. Acesso em: 01 jun. 2021.

GONÇALVES, S. C. L. Projeto ALIP (Amostra Linguística do Interior Paulista) e banco de dados Iboruna: 10 anos de contribuição com a descrição do português brasileiro. **Estudos Linguísticos**, v. 48, n. 1, p. 276-29, abr. 2019. Disponível em: https://revistas.gel.org.br/estudos-linguisticos/article/view/2430/1503. Aceso em: 25 abr. 2019.

GRIES, S.T.; BEREZ, A. L. Linguistic annotation in/for corpus linguistics.*In*: IDE, N.; PUSTEJOVSKY, J. **Handbook of linguistic annotation**. Dordrecht: Springer, 2017. p.379-409.

GUEDES, S. Emprego do artigo definido em situação de contato dialetal: um estudo da fala de migrantes paraibanos em São Paulo. **Domínios de Lingu@gem**, Uberlândia, v. 13, n. 4, p. 1401–1432, 2019. https://doi.org/10.14393/DL40-v13n4a2019-4-v13n4a2019-4. Disponível em: https://seer.ufu.br/index.php/dominiosdelinguagem/article/view/46873. Acesso em: 23 dez. 2022.

HIRSCHBERG, J.; MANNING, C. D. Advances in natural language processing. **Science**, v. 349, n. 6245, p. 261-266, 2015. Disponível em: https://www.science.org/doi/10.1126/science.aaa8685. Acesso em: 12 jul. 2021.

HONNIBAL, M. *et al*. **spaCy: Industrial-strength Natural Language Processing in Python**. 2020. Disponível em: https://zenodo.org/record/4091419/export/hx#.YXKr057MLIU. Acesso em: 10 nov. 2020.

HOTHORN, T.; SEIBOLD, H.; ZEILEIS, A. **partykit: A Toolkit for Recursive Partytioning**. 2022. Disponível em: https://cran.r-project.org/web/packages/partykit/index.html. Acesso em: 11 out. 2022.

HOVY, E.; LAVID, J. Towards a ‘science’of corpus annotation: a new methodological challenge for corpus linguistics. **International journal of translation**, v. 22, n. 1, 2010. Disponível em: https://url.gratis/Pvic9q. Acesso em: 30 mai. 2021.

IDE, N. Introduction: The handbook of Linguistic Annotation. In: IDE, N.; PUSTEJOVSKY, J. **Handbook of linguistic annotation**. Dordrecht: Springer, 2017 p. 1-21.

JURAFSKY, D. MARTIN, J. H. Introduction. *In*: JURAFSKY, D. MARTIN, J. H. **Speech and language processing: an introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**.2ª Ed. Upper Saddle River: Pearson Prentice Hall, 2009. p. 1-16.

JUSTICE as a linguistic matter. Conferência apresentada por William Labov. [s.l., s.n], 2020. 1 vídeo (1h 06min 33s). Publicado pelo canal da Associação Brasileira de Linguística. Disponível em: https://www.youtube.com/watch?v=cr5tyw8_gT0&t=2231s. Acesso em: 23 mai. 2020.

KENDALL, T. Data in the Study of Variation and Change. CHAMBERS, J. K.; SCHILLING, N. (Eds.). **The handbook of language variation and change**. 2ª Ed. Malden: Wiley-Blackwell, 2013. p. 38-56.

KENDALL, T. On the History and Future of Sociolinguistic Data. **Language and Linguistics Compass**, v. 2, n. 2, p. 332-351, 2008. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/j.1749-818X.2008.00051.xAcesso em 20 mai. 2019.

KENDALL, T.; FRENCH, A. Digital Audio Archives, Computer-Enhanced Transcripts, and New Methods in Sociolinguistic **Analysis. Digital Humanities**, Paris, Jul., p. 110-112, 2006. Disponível em: https://slaap.chass.ncsu.edu/pdfs/DH2006_pp110-112.pdf. Acesso em: 20 mai. 2019.

KHUN, D.; ABARCA, E.; NUNES, M. G. V. **Corpus NILC - Situação em Maio/2000**. São Carlos: ICMC-USP, 2000 (Relatório Técnico NILC-TR-00-07).

KÜBLER, S.; ZINSMEITER, H. Linguistic Annotation. *In*: KÜBLER, S.; ZINSMEITER, H. **Corpus linguistics and linguistically annotated corpora**. Londres: Bloomsburry Publishing, 2015. p. 43-156.

KUMAR, D.; GURPREET, S. J. Part of speech taggers for morphologically rich indian languages: a survey. **International Journal of Computer Applications**, v. 6, n.5, p. 32-41, 2010.
KUN, M. *et. al*. **caret: Classification and Regression Training**. 2022. Disponível em: https://cran.r-project.org/web/packages/caret/index.html. Acesso em: 17 jan. 2023.

LABOV, W. **Padrões sociolinguísticos**. Trad. Marcos Bagno, Maria Marta Pereira Scherre, Caroline Rodrigues Cardoso. São Paulo: Parábola Editorial, 2008. [1972]

LABOV, W. **The social stratification of English in New York city**. Cambridge University Press, 2006. [1966]

LABOV, W. **Principles of linguistic change: Social factors**. Malden: Blackwell, 2001. v.2

LABOV, W. **Principles of linguistic change: Internal Factors**. Malden: Blackwell, 1999 v.1.

LABOV, W. Field methods of the project on linguistic change and variation. **Sociolinguistic working paper**, n.81, p. 1- 43, 1981. Disponível em: https://eric.ed.gov/?id=ED250938. Acesso em: 13 jul. 2021.

LABOV, W. **Where Does the Linguistic Variable Stop? A Response to Beatriz Lavandera**. Working Papers in Sociolinguistics, n. 44, p. 1-17, 1978. Disponível em: https://eric.ed.gov/?id=ED157378. Acesso em: 26 out. 2021.

LABOV, W. Contraction, deletion, and inherent variability of the English copula. **Language**, v. 45, n. 4, p. 715-762, 1969. Disponível em: https://www.jstor.org/stable/412333. Acesso em: 13 jul. 2021.

LEECH, G. Adding Linguistic Annotation. *In*: WYNNE, M. (Ed.) *Developing Linguistic Corpora—A Guide to Good Practice*. [s. l.] Arts and Humanities Data Service, 2004. Disponível em: https://users.ox.ac.uk/~martinw/dlc/chapter2.htm. Acesso em: 23 set 2020.

LEECH, G. Introducing corpus annotation. *In*: LEECH, G.; GARSIDE, R. E.; MCENERY, T. **Corpus annotation: Linguistic information from computer text corpora**. 2ªEd. Londres: Routledge, 2013. p. 1-18.
LINGUISTIC DATA CONSORTIUM. **Data management**. Disponível em: https://www.ldc.upenn.edu/data-management. Acesso em: 29 ago. 2020.

OLIVEIRA, A. J. **PORTAL - Variação Linguística no Português Alagoano**. 2017. Disponível em: https://www.portuguesalagoano.com.br/p/inicio.html?m=1. Acesso em: 10 ago. 2021.

LYON, L. Transparency: the emerging third dimension of Open Science and Open Data. **Liber quarterly**, v. 25, n. 4, p. 153-171, 2016. Disponível em: https://liberquarterly.eu/article/view/10759. Acesso em: 19 set 2021.

MACHADO VIEIRA, M. S. *et. al*. **Plataforma da Diversidade Linguística Brasileira**. Projeto apresentado à Pró-Reitoria de Pós-Graduação e Pesquisa da UFRJ e à Fundação Universitária José Bonifácio, em razão do Edital BNDES - Chamada Pública para seleção de propostas no âmbito da iniciativa Resgatando a História No. 01/2021, agosto de 2021.

MELLO, H. Trabalhando com dados de fala: a experiência do Projeto C-Oral-Brasil. *In*: BRESCANCINI, C. R. **Projeto VARSUL: Variação Linguística no Sul do País 36 anos**. Porto Alegre: Zouk Editora. p. 41-69.

MENDES, R.B. **SP2010 – Construção de uma amostra da fala paulistana**. Projeto regular apresentado à FAPESP (Processo FAPESP 2011/09278-6). 2011. Disponível em http://projetosp2010.fflch.usp.br/producao-bibliografica. Acesso em 25 set. 2021.

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. Disponível em: https://ri.ufs.br/jspui/handle/123456789/5694. Acesso em: 10 out. 2021.

Oez (2018) - OEZ, M. A guide to the documentation of the Beth Qustan dialect of the Central Neo-Aramaic Language Turoyo. **Language Documentation and Conservation**, v. 12, p. 339-358, 2018. Disponível em: https://scholarspace.manoa.hawaii.edu/handle/10125/24773. Acesso em: 15 set 2020.

OLIVEIRA JR, M. NURC Digital: Um protocolo para a digitalização, anotação, arquivamento e disseminação do material do Projeto da Norma Urbana Linguística Culta (NURC). **CHIMERA. Romance Corpora and Linguistic Studies**, v. 3, n.2, 2016. Disponível em: https://revistas.uam.es/index.php/chimera/article/view/6519/6908. Acesso em: 19 jun. 2019.

OTHERO, G. A.; AYRES, M. R. Anotação morfológica automática de corpus de língua falada: desafios ao Aelius. **Texto Livre: linguagem e tecnologia**, Belo Horizonte, 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.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. P. F. **e-Dictor**. Versão 1.0 beta 10, 2013. Programa de Computador. Disponível em: https://edictor.net/download. Acesso em: 05 out. 2021.

PALMER, M.; XUE, N. Linguistic annotation. *In*: CLARK, A.; FOX, C.; LAPPIN, S. **Handbook of Computational Linguistics and Natural Language Processing**. Malden: Wiley-Blackwell, 2010. p. 238-270.

PARDO, T. A. S.; NUNES, M. G. V. DiZer an Automatic Discourse Analyzer for Brazilian Portuguese.XVII BRAZILIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE-SBIA.2004. São Luís. **Anais [...]**. São Luís. 1-10. Disponível em: https://sites.icmc.usp.br/taspardo/CTDIA06-PardoNunes.pdf. Acesso em 10 out. 2021.

PAROUBEK, P. Evaluating Part-of-Speech Tagging and Parsing: On the evaluation of automatic parsing of natural language. In: Dybkjær, L., Hemsen, H., Minker, W. (eds) **Evaluation of Text and Speech Systems. Text, Speech and Language Technology**, vol 37. Dodrecht: Springer, 2007. p. 99-124. https://doi.org/10.1007/978-1-4020-5817-2_4. Disponível em: https://link.springer.com/chapter/10.1007/978-1-4020-5817-2_4. Acesso em 19 set. 2020.

PRESEEA: Corpus del Proyecto para el estudio sociolingüístico del español de España y de América. (2014) Alcalá de Henares: Universidad de Alcalá. Disponível em: http://preseea.linguas.net. Acesso em: 15 set. 2020.

RADMAKER, A. *et al.* Universal Dependencies for Portuguese. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling). 2017. Pisa, Itália. **Proceedings […]**. Pisa. 197-206. Disponível em: http://aclweb.org/anthology/W17-6523. 21 set 2021.

RIBEIRO, C. C. S. **Deslocamento geográfico e padrões de uso linguístico: a variação entre as preposições locativas em ~ ni na comunidade de práticas da Universidade Federal de Sergipe**, 2019. 80 f. Dissertação (Mestrado em Letras). Centro de Educação e Ciências Humanas, Universidade Federal de Sergipe.

RODRIGUES, R.; SOUZA, J. W. C.; SANTOS, R. L. S. Descrição linguística e aprendizado de máquina. **Cadernos De Estudos Linguisticos**, v. 64, P. 1-15, 2022. DOI: 10.20396/cel.v64i00.8666995. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/cel/article/view/8666995. Acesso em: 15 mar. 2023.

SALOMÃO, A.C. Variação e mudança linguística: panorama e perspectivas da Sociolinguística Variacionista no Brasil. **Fórum Linguístico**, v. 8, n.2, p. 187-207, 2011. Disponível em: https://periodicos.ufsc.br/index.php/forum/article/view/1984-8412.2011v8n2p187. Acesso em: 23 ago. 2019.

SANTANA, R. R **Tipos de tipo em uma comunidade de práticas universitária. 2019.** Dissertação (Mestrado em Letras) – Universidade Federal de Sergipe, São Cristóvão, 2019. Disponível em: https://ri.ufs.br/handle/riufs/11481. Acesso em: 12 jan. 2021.

SANTOS, D.; ROCHA, P. Evaluating CETEMPúblico, a free resource for Portuguese. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. 2001. Toulouse. **Proceedings [...]**. Toulouse, 2001. p. 442-449.

SARDINHA, T. B. Lingüística de Corpus: Histórico e problemática. **DELTA.**, São Paulo, v. 16, n.2, p. 323-367, 2000. Disponível em: https://www.scielo.br/j/delta/a/vGknQkZQGsGYbrQfKmTZY4s/?lang=pt. Acesso em 15 abr. 2021.
SARDINHA, T.B. **Linguística de corpus**. Barueri: Editora Manole Ltda, 2004.

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

SEDRINS, A. P.; PEREIRA, D. K.; SILVA, C. R. A função sintática e o licenciamento de artigos definidos diante de antropônimos e de possessivos pré-nominais: um estudo com dados de fala em Carnaíba – Pernambuco. **Domínios de Lingu@gem**, Uberlândia, v. 13, n. 3, p. 1266–1295, 2019. https://doi.org/10.14393/DL39-v13n3a2019-17. Disponível em: https://seer.ufu.br/index.php/dominiosdelinguagem/article/view/42060. Acesso em: 24 dez. 2022.

SILVA, F.C.C.; SILVEIRA, L. O ecossistema da Ciência Aberta. **Transinformação**, v.31, e190001, 2019. http://dx.doi.org/10.1590/2318-889201931e190001. Disponível em: https://www.scielo.br/pdf/tinf/v31/2318-0889-tinf-31-e190001.pdf. Acesso em: 03 jul. 2020.

SILVA, E. V. Bancos de dados sociolinguísticos em português. **Idioma**, Rio de Janeiro, nº. 29, p. 168-180, 2º. Sem. 2015. Disponível em: http://www.institutodeletras.uerj.br/idioma/numeros/29/Idioma29_a04.pdf. Acesso em: 23 abr. 2019.

SILVA, G. M. O. Coleta de dados. In: MOLLICA, M. C.; BRAGA, M. L. (Orgs). **Introdução à Sociolinguística**. São Paulo: Contexto, 2004. p. 117-134.

SILVA, J. M. S. **Variação no preenchimento da posição determinante antes de possessivos pré-nominais: padrões dialetais e contatos**. 2020. Dissertação (Mestrado em Letras) – Universidade Federal de Sergipe, São Cristóvão, 2020.

SILVA, L. F. L. Desenvolvimento do conector 'na hora que' na Língua Portuguesa: uma análise qualitativa sob uma perspectiva construcional. **SIGNO Y SEÑA**, v. 32, p. 123-136, 2017. Disponível em: https://url.gratis/tWhoyP. Acesso em: 16 set. 2021.

SILVA, L. S. **Análise acústica ou de oitiva? Contribuições para o estudo da palatalização em Sergipe**. Dissertação (Mestrado em Estudos Linguísticos) – Universidade Federal de Sergipe, 2021.

SILVA, R. G.; COELHO, I. M. W. S. Metodologia de criação de um banco de dados linguísticos: desafios e contribuições para o processo de ensino-aprendizagem. Educitec- **Revista de Estudos e Pesquisas sobre Ensino Tecnológico**, v. 6, p. 1-15, 2020. Disponível em: http://200.129.168.14:9000/educitec/index.php/educitec/article/view/905. 10 out. 2021.

SINCLAIR, J. Current issues in corpus linguistics. *In*: SINCLAIR, J.; CARTER, R. **Trust the text: Language, corpus and discourse**. Londres: Routledge, 2004b. p. 185-193.

SINCLAIR, J. Intuition and annotation–the discussion continues. *In*: AIJMER, K.; ALTENBERG, B. (Eds.) **Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora** (ICAME 23). Göteborg: Brill, 2004. p.39-59. Disponível em: https://brill.com/view/title/30097. Acesso em 30 set. 2021.

SIQUEIRA, M. Efeitos do contato entre normas na variação linguística: a presença de artigo definido antecedendo possessivos no falar universitário da UFS. **Revista Porto das Letras**, v.6. n.1, p.-8-33, 2020. Disponível em: https://sistemas.uft.edu.br/periodicos/index.php/portodasletras/article/view/8324. Acesso em 20 dez 2022.

SOUSA, M. D. A. F. et al. Lidando com grande volume de dados: o processo de sistematização e divulgação de uma amostra do banco de dados Falares Sergipanos. Disponível em: <osf.io/8xdzc>. Acesso em 08 dez 2022.

SOUSA, M. D. A. F.; SOUZA., V. R. A. Transcrição e anotação de dados linguísticos usando as ferramentas ELAN e LancsBox. **Domínios de Lingu@gem**, v.16, n. 3, p.1173-1202, 2022. DOI https://doi.org/10.14393/DL51-v16n3a2022-10 Disponível em: https://seer.ufu.br/index.php/dominiosdelinguagem/article/view/62447. Acesso em 20 set 2022.

TAGLIAMONTE, S. A. **Variationist sociolinguistics: Change, observation, interpretation.** Malden: Wiley-Blackwell, 2012.

TAGLIAMONTE, S. A. Data, data and more data. *In*: TAGLIAMONTE, S. A. **Analysing sociolinguistic variation**. Cambridge: Cambridge University Press; 2006. p.50-69.

TAVARES, M. A.; MARTINS, M. A. O banco de dados Fala-Natal: uma agenda de trabalho. *In*: FREITAG, R. K. (Org.). **Metodologia de Coleta e Manipulação de Dados em Sociolinguística**. São Paulo: Blücher, 2014. p. 71-78. Disponível em: https://url.gratis/fUwn0w. Acesso em: 05 ago. 2020.

VANN, R. E. Best Practices for Information Architecture, Organization, and Retrieval in Digital Language Archives within University Institutional Repositories. *In*: ZAVALINA, O. L.; CHELLIAH, S. L. (Eds.) International Workshop on Digital Language Archives, LanArc2021, 2021, Barcelona. **Proceedings [...]** [s. l.] University of Texas, 2021. p. 36-39. Disponível em: http://hdl.handle.net/2142/111675. Acesso em: 04 out. 2021.

WEINREICH, U.; LABOV, W.; HERZOG, M. Empirical Foundations for a Theory of Language Change. In: LEHMANN, W. P.; MALKIEL, Y. (EDS.) **Directions for Historical Linguistics: A Symposium**. Austin: The University of Texas Printing Division, 1968. p. 95-188. Disponível em: https://liberalarts.utexas.edu/lrc/resources/books/directions/index.php. Acesso em: 02 jan 2021.
WICKHAM, H.tidyverse: **Easily Install and Load the 'Tidyverse'**. 2022. Disponível em: https://cran.r-project.org/web/packages/tidyverse/index.html. Acesso em: 11 out. 2022.

WILKINSON, M., et al. The FAIR Guiding Principles for scientific data management and stewardship. **Sci Data**, v.3, n. 160018, 2016. DOI https://doi.org/10.1038/sdata.2016.18. Disponível em: https://www.nature.com/articles/sdata201618. Aceso em: 30 jun. 2020.

WITTEN, H. I.; FRANK, E.; HALL, M. A. Ensemble learning. *In*: WITTEN, H. I.; FRANK, E.; HAL, M. A. **Data Mining: Practical Machine Learning tools and techniques**. 3ª Edição. Morgan Burlington: Kauffman Publishers, 2011. p. 351-374. Disponível em: https://www.wi.hs-wismar.de/~cleve/vorl/projects/dm/ss13/HierarClustern/Literatur/WittenFrank-DM-3rd.pdf. Acesso em: 23 jan. 2023

APÊNDICES

APÊNDICE A – Termo de autorização

Universidade Federal de Sergipe
Condomínio de Laboratórios Multiusuários de Informática e Documentação (LAMID)
Av. Marechal Rondon, s/n, Jd. Rosa Elze
São Cristóvão/SE - CEP 49100-000

Termo de autorização de uso

Para que os usuários possam receber os dados requisitados, eles deverão preencher e concordar com as condições estabelecidas neste termo de autorização de uso.

Eu, _____ (nome completo), _____ (afiliação), doravante denominado (a) como usuário(a), concordo com os termos e condições dispostos nesta autorização para utilização dos *corpora* nomeados abaixo como "*corpora* recebidos".

O(A) usuário(a) receberá as mídias referentes aos *corpora* recebidos contendo _____ (descrição dos arquivos). O usuário concorda em utilizar os *corpora* recebidos somente para propósitos não-comerciais em pesquisa, educação e desenvolvimento de tecnologias. No caso de o uso dos *corpora* resultar em desenvolvimento de um produto comercial, o usuário deverá formalizar a filiação do LAMID como recebedor de lucros provenientes de tal produto, isentando o LAMID de quaisquer ônus advindos de tal comercialização. A utilização de excertos dos dados em artigos, relatórios e outros documentos descrevendo os resultados do uso não-comercial dos *corpora* recebidos deve ser limitada. O usuário(a) não poderá copiar, redistribuir, transmitir, publicar ou usar os dados dos *corpora* recebidos para qualquer outra finalidade.

É obrigatório que o(a) usuário(a) faça a devida referência dos dados dos *corpora* recebidos em publicações científicas toda vez que eles forem citados.

O LAMID, a Universidade Federal de Sergipe bem como os coordenadores das pesquisas pelas quais foram gerados os dados se isentam de qualquer responsabilidade sobre problemas ocasionados pela incompatibilidade dos dados com os propósitos de pesquisa pelos quais os *corpora* recebidos foram requisitados pelo(a) usuário(a).

O(A) usuário(a) deverá assinar este documento e enviá-lo via e-mail para: lamid@academico.ufs.br.

São Cristóvão, _____ de 20_____.

_____
Assinatura do usuário

APÊNDICE B– Protocolo

PROTOCOLO PARA ANOTAÇÃO LINGUÍSTICA E GERENCIAMENTO DE DADOS DO BANCO DE DADOS FALARES SERGIPANOS

Área do conhecimento: Sociolinguística

Recursos a serem utilizados:
- Software: ELAN; Google Colaboratory; Jupyter; Google Drive.
- Linguagem de programação: Python, biblioteca spaCy.

Autora: Marta Deysiane Alves Faria Sousa.

Resumo: Este protocolo tem por objetivo descrever os procedimentos para a anotação linguística e o gerenciamento de dados do banco de dados Falares Sergipanos. Estão descritas instruções para: i) encontrar as normas de utilização do ELAN e de transcrição ortográfica; ii) fazer anotação POS (gramatical) nos dados das amostras utilizando a biblioteca spaCy do Python; iii) alinhar transcrição com a anotação linguística no ELAN; iv) documentar os metadados dos arquivos; v) documentar os metadados de amostras coletadas; vi) hierarquizar os arquivos nas pastas do Repositório GELINS. No caso de replicação deste protocolo, para se ter acesso aos links de documentos do GELINS, deve ser feito um pedido por e-mail para: lamid.ufs@gmail.com.

Palavras-chave: POS tagging. Dados de Fala. spaCy. Sociolinguística. Falares Sergipanos.

1. Antes de começar todo o processo

1. Tenha certeza de que o áudio está claro, ou seja, de que as falas dos indivíduos estejam em um bom volume e que, na presença de ruídos, estes não impeçam o entendimento das falas.
2. Caso o áudio não seja audível o suficiente, existe a possibilidade de melhorá-lo utilizando o software Audacity. A partir deste tutorial, é possível aprender a mexer no áudio.
3. Tenha a ficha social em mãos. Será por meio dela que os metadadados da amostra serão redigidos.
4. Tenha as normas de transcrição para lhe ajudar a transcrever. Lembre-se de nomear todos os arquivos da mesma forma.

2. Transcrição dos dados

Os procedimentos de utilização do ELAN em língua portuguesa para o banco de dados Falares Sergipanos estão disponíveis aqui. Observe que o artigo servirá apenas para referência de como usar o ELAN. Para realizar a transcrição ortográfica utilize as normas de transcrição.

3. Anotação linguística:

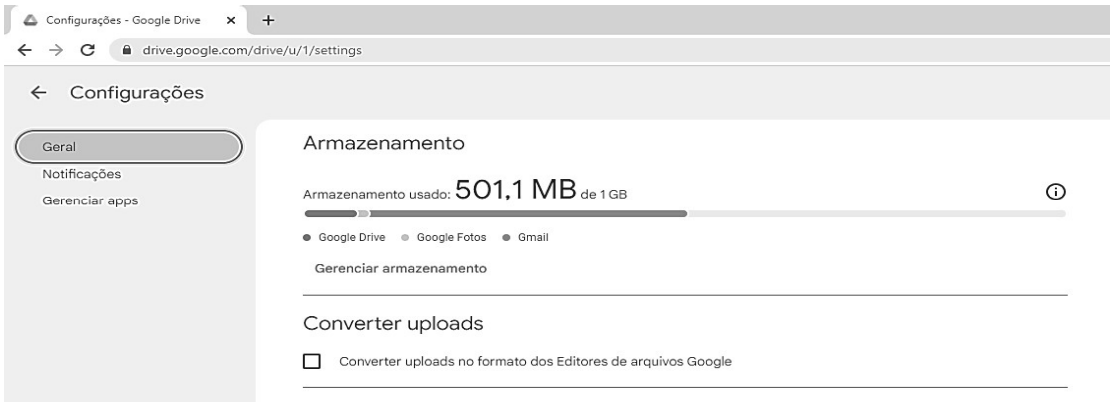
Existem duas formas de realizar a anotação linguística dos dados utilizando a biblioteca spaCy da linguagem Python: de forma *online* pelo Google Colaboratory ou *offline* por meio da instalação de uma IDE.

3.1 Anotação linguística online:

1. Faça uma conta no *Google.* Utilize o *Drive* (nuvem do *Google*) de sua conta para salvar uma cópia dos arquivos em .txt das transcrições em uma única pasta no Google Drive. Caso o Google

Drive salve seus arquivos em formato gdoc, você deve desabilitar a função de salvamento em formato dos *Editores de arquivos Google* automático. Conforme figura abaixo, em que a parte converter uploads está desabilitada.

Figura 1 – Desabilitando o salvamento em formato .gdoc.



Fonte: elaboração própria.

2. Acesse o Google Colaboratory (doravante Colab), utilizando a sua conta pessoal.

3. Clique em "Arquivo">> "Novo *notebook*" (figura 2).

Figura 2 – Criando um *Notebook* no Google Colaboratory.

Fonte: elaboração própria.

4. Aparecerá a tela da figura 3. A linha com os colchetes é uma linha de códigos. Clique uma vez com o cursor nela e será possível digitar o código. "+Código" é a opção para acrescentar mais uma linha de código.

Figura 3 – Digitando códigos no Google Colaboratory.

Fonte: elaboração própria.

5. Digite os 5 códigos (quadro 1), cada um em uma linha de comando diferente para instalar o spaCy (*a*, *b* e *c*) e as bibliotecas complementares (*d* e *e*) que serão utilizadas. Para rodar os códigos, aproxime o cursor dos colchetes e clique no ícone com formato de *play* (figura 4). Rode um código por vez.

Quadro 1 – Códigos para instalação das bibliotecas necessárias

```
a) pip install -U spacy
b) pip install -U spacy-lookups-data
c) !python -m spacy download pt_core_news_lg
d) pip install pandas
e) pip install wasabi
```

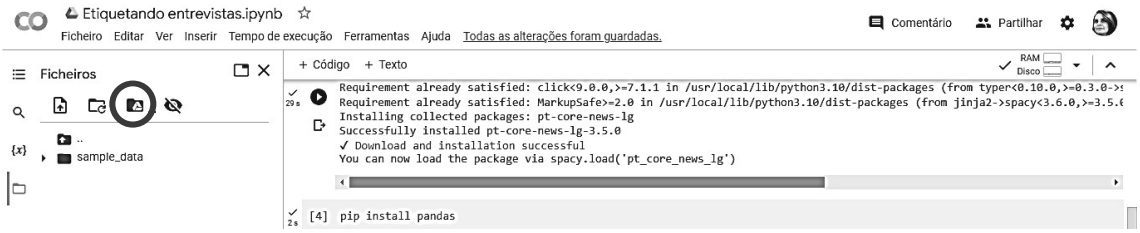
Fonte: elaboração própria

Figura 4 – Como rodar os códigos no Google Colaboratory.

Fonte: elaboração própria.

6. Após rodar os códigos, clique na pasta, no canto esquerdo do Colab. Em seguida, uma tela como na figura 5 aparecerá, clique na pasta com o ícone do Google Drive (parte destacada na figura 5) para conectar o Colab com o seu Drive.

Figura 5 – Conectando o Google Colaboratory ao Drive.



Fonte: elaboração própria.

7. Liste os arquivos para serem anotados adicionando uma nova linha de código e escrevendo o código: **nomes_arq = !ls '/content/drive/MyDrive/etiquetagem 17 05 2023/'** (figura 6). A parte entre aspas simples, deve ser alterada para o caminho da pasta onde os arquivos estão salvos. Esta parte do código de etiquetagem teve a colaboração de Kevenny de Jesus Santos.

Figura 6 – Código para trabalhar com todos os arquivos.

```
nomes_arq = !ls '/content/drive/MyDrive/Testes/'
```

Fonte: elaboração própria.

8. Abra uma nova linha de comando e digite os códigos do quadro 2 para realizar a etiquetagem automática das entrevistas. Veja a figura 7 para repetir a disposição deles na linha de comando. Observe que há uma tabulação antes dos códigos (*i*, *j*, *k*, *l*, *m*, *n*, *o*, *p*) e dois "enter" entre os códigos *m* e *n*. Como se pode ver pelo código (figura 7), a etiquetagem feita aqui é a POS.

Quadro 2 – Código para etiquetagem POS.

```
i) import spacy
j) nlp = spacy.load('pt_core_news_lg')
k) for nome in nomes_arq:
l)         nome_refatorado = nome.replace("'", "")
m)         arquivo = f'/content/drive/MyDrive/etiquetagem 17 05 2023/{nome_refatorado}'

n)          conteudo_arquivo= open(arquivo).read()
o)          doc = nlp(conteudo_arquivo)
p)          pos = [(token.orth_, token.pos_) for token in doc]
```

Fonte: elaboração própria.

Figura 7 – Visualização do código para etiquetagem POS no Google Colaboratory.

```
import spacy
nlp = spacy.load('pt_core_news_lg')
for nome in nomes_arq:
  nome_refatorado = nome.replace("'", "")
  arquivo = f'/content/drive/MyDrive/etiquetagem 17 05 2023/{nome_refatorado}'

  conteudo_arquivo= open(arquivo).read()
  doc = nlp(conteudo_arquivo)
  pos = [(token.orth_, token.pos_) for token in doc]
```

Fonte: elaboração própria.

9. Ainda na mesma célula, pressione “Enter” duas vezes e salve os arquivos etiquetados seguindo o código escrito no quadro 3 e representado na figura 8. Observe que os códigos *q*, *r* e *s* estão com uma tabulação, alinhados com o código *p*, já o código *t* tem duas tabulações. Esta parte do código teve a colaboração de Tulio Sousa de Gois.

Quadro 3 – Códigos para salvar os arquivos etiquetados.

```
q)         nome_refatorado = nome_refatorado.replace('.txt', '_pos_sp.txt')
r)         novo_arquivo = open(nome_refatorado, 'w+')
s)         for token in doc:
t)                   novo_arquivo.write(f"({token.text} {token.pos_})\n")
```

Fonte: elaboração própria.

Figura 8 – Código para renomear o arquivo etiquetado e salvá-lo.

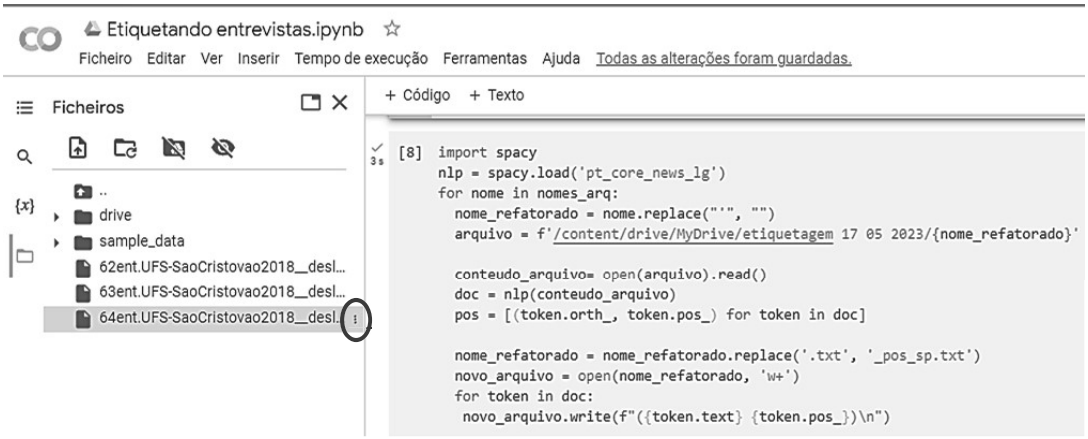
```
pos = [(token.orth_, token.pos_) for token in doc]

nome_refatorado = nome_refatorado.replace('.txt', '_pos_sp.txt')
novo_arquivo = open(nome_refatorado, 'w+')
for token in doc:
 novo_arquivo.write(f"({token.text} {token.pos_})\n")
```

Fonte: elaboração própria.

No código da figura 8, o arquivo é renomeado indicando que a anotação é gramatical (pos) e que o etiquetador usado é o spaCy (sp). Após rodar o código, os arquivos ficam disponíveis na lateral do ambiente de execução, abaixo de *sample data* (figura 9), para baixá-los, clique duas vezes em cada um, ou nos três pontos ao lado de cada um e clique na opção de fazer o *download*.

Figura 9 – Disposição dos arquivos.

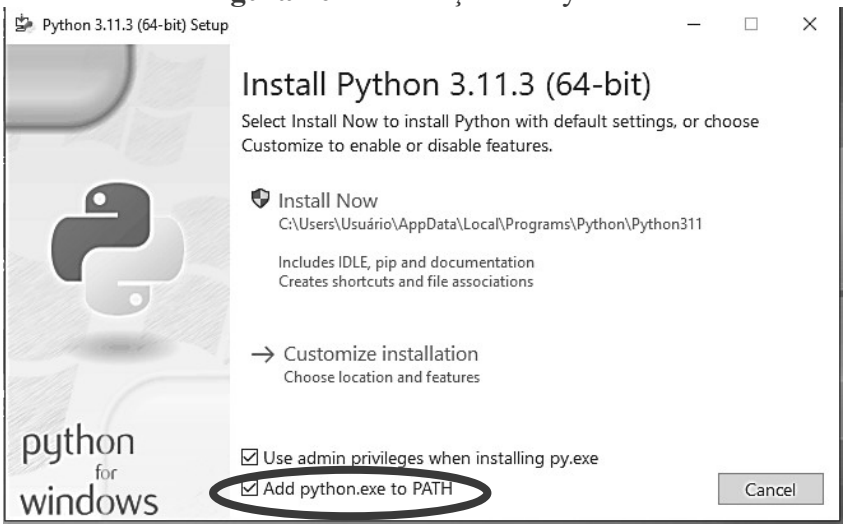


Fonte: elaboração própria.

3.2 Anotação linguística off-line

1. Instale a linguagem Python de acordo com o sistema operacional do seu computador. Lembre-se de marcar a opção "Add python.exe to PATH" Seguindo a configuração padrão indicada pelo instalador, conforme figura 10:

Figura 10 – Instalação do Python.



Fonte: elaboração própria.

2. Instale uma IDE de Python. No caso dos exemplos deste protocolo, é instalado o Jupyter. Siga os procedimentos de configuração indicados pelo instalador.

3. Abra o prompt de comando do windows (cmd) e instale o jupyter escrevendo: **pip install jupyter**.

4. Quando a instalação estiver concluída, a última linha do *prompt* de comando estará conforme a figura 11 abaixo.

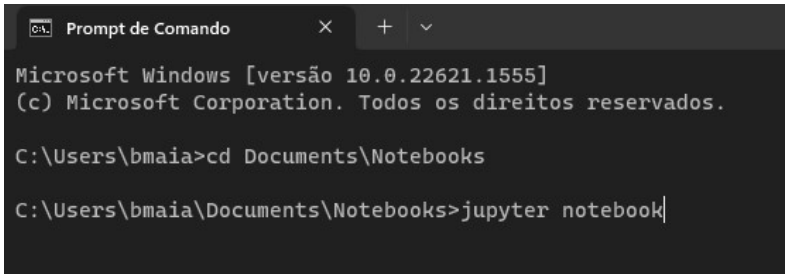
Figura 11 – Instalação do Jupyter concluída.

Fonte: elaboração própria.

5. Crie uma pasta para salvar seus *notebooks* e os arquivos gerados. No exemplo, salvamos na pasta "Documentos" dentro da pasta "Usuários". Copie o caminho desta pasta, clicando com o botão direito e selecionando a opção "Copiar Caminho".

6. Abra o *prompt* de comando, digite "cd" e o caminho da pasta. No exemplo da figura 12, a pasta foi criada em "Documentos" dentro da pasta "Usuários", logo, bastou digitar "cd Documents\Notebooks" para a pasta ser aberta. Em seguida, digite "jupyter notebook".

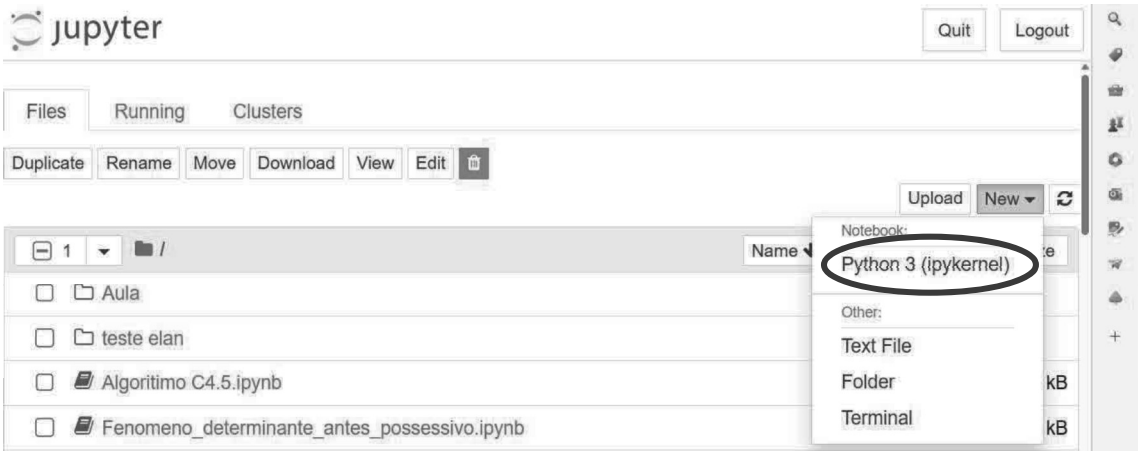
Figura 12 – Iniciando o *notebook* do Jupyter.



Fonte: elaboração própria.

7. Assim que ele carregar, abrirá o Jupyter *Notebook* em seu navegador de internet padrão. No caso, do sistema do exemplo, é o Microsoft Edge. Clique em "New", "Python 3 (ipykernel)" para começar o ambiente de execução, conforme a figura 13.

Figura 13 – Ambiente de execução do Jupyter.



Fonte: elaboração própria.

8. A tela do Jupyter (figura 14) é similar à do Colab. A linha com os colchetes é uma linha de comando. Para adicionar mais uma linha, basta clicar no ícone "+". Repita os códigos do passo 5 da seção 3.1 para instalar as bibliotecas necessárias para etiquetar as entrevistas.

Figura 14 – Tela inicial do Jupyter.

Fonte: elaboração própria.

9. Para importar os arquivos das entrevistas para o Jupyter, abra uma nova linha de comando, e digite o código da figura 15. A parte em entre as aspas, dentro dos parênteses deve ser substituída pelo caminho da pasta onde se encontram os arquivos do usuário.

Figura 15 – importação dos arquivos para trabalho.

```
In [16]: import os
         os.listdir("C:/Users/Bmaia/Documents/Notebooks/Entrevistas Deslocamentos 2019 do Drive")
```

Fonte: elaboração própria.

10. Para etiquetar e salvar os arquivos, abra uma nova linha de comando e repita os códigos dos passos 8 e 9 da seção 3.1. Os arquivos etiquetados ficarão disponíveis na pasta criada para salvar os *Notebooks*.

4. Alinhamento da etiquetagem com o ELAN.

O alinhamento da etiquetagem às entrevistas do ELAN é feito de maneira individualizada. Para cada entrevista, os procedimentos abaixo devem ser repetidos para cada trilha.

1. Abra a entrevista já transcrita utilizando o ELAN. Exporte, primeiramente, a trilha do documentador como "texto delimitado por tabulador".

2. Em seguida, nas opções de saída, marque os itens listados abaixo como na figura 16:

- ✓ Adicionar o tempo de início da mídia principal nos tempos de anotação
- ✓ *Exclude tier names from output.*
- ✓ *Exclude participant names from output.*
- ✓ Incluir coluna de tempo para: Tempo inicial e Tempo final.
- ✓ Incluir formato de tempo: hh:mm:ss

Figura 16 – Janela de exportação das trilhas.

Fonte: elaboração própria.

3. Clique em “OK”.

4. Renomeie o arquivo e o salve na mesma pasta da entrevista em formato .**csv**.

5. Acesse o *Notebook* do Jupyter colocando o caminho da pasta onde o arquivo com a trilha está salvo, ou seja, repita o procedimento 6 da seção 3.2. Fazendo isso, a saída com a trilha da entrevista etiquetada ficará na mesma pasta, facilitando a organização. Para realizar a operação a partir do Colab, salve a entrevista no Drive.

6. Repita os procedimentos 5 da seção 3.1 para instalação das bibliotecas necessárias para etiquetar a trilha tanto para o Jupyter quanto para o Colab.

7. Após a importação das bibliotecas, abra uma nova linha de comando e digite o código do quadro 4, seguindo a mesma disposição da figura 17. Esse procedimento é igual para os dois ambientes de desenvolvimento (Jupyter e Colab). Lembre-se de colar o caminho do arquivo no código dentro dos parênteses do código *open*. Essa parte do código foi feita com a colaboração de Túlio Sousa de Gois.

Quadro 4 - Código para etiquetagem de trilhas para o alinhamento no ELAN.

```
import spacy
import pandas as pd
import csv

colunas = ['inicio', 'fim', 'trilha', 'etiquetagem']
linhas = []


nlp = spacy.load('pt_core_news_lg')
arquivo = open('DOCT.csv', encoding = 'utf-8')
conteudo_arquivo = csv.reader(arquivo, delimiter='\t')

for dado in conteudo_arquivo:
    inicio = dado[1]
    fim = dado[2]
    trilha = dado [3]
    etiquetagem = ''
    doc = nlp(trilha)

    for token in doc:
        etiquetagem += (f"{token.orth_} {token.pos_} ")
    linhas.append([inicio, fim, trilha, etiquetagem])
```

Fonte: Elaboração própria.

Figura 17 – Etiquetagem de trilhas para alinhamento com ELAN.

Fonte: elaboração própria.

8. Salve o arquivo (figura 18) como .**csv** usando o código do Quadro 5 em uma linha de comando nova. O nome da trilha deve ser o mesmo nome que consta do arquivo do ELAN, o ideal é que se finalize o procedimento todo para cada uma delas antes de iniciar uma nova entrevista.

Quadro 5 - Código para arquivamento da trilha etiquetada.

```
df_etiquetagem = pd.DataFrame(linhas, columns = colunas)

df_etiquetagem.to_csv("arquivo1.csv", encoding = 'UTF-8')
print(df_etiquetagem)
```

Fonte: elaboração própria.

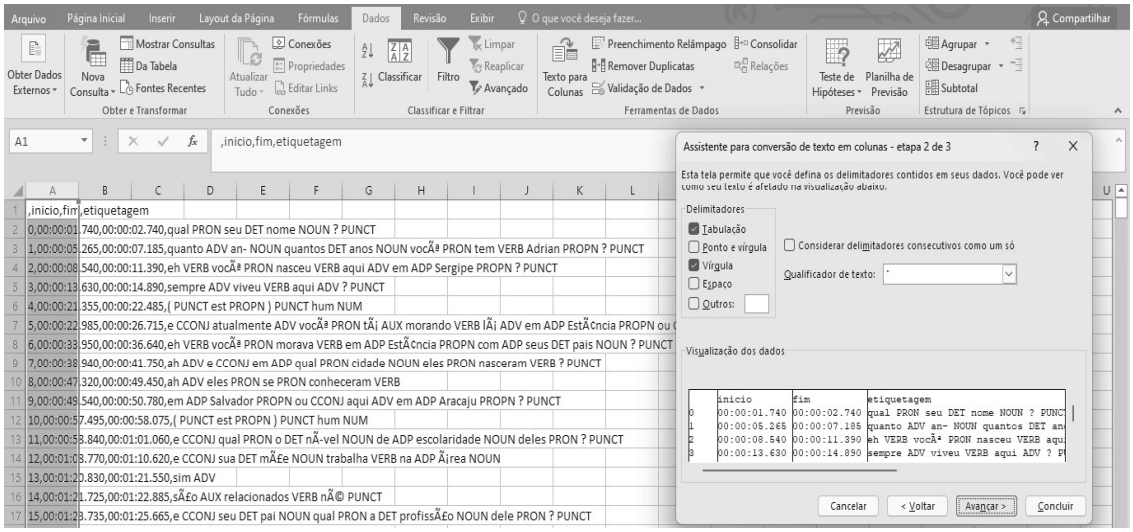
Figura 18 – Arquivamento da trilha.

```
In [19]: df_etiquetagem = pd.DataFrame(linhas, columns = colunas)
         df_etiquetagem.to_csv("trilha11-etq.csv", encoding = 'utf-8')
```

Fonte: elaboração própria.

11. O texto de saída, estará em formato csv. Abra-o, selecione a primeira coluna e clique em "Dados">> "Texto para Colunas". Na janela que se abrir, clique em "Avançar". Em "Delimitadores", selecione "Tabulação" e "Vírgula". Clique em "Concluir" (figura 19). Clique no ícone de disquete para salvar o arquivo. Uma janela se abrirá perguntando se deseja salvar mesmo assim, clique em "Sim". Feche o arquivo. Se abrir uma janela perguntando "Deseja salvar as alterações?" Clique em "Não".

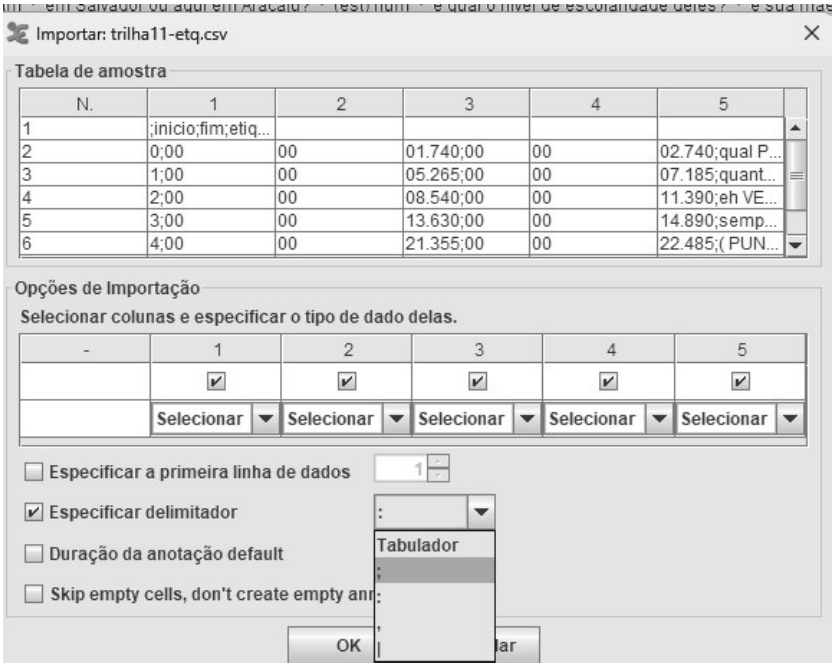
Figura 19 – Configuração do arquivo de saída.



Fonte: elaboração própria.

12. Abra novamente a entrevista escolhida no ELAN. Em seguida, clique em "Importar">> "csv/Arquivo delimitado por tabulador". A janela exposta na figura 20 se abrirá. Observe que, na primeira linha, na coluna 1, os títulos estão separados por ponto e vírgula (;), logo, clique em "especificar o delimitador" e selecione aquele que melhor se adequa aos dados, sendo, no caso do exemplo, o ponto e vírgula (;).

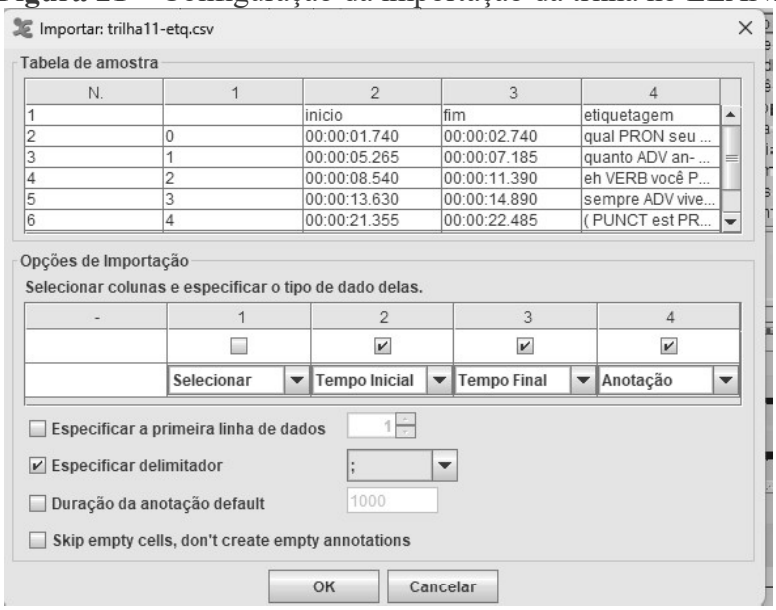
Figura 20 – Ajuste na trilha etiquetada para importação no ELAN.



Fonte: elaboração própria.

13. Desmarque a primeira coluna e selecione as outras, conforme figura 21. Abaixo de cada uma marque o que elas representam na seguinte ordem: “Tempo Inicial”, “Tempo Final” e “Anotação”.

Figura 21 – Configuração da importação da trilha no ELAN.



Fonte: elaboração própria.

14. Clique em OK e repita todo o processo para a segunda trilha. Repita toda a seção 4 para etiquetar e alinhar as outras entrevistas.

5. Escrita dos metadados

5.1 Metadados de cada arquivo

Todos os arquivos referentes às coletas devem possuir metadados escritos em formato .txt. Siga o exemplo do arquivo dos metadados de um arquivo de áudio abaixo para fazer a documentação (quadro 6.)

Quadro 6 - Informações de metadados.

HEAD>
<META name="DC.title" content="01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24"/> **(Nome do arquivo)**
<META name="DC.language" content="Portuguese" **(Língua)**
<META name="DC.description" content="Entrevista sociolinguística"/> **(Descrição do arquivo)**
<META name="DC.subject" content="Assuntos pessoais seguindo moldes do roteiro de entrevista laboviano."/> **(Assunto do arquivo)**
<META name="DC.contributor" content="Raquel Meister Ko. Freitag"/> **(Quem participou do trabalho, primeiramente a orientadora, seguida de quem coletou, quem transcreveu e quem revisou. Para cada colaborador, deve-se escrever a especificação novamente: <META name = "DC.contributor" content= "")**
<META name="Dc.date" content="2018"/> **(Ano da coleta do arquivo)**
<META name="Dc.type" content="audio"/> **(Tipo do arquivo)**
<META name="DC.format" content="wav"/> **(Formato do arquivo)**
<META name="DC.publisher" content="Condomínio de Laboratórios Multiusuários de Informática e Documentação (LAMID)"/> **(Onde estará publicado)**
<METAname="DC.identifier"
content="https://drive.google.com/drive/u/4/folders/1uZQcKjIvAzDxQn4YuNpbhmehY
D22uFvQ"/> **(URL do arquivo da pasta do Drive do Lamid)**
<META name="DC.relation" content="01ent.UFS-SaoCristovao2018__desl. I_final_lui.ms.24"/> **(Arquivo relacionado)**
<META name="DC.coverage" content="Brazil"/> **(País)**
<META name="DC.coverage" content="Sergipe/> **(Estado)**
<META name="DC.rights" content="Access limited to members."/> **(Direitos)**
</HEAD>

Fonte: elaboração própria.

5.2. Metadados da Amostra

Todas as amostras devem ter uma planilha para identificação dos informantes e seus metadados. Dessa tabela, devem constar: identificação da entrevista, iniciais do informante, idade, escolaridade, naturalidade, entre outras informações a depender do tipo da amostra, por isso a ficha social deve estar completa. **Esta planilha** é um exemplo feito com base na amostra Deslocamentos 2019.

6. Salvaguarda

1. Após etiquetar todas as entrevistas e realizar a escrita dos metadados, acesse o Drive do Lamid. Vá na pasta Repositório Gelins>> Banco de Dados Falares Sergipanos.

2. Caso a amostra etiquetada seja do tipo Deslocamentos, faça o upload dos arquivos da amostra da seguinte maneira:

>Nome do Deslocamento (Deslocamentos 2019, por exemplo)
>Deslocamento 1
>Início
> Informante X: Criar uma pasta para cada informante colocando todos os arquivos do informante em uma mesma pasta
>Final
>Informante Y: Criar uma pasta para cada informante colocando todos os arquivos do informante em uma mesma pasta

3. Caso a amostra tenha sido realizada por outro tipo de coleta, os arquivos devem ser hierarquizados de forma que todos os arquivos de um informante fiquem em uma mesma pasta.

ANEXOS

ANEXO A – RELATÓRIO DO LANCSBOX

Report

Introduction

This research report was automatically produced by #LancsBox (Brezina et al., 2015, 2018, 2020), a corpus analysis tool developed at Lancaster University. It uses cutting-edge technology and statistical sophistication (Brezina 2018) to analyze and visualize corpus data. For more information and tips on research report writing see the Research Report Guide.

Method

Data

The study analyzed the following corpus:

Table 1. Corpus used

Name	Language	Texts	Tokens	Additional information
Corpus 1	Portuguese	64	458,792	Types: 16,091 Lemmas: 15,078

In the study, 1 corpus was used of the total size of 458,792 running words (tokens) in 64 texts. A full description of the corpora is available in data\tsv\corpora.

Procedure

#LancsBox (Brezina et al., 2015, 2018, 2020) software package was employed to analyse the data. The following tools from the package were used: KWIC, GraphColl and Whelk. The KWIC tool generates a list of all instances of a search term in a corpus in the form of a concordance. The GraphColl tool identifies collocations and displays them in a table and as a collocation graph or network. The Whelk tool provides information about how the search term is distributed across corpus files. The following search terms were used: "meu", "minha", "meus", "minhas", "seu", "sua", "seus", "suas", "teu", "tua", "teus" and "tuas".

Results

Specific searches: Concordances and contexts

Search term "meu" in Corpus 1

The search term *meu* occurs 1114 times (24,281 per 10k) in Corpus 1 in 64 out of 64 texts. The distribution of this search term in the individual texts can be seen in Table 2; top 25 and bottom 25 texts according to the relative frequencies of the search term are shown. The full data set is available in data\csv\whelk\whelk_001_001.csv. Table 3 displays a random sample of 10 concordance lines, showing the most immediate contexts in which the search term is used.

Table 2. Distribution of the search term *meu* in Corpus 1

ANEXO B – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Estamos convidando-o a participar como voluntário de uma pesquisa de campo a ser realizada por meio da gravação de situações de interação.

A coleta será realizada com o objetivo de desenvolvermos um trabalho acadêmico vinculado ao Programa de Pós-graduação em Letras.

A entrevista coletada ficará disponível no banco de dados do Grupo de Estudos em Linguagem, Interação e Sociedade – GELINS; para ser utilizada em pesquisas futuras. Serão resguardadas todas as informações de identificação de forma que se mantenha o anonimato.

Não será cobrado nada, não haverá gastos e não estão previstos ressarcimentos ou indenizações.

Você poderá solicitar esclarecimento sobre a pesquisa em qualquer etapa do estudo. Você é livre para recusar-se a participar, retirar seu consentimento ou interromper a participação na pesquisa a qualquer momento, seja por motivo de constrangimento e/ou outros motivos. A sua participação é voluntária e a recusa em participar não irá acarretar qualquer penalidade ou perda de benefícios.

Desde já, agradecemos sua atenção e participação e colocamo-nos à disposição para maiores informações.

Consentimento para participação

Eu, _____, idade:_____, estado civil: _____, RG:_____, estou de acordo com a participação no estudo descrito acima. Eu fui devidamente esclarecido quanto os objetivos da pesquisa, aos procedimentos aos quais serei submetido e os possíveis riscos envolvidos na minha participação. Os pesquisadores me garantiram disponibilizar qualquer esclarecimento adicional a que eu venha a solicitar durante o curso da pesquisa e o direito de desistir da participação em qualquer momento, sem que a minha desistência implique qualquer prejuízo à minha pessoa ou à minha família, sendo garantido anonimato e o sigilo dos dados referentes à minha identificação, bem como de que a minha participação neste estudo não me trará nenhum benefício econômico. Ao mesmo tempo, libero a utilização de minha entrevista para fins científicos e de estudos (livros, artigos, slides e transparências), em favor dos pesquisadores, obedecendo ao que está previsto na Resolução do CNS nº 196/96. Autorizo também que a minha interação fique disponível no banco de dados acima referido para ser utilizada em pesquisas futuras.

_____, _____ de _____

Assinatura do (a) participante: _____

Assinatura do (a) pesquisador (a): _____

Assinatura do (a) coordenador (a)/ orientador (a): _____

ANEXO C – PARECER SOBRE DOCUMENTO DE AUTORIZAÇÃO DE ACESSO

UNIVERSIDADE FEDERAL DE SERGIPE
SISTEMA INTEGRADO DE PATRIMÔNIO, ADMINISTRAÇÃO E CONTRATOS
EMITIDO EM 05/08/2022 10:41

Documento nº. 23113.034464/2022-47

Tipo: MEMORANDO ELETRÔNICO

DESPACHO

Prezados,

O documento trata de obrigações assumidas pelo solicitante das informações. Nesse aspecto nada a opor aos termos propostos.

Sugiro , contudo , verificar os termos necessários para os proprietários ou detentores de direitos dos dados solicitados , eis que poderão, conforme o caso , demandar autorização de acesso.

Att

(Assinado eletronicamente em 05/08/2022 08:53)
PAULO CELSO REGO LEO
PROCURADORIA GERAL (11.03.07)

ANEXO D – PARECER SOBRE O PEDIDO DE REGISTRO DE SOFTWARE

AGITTE | ufs

UFS

UNIVERSIDADE FEDERAL DE SERGIPE

AGÊNCIA DE INOVAÇÃO E TRANSFERÊNCIA DE TECNOLOGIA – AGITTE

AGITTE | ufs

Parecer de Avaliação de Programa de Computador

Título do Software:	**JKM Soluções em Dados**
Código Notificação:	**NID351-2022**

Qual o propósito da criação do Programa de Computador?

Aplicado à educação/ pesquisa, o software apresenta funcionalidades úteis para preservação, organização e disseminação de dados.

Potencialidade do Programa de Computador e a Transferência de Tecnologia. Comente:

- Existe software similar? Sim. Porém uma das vantagens é organização de dados de maneira rápida, automática e de baixo custo.

- Existe potencial de comercialização? Sim. Devido ao seu status de desenvolvimento e sua aplicação, o software pode ser útil para instituições públicas e privadas de ensino, especialmente ensino superior.

- Existe interesse para a sociedade? Sim.

- Há necessidade de ajustes? Sim.

Recomenda-se a criação e registro de uma logomarca para identidade visual do software. Acesse: Registre sua Marca

Parecer Final do Relator:

Favorável (x) Favorável com ajustes ( ) Não favorável no estágio atual ( )

Desfavorável ( )*

**O processo de depósito/registro junto ao INPI poderá ser orientado pela CINTTEC-UFS, porém não será apropriado o conhecimento pela instituição, podendo seus autores procederem como inventores independentes.*

Justificativa:
O software atende todos os requisitos necessários e solicitados, bem como, apresenta um caráter inovador. Dessa forma apresentamos parecer favorável ao registro do pedido junto ao INPI.

Data	11/05/2023
Assinatura	**Antônio Martins de Oliveira Júnior** *COORDENADOR - TITULAR*