

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS
DEPARTAMENTO DE LINGUÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

José Roberto Homeli da Silva

**Classificação textual de narrativas de indivíduos com
Doença de Alzheimer ou Déficit Cognitivo Leve**

— Versão Corrigida —

São Paulo
2024

JOSÉ ROBERTO HOMELI DA SILVA

**Classificação textual de narrativas de indivíduos com
Doença de Alzheimer ou Déficit Cognitivo Leve**

— Versão Corrigida —

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Linguística da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo como requisito para obtenção do título de mestre.

Orientador: Prof. Dr. Marcos Lopes

São Paulo
2024

ENTREGA DO EXEMPLAR CORRIGIDO DA DISSERTAÇÃO/TESE

Termo de Anuência do (a) orientador (a)

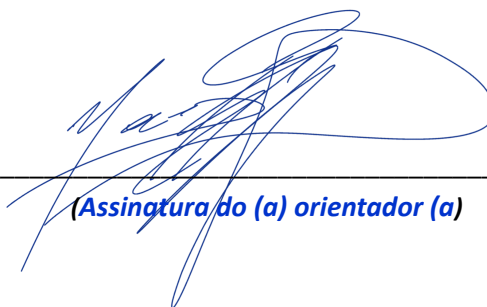
Nome do (a) aluno (a): José Roberto Homeli da Silva

Data da defesa: 15/01/2024

Nome do Prof. (a) orientador (a): Marcos Lopes

Nos termos da legislação vigente, declaro **ESTAR CIENTE** do conteúdo deste **EXEMPLAR CORRIGIDO** elaborado em atenção às sugestões dos membros da comissão Julgadora na sessão de defesa do trabalho, manifestando-me **plenamente favorável** ao seu encaminhamento ao Sistema Janus e publicação no **Portal Digital de Teses da USP**.

São Paulo, 15/03/2024.



(Assinatura do (a) orientador (a))

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catalogação na Publicação
Serviço de Biblioteca e Documentação
Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo

S586c Silva, José Roberto Homeli da
Classificação textual de narrativas de indivíduos com Doença de Alzheimer ou Déficit Cognitivo Leve / José Roberto Homeli da Silva; orientador Marcos Fernando Lopes - São Paulo, 2024.
91 f.

Dissertação (Mestrado)- Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. Departamento de Linguística. Área de concentração: Semiótica e Lingüística Geral.

1. LINGÜÍSTICA COMPUTACIONAL. 2. TRATAMENTO AUTOMÁTICO DE TEXTOS E DISCURSOS. I. Lopes, Marcos Fernando, orient. II. Título.

Agradecimentos

gostaria de expressar minha gratidão às pessoas que estiveram por perto e me ajudaram no caminho até este capítulo da minha jornada acadêmica; mais ainda: pessoal - cheguei aqui!

agradeço

ao meu orientador, marcos,
pelo seus suportes incansáveis - acadêmicos, pessoais, humanos
por sua dedicação inspiradora à ciência

a máquina do mundo se entreabriu

às companhias que sempre estiveram por perto:
isabella, pelos momentos de compartilhamento
cláudia, pelo engajamento com a arte e a academia
fernando, pelo ouvido à distância

olha, repara, ausculta: essa riqueza

ao fernando, meu amor,
aquele que, de repente,
do meio de um turbilhão de afazeres e sentimentos,
surgiu a me envolver num abraço perfeito
e a me mostrar que havia como
sobreviver
vivendo
livre
leve

sem emitir um som que fosse impuro

à minha mãe,
aquela que me mostrou a entrega
que me mostrou a resiliência
que me mostrou a dedicação que se deve ter com o amor
BRAVA! sempre

que [vai] pelos caminhos demonstrando

*hoje e sempre,
em especial,
ao meu pai*

seguia vagaroso, de mãos pensas.

Sumário

Resumo • *vii*

Abstract • *viii*

Lista de Figuras • *ix*

Lista de Tabelas • *xi*

- 1 Introdução à pesquisa** • 1
 - 1.1 Objetivos • 4
 - 1.2 Sobre a Doença de Alzheimer • 5
 - 1.3 Sobre o Comprometimento Cognitivo Leve • 6

- 2 Revisão da literatura** • 9
 - 2.1 Kumar et al. (2021) • 9
 - 2.2 Dreisbach et al. (2019) • 11
 - 2.3 Boyé, Tran e Grabar (2014) • 11
 - 2.4 Beltrami et al. (2018) • 13
 - 2.5 Hernández-Domínguez et al. (2016) • 14
 - 2.6 Santos et al. (2017) • 15
 - 2.7 Karlekar, Niu e Bansal (2018) • 17
 - 2.8 Vincze et al. (2016) • 18
 - 2.9 Abrisqueta-Gomez et al. (2004) • 19
 - 2.10 Nitrini, Caramelli et al. (2005) • 21
 - 2.11 Nitrini, Brucki et al. (2021) • 22
 - 2.12 Steiner et al. (2017) • 23
 - 2.13 Frota et al. (2011) • 24
 - 2.14 Toledo et al. (2018) • 25

- 3 Métodos** • 27
 - 3.1 Materiais (*corpora*) • 27
 - 3.1.1 *Cinderella* • 28
 - 3.1.2 *Dog e Lucia* • 33
 - 3.1.3 *Wallet* • 35
 - 3.1.4 Sobre os padrões de transcrição • 36
 - 3.1.5 Dados demográficos • 37
 - 3.2 Procedimentos • 37

3.2.1	Tratamento dos dados	• 39
3.2.2	Classificador Bayesiano Ingênuo Multinomial	• 41
3.2.3	Rede de propagação para frente	• 42
3.2.4	Rede BiLSTM	• 44
3.2.5	DistilBERT: modelo de linguagem baseado em BERT	• 46
3.3	Fenômenos linguísticos: hesitações	• 47
3.3.1	Anotação de preenchedores	• 48
3.3.2	Anotação de repetições completas	• 49
3.3.3	Anotação de repetições incompletas	• 50
3.4	Instrumentos	• 50
4	Resultados e discussão	• 51
4.1	Análise descritiva de dados	• 51
4.2	Resultados com classificador bayesiano	• 54
4.3	Resultados com rede de propagação para frente	• 55
4.3.1	Configuração A	• 56
4.3.2	Configuração B	• 59
4.3.3	Configuração C	• 62
4.4	Resultados com rede BiLSTM	• 64
4.5	Resultados com DistilBERT	• 65
4.6	Considerações acerca dos resultados dos classificadores	• 67
4.7	Resultados das análises de fenômenos linguísticos	• 69
4.7.1	Preenchedores	• 69
4.7.2	Repetições completas	• 71
4.7.3	Repetições incompletas	• 73
4.7.4	Repetições completas e incompletas	• 75
4.8	Discussão de análises de fenômenos linguísticos	• 77
5	Conclusão	• 80
5.1	Breve sumarização dos resultados	• 81
5.2	Contribuições principais	• 82
5.3	Posição dos achados frente à literatura	• 83
5.4	Limites desta pesquisa	• 85
5.5	Pesquisas futuras	• 86
	Referências	• 88

Resumo

SILVA, J. R. H. *Classificação textual de narrativas de indivíduos com Doença de Alzheimer ou Déficit Cognitivo Leve*. Dissertação (Mestrado em Linguística). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2024.

Os diagnósticos de Doença de Alzheimer (DA) e de Comprometimento Cognitivo Leve (CCL) são processos, por vezes, custosos (para o indivíduo ou para o orçamento público) por envolverem recursos materiais e humanos, como consultas e exames médicos; além disso, podem ser obtidos quando essas condições neurodegenerativas já se encontram em estado avançado. No entanto, quanto mais precoce for o diagnóstico, melhor qualidade de vida o indivíduo tenderá a ter. Assim, para isso e para além de marcadores biológicos, existem outras pistas que o corpo humano fornece, cujos surgimentos e primeiros estágios o indivíduo ou seus familiares podem até não notar, mas são bons caminhos para se tentar identificar o início dessas doenças. Dentre essas pistas, estão alterações no campo da linguagem — e é da classificação automática delas, bem uma análise linguística pormenorizada delas que esta pesquisa se ocupa. Para isso, foi feita (a) uma revisão da literatura acerca de *corpora*, tarefas, técnicas e resultados que tangenciam nosso escopo; (b) a escolha de um *corpus* para trabalho, chamado *Datasets of Neuropsychological Language Tests in Brazilian Portuguese* (DNLTP-BP), além da realização de uma breve limpeza, organização dos dados; (c) elaboração, por meio de técnicas de Processamento de Linguagem Natural, de quatro modelos de classificação textual, um bayesiano ingênuo, outros dois calcados em redes neurais artificiais recorrentes (uma de propagação para frente e outra bidirecional, BiLSTM), além de um modelo baseado em *transformers* adaptado ao Português Brasileiro, DistilBERT. Os resultados mostram que três dos modelos cumpriram a tarefa satisfatoriamente, conseguindo cobrir bem métricas de avaliação da tarefa de classificação, sobretudo na distinção entre DA vs. controle e CCL vs. controle. Uma análise paralela que conduzimos foi a da incidência de fenômenos linguísticos de hesitação, a saber: pausas preenchidas, gaguejamentos, falsos inícios e repetições hesitativas. Por parte dessa análise, notar-se-á novamente que o grupo-controle apresenta a menor incidência deles em relação aos outros dois grupos.

Palavras-chave: Linguística Computacional. Classificação textual. Doença de Alzheimer. Déficit Cognitivo Leve.

Abstract

SILVA, J. R. H. *Text classification of narratives produced by individuals with Alzheimer's Disease or Mild Cognitive Deficit*. Dissertação (Mestrado em Linguística). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2024.

The diagnoses of Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI) are sometimes costly processes (either for the individual or for public budgets) as they involve material and human resources, such as medical consultations and exams. Additionally, they may be obtained when these neurodegenerative conditions are already in an advanced state. However, the earlier the diagnosis, the better the individual's quality of life tends to be. Hence, in addition to biological markers, there are other clues provided by the human body, whose emergence and initial stages the individual or their family may not notice but are promising paths to try to identify the onset of these diseases. Among these clues are changes in the field of language, and it is the automatic classification of these changes, as well as a detailed linguistic analysis of them, that this research addresses. To achieve this, the study carried out: (a) A literature review on corpora, tasks, techniques, and results related to our scope. (b) The selection of a corpus for the study, named *Datasets of Neuropsychological Language Tests in Brazilian Portuguese* (DNLT-BP), along with a brief pre-processing routine of the data. (c) The development of four text classification models using Natural Language Processing techniques: a naïve Bayesian model, two based on recurrent artificial neural networks (one feed-forward and one bidirectional, BiLSTM), and a model based on transformers adapted to Brazilian Portuguese, DistilBERT. The results show that three of the models successfully performed the task, achieving good evaluation metrics, particularly in distinguishing between AD vs. control and MCI vs. control. A parallel analysis was conducted regarding the occurrence of linguistic hesitation phenomena, namely filled pauses, stuttering, false onsets, and hesitant repetitions. From this analysis, it is noteworthy that the control group has the lowest incidence of these phenomena compared to the other two groups.

Keywords: Natural Language Processing. Text Classification. Alzheimer's disease. Mild Cognitive Impairment.

Lista de Figuras

1.1	Dinâmica de biomarcadores na cascata patológica da doença de Alzheimer. O β -amiloide ($A\beta$) é identificado pelo fluido cerebrospinal (CSF) $A\beta_{42}$ ou pela imagem de amiloides por tomografia por emissão de pósitrons (PET). A lesão neuronal e disfunção mediadas por tau são identificadas pelo tau no CSF ou pelo PET com fluorodesoxiglicose. A estrutura cerebral é medida por ressonância magnética estrutural. (Legenda original traduzida.) Fonte: Savonenko et al. (2015, p. 323)	8
3.1	Ambos os gráficos mostram a quantidade de sentenças presente no <i>corpus</i> em cada grupo de análise (originários dos <i>corpora Cinderella e Dog</i>): Doença de Alzheimer, Comprometimento Cognitivo Leve e grupos-controle. O primeiro mostra a quantidade inicial em cada grupo; o segundo mostra a quantidade após o corte no grupo-controle.	40
3.2	Exemplo com uma camada da arquitetura de uma rede neural recorrente de propagação para frente (unidirecional).	43
3.3	Exemplo de três passos, bidirecionais, da arquitetura de uma rede BiLSTM por Cui et al. (2020).	45
4.1	Matriz de confusão com valores absolutos na tarefa de classificação (modelo bayesiano ingênuo <i>Multinomial</i>) entre DA (o), grupo-controle (1) e CCL (2). <i>Corpus</i> desbalanceado.	55
4.2	Matriz de confusão com valores absolutos na tarefa de classificação (modelo bayesiano ingênuo <i>Multinomial</i>) entre DA (o), grupo-controle (1) e CCL (2). <i>Corpus</i> balanceado.	56
4.3	Gráficos de acurácia (4.3a) e perda (4.3b) no treinamento e na validação do modelo ao se submeter os dados de DA e do grupo-controle. Configuração A.	57
4.4	Gráficos de acurácia (4.4a) e perda (4.4b) no treinamento e na validação do modelo ao se submeter os dados de CCL e do grupo-controle. Configuração A.	58
4.5	Gráficos de acurácia (4.5a) e perda (4.5b) no treinamento e na validação do modelo ao se submeter os dados de DA e de CCL. Configuração A.	59
4.6	Gráficos de acurácia (4.6a) e perda (4.6b) no treinamento e na validação do modelo ao se submeter os dados dos três grupos: DA, CCL e grupo-controle. Configuração A.	59
4.7	Gráficos de acurácia (4.7a) e perda (4.7b) no treinamento e na validação do modelo ao se submeter os dados de DA e do grupo-controle. Configuração B.	60
4.8	Gráficos de acurácia (4.8a) e perda (4.8b) no treinamento e na validação do modelo ao se submeter os dados de CCL e do grupo-controle. Configuração B.	60

4.9	Gráficos de acurácia (4.9a) e perda (4.9b) no treinamento e na validação do modelo ao se submeter os dados de DA e de CCL. Configuração B.	61
4.10	Gráficos de acurácia (4.10a) e perda (4.10b) no treinamento e na validação do modelo ao se submeter os dados dos três grupos: DA, CCL e grupo-controle. Configuração B.	61
4.11	Gráficos de acurácia (4.11a) e perda (4.11b) no treinamento e na validação do modelo ao se submeter os dados de DA e do grupo-controle. Configuração C.	62
4.12	Gráficos de acurácia (4.12a) e perda (4.12b) no treinamento e na validação do modelo ao se submeter os dados de CCL e do grupo-controle. Configuração C.	63
4.13	Gráficos de acurácia (4.13a) e perda (4.13b) no treinamento e na validação do modelo ao se submeter os dados de DA e de CCL. Configuração C.	63
4.14	Médias de acurácia do desempenho do modelo <i>distilbert-portuguese-cased</i> , baseado no BERTimbau, para os grupos de estudo dois a dois tanto para o <i>dataset</i> com texto bruto (br.) quanto para o anotado (an.).	67
4.15	Comparação entre médias e erros-padrão das razões do número preenchedores sobre o número total de tokens por narrativa para cada grupo.	70
4.16	Comparação entre médias e erros-padrão das razões do número repetições completas sobre o número total de tokens por narrativa para cada grupo.	72
4.17	Comparação entre médias e erros-padrão das razões do número repetições incompletas sobre o número total de tokens por narrativa para cada grupo.	74
4.18	Comparação entre médias e erros-padrão das razões do número repetições totais (completas e incompletas) sobre o número total de tokens por narrativa para cada grupo.	76

Lista de Tabelas

2.1	Médias obtidas antes e depois das alterações realizadas no <i>corpus</i> Cinderella.	16
3.1	Correlação entre as informações sociodemográficas dos participantes do grupo de indivíduos com DA (* significa que não processamos a métrica por haver lacunas nos dados).	38
3.2	Correlação entre as informações sociodemográficas dos participantes do grupo de indivíduos com CCL (* significa que não processamos a métrica por haver lacunas nos dados).	38
3.3	Correlação entre as informações sociodemográficas dos participantes do grupo-controle (* significa que não processamos a métrica por haver lacunas nos dados).	38
3.4	Correlação entre as informações sociodemográficas dos participantes de todos os grupos de análise (* significa que não processamos a métrica por haver lacunas nos dados).	39
3.5	Três cenários de testes da rede de propagação para frente submetendo os grupos de análise dois a dois e também os três simultaneamente. Variações contam o <i>corpus</i> balanceado e desbalanceado, bem como a entropia cruzada ser categorial ou binária. A configuração C para os três grupos (com entropia cruzada binária) não é possível.	44
4.1	Tabela que apresenta uma série de métricas calculadas durante o pré-processamento dos <i>corpora</i> em três momentos: a partir do texto das frases tal qual transcritas, lematizadas e stemizadas. (a): Média de tamanho de sentença (em palavras); (b): Desvio padrão de (a); (c): Erro padrão de (b); (d): Riqueza lexical (%); (e): Incidência de hápax legômena (%); (f) Média de palavras por narrativa; (g): Média da porcentagem de repetições de palavra por sentença (%).	52
4.2	Medidas de acurácia, precisão, cobertura e F1 do desempenho do classificador bayesiano frente a todos os agrupamentos possíveis dos grupos de análise.	54
4.3	Tabela para facilitar a movimentação entre os gráficos com resultados da rede de propagação para frente, organizando-os por grupos e configuração aplicada.	57
4.4	Resultados da rede BiLSTM para o <i>dataset</i> com texto bruto por 4 épocas para DA e CCL.	64
4.5	Resultados da rede BiLSTM para o <i>dataset</i> com texto bruto por 4 épocas para DA e CTR.	64

4.6	Resultados da rede BiLSTM para o <i>dataset</i> com texto bruto por 4 épocas para CCL e CTR.	64
4.7	Resultados da rede BiLSTM para o <i>dataset</i> com texto anotado por 4 épocas para DA e CCL.	64
4.8	Resultados da rede BiLSTM para o <i>dataset</i> com texto anotado por 4 épocas para DA e CTR.	65
4.9	Resultados da rede BiLSTM para o <i>dataset</i> com texto anotado por 4 épocas para CCL e CTR.	65
4.10	Resultados de acurácia do desempenho do modelo <i>distilbert-portuguese-cased</i> , baseado no BERTimbau, para os grupos de estudo dois a dois tanto para o <i>dataset</i> com texto bruto (br.) quanto para o anotado (an.) por dez execuções do código acompanhados da média e desvio padrão.	66
4.11	Médias e erros-padrão das razões entre o número de preenchedores e o número total de tokens por narrativa para cada grupo.	70
4.12	Resultados do teste U de Mann-Whitney e valor- <i>p</i> ao se submeter as proporções de preenchedores sobre tokens na narrativa de cada indivíduo a cada dois grupos.	71
4.13	Médias e erros-padrão das razões entre o número de repetições completas sequenciais e o número total de tokens por narrativa para cada grupo.	72
4.14	Resultados do teste U de Mann-Whitney e valor- <i>p</i> ao se submeter as proporções de preenchedores sobre tokens na narrativa de cada indivíduo a cada dois grupos.	73
4.15	Médias e erros-padrão das razões entre o número de repetições incompletas sequenciais e o número total de tokens por narrativa para cada grupo.	73
4.16	Quantidade de narrativas em que não apareceram repetições incompletas e de narrativas em que apareceram uma ou mais repetições incompletas em cada grupo.	74
4.17	Teste de qui-quadrado para repetições incompletas, considerando a divisão em cada grupo de narrativas sem presença do fenômeno em oposição àquelas em que ele ocorre uma ou mais vezes.	75
4.18	Médias e erros-padrão das razões entre o número de repetições totais (completas e incompletas) sequenciais e o número total de tokens por narrativa para cada grupo.	76
4.19	Teste de qui-quadrado para repetições completas e incompletas.	77

Introdução à pesquisa

Para aumentar as chances de uma pessoa desfrutar de uma vida com mais qualidade caso ela que venha a sofrer os efeitos neurodegenerativos da doença de Alzheimer (DA) ou do Comprometimento Cognitivo Leve (CCL)¹, é imprescindível que a doença seja identificada o quanto antes. Uma vez que essas doenças não têm uma cura ou altas chances de reversão (Boyé, Tran e Grabar 2014, pp. 412–413), uma boa medida a ser tomada é o diagnóstico precoce, enquanto os sintomas são parcos ou pouco pronunciados, para que alguns tratamentos e terapias sejam aplicados a fim de estabilizá-los em um estágio inicial por mais tempo.

No âmbito médico, há evidências biológicas que podem ser coletadas a fim de identificar a instauração das doenças no ser humano (Hernández-Domínguez et al. 2016, p. 10). No entanto, são formas de diagnóstico custosas e tendem a ser buscadas quando a doença está já em um estágio mais avançado. Por outro lado, sabe-se que alterações cognitivas — dentre elas, alterações na linguagem — aparecem nos estágios iniciais das doenças e são uma das pistas que podem ser usadas para o diagnóstico precoce. Além disso, com a ajuda de ferramentas baseadas em Inteligência Artificial (IA), essas pistas podem fornecer novas fontes de suspeitas e confirmação para a composição do diagnóstico por

¹A princípio, vínhamos adotando a nomenclatura Déficit Cognitivo Leve (DCL) como tradução de *Mild Cognitive Impairment* (MCI); entretanto, durante a defesa desta dissertação, foi-nos sugerido que adotássemos Comprometimento Cognitivo Leve (CCL), sendo esse o termo mais atual na literatura em língua portuguesa. Dessa forma, passamos a empregar esse termo na presente Versão Corrigida deste trabalho, com exceção do título, resumo e palavras-chave, que, conforme o regramento da Pós-Graduação, não podem ser alterados após a entrega da Versão Original.

parte do corpo médico, contribuindo para reduzir custos para os pacientes. As intervenções para a leitura dessas pistas podem ser feitas com base em materiais de diversos tipos. Aqui, abordaremos os textuais, ou, mais especificamente, falas transcritas de autoria dos próprios indivíduos – sem deixar de mencionar que prontuários padronizados (Kumar et al. 2021; Wu et al. 2019) também são instrumentos importantes para propósitos similares.

A partir de uma seleção dos quatro *corpora* compilados nos *Datasets of Neuropsychological Language Tests in Brazilian Portuguese*² (DNLT-BP), neste trabalho, buscamos expandir, para o português brasileiro, o ferramental disponível para diagnóstico de DA e de CCL fazendo uso de recursos de aprendizado de máquina para automatizar essa tarefa. Essa proposta tem como premissa justamente o fato de a linguagem em pessoas desses grupos começar a decair em diversos aspectos linguísticos conforme as condições se manifestam e se agravam (Karlekar, Niu e Bansal 2018, p. 701; Vincze et al. 2016, p. 181).

Nos quatro *corpora*, estão transcritas, como parte das baterias de testes que os compõem, as falas de indivíduos saudáveis e com DA ou CCL. As tarefas dadas aos participantes são narrações elaboradas a partir de um estímulo como, por exemplo, a história da Cinderela, ou de um apoio visual (como narrar uma história com base em uma sequência de imagens).

Sucintamente, o processamento desses dados será feito com base numa versão pré-processada, por nós mesmos, de dois dos *corpora* do conjunto mencionado. Com essa filtragem e organização dos dados, eles serão submetidos a quatro modelos de teste, sendo o primeiro deles um bayesiano³ de nossa autoria; seguido pelo modelo `Sequential` da biblioteca `keras`⁴, como representante de uma rede neural de propagação para frente; além de uma rede `BiLSTM`⁵, também implementada via `keras`; e, finalmente, um modelo mais robusto, baseado em

²<https://github.com/nilc-nlp/DNLT-BP>

³O modelo bayesiano baseia-se na teoria de probabilidade condicional, em que a inferência é feita através do teorema de Bayes.

⁴A biblioteca `keras` foi escolhida para implementação do modelo devido à sua facilidade de uso e flexibilidade para construção de redes neurais artificiais. Além disso, trata-se da arquitetura que desejávamos com esse teste: de propagação para frente, isto é, informação flui em uma direção, da entrada para a saída apenas.

⁵Ao contrário do modelo sequencial, este modelo é alimentado com informações em dois tipos de camada: da entrada para a saída e vice-versa.

arquitetura *transformer*, o *distilbert-portuguese-cased*, proveniente do BERT (*Bidirectional Encoder Representations from Transformers*)⁶.

Com esse cenário no horizonte, desenham-se os objetivos expressos a seguir neste Capítulo. Além disso, apresenta-se também uma breve contextualização sobre como se comporta a linguagem humana em face da DA e do CCL. A seguir, no Capítulo 2 – *Revisão da literatura*, estão organizadas as principais contribuições que uma série de artigos e *proceedings* tiveram para amadurecimento desta pesquisa — seja como inspiração, influência direta ou contraste. Essa revisão não tem como objetivo ser exaustiva a respeito do estado da arte da classificação textual computacional, mas, sim, explorar cenários com que outros pesquisadores trabalharam e que contam com interseções ou distinções frente à nossa pesquisa; para isso, resenhamos brevemente os pontos de interesse das publicações selecionadas, trazendo também alguns comentários próprios. No Capítulo 3 – *Métodos*, exploramos em mais detalhes os *corpora* escolhidos (e o motivo de o terem sido), a composição deles, cobrindo as tarefas por que passaram os indivíduos e a forma como foram transcritas as falas; por fim, introduzimos os procedimentos adotados para análise dos dados e modelos adotados. No Capítulo 4 – *Resultados e discussão*, apresentamos uma sequência de dados estatísticos colhidos dos dados (como médias de palavras por sentença, de repetições de palavras por sentença; riqueza vocabular; taxa de *hapax legomena*), além de abordarmos os resultados do desempenho dos modelos (bayesiano, rede neural de propagação para frente, rede BiLSTM (bidirecional), modelo *transformers*), seguidos por uma análise linguística de diferentes manifestações de fenômenos de hesitação observados nos *datasets*. Ao final, temos o Capítulo 5 – *Conclusão*, em que elencamos pontos de confluência para com a literatura, as nossas contribuições por meio da atual pesquisa, além de limitações que enfrentamos e eventuais maneiras por meio das quais pesquisas futuras poderiam se ocupar a fim de superá-las.

⁶A arquitetura *transformer* é conhecida por sua eficiência no processamento de sequências e sua capacidade de capturar relações de longo alcance. Modelos com essa arquitetura são pré-treinados com quantidades substanciais de dados, proporcionando robustez e desempenho. Por outro lado, é um tipo de modelo que costuma requerer um poder computacional significativo a depender da tarefa e da quantidade de dados administrada.

1.1 Objetivos

Na pesquisa em curso, em linhas gerais, detemo-nos em dois principais objetivos, sendo um deles voltado ao tratamento dos dados focado no desenvolvimento de modelos que sejam capazes de reproduzir os diagnósticos recebidos pelos indivíduos de cada grupo (DA, CCL ou controle, sendo este último composto por indivíduos com diagnóstico negativo para DA ou CCL).

O segundo objetivo é a análise dos resultados gerados por esses modelos confrontando-os com execuções semelhantes relatadas na literatura.

Mais especificamente, o foco com o primeiro objetivo é a caracterização de um método para descrever linguisticamente as doenças e como se manifesta a linguagem nelas nos aspectos textuais passíveis de análise disponíveis no *corpus* – como o uso do vocabulário, contagens de palavras, repetições de palavras ou fragmentos de palavras, e tamanho médio das sentença. Ou seja, visamos a encontrar e usar os indícios linguísticos mais proeminentes na distinção entre a fala de uma pessoa saudável e a fala de uma pessoa afetada por alguma das duas patologias consideradas. Isso é feito por meio de duas classificações: uma mais abrangente, global, em que constem fatos relatados na literatura e por nós referentes ao comportamento da linguagem na presença das doenças, e uma específica, elaborada a partir dos resultados gerados pelos recursos da Linguística Computacional quando expostos os dados sensíveis aos modelos classificadores.

A seguir, com o segundo objetivo (necessário também para alimentar o primeiro), vem a necessidade de testar os modelos elaborados a fim de avaliar seus resultados, bem como estabelecer um comparativo com outros trabalhos publicados que tiveram objetivos de classificação semelhantes, mas que se valeram de um ferramental ou de premissas distintas.

Além desses objetivos aqui estabelecidos para serem conquistados na presente pesquisa, vale ressaltar como um objetivo complementar paramentar pesquisas e profissionais da saúde, como fonoaudiólogos, que lidem com a linguagem de pessoas com DA e CCL, favorecendo:

- O aumento da confiabilidade e precisão de diagnósticos;

- A criação de uma eventual ferramenta, para usuários finais, baseada nos nossos métodos, a fim de fornecer uma expectativa de que possam ou não se enquadrar como potenciais portadores de alguma das duas condições e possam procurar, mais direcionadamente, por auxílio médico;
- Recursos para a promoção de uma compreensão mais aprofundada sobre a expressão linguístico-discursiva de pessoas desses grupos.

Assim senso, a fim de ajudar na aplicação da metodologia aqui proposta de maneira a atingir estágios iniciais das doenças, ajudando a compor o diagnóstico precocemente, uma hipótese que se faz necessária é que, não apenas a DA, mas principalmente o CCL apresente distinções linguísticas suficientes para que o modelo aprenda as marcações que acometem a linguagem e ajude a classificar o grupo corretamente com um desempenho satisfatório.

1.2 Sobre a Doença de Alzheimer

Nesta Seção e na seguinte, [1.3](#), apresentamos brevemente as manifestações da DA e do CCL, bem como algumas das referências nas quais nos embasamos para essa pesquisa.

Da perspectiva da neuroanatomia, Patestas e Gartner ([2006](#), p. 358) resumizam a DA da seguinte forma:

[...] é causada por alterações patológicas, incluindo emaranhados neurofibrilares, placas neuríticas e degeneração neuronal, que aparecem inicialmente nas ilhas de células piramidais (da camada II) do córtex entorrinal. A partir daí, a degeneração se espalha para a zona CA1 do hipocampo propriamente dito e depois retorna para as camadas mais profundas do córtex entorrinal. Consequentemente, essa degeneração neuronal dificulta o fluxo normal de informações através da formação hipocampal. Confusão e déficits na função executiva ocorrem após o espalhamento adicional dos emaranhados neurofibrilares para o polo temporal e córtex pré-frontal. A patolo-

gia subicular ocorre aproximadamente ao mesmo tempo em que os emaranhados neurofibrilares invadem o neocórtex temporal.

Essa doença é uma condição de demência que não conta com uma cura compactuada pela comunidade científica, porém há formas, medicamentosas ou não, de retardá-la ou atenuar os efeitos que afetam a vida dos indivíduos em que se manifesta (Steiner et al. 2017). Para tanto, o quanto antes forem identificados indícios de que uma pessoa apresentará a doença, uma melhor qualidade de vida poderá ser oferecida para a pessoa, bem como uma redução de custos com tratamentos se torna mais factível (Frota et al. 2011; Nitrini, Brucki et al. 2021).

Dentre esses indícios, estão manifestações de alterações sobretudo em dois aspectos: marcadores biológicos e habilidades cognitivas em declínio. Em busca de diagnóstico, profissionais podem avaliar ambos os aspectos, sendo o primeiro de grande eficácia em estágios mais tardios da doença, enquanto o segundo vem se tornando cada vez mais buscado a fim de se estabelecer um diagnóstico precoce (Hernández-Domínguez et al. 2016, p. 10).

As alterações cognitivas se mostram uma pista confiável para o diagnóstico, uma vez que elas tendem a se manifestar antes de biomarcadores. A memória costuma ser um dos primeiros aspectos a manifestar problemas, enquanto habilidades linguísticas não enfrentam ainda sinais claros de deterioração. Após os estágios iniciais, “com a evolução dos problemas cognitivos e de linguagem, os problemas de comunicação tornam-se irreversíveis e impactam severamente o cotidiano das pessoas afetadas e de suas famílias” (Boyé, Tran e Grabar 2014, pp. 412–413).

1.3 Sobre o Comprometimento Cognitivo Leve

De acordo com Negash et al. (2007), o CCL se refere ao momento em que um indivíduo começa a desenvolver sinais de demência em uma idade em que se espera um desempenho cognitivo normal. Essa debilitação precoce acontece junto do enfraquecimento da memória, de dificuldades de aprendizagem explícita — por exemplo, decorar uma lista, narrações de uma história recentemente apresentada — e aprendizagem implícita — ou seja, quando se aprende algo

sem uma instrução dedicada, como a reprodução de movimentos, aquisição da língua materna e seus padrões —, embora esta última ainda necessite de comprovação científica mais extensa.

O CCL pode ser dividido em dois tipos: amnésico, em que a memória é o principal fator que sofre perdas, e não amnésico, em que outros domínios cognitivos são afetados. Além disso, cada um deles pode ser classificado como de domínio único ou múltiplo a depender, novamente, dos domínios cognitivos afetados (Beltrami et al. 2018). Qualquer um dos tipos pode evoluir para um cenário de demência, sendo casos de memória episódica um dos preditores. Em qualquer um dos casos, sendo um fator de degradação da linguagem, alterações linguísticas se tornam um ponto de interesse de estudo para entender quais são e em que medida aparecem as manifestações dessa condição na linguagem. A maior parte dos estudos, portanto, foca em avaliar “a habilidade verbal, a memória e a aprendizagem verbal, a fluência verbal para nomeação, categorias ou com base em letras, a memória verbal episódica” (Beltrami et al. 2018, p. 2).

O CCL entra nas análises da pesquisa porque é considerado frequentemente um estágio anterior à Doença de Alzheimer em termos de presença e intensidade de sintomas. Entre estes está o aparecimento de disfunções na produção de fala até nove anos antes de um diagnóstico médico de fato (Vincze et al. 2016, p. 181; Negash et al. 2007, p. 885), ou seja, durante o CCL, começam a aparecer as disfunções que tendem a se intensificar em caso de agravamento do caso para a doença de Alzheimer. Também devido ao fato de o CCL se estabelecer como esse estágio inicial da DA (Steiner et al. 2017) é que se faz premente sua identificação precoce, aumentando as chances de, ao menos, retardar o estabelecimento de graus de demência mais intensos como a DA. Esse agravamento em termos de estágios de demência é representado na Figura 1.1, apresentada por Savonenko et al. (2015, p. 323), em que notamos a progressão de acordo com os principais biomarcadores acometidos pelas doenças.

Com um fator de conversão de CCL para DA de cerca de 10% a 40% ao ano — contra 1% a 2% de desenvolvimento de DA em idosos saudáveis (Teixeira et al. 2012, p. 175) —, é justificável a importância de se tentar mapear marcadores que indiquem a manifestação de tal condição. Dentre esses marcadores, estão

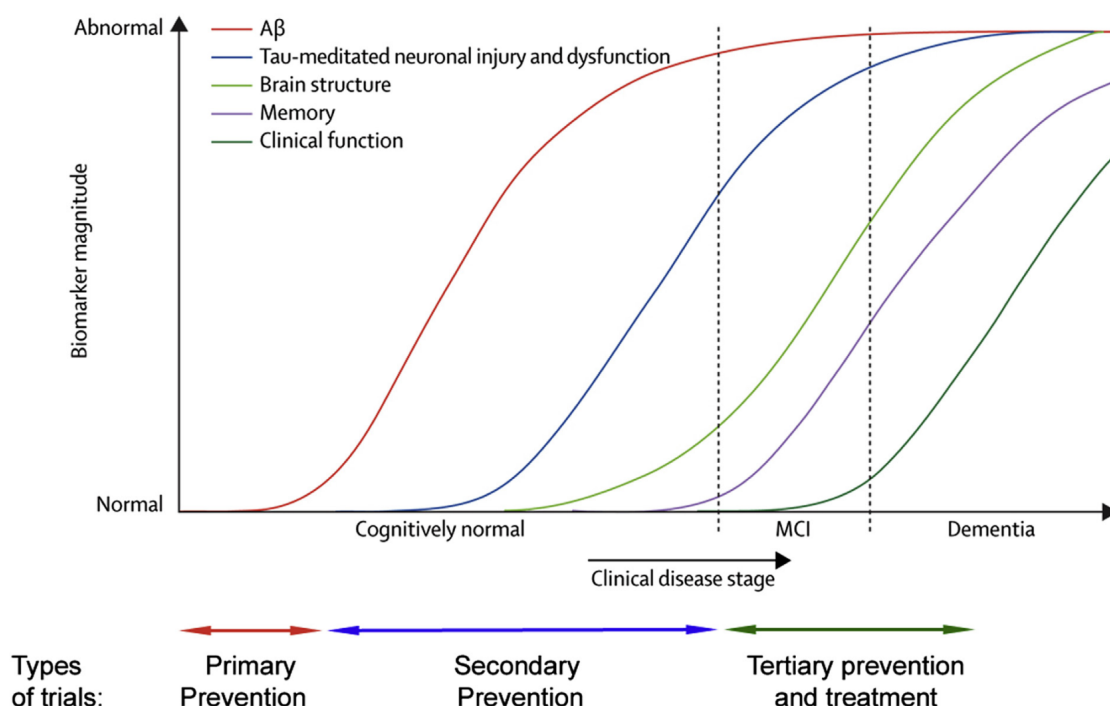


Figura 1.1: Dinâmica de biomarcadores na cascata patológica da doença de Alzheimer. O β -amiloide ($A\beta$) é identificado pelo fluido cerebrospinal (CSF) $A\beta_{42}$ ou pela imagem de amiloides por tomografia por emissão de pósitrons (PET). A lesão neuronal e disfunção mediadas por tau são identificadas pelo tau no CSF ou pelo PET com fluorodesoxiglicose. A estrutura cerebral é medida por ressonância magnética estrutural. (Legenda original traduzida.) Fonte: Savonenko et al. (2015, p. 323)

fenômenos linguísticos que podem servir de pista deste início de decaimento cognitivo já durante a fase do CCL (Frota et al. 2011, pp. 147–150), a fim de antecipar a busca por médicos e para diagnóstico.

Assim, uma vez que nosso *corpus* conta com transcrições realizadas a partir da fala de indivíduos que compõem os três grupos (Alzheimer, CCL e controle) tendo passado pelas mesmas tarefas, um de nossos objetivos é construir um classificador capaz de distinguir todos os três, além de ser capaz também de notar essa gradação no estágio de desgaste linguístico refletido como sintoma de cada diagnóstico.

Revisão da literatura

Neste capítulo, detenho-me em resenhar brevemente alguns artigos cujas contribuições foram relevantes para o andamento da pesquisa. Os artigos escolhidos serão apenas parcialmente comentados, uma vez que, por vezes, tratam de mais temas de análise computacional de dados que não linguísticos, como exames de imagem por ressonância magnética e aspectos sociodemográficos. O objetivo deste capítulo é situar brevemente o leitor a respeito de técnicas de análise iguais ou diferentes em relação às nossas; tipos de materiais estudados; tipos de recorte de grupos de indivíduos de interesse; entre outras abordagens metodológicas adotadas por diversos autores em seus projetos. Assim sendo, não se trata de uma revisão exaustiva, mas um pontual levantamento de tópicos caros à nossa pesquisa seja pela semelhança ou mesmo pela diferenciação.

2.1 Kumar et al. (2021)

Kumar et al. (2021) exploram em seu texto, que se pretende uma revisão sistemática de literatura, estudos publicados numa janela de dez anos, de 2010 a 2020, acerca de aplicações de “aprendizado de máquina em dados clínicos e de Prontuários Eletrônicos do Paciente (PEP) para identificar fatores de predição de risco para progressão da demência da doença de Alzheimer”. As diretrizes seguidas pelos autores do estudo se resumem nas seguintes perguntas:

1. Que tipo de métodos de AM têm sido usados para a detecção do surgimento da demência na DA e para prever a trajetória da progressão da doença?
2. Que tipos de dados derivados de PEPs e fatores de risco (como fisiológico, genético, demográfico) foram usados como recursos para modelagem preditiva?
3. Quais são os focos de pesquisa dos artigos revisados que usam métodos de AM em dados derivados de PEP para modelar e prever a progressão da demência na DA?

Quanto às técnicas de AM aplicadas para solucionar tarefas de modo a paramentar a capacidade de decisão clínica, os autores separam os artigos nos seguintes grupos: métodos de regressão, máquina de vetores de suporte (SVM), árvore de decisão, redes bayesianas, redes neurais e diversas técnicas de processamento de língua natural (Kumar et al. 2021, p. 5).

Entre os artigos revisados, houve apenas quatro (6%) que levaram em consideração anotações clínicas em suas análises, portanto, sendo os únicos que se valeram do PLN e suas técnicas. Ainda assim, vejamos as técnicas de que se valeram esses artigos na análise da produção linguística.

Segundo os autores, essa escassa produção que considerava as anotações clínicas acerca do acompanhamento realizado por equipes médicas com os pacientes não contava com transcrições feitas de modo padronizado. Assim, esses textos não poderiam ser submetidos a análises de algoritmos de AM clássicas como as mencionadas acima. Asseguram que técnicas convencionais de *word embeddings* — mencionadas em Wu et al. (2019, p. 460) de acordo com a frequência que encontraram ao fazer outra revisão sistemática: Word2Vec, com predominância de 74,1%, e GloVe, com 9,9% — têm sido menos eficazes na extração de informação relevante se comparadas com métodos de aprendizagem profunda. Ainda nesse tópico, por fim, sugerem que novas publicações deveriam seguir padronizações de *corpora* de PEPs, além da maior divulgação de trabalhos em andamento e publicações num contexto entre instituições, para aumentar a reprodutibilidade de métodos e resultados.

2.2 Dreisbach et al. (2019)

Nesse artigo, os autores elaboram uma revisão sistemática cujo objetivo é analisar o estado da arte com relação ao uso de PLN na extração de sintomas a partir de textos eletrônicos de autoria do paciente. As fontes foram 21 artigos (selecionados a partir dos portais PubMed e EMBASE) que contavam com estudos feitos a partir de buscas em bases de dados como comunidades e fóruns virtuais orientados ao relato e ao esclarecimento de dúvidas acerca de doenças e sintomas, além de redes sociais, como o Twitter e o Reddit.

Os artigos por eles selecionados foram classificados em três grupos de acordo com a abordagem, sendo elas: uso de PLN (sendo 14 artigos nessa condição), uso de mineração de texto (com 6 representantes) e ambas (contando um artigo). Os autores questionam a eficácia de se usar extração automática de sintomas em comparação com formas manuais devido à “falta de léxicos de sintomas formais e leigos ou coloquiais padrão-ouro” (p. 43). Isso fez com que, por exemplo, um dos softwares usados em mais de um dos estudos analisados, Treato, não performasse de maneira consistente neles.

Os autores endossam a ideia de automatizar tarefas simples para escalar o que em mãos humanas seria muito custoso quanto à raspagem de texto. Essas automatizações devem ser personalizadas de acordo com a plataforma on-line com que se trabalha (p. 44). Eles notaram que a proporção de estudos que se pararam de ferramentas de mineração textual (como “frequência de palavras e análise de sentimento”, em oposição ao PLN) foi maior (28,6%) em comparação com outro estudo sistemático anterior que realizaram focado em PEPs (Koleck et al. 2019), no qual essa porcentagem chegou a 7,4% apenas. Portanto, a extração de informação de dados clínicos é mais precisa e direta em se tratar de PEPs, documentação em que a organização de dados é mais bem padronizada.

2.3 Boyé, Tran e Grabar (2014)

O objetivo dos autores com esse artigo é “estudar a linguagem de pacientes com DA produzida em contexto conversacional com um interlocutor familiar

[...] [por] ser mais relevante para estudar a especificidade da linguagem na DA” (Boyé, Tran e Grabar 2014, p. 414). Isso é feito usando métodos e ferramentas de PLN. Para usar esses recursos, as transcrições foram feitas com ortografia correta, mas sem deixar de contar com marcações de pausas, hesitações e disfluências. Dentre os recursos estão Transcriber¹, usado para transcrever; TreeTagger² e Flemm³, para, respectivamente, anotar e corrigir a análise gramatical sintaticamente; e DériF⁴, para análise de lemas.

A análise de interação verbal é dada sob a óptica dos turnos de fala e tempo de fala, incluindo sobreposições. A análise do discurso conversacional estudada se vale de traços da fala (como pausas vazias e não vazias; correções, disfemias e autocorreções), lexicais (como informatividade da sentenças — pode se tratar apenas de interjeições, respostas sim/não —, diversidade lexical — considera a quantidade de tipos de lemas —, complexidade morfológica, frequência lexical), e sintáticos (comprimento médio de enunciados; sentenças interpoladas e discurso indireto; quantidade de lemas de verbos e de pronomes pessoais).

Entre as características lexicais consideradas nas falas do grupo com DA e o controle, observaram-se como relevantes para o diagnóstico da doença (isto é, quando os grupos diferiam por 20% ou mais em determinado quesito) as seguintes (Boyé, Tran e Grabar 2014, p. 418, tradução nossa):

o número de palavras (maior no grupo-controle), o número de enunciados *Sim/Não* (maior no grupo DA), a diversidade lexical (maior no grupo-controle), a razão entre lemas/ocorrências total (maior no grupo-controle), as frequências médias nos *corpora* estudados (maior no grupo DA porque o léxico é mais redundante) e na web (maior no grupo DA porque os pacientes com DA usam palavras mais comuns e frequentes). As características de informatividade e complexidade morfológica são comparáveis em ambos os grupos. Quanto à complexidade morfológica, vários afixos comuns (como *-ment*, *-tion*, *dé-*, *re-*

¹<http://trans.sourceforge.net/en/presentation.php>

²<https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

³<https://www.cnrtl.fr/outils/flemm/>

⁴<https://www.cnrtl.fr/outils/DeriF/>

[do francês]) não são processados atualmente pelo Dérif. Assumimos que o tratamento deles pode alterar o impacto dessa característica.

Quanto às sintáticas, temos: “a duração média dos enunciados (maior no grupo-controle), as orações interpoladas e fala relatada (maior no grupo-controle), pronomes pessoais (maior no grupo DA), número de verbos (maior no grupo-controle)”.

As hipóteses e os métodos que adotamos em nossa pesquisa, que será detalhada nos capítulos a seguir, vão se mostrar em consonância com os de Boyé, Tran e Grabar (2014), uma vez que não somente a natureza dos *corpora* bem como o método como foram transcritos se assemelham em larga escala, mas também os resultados relativos aos aspectos analisados por meio de recursos de PLN em pessoas com DA poderão ser comparados mais adiante de modo complementar.

2.4 Beltrami et al. (2018)

Neste artigo, os autores relatam o estudo que conduziram com o objetivo de diferenciar, via processamento de narrativas de cada indivíduo, o grupo-controle ($n = 48$) e os grupos clínicos (incluindo CCL amnésico ($n = 16$), CCL de múltiplo domínio ($n = 16$) e demência precoce ($n = 16$)). Cada indivíduo produziu duas narrativas (baseadas nos seguintes estímulos: “Você poderia, por favor, descrever um dia de trabalho comum?” e “Você poderia, por favor, descrever o último sonho de que se lembra?”), bem como uma descrição a partir de uma imagem de uma sala de estar.

Além de cada indivíduo ter sido submetido a uma série de testes neuropsicológicos, com a gravação de suas falas em mãos, os pesquisadores as analisaram do ponto de vista acústico, tendo podido gerar medidas acerca de traços como a duração de segmentos de silêncio e o tempo de fonação padronizado. Quanto aos parâmetros medidos a partir da descrição — de maior relevância para nossa pesquisa —, os autores chegam a conclusões como a marcadamente menor riqueza lexical no lado dos grupos clínicos.

Em mais detalhes e filtrando os resultados que os pesquisadores obtiveram com traços similares e relevantes aos analisados durante nossa pesquisa, isto é, os traços lexicais e sintáticos investigados, pode-se destacar entre os primeiros um evidente empobrecimento global na produção, além de a “densidade de conteúdo (ou seja, a proporção de palavras de classe aberta sobre palavras de classe fechada) ser consistentemente reduzida, especialmente para a tarefa de descrição de imagem” (Beltrami et al. 2018, p. 5). Quanto aos traços sintáticos, os autores chegam a conclusões semelhantes, agora notando pobreza nas estruturas das sentenças: “estruturas sintáticas produzidas pelo grupo clínico, apesar de gramaticalmente corretas e coerentes, contêm relações menos complexas entre as frases e menos estruturas subordinadas” (Beltrami et al. 2018, p. 6).

2.5 Hernández-Domínguez et al. (2016)

Nesse texto, os autores se propõem a apresentar métodos automatizados para se analisar descrições de objetos comuns textuais transcritas de modo a determinar se o indivíduo tem doença de Alzheimer ou não. Eles argumentam em prol da análise de discurso livre em detrimento de testes, que podem gerar resultados menos conclusivos baseados em tarefas que podem ser afetadas por fatores como prática ou nervosismo (como nomear objetos em uma imagem ou listar palavras com a mesma inicial). A fim de comparar resultados, os autores realizam o processamento de um *corpus* de descrições de objetos feitas por pacientes espanhóis com base em um classificador SVM (máquina de vetores de suporte) treinado por eles; este mesmo *corpus* foi processado por uma rede bayesiana por outros autores e o resultado deles é o alvo de comparação.

Os traços analisados foram 7 (Hernández-Domínguez et al. 2016, p. 12): taxas de verbos, substantivos, preposições e conjunções (medidas a cada 100 palavras); taxa de verbos secundários (número de verbos secundários dividido pelo número de verbos); índice W de Brunet (para determinar riqueza vocabular); e estatística R de Honoré (para medir riqueza lexical com base em *hapax legomena*).

As análises foram divididas em dois experimentos: as descrições produzidas foram classificadas de acordo com o item transcrito em animadas (como “ca-

chorro”) e inanimadas (como “carro”); o primeiro experimento contou apenas com essa separação do *corpus* nesses dois grupos, tendo sido medidos 14 traços (os mesmos 7 acima, uma vez em cada grupo); o segundo experimento se valeu do agrupamento por indivíduo, tendo sido avaliados os 7 traços.

Dois *kernels* de svm foram usados: um linear (modelo com o qual o segundo experimento teve uma melhor acurácia, de 88%) e um de função de base radial (RBF) (com o qual o primeiro experimento teve uma melhor acurácia, de 86%). Como conclusões a que chegaram os pesquisadores, destaco a seguinte (Hernández-Domínguez et al. 2016, p. 14, tradução nossa):

O classificador do segundo experimento tem uma sensibilidade levemente maior (2% a mais), o que significa que ele tem uma tendência menor de deixar os participantes com DA passarem despercebidos. [...] Com isso, concluímos que para os traços linguísticos considerados, não há necessidade de separar as descrições dos participantes em categorias animadas e inanimadas.

2.6 Santos et al. (2017)

No artigo, os autores detalham um trabalho com narrativas e descrições produzidas por indivíduos saudáveis (controle) e com CCL e se valem de redes complexas para que classificadores, por meio de aprendizado de máquina, distingam ambos os grupos (Santos et al. 2017). Os *corpora* que analisaram foram o DementiaBank (contém descrições em inglês), o Cinderella e o ABCD (Arizona Battery for Communication Disorders of Dementia) (ambos contêm narrativas em português). Sobre o segundo *corpus*, vale notar que foram submetidas aos testes duas versões diferentes: na segunda, houve um trabalho de revisão que, em geral, removeu uma série de tokens dos textos, como hesitações, repetições, emissões não relacionadas à narração, vogais prolongadas; por outro lado, sujeitos omitidos foram inseridos. Esse tratamento, notavelmente custoso, casou as mudanças nas médias de sentenças por narrativa e palavras por sentença apresentadas na tabela 2.1 (Santos et al. 2017, p. 1288).

Cinderella	Sem revisão		Com revisão	
	CCL	Controle	CCL	Controle
Média de sentenças por narrativa	29,9	30,8	31,4	45,1
Média de palavras por sentença	13,03	12,17	10,91	8,17

Tabela 2.1: Médias obtidas antes e depois das alterações realizadas no *corpus* Cinderella.

Após um pré-processamento dos textos, os pesquisadores usaram um modelo de coocorrência de palavra “porque a maioria das relações sintáticas ocorre entre palavras vizinhas [...] Cada palavra distinta se torna um nó e palavras adjacentes no texto são conectadas por uma aresta” (Santos et al. 2017, tradução nossa). Para lidar com o problema de o tamanho dos *corpora* adotados ser reduzido, adaptaram “a abordagem de induzir redes linguísticas a partir de *word embeddings* [...] [dessa forma,] para cada par de palavras no texto que não estivessem conectadas, uma aresta era criada caso seus vetores tivessem uma similaridade de cosseno maior que certo limiar” (Santos et al. 2017, pp. 1286–7).

Os traços utilizados na avaliação dos modelos foram três: métrica topológica de redes de coocorrência (aqui, os *word embeddings* são gerados a nível de caractere, havendo um “bag of character *n*-gram”, o que, segundo pesquisa dos autores, melhora a avaliação sintática em comparação com modelos tradicionais), traços linguísticos (a extração de traços foi feita com a ferramenta Coh-Metrix⁵, para inglês, e Coh-Metrix-Dementia, para português) e representações de *bag of words*. Os algoritmos de aprendizado de máquina usados na classificação foram o classificador bayesiano ingênuo gaussiano (GNB), *k*-Nearest Neighbor (*k*-NN), Máquina de Vetores de Suporte (SVM), funções de bases linear e radial (RBF) e Random Forest (RF).

Quanto aos resultados, destacam-se alguns pontos: enriquecer a rede com *word embeddings* tende a gerar melhores resultados do que não fazê-lo; SVM apresenta os melhores resultados de acurácia. A respeito dos *corpora* em específico, DementiaBank e Cinderella não revisado tiveram melhor acurácia com as redes com *word embeddings* (62% e 65%, respectivamente); ABCD performou melhor com *bag of qwords* (75%); por fim, Cinderella revisado tem maior acurá-

⁵<http://cohmetrix.com/>

cia com as métricas linguísticas (72%), indicando a eficácia da revisão para uma classificação mais precisa (Santos et al. 2017, pp. 1291–2).

2.7 Karlekar, Niu e Bansal (2018)

Nesse trabalho, os pesquisadores se valem de redes neurais de dois tipos a fim de classificar transcrições como provenientes ou não de pacientes com doença de Alzheimer. As redes usadas são as convolucionais (CNN) e as recorrentes (RNN, no caso, a utilizada é uma LSTM); além disso, foi testada uma combinação de ambas. A CNN consiste em (Karlekar, Niu e Bansal 2018, tradução nossa):

aplicar uma camada de *embedding* e uma convolucional, seguidas por uma camada *maxpooling*. Os traços convolucionais são obtidos ao se aplicar filtros de tamanhos de janela variáveis a cada janela de palavras. O resultado, então, passa por uma camada *softmax*, a qual retorna as probabilidades sobre as duas classes.

Já a LSTM, para trabalhar com dependências mais distantes com melhor eficácia, consiste em uma “camada de *embedding* seguida por uma LSTM. O estado final, contendo informações de toda a sentença, é alimentado a uma camada totalmente conectada seguida por uma camada *softmax* para obter as probabilidades de retorno” (Karlekar, Niu e Bansal 2018, p. 702). Devido ao caráter complementar, foi montada uma arquitetura híbrida em que o texto fosse processado pela CNN e pela RNN, nessa ordem.

Os autores não encontraram uma diferença estatisticamente significativa em termos de gênero ao opor os indivíduos com Alzheimer de cada grupo, tendo, por exemplo, ambos os grupos as mesmas categorias gramaticais em ordem de frequência.

Como forma de garantir uma clara compreensão dos traços analisados de modo a evidenciar a diferença entre a produção linguística de cada grupo, os autores investiram em técnicas de visualização, sendo elas Activation Clustering e First Derivative Saliency. Dentre outras conclusões, essas técnicas ajudaram a atestar que indivíduos com DA tendem a empregar mais preenchedores (como

“uh” e “um”) e a iniciar *clusters* com conjunções coordenativas em comparação com indivíduos saudáveis, enquanto determinantes têm um papel mais marcante na classificação dos *clusters* na fala de grupos-controle (Karlekar, Niu e Bansal 2018, p. 705).

2.8 Vincze et al. (2016)

Vincze et al. (2016) se baseiam em traços morfológicos para distinguir entre pacientes húngaros já diagnosticados com distúrbio cognitivo leve e o grupo-controle. O *corpus* foi baseado na tarefa de descrição de três itens: os 84 indivíduos assistiram a dois filmes para, em seguida, elaborarem descrições, nesta ordem: o primeiro filme, o dia anterior deles e então o segundo filme; assim, testando o fator memória afetado pelo CCL. Fenômenos fonológicos (como apagamentos) foram mantidos; pausas e outras marcas de discurso foram representadas na transcrição também.

A ferramenta de pré-processamento linguística para o húngaro Magyarlanc⁶ foi utilizada para análises sintática e morfológica. As análises levaram em conta 4 conjuntos de traços com base na produção ou nos indivíduos: os de fala espontânea (relacionados a pausas e hesitações, por exemplo), morfológicos (como número e taxa de classes gramaticais), semânticos (como número e taxa de termos relacionados à memória, por exemplo, algo como “não me lembro”) e demográficos (gênero, idade e educação).

Estatisticamente, foram analisados os traços que se mostraram mais proeminentes na distinção dos indivíduos entre os dois grupos. Assim sendo, com os traços listados junto à significância (valor-*p*), nota-se que a maioria deles se mostrou relevante para a distinção, sobretudo na segunda tarefa em que o indivíduo narra seus feitos do dia anterior.

Os pesquisadores usam SVMs como técnica de aprendizado de máquina e, além disso, usam validação cruzada do tipo *leave-one-out*, devido ao tamanho restrito do *corpus* com que lidaram — o que também é válido de ser empregado

⁶<https://github.com/zsibritajanos/magyarlanc>

na pesquisa em curso pois também conta com um *corpus* que não ultrapassa centenas de indivíduos.

Por fim, a acurácia que o modelo húngaro atingiu no diagnóstico foi de 69,1%. Uma tentativa de remover os traços que não se mostraram estatisticamente significantes se mostrou exitosa ao elevar a acurácia a 75%, levando a crer que “alguns [dos] traços originais são supérfluos e só confundiram o sistema” (Vincze et al. 2016, p. 185). Analisando alguns erros, os autores notaram que pacientes com CCL cujas sentenças eram curtas, devido ao número reduzido de pausas e hesitações, tendiam a ser classificados como controle; por outro lado, os “indivíduos saudáveis que falavam mais e também hesitavam mais [e cujo] uso de pronomes e conjunções também era mais similar ao dos pacientes com CCL” recebiam um diagnóstico falso positivo para a condição.

2.9 Abrisqueta-Gomez et al. (2004)

O artigo apresenta um estudo longitudinal sobre os benefícios de um programa de reabilitação neuropsicológica (PRN) em pacientes com doença de Alzheimer em fase inicial a moderada. O objetivo do estudo foi avaliar a duração do benefício do PRN em termos de melhora cognitiva, estabilização funcional e redução dos problemas comportamentais nos pacientes. Os resultados mostraram que após o primeiro ano do PRN houve uma melhora cognitiva, estabilização funcional e redução dos problemas comportamentais nos pacientes. No entanto, observou-se que essa melhora não se estendeu para o segundo ano, mostrando a doença sua característica progressiva. O estudo sugere que o PRN pode ser uma intervenção útil para melhorar a qualidade de vida dos pacientes com doença de Alzheimer em fase inicial a moderada, mas que os benefícios podem ser limitados no longo prazo.

A respeito da doença de Alzheimer em si, os autores a apresentam como uma doença neurodegenerativa que afeta principalmente a memória e que também pode afetar outras funções cognitivas, como a linguagem, a atenção e o raciocínio. A doença é progressiva e pode levar a uma perda gradual da independência e da capacidade de realizar atividades da vida diária. Além disso,

a doença pode causar problemas comportamentais, como agitação, agressividade e depressão, que podem afetar a qualidade de vida dos pacientes e de seus cuidadores.

O PRN consiste em uma intervenção terapêutica que visa melhorar a qualidade de vida de pacientes com doenças neurológicas, como a doença de Alzheimer. O programa utiliza técnicas de treinamento cognitivo e atividades da vida diária para trabalhar a memória, a atenção, a linguagem e outras funções cognitivas preservadas nos pacientes. O objetivo do PRN é ajudar os pacientes a manter ou melhorar suas habilidades cognitivas e funcionais, retardando a progressão da doença e reduzindo os problemas comportamentais associados. Os três indivíduos acompanhados participam da avaliação em três momentos: antes da intervenção do PRN (T₁), depois de 12 (T₂) e 24 (T₃) meses após o PRN.

Dentre os aspectos de análise do PRN, estão incluídas nos testes tarefas que avaliam a orientação temporal e espacial, atenção e concentração, por exemplo. Além dessas, há também um foco em tópicos linguísticos, a saber: “tarefa de fluência verbo-semântica (animais) e uma tarefa de fluência fonológica (palavras iniciadas com a letra F)” (Abrisqueta-Gomez et al. 2004, p. 779); nos resultados também aparecem avaliações a respeito de compreensão, nomeação, leitura, escrita e cópia de sentença.

Os resultados do estudo mostraram que após o primeiro ano do programa de reabilitação neuropsicológica (PRN), houve uma melhora cognitiva, estabilização funcional e redução dos problemas comportamentais nos pacientes com doença de Alzheimer em fase inicial a moderada. No entanto, observou-se que essa melhora não se estendeu para o segundo ano, mostrando que a doença manteve sua característica progressiva. Portanto, o estudo sugere que o PRN pode ser uma intervenção útil para melhorar a qualidade de vida dos pacientes com doença de Alzheimer em fase inicial a moderada, mas que os benefícios podem ser limitados no longo prazo.

Quanto aos resultados das análises de aspectos linguísticos, as variações entre as avaliações ocorreram da seguinte forma: de T₁ para T₂, o teste de fluência verbo-semântica (animais) piorou a performance; por outro lado, o de fluência fonológica (palavras iniciadas com a letra F), o de compreensão e

o de leitura melhoraram; os demais apresentaram manutenção do resultado. De T2 para T3, ambos os testes de fluência apresentaram declínio; e os demais, manutenção.

2.10 Nitrini, Caramelli et al. (2005)

O artigo do Departamento Científico de Neurologia Cognitiva e do Envelhecimento da Academia Brasileira de Neurologia tem como objetivo estabelecer condutas padronizadas, normas, recomendações ou sugestões para o diagnóstico clínico de doença de Alzheimer no Brasil. A necessidade de condutas autorais para a população brasileira é indispensável, dada a desigualdade da sociedade em diversos níveis, tais como educacional e financeiro. Para isso, foram avaliados sistematicamente artigos sobre o diagnóstico de DA no Brasil disponíveis no PUBMED ou LILACS. A análise do estado-da-arte do diagnóstico no país se deu por meio da busca de (Nitrini, Caramelli et al. 2005, p. 721):

questionários, escalas e testes que já haviam sido aplicados no Brasil para o diagnóstico de DA. Obedeceu-se ao princípio de que este diagnóstico pode ser feito, na maioria das vezes, pelo exame clínico (com a exclusão de outras possibilidades diagnósticas mediante exames complementares), e que este exame deve ser de aplicação simples e breve. Quando o exame do médico for insuficiente para estabelecer o diagnóstico, deve ser complementado por avaliação neuropsicológica especializada.

Os autores trazem uma seção sobre “Avaliação da linguagem” (Nitrini, Caramelli et al. 2005, p. 723) que aborda a importância da avaliação da linguagem no diagnóstico da doença de Alzheimer. O texto destaca que, nos estágios iniciais da doença, a pessoa pode apresentar problemas semântico-lexicais e dificuldades semântico-discursivas na interpretação de metáforas, provérbios, moral de histórias e material humorístico. Já nos estágios intermediários, podem ocorrer alterações similares às da afasia de Wernicke ou afasia transcortical sensorial. O artigo apresenta recomendações práticas para a avaliação da lin-

guagem, sugerindo o uso de testes específicos, como o Teste de Nomeação de Boston, o de nomeação de objetos reais do ADAS-Cog ou o de nomeação de oito figuras do NEUROPSI.

2.11 Nitrini, Brucki et al. (2021)

Os objetivos do artigo são realizar uma revisão narrativa da origem da Bateria Breve de Rastreio Cognitivo (BBRC), relatar todos os estudos que utilizaram o Teste de Memória de Figuras (TMF) da BBRC e demonstrar que essa é uma bateria útil para regiões cuja população possui formação educacional heterogênea, como é o caso do Brasil. Para tal, foi realizada uma busca nas bases de dados PubMed, SciELO e LILACS utilizando os termos “Brief Cognitive Screening Battery” e “Brief Cognitive Battery”. Foram obtidos um total de 49 artigos no PubMed, 32 na SciELO e 28 na LILACS. Após a exclusão de artigos duplicados, totalizaram-se 54 publicações; mais cinco estudos foram incluídos com base no conhecimento prévio dos autores. Vinte e quatro artigos foram relacionados ao impacto da educação no desempenho, à precisão, pontuações de corte e estudos normativos.

Os estudos selecionados foram apresentados em ordem cronológica, de acordo com o ano de publicação e o tipo de estudo, e foram classificados em dois tipos: aqueles que demonstraram baixo ou nenhum impacto da educação ou que estavam relacionados à validade de construto ou a estudos normativos; e aqueles que utilizaram a BBRC, particularmente o TMF, para confirmar ou excluir o comprometimento da memória.

Os autores relatam que 37 artigos utilizaram o TMF da BBRC em estudos clínicos em diferentes ambientes, de ambulatórios de clínicas especializadas a estudos epidemiológicos e na avaliação de indivíduos residentes nas margens de rios da bacia amazônica, e sempre foi considerado de fácil aplicação. Além disso, os autores relatam que a recordação diferida do TMF mostrou a melhor precisão para o diagnóstico de demência com uma pontuação de corte de ≤ 5 em diferentes níveis educacionais.

Os autores concluem que o TMF da BBRC é uma ferramenta simples e rápida para o diagnóstico de demência em populações com heterogeneidade educacional.

2.12 Steiner et al. (2017)

O objetivo deste artigo é ajudar os clínicos gerais a atuar na detecção do risco de desenvolver alguma condição de demência em idosos, com foco no comprometimento cognitivo leve e na sua progressão para a doença de Alzheimer. O texto apresenta uma revisão da literatura sobre o tema, com o objetivo de fornecer informações úteis para profissionais de saúde que trabalham com pacientes idosos. O texto destaca a importância da detecção precoce do CCL e da DA para permitir o início de tratamentos e intervenções que possam retardar a progressão da doença e melhorar a qualidade de vida do paciente e de seus cuidadores.

De acordo com os autores, o CCL é caracterizado por alterações cognitivas que não interferem significativamente nas atividades diárias do indivíduo. Os sintomas podem incluir dificuldade em lembrar informações recentes, problemas de atenção e concentração, dificuldade em tomar decisões e realizar tarefas complexas, entre outros. O CCL é considerado um estágio inicial da DA, e muitas pessoas com CCL eventualmente desenvolvem a DA.

São apresentadas informações sobre critérios para o diagnóstico do MCL, dentre elas, temos que “o indivíduo não é normal nem portador de demência; evidência de declínio cognitivo medido objetivamente ou com base na percepção subjetiva combinada com comprometimento cognitivo objetivo; preservação ou comprometimento mínimo da vida básica e de atividades instrumentais complexas” (Steiner et al. 2017, p. 652).

Os autores mencionam que identificar indivíduos com CCL e analisar as comorbidades associadas podem ser oportunidades para direcionar futuras intervenções e prevenir a progressão para a DA. O texto não especifica quais são essas comorbidades, mas apresenta informações sobre fatores de risco para a DA em geral, que podem ser relevantes para pacientes com CCL. Esses

fatores incluem idade avançada, história familiar de DA, presença de certas variantes genéticas, baixa escolaridade, tabagismo, sedentarismo, obesidade, hipertensão arterial, diabetes mellitus e depressão. É importante ressaltar que a presença desses fatores de risco não significa necessariamente que um indivíduo desenvolverá a DA, mas eles podem aumentar o risco de progressão da doença em pacientes com CCL.

2.13 Frota et al. (2011)

O objetivo desse consenso é recomendar novos critérios para diagnóstico de demência e doença de Alzheimer no Brasil. O diagnóstico é designado quando há sintomas cognitivos ou comportamentais que interferem na capacidade de trabalhar ou realizar atividades habituais, representam declínio em relação aos níveis pré-mórbidos de funcionamento e desempenho e não podem ser explicados por delírio ou doença psiquiátrica importante. A nova proposta exige comprometimento funcional e cognitivo, sendo este último atingido por dois dos cinco domínios a seguir: memória, função executiva, linguagem, habilidade viso-espacial e mudança de personalidade. Reproduzimos aqui, não exaustivamente, alguns dos critérios e apontamentos apresentados, sobretudo relacionados à linguagem (Frota et al. 2011, pp. 147–150):

- Discurso (expressão, compreensão, leitura e escrita), com sintomas que incluem: dificuldade em encontrar e/ou compreender palavras, erros na fala e na escrita e troca de palavras ou fonemas, não explicáveis por comprometimento sensorial ou motor.
- Apresentação não amnésica (deve haver outro domínio afetado): fala (lembrar palavras).
- Evidência de comprometimento em um ou mais domínios cognitivos, normalmente incluindo memória, obtida por meio de avaliação que abrange os seguintes domínios cognitivos: memória, função executiva, fala e habilidades viso-espaciais ou exame neuropsicológico.

Outra contribuição dos autores é a segmentação da DA em fases, sendo o CCL uma delas:

- Fase pré-clínica;
- Comprometimento cognitivo leve causado por DA;
- Demência.

2.14 Toledo et al. (2018)

Os autores deste artigo começam por destacar a crescente população idosa, a relevância do envelhecimento como um problema de saúde, e a associação entre idade e incidência de demência, com foco na doença de Alzheimer (DA). A importância dos distúrbios de linguagem nas fases iniciais da DA é ressaltada, especialmente através da análise de discurso, com ênfase nos níveis macrolinguísticos. Eles revisitam a literatura para evidenciar as perturbações linguísticas no CCL e destacam a necessidade de identificação precoce desses traços. A teoria de Kintsch e van Dijk é mencionada para apoiar o modelo de análise de microestrutura e macroestrutura no estudo do discurso em DA. Eles fazem uso da narrativa da Cinderela (um dos *corpora* com que trabalhamos em nossa pesquisa). O estudo é justificado pela busca por ferramentas que facilitem a observação de resultados de intervenções clínicas na linguagem em demência, com uma hipótese específica sobre métricas diferenciadoras entre grupos.

A amostra incluiu 60 indivíduos divididos em três grupos: DA leve, CCL amnésico e um grupo-controle saudável. A avaliação do discurso utilizou um livro com a história da Cinderela, sendo os participantes instruídos a narrá-la em suas palavras. As narrativas foram gravadas, transcritas manualmente e editadas para análise. A análise dos dados envolveu o uso do SPSS 14.0 (*Statistical Package for Social Sciences*) e do Coh-Metrix-Dementia para extração de métricas computacionais. Um manual de notas de proposições em sentenças foi utilizado em três fases: remoção de palavras específicas, segmentação do texto em frases e marcação de proposições narrativas. A concordância entre avaliadores foi verificada pelo índice Kappa. 28 proposições foram agrupadas

em quatro componentes narrativos. As macrocaracterísticas do discurso foram analisadas quanto à informatividade, coerência global e modalização. O estudo buscou identificar marcadores distintos entre os grupos e contribuir para a compreensão das características discursivas em contextos de Alzheimer e CCL.

Sobre informatividade e estrutura narrativa, pacientes com DA leve apresentaram discursos menos informativos em comparação com CCL amnésico e controle. A estrutura narrativa revelou desempenho semelhante entre CCL amnésico e controle, superando DA leve. Quando a coerência global e modalização, diferenças foram encontradas na média de características entre sentenças adjacentes e similaridade entre pares de sentenças, destacando CCL amnésico com os valores mais baixos. Indivíduos com DA leve mostraram maior desvio padrão entre sentenças. A densidade total de ideias revelou menor quantidade em DA leve. Além disso, DA leve apresentou maior produção de sentenças vazias e modalizações em comparação com CCL amnésico e controle.

Limitações incluem a necessidade de uma amostra de participantes mais extensa, considerando a complexidade da atividade discursiva, e a recomendação de ambiente adequado para análise acústica.

Métodos

Na Seção 3.1 deste capítulo, apresentamos os *corpora* utilizados, incluindo suas composições, contexto de coleta e normas de transcrição. A seguir, a Seção 3.2 ocupa-se de descrever os processos e as ferramentas empregadas na pesquisa; assim, apresentam-se os tratamentos textuais que foram feitos nos conjuntos de dados para fins de pré-processamento (como tokenizações e remoção de *stopwords*), na Subseção 3.2.1. Com o texto dos *corpora* estabilizado, apresentamos os modelos de classificadores que foram confeccionados e selecionados por nós em 3.2.2 (um classificador bayesiano ingênuo), em 3.2.3 (uma rede neural unidirecional), em 3.2.4 (uma rede neural bidirecional) e em 3.2.5 (um modelo baseado em *transformers*). Abordamos aqui também nosso embasamento teórico linguístico para a análise da incidência de fenômenos apresentados pelos grupos clínicos na Subseção 3.3. Por fim, a última Seção, 3.4, menciona os instrumentos tecnológicos de que dispusemos para realização de todas as nossas análises.

3.1 Materiais (*corpora*)

Os *corpora* empregados nesta pesquisa são os *Datasets of Neuropsychological Language Tests in Brazilian Portuguese* (DNLT-BP)¹, sua coleta e compilação é descrita no endereço eletrônico do GitHub e também mencionada em Casanova et al. (2020). No conjunto, há quatro *corpora* distintos com diferentes quantidades de dados e balanceamento nem sempre proporcional.

¹Disponíveis em <https://github.com/nilc-nlp/DNLT-BP>.

Todos os *corpora* incluem narrações realizadas por indivíduos em contexto clínico, compondo três diferentes grupos de indivíduos diagnosticados com Doença de Alzheimer, indivíduos com Comprometimento Cognitivo Leve e indivíduos saudáveis integrando os grupos-controle. Caracterizando brevemente cada *corpus*, as Subseções de 3.1.1 a 3.1.3 a seguir descrevem a tarefa executada em cada um, a composição e distribuição de indivíduos entre os grupos e a instituição responsável pela coleta e transcrição conforme descrito no site (ver nota 1) e em Casanova et al. (2020). Outros dados comuns a todos, como estado dos *corpora* e padrões de transcrição, encontram-se na Subseção 3.1.4.

3.1.1 *Cinderella*

Este *corpus* é composto pelas narrativas de 60 indivíduos, sendo 20 com DA (de 68 a 86 anos de idade; de 3 a 20 anos de escolaridade; Mini Exame do Estado Mental (MEEM) de 18 a 29), 20 com CCL amnésico (de 63 a 83 anos de idade; de 4 a 20 anos de escolaridade; MEEM de 26 a 30) e 20 no grupo-controle (de 61 a 95 anos de idade; de 6 a 15 anos de escolaridade; MEEM de 23 a 30). Os diagnósticos foram obtidos na Faculdade de Medicina da Universidade de São Paulo (FMUSP). Essas narrativas foram coletadas com base na história da Cinderela, presente no imaginário comum, mas, além desse conhecimento prévio, os indivíduos dispunham de um livro contendo 23 ilustrações de modo a apresentar visualmente a sequência da história. O entrevistador solicitava a narração da história para o indivíduo; mais tarde, o anotador transcrevia a fala e pontuava a narração conforme a quantidade de unidades de informação esperadas (essa pontuação, no entanto, não consta no repositório).

De modo a ilustrar essas transcrições, vejamos o primeiro texto de cada grupo.

0_Cinderella (DA):

bom aqui esse é o príncipe né . ele tinha um cavalo bonito . e essa é a cinderela . aí ela gostava do cavalo . ela ficava tratando do cavalo . então essa é a a essa é a mulher dele . ela tinha essas filhas moças né . e essa é a cinderela . ela era muito bonita . então as moças tinha

inveja dela e judiavam dela . aí puseram ela pra fazer serviço fazer limpeza fazer faxina . aí punha ela pra limpar o chão . elas um balde de água . ela ta fazendo faxina continua . e as outras acho que estão dando ordem . estão olhando pro trabalho dela . aqui ela tá com com a vassoura . aqui tem umas tem umas cartas que vem né que vem . ela tá tirando uma carta aqui pra ela . essa aqui é a é a mãe das das meninas . e elas tão olhando pra ela . que que é aqui ta fazendo o que é um vestido para a cinderela . ela já com o vestido . aí ela queria tirar o vestido dela né . tava muito bonita . ela queria tirar o vestido . a cinderela tava chorando . aqui tava triste . o que que ela fazendo aqui estava medindo o vestido ver se servia para ela . aí ela tirou mesmo né . não não tô achando . ah o sapato ela perdeu o sapato . aí ele achou o sapato . ele pegou o sapato da outra e queria entra queria calçar né . e ela ficou com um pé descalço . aí ele calçou nela o sapato né . e eles ficaram ficaram juntos pro resto da vida .

20_Cinderella² (CCL):

era uma vez uma uma menina uma garota né que vivia numa castelo com o pai . e ela gostava muito de animais . e ela estava ahn fazendo um passeio a cavalo . e ela morava num castelo . e ela tinha assim uma vida muito livre né . agora acontece que ela ela perdeu o pai dela num acidente . e não não foi o pai quem ela perdeu não quem ela perdeu foi a mãe . porque eu me lembro o pai que casou novamente né . então ahn ela ela perdeu a mãe ficou ficou só com o pai . e o pai era jovem então resolveu resolveu se casar e não soube escolher bem a noiva . e ai encontrou uma noiva num castelo perto do seu de uma senhora que que tinha três filhas . e ela queria que uma das filhas ao menos uma das filhas se casasse com o príncipe . então quando esse vizinho enviuvou ele foi ahn a essa mãe foi levar as filhas para ele conhecer . e ele acabou não

²Notamos que alguns poucos termos parecem incorretos, como ‘madrastra’ e ‘madratra’ nesta transcrição. No entanto, os arquivos de áudio dessas entrevistas não estão disponíveis no repositório para conferência; então, manteremos essas formas tal qual estão, uma vez que podem ser comprometimentos fonológicos que alteraram-nas.

agora eu confundi tudo né . ahn ele acabou se casando . e a filha dele foi morar no castelo . so que a a a sogra espizinhava muito a menina fez a menina fazer trabalho pesado que ela não estava acostumada né . então ela passou a fazer trabalhos muito pesados que deveriam ser feitos para as filhas né pra trabalhar junto . mas a mãe não deixava . ela tinha que servir as filhas dela . ela foi uma empregada mesmo né . e um dia ela estava ela estava varrendo o jardim e e veio uma uma carta uma carta lacrada . e ela então levou pra pra patroa né . e e a patroa abriu e viu que era um convite para um um baile no castelo . e a a cinderela ficou muito triste porque ela além de não ser convidada ela ela foi muito discriminada pela pela pela senhora né . que o serviço dela não era frequentar salões era fazer limpeza fazer faxina . e ela ficou muito triste . e ela tinha uns uns uns uns amiguinhos . como se trata de de fantasia eles até falavam né falavam escutaram e depois eles vão ajuda-la . mas naquele momento ela ficou muito triste porque ela não tinha como ir ao castelo . e então essa tristeza assim tomou conta dela . ela perdeu assim o o o pique de tudo chorava muito e ficou muito triste se isolava muito . mas a madrastra a ma a a a madratra e as duas filhas ahn faziam de proposito pra ela ser escravizada . sujavam a roupa novamente . se ela fazia alguma coisa elas desfaziam . então queriam assim torna-la uma inútil né . e ela fazendo com muita humildade com dedicação . ela fazia ela não se revoltava não . e olha chorava quando ela ia pro pro pro bosque . então era uma menina triste né . e um dia ela estava la no bosque e veio a fada madrinha porque antigamente tinha isso de fada madrinha né tinha de fada madrinha e foi tirar umas medidas dela . ela ficou assustada mas tirar minha medida porque . dai a fada madrinha falou . olha o que eu tenho pra você . e mostrou uma carruagem que a carruagem foi feita pelos amiguinhos dela o ratinho aqueles animaizinhos . olha aqui ta vendo . eles fizeram parece que foi de uma abobora não tenho muita certeza . ahn fizeram a carruagem é parece que fizeram

a carruagem para ela . e a madrinha foi assim mostrar para ela dizendo que não ficasse triste porque ela ia ao baile de qualquer forma . então no dia do baile ela a madrinha deu o vestido para ela . ela se vestiu entrou na carruagem e foi para o castelo . chegando lá ela foi recebida pelo príncipe assim que admirou muito a beleza dela né . e ahn pediu a ela se ela poderia conceder uma dança para ele né porque antigamente era assim né . então ela muito vaidosa va vaidosa no sentido de de se sentir assim é privilegiada porque as as irmãs não foram mas ela foi chamada para dançar né . só que ela tinha um horário para voltar para casa pelo fato dela ser uma serviçal né . ela sabia disso não que a madrasta tivesse falado porque ela não sabia que ela ia ao baile né . ela tinha que voltar meia noite em ponto a madrinha avisou bem meia noite em ponto senão você não tem mais a carruagem o seu sonho acaba né . e quando estava faltando alguns minutinhos pra meia noite ela saiu correndo saiu correndo . o príncipe não estava com ela . e ela perdeu um sapatinho na escadaria . perdeu um sapatinho mas ela foi correndo para casa de medo de perder a carruagem que iria desaparecer né . o príncipe foi atrás dela mas a carruagem foi embora foi levá-la né porque o horário já estava bem apertado . então uma a a a madrasta fechou a porta da entrada . e deixa eu ver é fechou a porta da entrada com chave . e ela veio a branca de neve veio pra pra entrar dentro de casa . dai ela percebeu que ela estava com um sapato só que ela não tinha nem percebido que ela tinha perdido um sapato . então dai ela percebeu né . ela percebeu entrou em casa continuou nos afazeres . mas o príncipe se encantou com ela dizia que ela era uma dama que ela era perfeita que ela era linda que ela tinha ahn assim que ela tinha uma linhagem muito bonita . e mas ele não sabia quem era ela né . então ele ficou uns dia pensando pensando pensando como eu vou fazer não sei nome não sei o castelo onde mora eu não sei nada . dai parece-me que um um mordomo teve uma ideia e sugeriu para ele que com aquele sapato que estava com ele que procurasse

na vila quem é que tinha o outro pé porque dai completaria e seria a a a noiva mesmo né . e e foi o que o príncipe fez . ele ele foi até fez o o assim organizou a a a os os ahn os empregados dele porque por ele ser príncipe ele tinha que ser protegido né . então e foram para a vila pra ver qual pesinho cabia no sapato . e a a a cinderela fazia parte da da colônia ali né . e ele experimentou em todas não achou . quando chegou na cinderela serviu certinho . então ele descobriu que a cinderela é que era a realmente a dona do sapato né . e dai ele propôs casamento pra ela levou pro castelo . e a a madrasta e mais as as filhas as filhas que eram muito feias ficaram muito revoltadas muito revoltadas mas continuaram no seu castelo . e ela foi para o castelo ser princesa .

40_Cinderella (grupo-controle):

a cinderela é tava com o pai né . depois ahn o pai morreu ela ficou sozinha . foi depois foi pra casa da da madrasta viu . uhn ta tinha as as duas duas filhas feias né e fizeram ela de faxineira . e ai ela elas sempre tiravam sarro na coitada da cinderela . ai ela recebeu uma carta por com um um convite pra ir pro baile . mas a madrasta não quis deixar ela ir pro baile . ai os amiguinhos delas né resolveram fazer um vestido bonito pra ela ir pro baile . mas a as as as irmãs feias estragaram o vestido pra ela não sair . ai ela ficou chorando né não sabia o que fazer . ai apareceu a fada ah fiz ohr fizeram viu mais um vestido bonito . e ai foi pro foi pro baile né com a carruagem né foi pro baile . ai chegando la e todo mundo admirou ela tudo direitinho . ta ai apareceu o príncipe beijou a mão dela . e ela e ela pa a tinha que sai a meia noite . mas como tava saiu correndo e perdeu o sapatinho . ai o príncipe achou o sapato e queria saber de quem era o sapato . ai andou procurando pa pelas moças pra saber quem qual é que servia o sapato né . uma dizia uma as irmãs dela as feias nenhum tudo serviu . ai chegou no pé dela . ela deu certinho

né . depois ele colocou o sapato serviu direitinho . ai se casaram foram muito felizes .

3.1.2 *Dog e Lucia*

Ambos os *corpora* foram estabelecidos a partir da *Battery of Language Assessment in Aging* (BALE), uma bateria de testes padronizada capaz de abarcar indivíduos de 60 a 90 anos de diferentes graus de educação (analfabeto, baixo ou alto).

O primeiro conjunto, de 106 transcrições — grupo-controle: 82 (de 60 a 80 anos de idade; de 0 a 19 anos de educação); DA: 12 (de 59 a 81 anos de idade; de 0 a 8 anos de educação); CCL: 12 (de 57 a 82 anos de idade; de 0 a 18 anos de educação; 9 casos de CCL amnésico e 3 casos de CCL de múltiplo domínio) —, provém da tarefa de descrever a história de um garoto que leva para casa um cachorro encontrado na rua e o esconde, mais tarde, convence sua mãe de mantê-lo. A base a partir da qual o participante deverá narrar essa história é visual, sendo composta por uma sequência de sete imagens. A narração é feita com as imagens disponíveis para o participante concomitantemente.

O segundo *corpus*, *Lucia*, contém 89 transcrições — grupo-controle: 80 (de 63 a 82 anos de idade; de 2 a 19 anos de educação); DA: 9 (de 68 a 81 anos de idade; de 4 a 8 anos de educação) —, dessa vez, porém, a tarefa é o participante recontar uma história que lhe foi apresentada de modo oral. Ele deverá recontá-la logo após ter sido exposto a ela.

Ambos foram coletados por pesquisadores da Escola de Humanidades da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS).

Vejam os exemplos das primeiras transcrições de cada grupo para o *corpus Dog*:

0_Dog (DA):

vou colocar nome no gurizinho tá pedro ia passear na rua . e encontrou um cachorrinho . aí ele levou pra casa . e pediu pro cachorrinho ficar quieto . e ele escondeu o cachorrinho dentro do guarda roupa pra mão dele não ver . aí quando a mãe dele viu o cachorrinho a mãe dele ficou surpresa . aí perguntou pra ele . e ele pediu pra ela

deixa eu ficar com ele . e ele disse que ia cuidar do cachorrinho . e ela disse que ele podia ficar .

12_Dog (CCL):

eu mesmo não tô entendendo nada aqui é um colégio . o que que é isso daqui isso aqui é um colégio . que que é isso aqui aqui o guri tá chamando o cachorro . aqui o guri tá correndo do cachorro . aqui ele tá não sei o que pro cachorro enfiando o dedo no nariz do cachorro . aqui ele perdeu a cabeça . aqui ele tá assustado . aqui ele tá falando com gurizinho . aqui parece que ele ta xingando o cachorro ou xingando o guri .

24_Dog (grupo-controle):

o menino ia passando na rua e viu um cachorrinho . e gostou do cachorrinho . aí chamou o cachorrinho . e o cachorrinho veio do lado dele . aí o cachorrinho acompanhou ele . e aí chegando em casa e escondeu o cachorrinho dentro do roupeiro pra sua mãe não ver . mas sua mãe achou o cachorrinho dentro do roupeiro . e brigou com o menino . e o menino pediu por favor pra ela . aí ela deixou . e tanto é que fez a casinha e tudo . ajudou ele a fazer a casinha pro cachorro .

Agora, as narrações dos primeiros participantes do *corpus Lucia*:

0_LUCIA (DA):

lucia é a pessoa né pego um taxi . e foi pra rodoviária . barbaridade sacanagem isso eu não consigo minha cabeça ta exatamente em função disso que eu to no médico ela foi pra rodoviária e da rodoviária ela fez o que não deu .

9_LUCIA (grupo-controle):

lúcia mora no interior do paraná . e aquele dia ela ia fazer uma entrevista pra emprego . conseguiu uma carona . mas no caminho o

pneu furo . daí resolveu pega um táxi . e começou a chover a chuva
acho que foi antes e ela resolveu pega o taxi pra chega na rodoviária .

3.1.3 *Wallet*

Este *corpus* é oriundo de um subtteste presente no *Arizona Battery for Communication Disorders* (ABCD), sendo composto de uma história que deve ser memorizada pelo entrevistado e contada em dois momentos: imediata e tardiamente. O intuito é avaliar a memória episódica. A história é sobre uma pessoa que perde sua carteira em um supermercado, deixa o estabelecimento e retorna para casa; em casa, recebe um telefonema sendo avisada que sua carteira foi encontrada. A história foi traduzida e adaptada para português brasileiro (PB) — não consta na documentação, porém, se o participante recebe a história oralmente ou se ele a lê. Dos 35 participantes, 23 tem CCL (com idades de 63 a 89 anos; de 7 a 20 anos de estudos; dois participantes com MEEM = 29; alguns indivíduos sem informações) e 12 são saudáveis (com idades de 55 a 69 anos; de 11 a 29 anos de estudos; MEEM de 25 a 30; alguns indivíduos sem informações); totalizando 70 produções. A execução do teste foi feita na FMUSP.

O_immediately_Wallet (CCL com narração imediata):

uma senhora fez as compras . e não percebeu que caiu a carteira .
quando ela se chegou ao caixa sentiu que não tinha dinheiro para
pagar . o que ela fez . ela separou as compras . e ficou desesperada .
quando recebeu o telefonema de uma criança avisando que tinha
achado sua carteira e ela ficou totalmente feliz .

O_delayed_Wallet (CCL com narração tardia):

uma senhora estava no supermercado fazendo compras . quando não
percebeu que caiu sua carteira no chão . quando se dirigiu ao caixa
para pagar a mesma percebeu que estava sem a carteira . entrou em
desespero . não conseguiu pagar a conta . deixou a despesa de lado .
foi quando uma criança ligou para ela . eu achei sua carteira . ela
ficou toda feliz .

23_immediately_Wallet (grupo-controle com narração imediata):

a senhora foi fazer compras . estava fazendo as compras . e depois ela foi pagar . na hora de pagar ela descobriu que ela estava sem a carteira . a carteira tinha caído no chão . ela fo não percebeu . e foi para casa . deixou as compras lá . chegando em casa o telefone tocou . ela pegou atendeu o telefone . e era uma menininha dizendo que achou a carteira dela .

23_delayed_Wallet (grupo-controle com narração tardia):

uma senhora foi fazer compras . pegou as compras . e quando chegou ao caixa caiu a carteira dela . ela não percebeu . foi pagar . não conseguiu pagar . foi embora pra casa . quando ela chegou em casa tocou a campá é tocou o telefone . ela foi atender . e era uma menininha dizendo que achou a carteira dela .

3.1.4 Sobre os padrões de transcrição

Em dois desses *corpora*, pode-se notar que apenas dois dos três grupos foram participantes: *Lucia* conta apenas com grupo-controle e DA, enquanto *Wallet* conta com grupo-controle e CCL. Portanto, a fim manter uma maior uniformidade nos nossos dados, selecionamos, para fins de análise, os *corpora Cinderella* e *Dog*, por serem compostos por integrantes dos três grupos nas mesmas tarefas. Ainda assim, haverá um desbalanceamento em termos de quantidade de participantes em cada grupo. Discutiremos mais sobre isso na Subseção 3.2.1.

Do ponto de vista da anotação das transcrições, temos, para todos os conjuntos, textos em que não há marcação de pausa ou intervenção do interlocutor. Repetições estão presentes (por exemplo, em 46_Cinderella (controle): “e ela tinha que ficar na festa até meia noite porque senão era era o tempo que ela podia ficar ali como princesa né”), bem como frases interrompidas (por exemplo, em 25_Dog (controle): “**vamo** compra uma casinha de cachorro **pra** coloca o”), termos transcritos tal qual se apresentam na fala oral (como *vamo* e *pra* no exemplo anterior), além de interjeições e preenchedores (por exemplo, em

12_delayed_Wallet (CCL): “ao chegar em casa **ahm** uma menina telefonou avisando que havia encontrado a carteira dela”). Segundo os condutores das pesquisas de cada *corpus*, a transcrição foi executada por anotadores especializados na tarefa. A anotação foi feita em texto corrido e em minúsculas; a única pontuação presente são pontos finais que separam as frases.

3.1.5 Dados demográficos

Além do contexto de coleta de cada projeto e dos próprios *corpora*, os pesquisadores disponibilizam dados demográficos de cada participante na página dos conjuntos de dados do GitHub. Com base nesses dados, existe a possibilidade de se traçar correlações entre as diversas informações que os autores fornecem, a saber: idade, anos de estudo, Mini Exame do Estado Mental (MEEM), Avaliação Clínica da Demência (*Clinical Dementia Rating*, CDR) e Exame Cognitivo de Addenbrooke-Revisado (*Addenbrooke’s Cognitive Examination–Revised*, ACE-R).

Nem todas as métricas estão presentes para todos os *corpora*. Além disso, em alguns dos conjuntos de dados, há lacunas para alguns participantes, mesmo que os demais tenham tido o registro do mesmo critério; nesses casos, não calculamos a correlação entre as informações em questão.

Os resultados são apresentados nas Tabelas de 3.1 a 3.4, representando os dados para, respectivamente, os grupos DA, CCL, controle e, por fim, a união de todos. Entretanto, esses dados não foram frutíferos para análises futuras, senão apenas por podermos definir que nesses dados não apresentaram correlações significativas. As informações que apresentaram alguma correlação direta (isto é, $r > 0.5$) ou inversa (isto é, $r < -0.50$) foram anos de estudo com MEEM para os pacientes com DA ($r = 0.5$) e idade com MEEM em correlação inversa ($r = -0.57$). Sendo que para tomarmos a correlação como forte esperaríamos valores mais extremos, ou seja, próximos de 1 ou -1 .

3.2 Procedimentos

A função de um classificador é mapear determinados dados em determinados conjuntos. Em linhas gerais, um modo de se realizar computacionalmente essa

Correlação	<i>Cinderella</i>	<i>Dog</i>	<i>Walet</i>	<i>Lucia</i>	Todos
Idade e anos de estudo	0.33	0.47	-	-0.02	0.44
Idade e MEEM	0.37	-	-	-	-
Idade e CDR	-	-*	-	-*	-
Idade e ACE-R	-	-	-	-	-
Anos de estudo e MEEM	0.50	-	-	-	-
Anos de estudo e CDR	-	-*	-	-*	-
Anos de estudo e ACE-R	-	-	-	-	-
MEEM e ACE-R	-	-	-	-	-

Tabela 3.1: Correlação entre as informações sociodemográficas dos participantes do grupo de indivíduos com DA (* significa que não processamos a métrica por haver lacunas nos dados).

Correlação	<i>Cinderella</i>	<i>Dog</i>	<i>Walet</i>	<i>Lucia</i>	Todos
Idade e anos de estudo	0.30	-0.13	-0.01	-	0.09
Idade e MEEM	0.23	-	-*	-	-
Idade e CDR	-	-*	-	-	-
Idade e ACE-R	-	-	-*	-	-
Anos de estudo e MEEM	0.30	-	-*	-	-
Anos de estudo e CDR	-	-*	-	-	-
Anos de estudo e ACE-R	-	-	-*	-	-
MEEM e ACE-R	-	-	-*	-	-

Tabela 3.2: Correlação entre as informações sociodemográficas dos participantes do grupo de indivíduos com CCL (* significa que não processamos a métrica por haver lacunas nos dados).

Correlação	<i>Cinderella</i>	<i>Dog</i>	<i>Walet</i>	<i>Lucia</i>	Todos
Idade e anos de estudo	-0.36	-0.20	0.29	-0.14	-0.21
Idade e MEEM	-0.06	-	-0.57	-	-
Idade e CDR	-	-0.06	-	-*	-
Idade e ACE-R	-	-	0.04	-	-
Anos de estudo e MEEM	0.23	-	0.38	-	-
Anos de estudo e CDR	-	0.26	-	-*	-
Anos de estudo e ACE-R	-	-	0.26	-	-
MEEM e ACE-R	-	-	0.47	-	-

Tabela 3.3: Correlação entre as informações sociodemográficas dos participantes do grupo-controle (* significa que não processamos a métrica por haver lacunas nos dados).

tarefa é determinando a probabilidade de certa classificação ocorrer de acordo com atributos encontrados nos dados que expressarão padrões (por exemplo, a classificação de sentimento de um comentário em redes sociais pode tomar como atributos emojis, xingamentos, uso de caixa alta ou baixa, entre outros, a fim de determinar a probabilidade de a mensagem ser negativa, neutra ou

Correlação	<i>Cinderella</i>	<i>Dog</i>	<i>Walet</i>	<i>Lucia</i>	Todos
Idade e anos de estudo	-0.01	-0.21	-0.19	-0.19	-0.14
Idade e MEEM	-0.11	-	-0.25	-	-
Idade e CDR	-	-0.20	-	-*	-
Idade e ACE-R	-	-	0.04	-	-
Anos de estudo e MEEM	0.46	-	0.19	-	-
Anos de estudo e CDR	-	0.46	-	-*	-
Anos de estudo e ACE-R	-	-	0.26	-	-
MEEM e ACE-R	-	-	0.46	-	-

Tabela 3.4: Correlação entre as informações sociodemográficas dos participantes de todos os grupos de análise (* significa que não processamos a métrica por haver lacunas nos dados).

positiva). Outra maneira de se classificar dados é por meio do treinamento com base em dados já mapeados — o que faremos usando uma rede neural.

Com esse contexto em foco, nesta seção, apresentam-se os procedimentos utilizados para pré-processamento — isto é, tratamento dos dados antes de submetê-los aos modelos de classificação empregados em 3.2.1 — e processamento textual — os modelos utilizados de fato, sendo eles um classificador bayesiano ingênuo, uma rede de propagação para frente, uma rede bidirecional (BiLSTM) e um modelo baseado em *transformers*, respectivamente, de 3.2.2 a 3.2.5.

3.2.1 Tratamento dos dados

Os *corpora* foram recolhidos diretamente do repositório do Núcleo Interinstitucional de Linguística Computacional (NILC) — grupo de pesquisa com pesquisadores de diferentes universidades, como a Universidade de São Paulo (USP), sobretudo representada pelo Instituto de Ciências Matemáticas e de Computação (ICMC), e a Universidade Federal de São Carlos (UFSCar) — na plataforma GitHub (ver nota 1). Todo o processamento descrito a partir daqui foi executado utilizando a linguagem de programação Python³ via Google Colab⁴.

A princípio, após selecionados os *corpora* como descrito em 3.1, eles foram unidos em um *dataframe* da biblioteca de análise de dados *pandas*⁵ e a limpeza dos dados se deu por meio da remoção de *stopwords* — com base na lista para

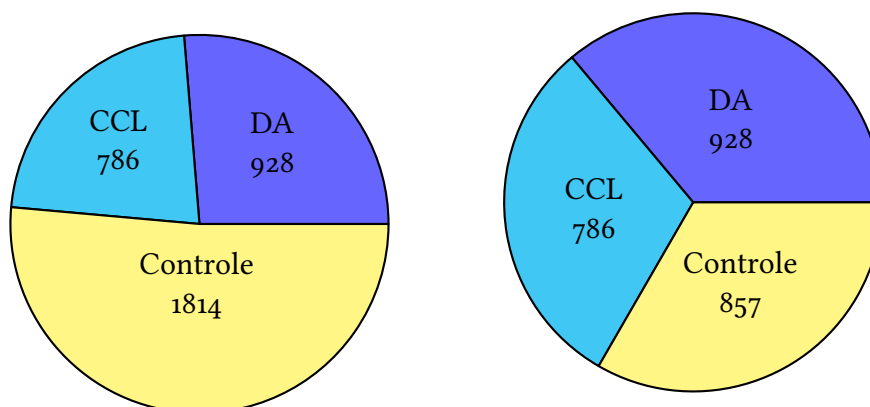
³<https://www.python.org/>

⁴<https://colab.research.google.com/>

⁵<https://pandas.pydata.org/>

o português da biblioteca *Natural Language Toolkit* (NLTK)⁶ com algumas alterações, como remoção de palavras de negação e inclusão de formas reduzidas como ‘pra’ (‘para’).

A seguir, com as sentenças tokenizadas, notou-se um desbalanceamento nos dados pendendo para o grupo-controle, como evidenciado nos gráficos da Figura 3.1. Para balancear essas quantidades e evitar enviesamentos nas análises mais tarde, reduzimos aleatoriamente as sentenças dos grupos-controle para 857 — a média dos volumes dos outros dois grupos — assim, temos quantidades de sentenças mais próximas entre os grupos, como se nota no gráfico da Figura 3.1b, sem alterar os dois grupos de maior interesse, uma vez que não se trata de um *corpus* extenso.



(a) Quantidade de sentenças desbalanceada.

(b) Quantidade de sentenças balanceada.

Figura 3.1: Ambos os gráficos mostram a quantidade de sentenças presente no *corpus* em cada grupo de análise (originários dos *corpora Cinderella* e *Dog*): Doença de Alzheimer, Comprometimento Cognitivo Leve e grupos-controle. O primeiro mostra a quantidade inicial em cada grupo; o segundo mostra a quantidade após o corte no grupo-controle.

A fim de remover informações morfossintáticas, que não nos seriam úteis na análise, para maior estabilidade dos dados, como marcação de número e desinências verbais, experimentamos reduzir as palavras nas sentenças tanto a lemas quanto a raízes. Para essas tarefas, foram escolhidas, respectivamente, o lematizador CoGrOO⁷ e o stemizador Removedor de Sufixos da Lingua Portuguesa (RSLP) do NLTK.

⁶<https://www.nltk.org/>

⁷<https://cogroo.sourceforge.net/>

Optamos, por fim, por seguir com os dados stemizados devido a uma métrica de erro padrão que nos indica maior confiabilidade nos dados dessa forma, como será justificado na Seção 4.1.

3.2.2 Classificador Bayesiano Ingênuo Multinomial

Classificadores bayesianos são fundamentados na Regra de Bayes para calcular a probabilidade condicional de ocorrência de determinado evento. Segundo Ferreira e Lopes (2019, pp. 161–184), essa regra “permite calcular a probabilidade condicional de X , dado Y , a partir da probabilidade condicional de Y , dado X , além das probabilidades não-condicionais de X e Y ”, o que é sumarizado na seguinte formulação.

$$P(X|Y) = \frac{P(X) \cdot P(Y|X)}{P(Y)} \quad (3.1)$$

Uma segunda noção importante para entender os princípios de um classificador bayesiano é a noção de independência dos eventos, uma vez que:

$$P(A \& B) = P(A) \cdot P(B) \quad \text{se } A \text{ e } B \text{ forem independentes} \quad (3.2)$$

Do contrário, caso haja uma relação de dependência entre os eventos, é necessário que a probabilidade de ocorrência de um evento ($P(B)$) seja restringida apenas aos casos em que o outro evento ($P(A)$) ocorre (portanto, $P(B|A)$) — o contrário também é válido —, o que se apresenta na Equação 3.3.

$$\begin{aligned} P(A \& B) &= P(A) \cdot P(B|A) \\ &= P(B) \cdot P(A|B) \end{aligned} \quad (3.3)$$

se A e B forem dependentes

A execução desse modelo se baseia em três passos principais, com base em métodos da biblioteca Scikit-learn⁸, e foi feita para ambos os cenários de balanceamento do *corpus* (Figura 3.1):

⁸<https://scikit-learn.org/>

1. Vetorizar os dados textuais.

Por meio do método `feature_extraction.text.CountVectorizer`;

2. Converter a matriz de vetores gerada em uma representação tf-idf.

Por meio do método `feature_extraction.text.TfidfTransformer`;

3. Aplicação do classificador bayesiano ingênuo multinomial.

Por meio do método `naive_bayes.MultinomialNB`).

No cenário de o *corpus* estar desbalanceado, aplicamos também o parâmetro `fit_prior=True`.

3.2.3 Rede de propagação para frente

Como comentado na abertura da seção, uma maneira de classificar dados é usar dados já mapeados em suas determinadas classes e, então, empregar métodos de aprendizado de máquina — no caso, utilizaremos o classificador bayesiano e diferentes redes neurais artificiais para a tarefa — para identificação de padrões a fim de aplicá-los na classificação de dados não mapeados. Entretanto, é necessário que se certifique de que a classificação esteja sendo feita de forma correta ou, melhor, esteja atingindo uma proporção de acertos e erros aceitável para os objetivos da aplicação da tarefa em questão.

Para avaliar a classificação, uma prática comum é dividir um *corpus* com a classificação correta disponível em dois conjuntos, sendo um para treinamento e outro para validação. Assim sendo, no primeiro, a rede neural irá tentar estabelecer padrões, em seguida, esse aprendizado é posto em prática classificando os dados do segundo conjunto para, por fim, verificar a consistência dos rótulos gerados com os corretos.

Na Figura 3.2⁹, temos uma ilustração de como funciona a arquitetura de uma rede recorrente. Cada componente da entrada é associado a um neurônio na camada de entrada da rede (*inputs*). O objetivo da propagação para frente é calcular a saída da rede neural com base em uma entrada específica. A saída da função de ativação (setas para uma nova camada) é a saída da camada atual.

⁹Fonte: <https://towardsdatascience.com/understanding-recurrent-neural-networks-the-preferred-neural-network-for-time-series-data-7d856c21b759>. Último acesso em 26/11/2023.

Esse processo é repetido para cada camada da rede (em *hidden layer*, pode haver mais camadas), da camada de entrada até a camada de saída. Ou seja, a saída de uma camada serve como entrada para a próxima camada, e o processo é repetido até que a saída final da rede seja alcançada. O gradiente nos neurônios ilustra o problema de gradientes evanescentes, o que causaria, por exemplo, a perda de uma determinada informação antiga (“contexto”) com o passar das camadas e sequências de neurônios — recuperá-la para processamento torna-se cada vez mais custoso computacionalmente.

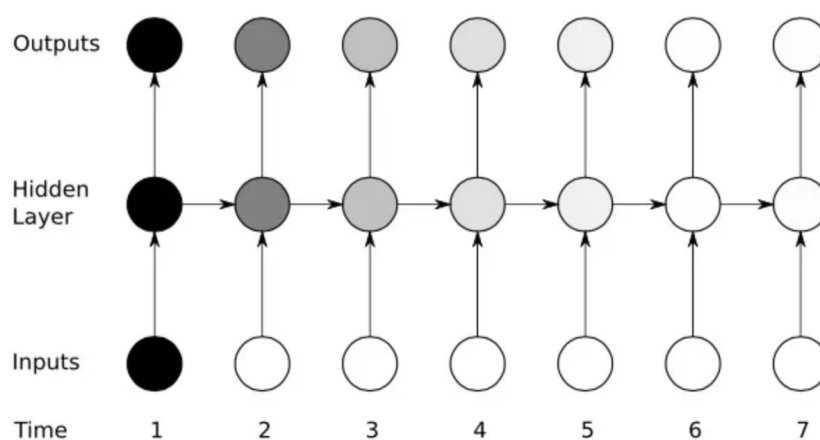


Figura 3.2: Exemplo com uma camada da arquitetura de uma rede neural recorrente de propagação para frente (unidirecional).

Optamos por um modelo sequencial (*Sequential*¹⁰), devido a sua melhor adaptabilidade com dados de natureza textual. Durante as execuções, testamos diferentes parâmetros como tamanho de lote (`batch_size`). Sem muitas diferenças significativas, mantivemo-lo com valor 64. Entretanto, a função de perda (`loss`) foi executada de maneiras distintas, como veremos a seguir.

Na Tabela 3.5, vemos três possíveis variações de configurações que foram aplicadas à rede e cujos resultados serão expostos na Seção 4.3. Os cenários incluem o balanceamento ou não do *corpus*, isto é, reduzindo, em B e C (Figura 3.1b), ou mantendo, em A (Figura 3.1a), o grupo-controle como apontado na Subseção 3.2.1. O outro aspecto em alternância é a entropia cruzada (`cross_entropy`), assumindo os valores categorial (`categorical_crossentropy`), em A e B, e binário (`binary_crossentropy`), em C (com exceção do teste com os três gru-

¹⁰https://keras.io/guides/sequential_model/

pos, por não ser aplicável). Esse último parâmetro é uma função de perda, sendo uma das responsáveis por determinar o quão bem a rede neural irá se adequar aos dados.

Configuração	Grupos dois a dois		Três grupos	
	Balanceado?	Entropia cruzada	Balanceado?	Entropia cruzada
A	Não	Categorial	Não	Categorial
B	Sim	Categorial	Sim	Categorial
C	Sim	Binária	-	-

Tabela 3.5: Três cenários de testes da rede de propagação para frente submetendo os grupos de análise dois a dois e também os três simultaneamente. Variações contam o *corpus* balanceado e desbalanceado, bem como a entropia cruzada ser categorial ou binária. A configuração C para os três grupos (com entropia cruzada binária) não é possível.

3.2.4 Rede BiLSTM

A BiLSTM (do inglês *Bidirectional Long Short-Term Memory*) é uma arquitetura de rede neural recorrente usada principalmente em tarefas de processamento de linguagem natural e sequências temporais. A BiLSTM é uma extensão da LSTM (*Long Short-Term Memory*), uma arquitetura de rede neural recorrente projetada para superar o problema de desaparecimento do gradiente, comum em redes neurais tradicionais.

A LSTM é capaz de aprender dependências de longo prazo em sequências temporais, tornando-se eficaz em tarefas que envolvem dados sequenciais. A BiLSTM melhora essa abordagem adicionando uma camada bidirecional. Isso significa que ela processa a sequência de entrada em duas direções: da esquerda para a direita (“para frente”, como a rede apresentada na Subseção 3.2.3) e da direita para a esquerda (“para trás”). Isso permite que a rede capture informações contextuais de ambos os lados de cada ponto na sequência, melhorando sua capacidade de entender o contexto global (Cui et al. 2020, p. 9).

Essa direcionalidade está representada na Figura 3.3. Em linhas gerais, x_n representa as informações entregues à rede (cada um pode representar cada *token* de uma sentença, por exemplo). Em seguida, esse *input* é operado pela rede não só em uma direção (para frente), mas também na direção contrária, o que garante maior retenção de “contexto” para o processamento de toda a

sequência, sendo o intuito principal um melhor desempenho final da tarefa desejada.

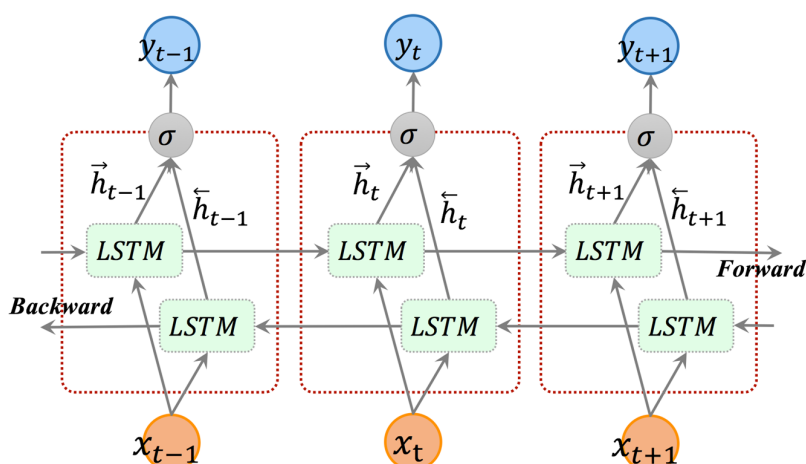


Figura 3.3: Exemplo de três passos, bidirecionais, da arquitetura de uma rede BiLSTM por Cui et al. (2020).

A estrutura bidirecional da BiLSTM é particularmente útil em tarefas de PLN, como análise de sentimento, reconhecimento de entidades nomeadas, tradução automática e outras em que o contexto de toda a sequência é importante para a compreensão correta do significado. A BiLSTM tem sido amplamente utilizada em uma variedade de aplicações onde a modelagem de sequências é crucial para o desempenho do modelo.

Em nossa pesquisa, elegemos essa rede como uma boa candidata para treinamento por estarmos lidando com uma tarefa de classificação baseada em sentenças. Para o processamento, usamos um conjunto de dados aleatorizado a partir dos *corpora Cinderella* e *Dog*. Para essa aleatorização, todas as narrativas foram segmentadas em sentenças acompanhadas da etiqueta correspondente ao grupo clínico ao qual o participante pertence.

Para executar nossos treinamentos, usamos a implementação padrão da biblioteca keras (`tf.keras.layers.Bidirectional`). Padronizamos o número de camadas em 64 (`tf.keras.layers.LSTM`); aumentar ou diminuir esse valor não causou mudanças significativas nos resultados. Seis execuções foram feitas variando entre dados brutos e anotados, além de passarmos as sentenças dos grupos dois a dois (ou seja, três combinações), binarizando a tarefa de classificação. Estabelecemos a recorrência de apenas quatro épocas para cada

execução, pois, como veremos nos resultados na Seção 4.4, a acurácia teve pouca ou nenhuma variação de uma época para outra.

3.2.5 DistilBERT: modelo de linguagem baseado em BERT

BERT (do inglês *Bidirectional Encoder Representations from Transformers*) é um modelo de linguagem pré-treinado desenvolvido pela Google Research (Devlin et al. 2018). Ele pertence à família de modelos baseados em *transformers*, que são arquiteturas de rede neural projetadas para processar dados sequenciais, como texto. O BERT foi introduzido em 2018 e se destacou por seu desempenho impressionante em uma variedade de tarefas de PLN.

A arquitetura de direções do BERT é bidirecional, tal qual as redes BiLSTM. Isso significa que ele também leva em consideração o “contexto” de todas as palavras em uma sentença, o que ajuda a capturar nuances e relações mais complexas entre as palavras.

O treinamento do BERT envolve uma fase prévia em que o modelo é treinado em grandes quantidades de dados textuais. Esse treinamento prévio permite ao BERT aprender representações contextualizadas de palavras e frases. Posteriormente, o modelo pré-treinado passa por etapas *fine-tuning* por meio de conjuntos de dados direcionados a tarefas específicas, como classificação de texto, extração de informações ou tradução automática.

Em nossa pesquisa, adotamos uma versão do BERT treinada especificamente para o português brasileiro, o BERTimbau¹¹, porém seguindo a implementação de um modelo mais leve dele para viabilizar nossa execução recorrente: o `distilbert-portuguese-cased`¹².

Adiantando um ponto que voltaremos a discutir no momento de apresentar os resultados na Seção 4.5, na metodologia de testes desse modelo, observamos que a acurácia, a cada execução com uma mesma configuração, variava significativamente. Nossa hipótese é de que, com a aleatorização dos dados antes do treinamento e validação, teríamos os dois conjuntos pouco consistentes em termos de representatividade devido à pequena quantidade de dados que temos.

¹¹<https://huggingface.co/neuralmind/bert-base-portuguese-cased>

¹²<https://huggingface.co/adalbertojunior/distilbert-portuguese-cased>

3.3 Fenômenos linguísticos: hesitações

Um dos pontos que adotamos para análise linguística foram as hesitações por estarem na maioria dos casos bem demarcadas nas transcrições das narrativas. Como referencial teórico adotamos Marcuschi (1999) e Marcuschi (2003), que apresenta as seguintes classificações de diferentes fenômenos de hesitação:

- i. pausas não preenchidas (PNP): silêncios prolongados que se dão como rupturas em lugares não previstos pela sintaxe (cortes de estruturas sintagmáticas) e pelo fluxo da fala.
- ii. pausas preenchidas (PP): ocorrências de marcadores de hesitação do tipo “éh”, “mm”, “ah”; alongamentos vocálicos com características hesitativas e marcadores conversacionais acumulados.
- iii. gaguejamentos (GA): repetições de unidades inferiores a um item lexical e pedaços de palavras iniciadas.
- iv. repetições hesitativas (RH): as repetições julgadas não significativas semanticamente, geralmente repetição de itens formais.
- v. falsos inícios (FI): todos os inícios de unidades sintáticas oracionais, que são iniciados com algum problema e refeitos ou retomados, o que distingue este tipo dos cortes oracionais que são construções abandonadas.

Como o autor sugere, hesitações são uma falha, um interrompimento na fluência do discurso e, a partir delas, podemos notar padrões de erro lexical, sintático ou mesmo semântico. E isso se torna um dos nossos grandes objetos de estudo justamente pelo fato de manifestações linguísticas como essas serem afetadas com a progressão das doenças, havendo distúrbios. Como apontado por Nitrini, Caramelli et al. (2005, p. 723), logo nos estágios iniciais da doença, problemas semânticos, lexicais e discursivos já são observados em relação a indivíduos saudáveis.

Aqui, colocamos as hesitações nesse lugar de análise por se encaixarem nesse escopo linguístico. No entanto, as pausas não preenchidas, um dos itens

propostos por Marcuschi, não foram analisadas, uma vez que na anotação dos transcritores não foram incluídas marcações de pausas seja por caractere, seja por indicação da duração da pausa. Assim sendo, esse item não entra nas nossas avaliações. Além dessa forma de pausa, temos as pausas preenchidas, que por sua vez podem ser encontradas na transcrição com frequência; nomearemos-as *preenchedores*.

Outros dois itens foram agrupados em um único objeto: gaguejamentos e falsos inícios, por se tratar de uma realização linguística incompleta para ambos, foram unidos sob o nome *repetições incompletas* — as ocorrências são repetições em que a palavra que precede outra consiste integralmente nos primeiros caracteres da palavra à direita. As repetições hesitativas serão consideradas *repetições completas* em oposição às incompletas; ou seja, no caso delas, consideramos uma repetição completa quando uma palavra corresponde integralmente à anterior. Resumidamente:

- i. Preenchedores: PP.
- ii. Repetições completas: RH.
- iii. Repetições incompletas: GA e FI.

Nas próximas subseções, detalhamos cada agrupamento de acordo com a forma com que foram anotados.

3.3.1 Anotação de preenchedores

Para identificar o que chamamos de preenchedores (as *pausas preenchidas*, de acordo com Marcuschi (1999, pp. 168–169)), fizemos uma inspeção manual de cada narrativa, substituindo as ocorrências de cada preenchedor por uma etiqueta (a saber: <FW>). Um exemplo dessa substituição pode ser visto a seguir, retirada do arquivo 42_cinderella.

- era uma vez um reino onde existia um castelo e viviam vivia o casal de reis **tá** . e eles não tinham filhos . mas tiveram uma menina muito bonita **né** .

- era uma vez um reino onde existia um castelo e viviam vivia o casal de reis <FW> . e eles não tinham filhos . mas tiveram uma menina muito bonita <FW> .

Fizemos esse procedimento manualmente por não termos de antemão uma lista exaustiva de quais seriam os preenchedores utilizados e de que formas poderiam estar grafados diferentemente. Assim sendo, aproveitamos para compartilhar a seguir a lista dos itens que consideramos preenchedores.

a, ah, ahn, ai, aí, bem, bom, dai, daí, e ai, e aí, e dai, e daí, e tal, eh, ehm, então, então tá, foi, ha, han, hanhan, heim, hein, hinhun, hm, hrum, hum, hun, hunhun, isso aí, ne, nossa, não é, né, oh, ohr, olha, oo, pegou, pois é, sei la, ta, tal, tal e coisa, tam, tchurtchu, tudo mais, tá, uhm, uhn, viu, ã, é, ó, óh

Entendemos que parte das ocorrências de determinados preenchedores (como “né”, “tá” e “uhm”) pode representar casos em que o sujeito está marcando uma função pragmática no discurso e não realmente um caso em que houve uma ruptura no processamento da linguagem que geraria uma hesitação como os preenchedores. Por exemplo, um preenchedor desses poderia estar agindo como um pedido de confirmação no qual o sujeito está eventualmente apontando algo nas imagens oferecidas como suporte visual (o que apenas propomos como uma possibilidade por não termos acesso a gravações visuais da entrevista para validação), buscando confirmar com o(a) entrevistador(a) se ele(ela) está compreendendo a narração ou averiguando se a narração está sendo aprovada. Ainda assim, para fins de padronização e por falta de mecanismos de verificação, consideramos todas as ocorrências de cada preenchedor na contagem — desde que em posições que poderiam agir como um preenchedor de fato.

3.3.2 Anotação de repetições completas

As repetições completas (*repetições hesitativas* nas palavras de Marcuschi (1999, p. 169)) foram anotadas automaticamente por meio de uma expressão regular¹³

¹³A expressão regular usada foi: '\b(\w+)\s+\1\b'

que verifica se uma *palavra* é idêntica à palavra que aparece imediatamente antes.

3.3.3 Anotação de repetições incompletas

As repetições incompletas agrupam duas definições de Marcuschi (1999, p. 169), os falsos inícios e os gaguejamentos. Elas também foram anotadas automaticamente; neste caso, o código verifica se os n caracteres que iniciam uma palavra correspondem à totalidade de caracteres da palavra anterior (como em “ela tinha **leva levado** a carta”). Adotamos $\{n \in \mathbb{N} \mid 1 \leq n \leq 4\}$, para que a correspondência seja de 1 a 4 caracteres. Além disso, fizemos uma exclusão manual dos artigos ‘a’ e ‘o’ e da conjunção ‘e’ por estarem inflando os dados dada a alta recorrência dessas palavras (como em “**e** ele saiu”).

3.4 Instrumentos

Para execução dos códigos e *notebooks* em Python que utilizamos em nossos processamentos, análises e geração de resultados, usamos o Google Colab¹⁴ na maior parte do tempo. Alguns treinamentos, no entanto, requeriam um poder de processamento mais robusto — para os quais o Google Colab levava muito tempo de execução ou era interrompido devido à limitação de recursos. Fizemos, então, uso de uma máquina virtual de alto desempenho via JupyterHub¹⁵ cujo acesso foi permitido para os fins da pesquisa pela empresa Alana AI¹⁶. A configuração usada foi a seguinte:

- GPU NVIDIA A100 PCIe
- AMD EPYC 7V13 (Milan)
- 4 GPU com 320GB de memória
- 96 vCpu – 880 GB RAM

¹⁴<https://colab.google/>

¹⁵<https://jupyter.org/hub>

¹⁶<https://alana.ai/>

Resultados e discussão

Neste capítulo, são apresentados e discutidos os resultados obtidos a partir das ferramentas de classificação automática, dos *corpora* a elas submetidos, bem como da análise da ocorrência de fenômenos linguísticos que se manifestam diferentemente nos grupos em estudo — conforme detalhados no Capítulo 3. Além disso, confrontamos nossos resultados com os encontrados na literatura — *cf.* Capítulo 2 — e que de alguma forma possam servir de parâmetro de comparação para esta pesquisa.

4.1 Análise descritiva de dados

Uma vez realizado o tratamento nos *corpora*, conforme exposto na Subseção 3.2.1, pudemos gerar uma série análises estatísticas. A Tabela 4.1 traz algumas delas, de modo a nos ajudar a nos familiarizar com os *corpora* sob outras perspectivas. A seguir, nesta seção, detemo-nos em comentá-las, interpretando suas principais decorrências.

As primeiras três medidas da Tabela 4.1 apresentam a média de palavras por sentença (a), seu desvio padrão (b) e o erro padrão (c). O primeiro dado indica uma maior concisão na construção de sentenças por parte do grupo DA. Uma hipótese é a presença de sentenças interrompidas logo nas primeiras palavras. Curiosamente, porém, o grupo CCL é o que apresenta as sentenças com maior extensão, superando o grupo-controle.

	apenas pré-processado			lematizado			stemizado		
	DA	CCL	CTR	DA	CCL	CTR	DA	CCL	CTR
(a)	4,93	6,05	5,18	5,00	6,26	5,30	5,05	6,14	5,28
(b)	3,62	3,50	3,37	3,66	3,62	3,42	3,61	3,51	3,36
(c)	0,59	0,59	0,59	0,66	0,66	0,66	0,57	0,57	0,57
(d)	25,09	25,97	16,38	17,19	17,14	10,27	26,72	27,98	18,04
(e)	14,13	14,81	8,06	8,49	9,25	4,77	15,18	16,00	9,14
(f)	142,84	148,50	92,06	145,06	153,72	94,27	146,53	150,78	93,87
(g)	6,52	6,89	5,23	8,33	9,67	7,80	5,84	6,49	4,85

Tabela 4.1: Tabela que apresenta uma série de métricas calculadas durante o pré-processamento dos *corpora* em três momentos: a partir do texto das frases tal qual transcritas, lematizadas e stemizadas. (a): Média de tamanho de sentença (em palavras); (b): Desvio padrão de (a); (c): Erro padrão de (b); (d): Riqueza lexical (%); (e): Incidência de hápax legômena (%); (f) Média de palavras por narrativa; (g): Média da porcentagem de repetições de palavra por sentença (%).

É também o grupo DA que exhibe maior variabilidade no tamanho dessas sentenças, como aponta o desvio padrão na linha (b), seguido pelo CCL e o grupo-controle, sendo esse último o que tende à maior estabilidade entre os três nessa medida.

Além disso, para os modelos apresentados nas Subseções 3.2.2 e 3.2.3 com resultados adiante, nas Seções 4.2 e 4.3, optamos por utilizar os dados stemizados, uma vez que apresentam uma maior confiabilidade, apresentando erro padrão (linha (c)) $SE_{stem} = 0,57$, menor que $SE_{lem} = 0,66$ dos dados lematizados.

Em seguida, na linha (d), estão calculadas as riquezas lexicais de cada grupo. Essa medida indica a variabilidade do vocabulário (quanto maior a medida, mais *rico* o vocabulário) por meio da razão entre palavras sem repetição (*types*) e número total de ocorrências (*tokens*), como esquematizado na Fórmula 4.1. Os resultados apontam uma proximidade entre os grupos DA e CCL (com vantagem para CCL nos dados apenas sem *stopwords* e stemizados e para DA nos lematizados), distanciados por cerca de 7 a 10 pontos percentuais do grupo-controle. Esse fato indica a possível existência de um fenômeno em comum entre os grupos clínicos, que não ocorre no grupo-controle, fazendo com que o total de *types* entre os primeiros seja proporcionalmente maior que os do último.

$$R_{\text{lex}} = \frac{\text{types}}{\text{tokens}} \quad (4.1)$$

Uma hipótese, que pode ser verificada futuramente, para esse possível fenômeno é o caso em que os grupos clínicos realizem mais evitações, palavras incompletas ou outros tipos de recursos que interrompam o discurso; esses termos podem gerar mais itens únicos (*types* com apenas um *token* de ocorrência) em uma mesma narração, o que tenderia a aumentar essa métrica de riqueza.

As porcentagens em (e), referentes a hápax legômena¹, foram calculadas de modo considerar a contagem de hápax legômena dividida pela contagem de palavras totais (*tokens*) de cada indivíduo, por fim, foi feita uma média desses valores por grupo. Com esses números, nota-se que o grupo-controle é o que tem a menor incidência de hápax legômena, seguido de DA e CCL, nessa ordem. Esse dado vai na mesma direção do anterior, o qual expressa uma significativa maior quantidade de *types* em comparação com o grupo-controle. Aqui, com os hápax, a diferença chega a ser de quase duas vezes. Assim, ambos apontam para uma maior variabilidade na *escolha* lexical por parte de DA e CCL.

É necessário, porém, levantar a possibilidade de esses dois dados (ainda sobre as linhas (d) e (e)) serem tendenciosos em algum nível, dado o tamanho restrito das narrativas, uma vez que em documentos curtos a chance de determinadas palavras aparecerem apenas uma vez (hápax) é alta; além de haver um valor *tokens* menor, fazendo com que a riqueza lexical aumente. Assim sendo, cabe analisar a extensão de cada narrativa como vemos em (f), onde se expressa a média de palavras por narrativa em cada grupo. Vê-se que cada indivíduo nos grupos clínicos tende a produzir narrativas entre 55% e 60% maiores que as dos indivíduos saudáveis. Novamente com valores que permitem agrupar DA e CCL pela proximidade, distanciando-os do grupo-controle. Portanto, sendo DA e CCL os grupos que contam com textos mais extensos, reduzem-se as chances de a hipótese da tendenciosidade ser válida, uma vez que, em comparação, o grupo com menores extensões não apresenta esse comportamento.

¹Aqui, adotamos o termo significando palavras de ocorrência única em um dado documento – no caso, cada narrativa.

Na linha (g), por fim, temos dados que indicam a razão entre a quantidade de *tokens* de cada *type* com, pelo menos, duas ocorrências (isto é, uma repetição) e a quantidade de tokens total em cada sentença. Com a média desses dados, temos um valor aproximado da porcentagem de palavras que se repete em cada sentença. E, assim, notamos que o grupo CCL tende a ter mais repetições que DA, o qual, por sua vez, incorre em mais repetições que o grupo-controle.

4.2 Resultados com classificador bayesiano

Nesta seção, apresentamos os resultados obtidos com a execução do classificador bayesiano ingênuo descrito na Subseção 3.2.2 tanto com *corpus* desbalanceado (quantidade de sentenças original) quanto balanceado (quantidade de sentenças do grupo-controle reduzida). Na Tabela 4.2, estão organizados os dados estatísticos a respeito do desempenho da classificação que esse modelo bayesiano apresentou para ambos os cenários de balanceamento.

Estado do <i>corpus</i>	Grupos	Acurácia	Precisão	Cobertura	Medida F1
Desbalanceado	DA, DLC, controle	0,61	0,64	0,48	0,47
	DA, controle	0,78	0,81	0,69	0,71
	CCL, controle	0,69	0,56	0,52	0,48
	DA, CCL	–	–	–	–
Balanceado	DA, DLC, controle	0,59	0,59	0,58	0,58
	DA, controle	0,75	0,75	0,75	0,75
	CCL, controle	0,68	0,68	0,68	0,68
	DA, CCL	0,69	0,69	0,68	0,68

Tabela 4.2: Medidas de acurácia, precisão, cobertura e F1 do desempenho do classificador bayesiano frente a todos os agrupamentos possíveis dos grupos de análise.

Se compararmos os desempenhos entre os cenários de balanceamento, notamos acurácia e precisão com melhores resultados no *corpus* desbalanceado; enquanto, no balanceado, são apresentadas cobertura e medida F1 com valores mais elevados.

Em termos de acurácia, os dados desbalanceados levam vantagem. No entanto, há alguns problemas a serem interpretados. A cobertura baixa nos dados desbalanceados mostra que se cometem mais erros que os dados balanceados, por exemplo, assumindo como falso um dado que, na verdade, é verdadeiro.

Por outro lado, nos dados de cobertura e precisão para o *corpus* balanceado, vemos números iguais ou próximos para ambas as medidas, ou seja, o número de casos verdadeiros positivos está relativamente baixo.

Por fim, temos, nas Figuras 4.1 e 4.2, o desempenho do modelo por meio das matrizes de confusão, exibindo números de erros e acertos na tarefa de classificação. Com o mapa de calor, podemos notar como o grupo-controle (identificado com o número 1) tende a ser mais facilmente distinto dos demais, seguido de DA (identificado com o número 0) e, por último, CCL com poucos acertos em relação aos outros grupos.

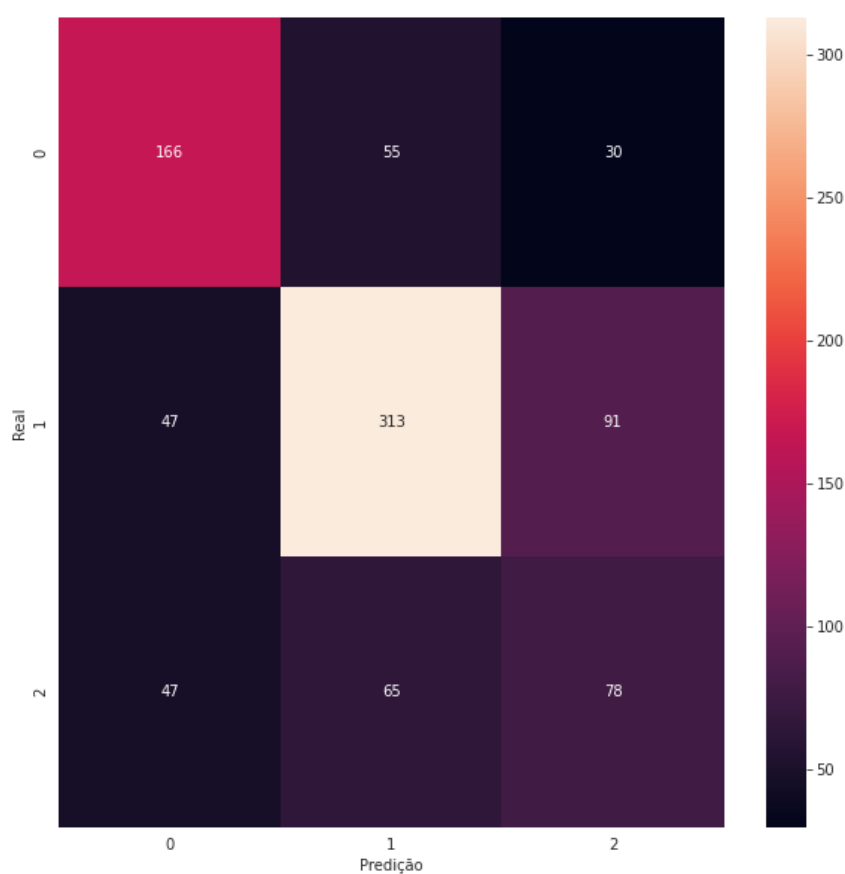


Figura 4.1: Matriz de confusão com valores absolutos na tarefa de classificação (modelo bayesiano ingênuo *Multinomial*) entre DA (0), grupo-controle (1) e CCL (2). *Corpus* desbalanceado.

4.3 Resultados com rede de propagação para frente

Passemos agora à análise dos resultados obtidos com a classificação via rede de propagação para frente estruturada no modelo *Sequential* da biblioteca *keras*.

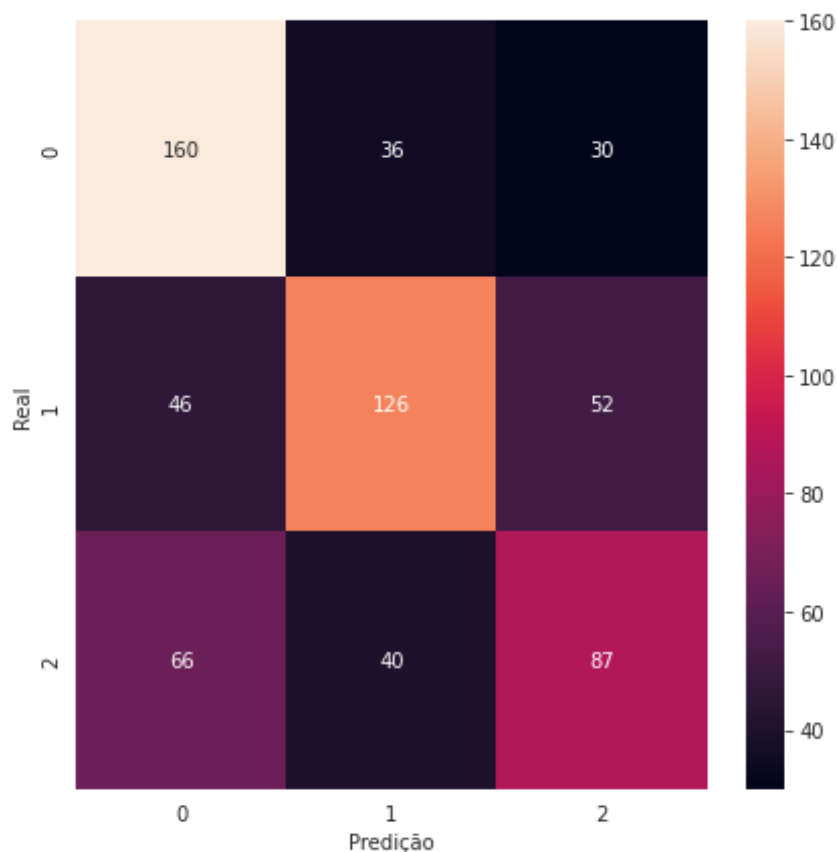


Figura 4.2: Matriz de confusão com valores absolutos na tarefa de classificação (modelo bayesiano ingênuo *Multinomial*) entre DA (0), grupo-controle (1) e CCL (2). *Corpus* balanceado.

Nas figuras a seguir, de 4.3 a 4.13, são apresentados pares de gráficos, em que os gráficos à esquerda, identificados com (a), mostram a acurácia do modelo tanto no treinamento (linhas pontilhadas) quanto na validação (linhas contínuas) dos dados — isto é, momento em que o aprendizado gerado no treinamento é testado e validado. Nos gráficos à direita, identificados com (b), apresentam-se as perdas, novamente, no treinamento e na validação. Ambos os tipos de gráfico são dados ao longo de épocas.

As próximas três Subseções ocupam-se de analisar esses dados gráficos segmentadas conforme os cenários estabelecidos na Tabela 3.5. Para melhor transitar entre os gráficos, consulte a Tabela 4.3.

4.3.1 Configuração A

Inicialmente, a fim de testar os dados desbalanceados, não fizemos nenhum tipo de restrição à quantidade de dados; ou seja, usamos o grupo-controle completo,

Grupos	A	B	C
DA, controle	4.3	4.7	4.11
CCL, controle	4.4	4.8	4.12
DA, CCL	4.5	4.9	4.13
DA, CCL, controle	4.6	4.10	–

Tabela 4.3: Tabela para facilitar a movimentação entre os gráficos com resultados da rede de propagação para frente, organizando-os por grupos e configuração aplicada.

tal qual mostra a contagem na Figura 3.1a. Além disso, vale lembrar que a entropia cruzada categorial foi aplicada como função de perda.

Na Figura 4.3a, a acurácia no treinamento do modelo, quando submetidos os grupos DA e controle, inicia-se por volta de apenas 55% na primeira época, mas logo se estabiliza, até a oitava época, por volta de 67%, para em seguida começar a crescer. Na validação, no entanto, temos duas regiões de estabilidade: antes da nona época, por volta de 65%, e depois dela, por volta de 75%.

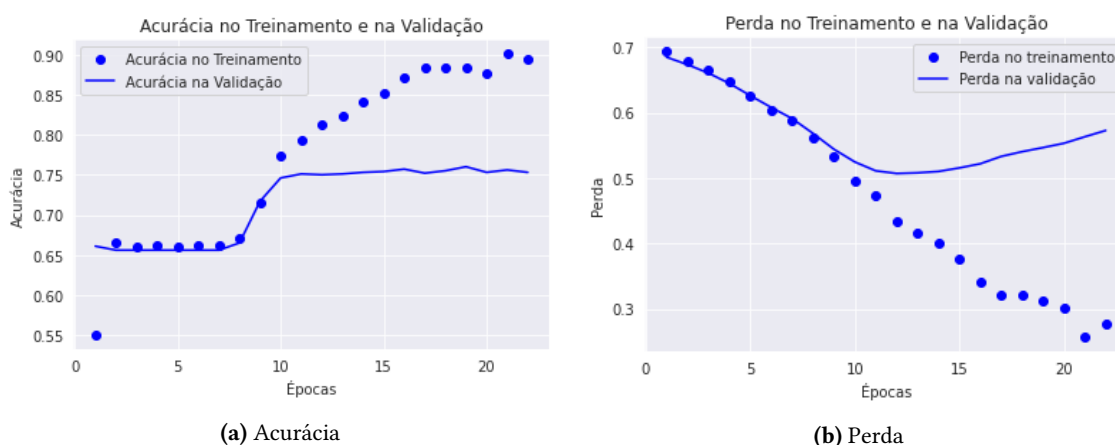


Figura 4.3: Gráficos de acurácia (4.3a) e perda (4.3b) no treinamento e na validação do modelo ao se submeter os dados de DA e do grupo-controle. Configuração A.

A acurácia do modelo ao contrastar CCL com grupo-controle (Figura 4.4a) nos mostra um crescimento constante no treinamento, tendo um estágio inicial (da segunda à nona época) relativamente estável por volta de 72%, após o qual a acurácia melhora quase linearmente até 90% na 22^a época. Por outro lado, a validação se mostra bastante estável ao longo das épocas se mantendo em torno de 70%.

Na Figura 4.5a, a acurácia indica o desempenho do modelo ao tentar distinguir os grupos de DA e de CCL. Aqui, o treinamento tem um crescimento

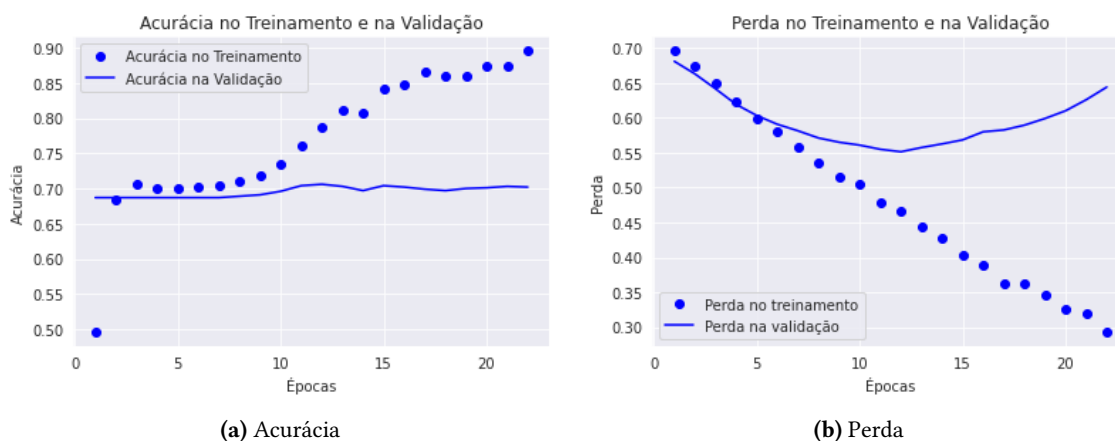


Figura 4.4: Gráficos de acurácia (4.4a) e perda (4.4b) no treinamento e na validação do modelo ao se submeter os dados de CCL e do grupo-controle. Configuração A.

relativamente constante ao longo do tempo; porém, desta vez, diferentemente dos dois testes anteriores, a acurácia demora mais épocas até atingir um nível de estabilidade: somente após a 21^a se mantém por volta de 65%. Essa porcentagem revela outra situação: ela representa também o pico do desempenho, ao passo que nos dois casos anteriores (DA vs. grupo-controle e CCL vs. grupo-controle) esse era um dos menores níveis alcançados, tendendo a terem uma acurácia de 70% ou mais. Uma interpretação para esse fato são as diferenças entre a fala de indivíduos saudáveis e aqueles acometidos por algum dos casos de demência aqui abordados e que afetam a produção linguística de maneira significativa. Por outro lado, buscar uma diferenciação entre os dois grupos clínicos se mostra uma tarefa mais desafiadora para o modelo, uma vez que essas alterações aparecem em maior ou menor quantidade ou intensidade, mas tendem a ser semelhantes.

Por fim, ao analisar o modelo computando os três grupos gera os resultados vistos na Figura 4.6. A Figura 4.6a aponta para um período de estabilidade na acurácia da segunda à oitava época tanto para o treinamento quanto para a validação, contando com níveis de 53% e 48%, respectivamente. A partir desse momento, os movimentos divergem: o treinamento segue em crescimento próximo a linear passando de 75% a partir da 21^a época; a validação cresce por algumas épocas, mas estabiliza-se atingindo no máximo 60% de acurácia.

Todos os gráficos de perda na configuração A seguem um padrão semelhante e com números semelhantes (com exceção da perda para os três grupos, em

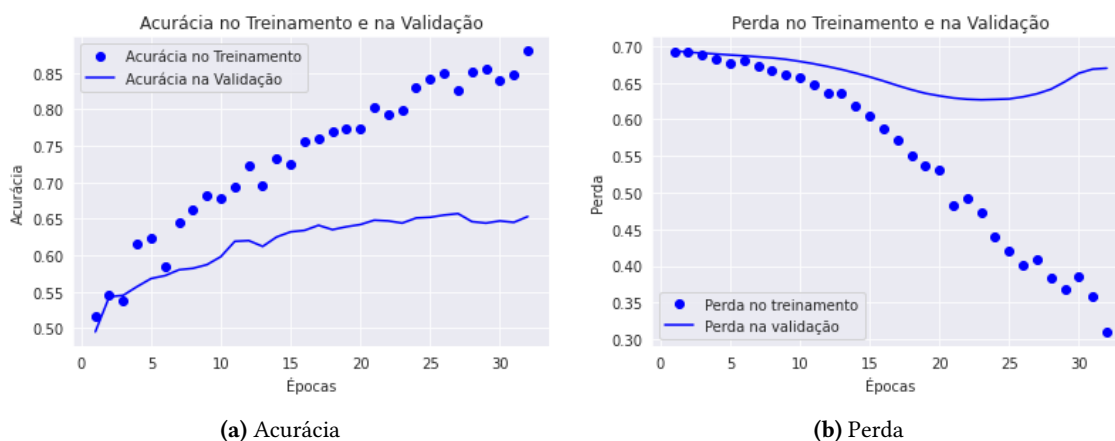


Figura 4.5: Gráficos de acurácia (4.5a) e perda (4.5b) no treinamento e na validação do modelo ao se submeter os dados de DA e de CCL. Configuração A.

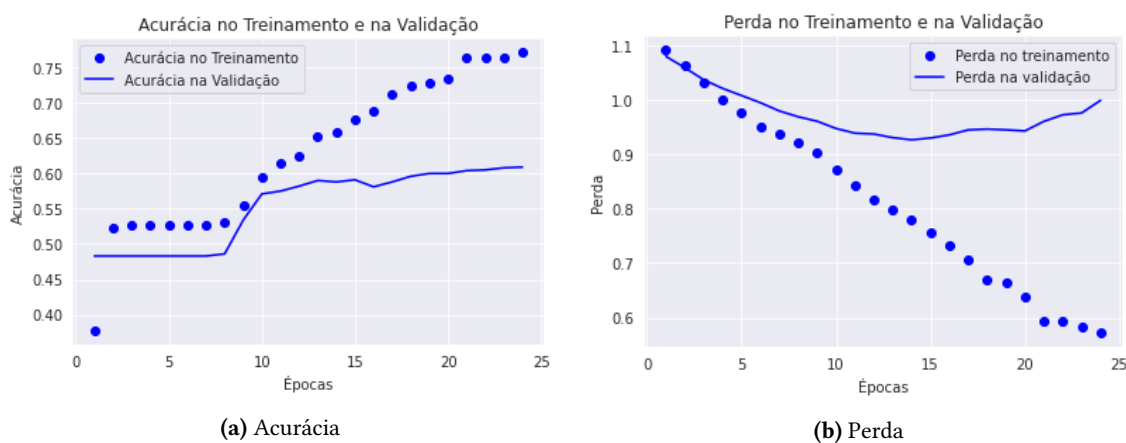


Figura 4.6: Gráficos de acurácia (4.6a) e perda (4.6b) no treinamento e na validação do modelo ao se submeter os dados dos três grupos: DA, CCL e grupo-controle. Configuração A.

que os valores são mais altos). De um modo geral, os treinamentos apresentam uma linearidade decrescente, condicionando um *overfitting* aos dados em algum momento. A validação, por sua vez, apresenta uma curva, apresentando queda, mas volta a aumentar ao longo do tempo. Vale notar a perda para os grupos DA e CCL (Figura 4.5b) que mostra a menor diminuição na perda entre os demais; o que confirma a maior dificuldade de o modelo distinguir entre os grupos clínicos.

4.3.2 Configuração B

Na configuração B, em oposição à A, implementamos a alteração nos dados que reduz o grupo-controle, a fim de balanceá-lo tomando como parâmetro os demais grupos. A entropia cruzada continua sendo categorial.

Analisemos os resultados da configuração B nos mesmos recortes de *corpora* em comparação com A. Nos grupos DA e controle (Figuras 4.3 e 4.7), passamos a ter um período de estabilidade da acurácia menor, atingindo cerca 70% na validação, enquanto o treinamento apresenta *overfitting* – 5% a menos do que em A.

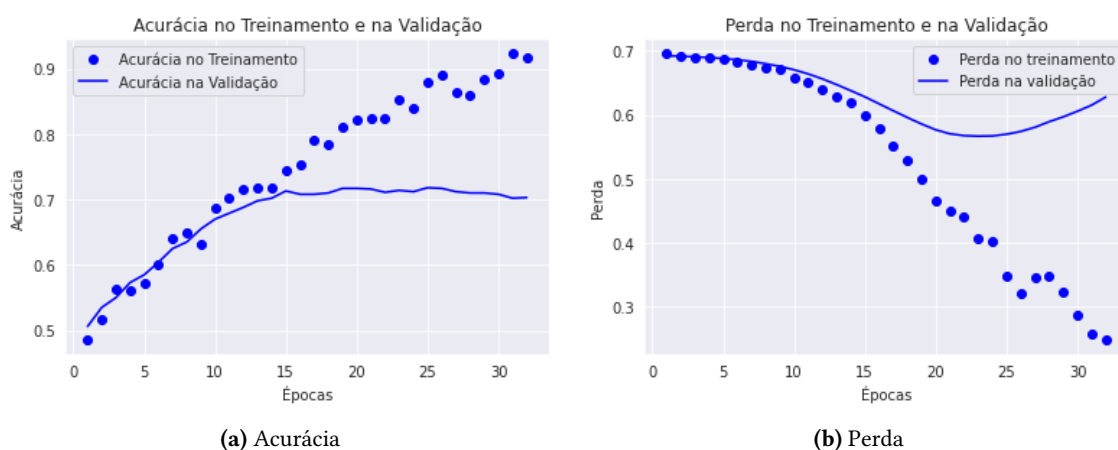


Figura 4.7: Gráficos de acurácia (4.7a) e perda (4.7b) no treinamento e na validação do modelo ao se submeter os dados de DA e do grupo-controle. Configuração B.

Nos grupos CCL e controle (Figuras 4.4 e 4.8), a queda no desempenho torna a aparecer, com uma estabilidade de 60%, enquanto o cenário A apresentava 70% de acurácia.

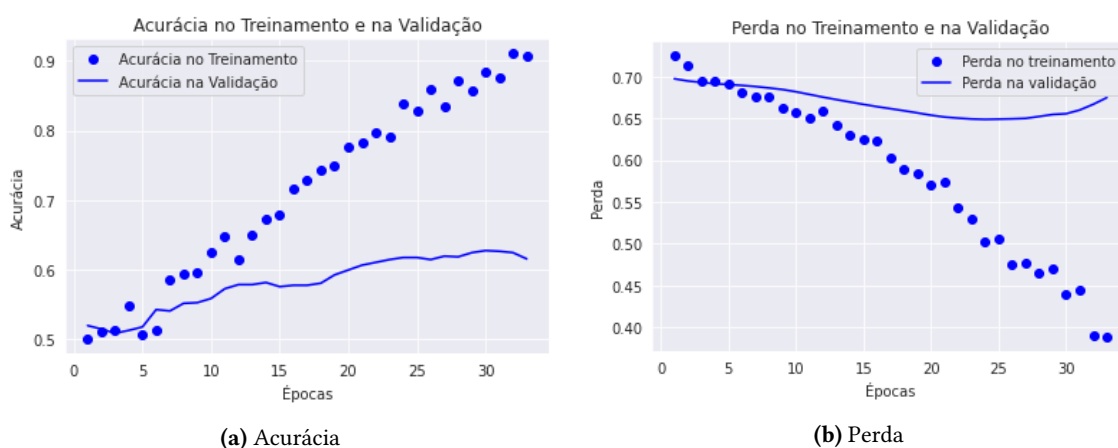


Figura 4.8: Gráficos de acurácia (4.8a) e perda (4.8b) no treinamento e na validação do modelo ao se submeter os dados de CCL e do grupo-controle. Configuração B.

Nos grupos DA e CCL (Figuras 4.5 e 4.9), os gráficos apresentam um comportamento similar, com um treinamento crescente até o *overfitting*, ao passo que a validação atinge, no máximo, por volta de 65%.

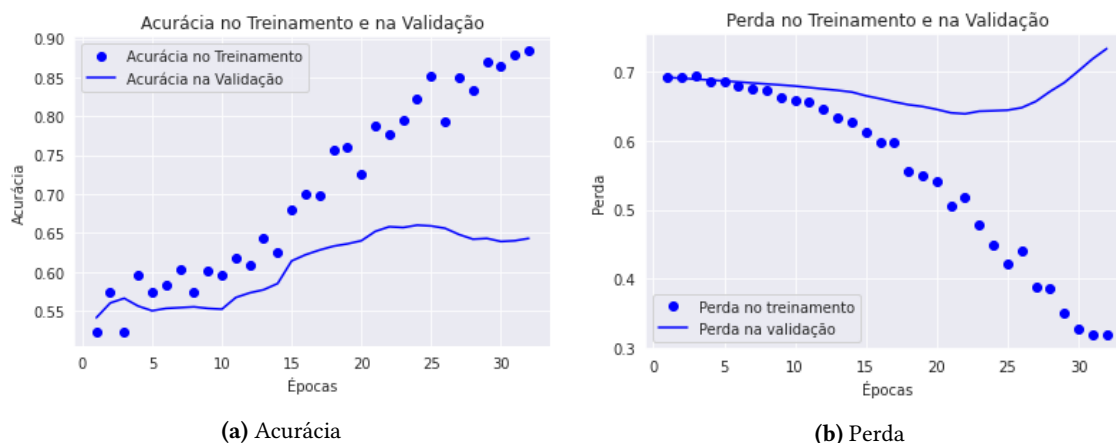


Figura 4.9: Gráficos de acurácia (4.9a) e perda (4.9b) no treinamento e na validação do modelo ao se submeter os dados de DA e de CCL. Configuração B.

Por fim, submetendo-se ao modelo os três grupos (Figuras 4.6 e 4.10), o que encontramos é, na configuração B, novamente um desempenho um pouco menor: enquanto, em A, superavam-se os 60% de acurácia, aqui, o máximo se encontra por volta de 55%.

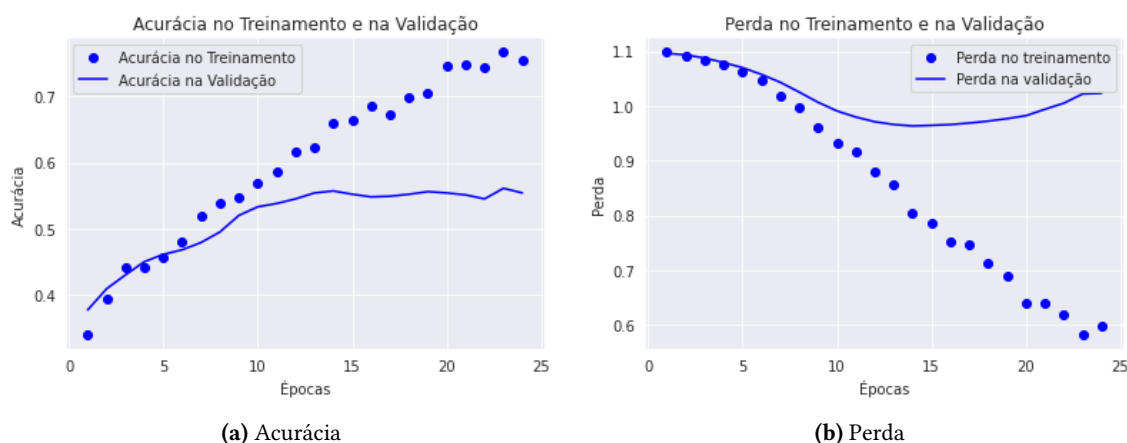


Figura 4.10: Gráficos de acurácia (4.10a) e perda (4.10b) no treinamento e na validação do modelo ao se submeter os dados dos três grupos: DA, CCL e grupo-controle. Configuração B.

Portanto, considerando as acurácias que o modelo apresentou sob as configurações A e B, notamos que em B o desempenho variava entre cerca de cinco ou dez pontos percentuais para baixo — com exceção do caso em que se submetem os grupos DA e CCL, pois não são influenciados pelo corte no grupo-controle; nessa dupla, o desempenho se manteve.

Como confirmação, se avaliarmos os gráficos de perda das ocorrências de ambas configurações, notamos que, em B — e em linha com o fato de a acurácia

ser mais baixa nesse cenário —, a perda é mais agressiva. Como exemplo, vejamos os casos dos grupos DA e controle: enquanto, em A (Figura 4.3b), atinge-se um momento em que a perda é de 0,5, em B (Figura 4.7b), seu mínimo é 0,58 antes de retomar o aumento.

Em suma, não é surpreendente que com a intervenção no *corpus* de modo a reduzir o tamanho do conjunto de sentenças do grupo-controle, o desempenho do modelo se mostre menos eficaz; essa é uma possibilidade quando se trata de aprendizado de máquina em que a quantidade de dados é fundamental para aprendizado de padrões. Assim sendo, o enviesamento poderia ser um problema em *corpora* maiores; aqui, uma hipótese preliminar é que a retirada de dados do grupo-controle diminuiu o desempenho na classificação quando esse grupo estava envolvido (CCL vs. controle, DA vs. controle, além dos três).

4.3.3 Configuração C

Agora, comparemos os resultados do modelo na configuração C, que se vale de uma entropia cruzada binária (`binary_crossentropy`) em vez de `categorical_crossentropy` como ocorreu nas configurações A e B; além de receber o *corpus* do grupo-controle com a redução para balanceamento.

Começando por DA e grupo-controle (Figuras 4.3, 4.7 e 4.11), observa-se comportamento e números bastante próximos dos resultados com a configuração B, apresentando acurácia com seu pico em torno de 70% (com A, ela atingia 5% a mais).

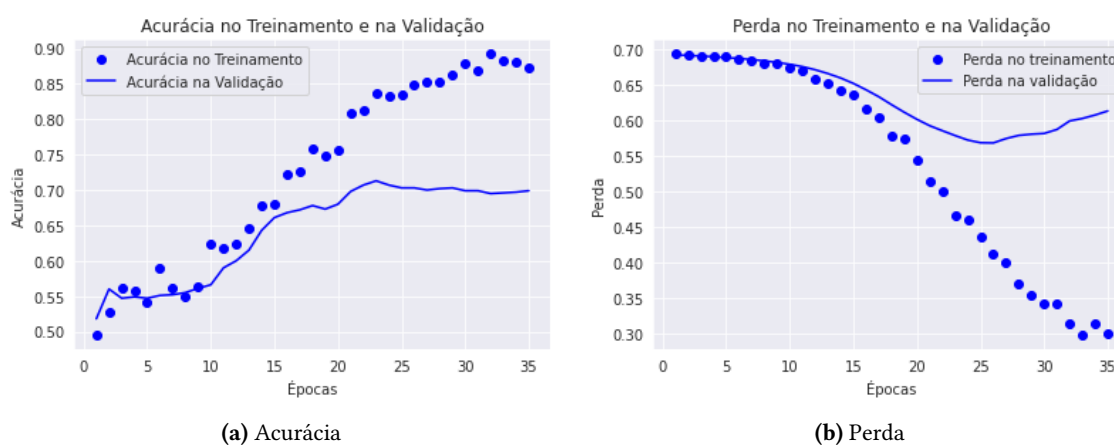


Figura 4.11: Gráficos de acurácia (4.11a) e perda (4.11b) no treinamento e na validação do modelo ao se submeter os dados de DA e do grupo-controle. Configuração C.

Novamente exibindo métricas próximas a B, a configuração C para os grupos CCL e controle (Figuras 4.4, 4.8 e 4.12) apresenta uma curva de aprendizado crescente com validação atingindo no máximo 64%.

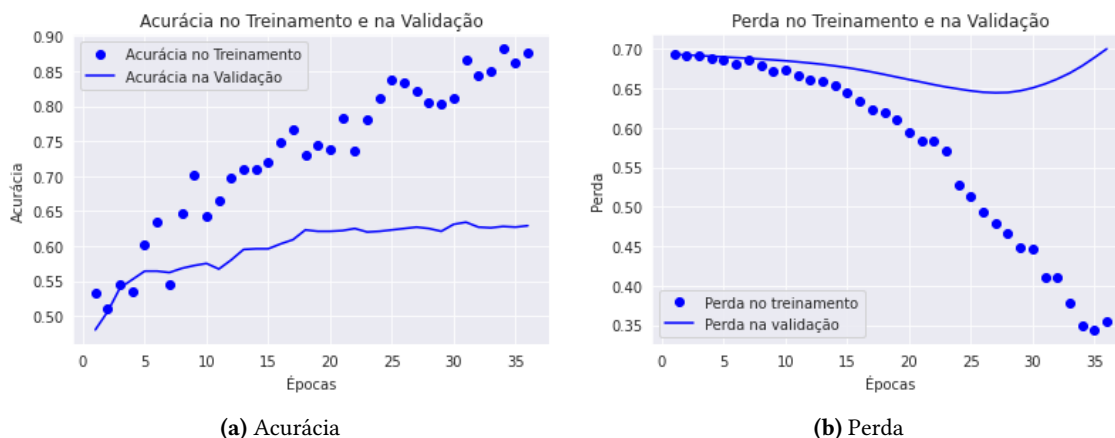


Figura 4.12: Gráficos de acurácia (4.12a) e perda (4.12b) no treinamento e na validação do modelo ao se submeter os dados de CCL e do grupo-controle. Configuração C.

Com exceção do fato de apresentar um início mais estável na acurácia (por volta de 54%), em C, os grupos DA e CCL (Figuras 4.5, 4.9 e 4.13) performam semelhantemente. A acurácia estaciona, a partir de 25 épocas, em torno de 65%.

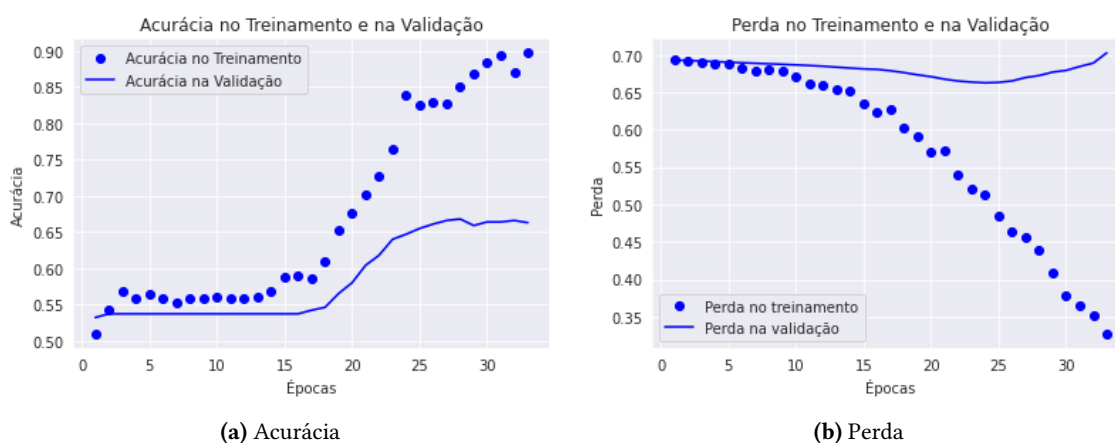


Figura 4.13: Gráficos de acurácia (4.13a) e perda (4.13b) no treinamento e na validação do modelo ao se submeter os dados de DA e de CCL. Configuração C.

Apesar de o desempenho em C também ser menor do que em A, a maior desvantagem aqui é não poder processar todos os três conjuntos de dados, prejudicando o intuito inicial de elaborar um classificador que distinga entre os três grupos de interesse.

4.4 Resultados com rede BiLSTM

Nas Tabelas 4.4, 4.5 e 4.6, estão os dados para cada par de grupo de análise com o processamento do texto na sua forma original; enquanto as Tabelas 4.7, 4.8 e 4.9 apresentam respectivamente a mesma divisão, porém agora tendo como entrada o texto anotado por nós. Cada Tabela contém dados de perda e acurácia tanto de treinamento quanto de validação por 4 épocas.

Época	Perda	Acurácia	Perda da Validação	Acurácia da Validação
1	-3.637	0.224	-6.318	0.204
2	-6.698	0.225	-9.043	0.204
3	-8.903	0.225	-11.685	0.204
4	-11.165	0.225	-14.199	0.204

Tabela 4.4: Resultados da rede BiLSTM para o *dataset* com texto bruto por 4 épocas para DA e CCL.

Época	Perda	Acurácia	Perda da Validação	Acurácia da Validação
1	-4.244	0.213	-5.632	0.257
2	-7.388	0.212	-7.965	0.257
3	-9.729	0.212	-10.134	0.257
4	-12.204	0.212	-12.267	0.257

Tabela 4.5: Resultados da rede BiLSTM para o *dataset* com texto bruto por 4 épocas para DA e CTR.

Época	Perda	Acurácia	Perda da Validação	Acurácia da Validação
1	-4.057	0.223	-5.899	0.214
2	-7.241	0.222	-8.276	0.214
3	-9.561	0.222	-10.644	0.214
4	-11.654	0.222	-12.808	0.214

Tabela 4.6: Resultados da rede BiLSTM para o *dataset* com texto bruto por 4 épocas para CCL e CTR.

Época	Perda	Acurácia	Perda da Validação	Acurácia da Validação
1	-14.447	0.212	-14.349	0.257
2	-16.584	0.212	-16.363	0.257
3	-18.681	0.212	-18.448	0.257
4	-21.129	0.212	-20.503	0.257

Tabela 4.7: Resultados da rede BiLSTM para o *dataset* com texto anotado por 4 épocas para DA e CCL.

A perda é uma medida de quão bem o modelo está se saindo. É um único valor escalar que representa a diferença entre os valores previstos e os valores

Época	Perda	Acurácia	Perda da Validação	Acurácia da Validação
1	-3.922	0.222	-4.891	0.221
2	-6.731	0.221	-6.994	0.221
3	-9.125	0.221	-8.978	0.221
4	-11.395	0.221	-10.880	0.221

Tabela 4.8: Resultados da rede BiLSTM para o *dataset* com texto anotado por 4 épocas para DA e CTR.

Época	Perda	Acurácia	Perda da Validação	Acurácia da Validação
1	-3.790	0.221	-5.161	0.221
2	-6.992	0.221	-7.482	0.221
3	-9.433	0.221	-9.679	0.221
4	-12.018	0.221	-11.821	0.221

Tabela 4.9: Resultados da rede BiLSTM para o *dataset* com texto anotado por 4 épocas para CCL e CTR.

reais. Normalmente, deveria ser positiva e ir sendo minimizada ao longo das épocas de treinamento, ou seja, caminhando em direção ao zero. No caso, a perda é negativa e se distancia de zero, ficando cada vez menor.

Os dados de acurácia mostram a proporção de amostras classificadas corretamente. Aqui, é bastante baixa, em torno de 22 a 23%.

Esses dados são inconclusivos por ora, uma vez que indicam que o modelo não está aprendendo bem ou que há um problema que não identificamos com a implementação dele. Poderia se tratar de um problema com os dados, porém é improvável, uma vez que os outros modelos — que tiveram desempenhos consistentes e, em geral, bons — tiveram como entrada os mesmos dados, provenientes da mesma fonte, apenas com as adaptações necessárias para os requisitos de cada modelo.

4.5 Resultados com DistilBERT

Os resultados do treinamento a partir do modelo DistilBERT são apresentados a seguir. Foram seis tipos diferentes de treinamento, sempre treinando os grupos de modo binário: DA e CCL, DA e controle, controle e CCL. Além disso, fizemos o treinamento tanto com os dados na forma do texto bruto (ou seja, tal qual foram transcritos) quanto com os dados anotados (isto é, com as etiquetas que

usamos para substituir tokens que representavam manifestação dos fenômenos linguísticos de hesitação, como <FW> para os preenchedores).

Durante a execução dos treinamentos, observamos que os resultados variavam para uma mesma configuração de treinamento; provavelmente, devido à aleatorização dos dados e quantidade restrita de dados, sobretudo quando se trata do conjunto de validação. Portanto, para termos resultados mais robustos fizemos dez execuções para cada uma das seis possíveis configurações, registrando a acurácia de cada uma. Por fim, calculamos a média e o desvio padrão para cada combinação — esses dados são todos apresentados na Tabela 4.10. Além desta Tabela, as médias com seus respectivos desvios padrões são esquematizadas graficamente na Figura 4.14.

Execução	Texto bruto			Texto anotado		
	DA e CCL	DA e CTR	CTR e CCL	DA e CCL	DA e CTR	CTR e CCL
1	0.71	0.73	0.70	0.70	0.78	0.69
2	0.73	0.78	0.71	0.72	0.76	0.74
3	0.72	0.77	0.68	0.70	0.80	0.68
4	0.70	0.79	0.71	0.74	0.79	0.67
5	0.73	0.78	0.72	0.75	0.79	0.68
6	0.68	0.78	0.70	0.73	0.77	0.66
7	0.70	0.77	0.70	0.70	0.78	0.72
8	0.71	0.78	0.69	0.71	0.77	0.70
9	0.72	0.78	0.72	0.71	0.76	0.68
10	0.74	0.76	0.67	0.69	0.79	0.69
Média	0.71	0.77	0.70	0.72	0.78	0.69
DP	0.02	0.02	0.02	0.02	0.01	0.02

Tabela 4.10: Resultados de acurácia do desempenho do modelo `distilbert-portuguese-cased`, baseado no BERTimbau, para os grupos de estudo dois a dois tanto para o *dataset* com texto bruto (br.) quanto para o anotado (an.) por dez execuções do código acompanhados da média e desvio padrão.

Observamos uma variação no desempenho do modelo em diferentes configurações de texto. Notavelmente, ambas as configurações que envolvem DA performaram melhor com os dados anotados, enquanto CCL com o grupo-controle tiveram desempenho levemente melhor com o texto bruto. De qualquer forma, surge aí um indício de que a informação adicional fornecida pela anotação é benéfica para o treinamento.

O grande destaque com esses dados não envolve a comparação entre texto bruto e anotado, que tiveram 0.01 de variação para cada par de análise, com

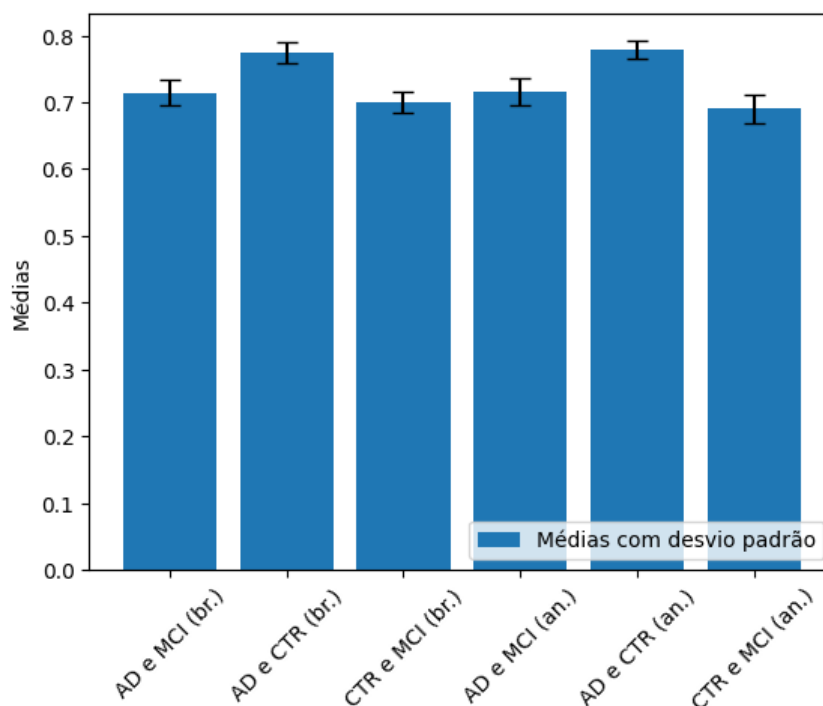


Figura 4.14: Médias de acurácia do desempenho do modelo distilbert-portuguese-cased, baseado no BERTimbau, para os grupos de estudo dois a dois tanto para o *dataset* com texto bruto (br.) quanto para o anotado (an.).

quase todos os desvios padrões a 0.02. O destaque recai sobre a performance muito mais alta na distinção de DA e grupo-controle. Isso seria de se esperar, uma vez que DA tende a apresentar maior incidência de fenômenos linguísticos em comparação com controle, que está no oposto do nosso *continuum*. Vale apontar, também, que o texto anotado pode ter garantido ainda melhor desempenho para esse par, visto que apresenta o menor desvio padrão.

Em termos de robustez do modelo, a baixa variação do desvio padrão em todas as configurações sugere um bom desempenho ao longo das várias execuções, com valores consistentes.

4.6 Considerações acerca dos resultados dos classificadores

Com base nos resultados obtidos, uma observação que se faz notar é o fato de que se está reproduzindo, sobretudo com a rede de propagação para frente e com o modelo baseado em BERT, uma situação semelhante à percepção clínica

em que se considera um *continuum* de empobrecimento lexical e sintático partindo dos indivíduos saudáveis (Beltrami et al. 2018, p. 10) — os quais, em situações narrativas sobretudo, mas também nas descritivas, não teriam a produção dessas e outras manifestações linguísticas (Abrisqueta-Gomez et al. 2004; Nitrini, Caramelli et al. 2005) prejudicadas cognitivamente, constituindo-se no grupo com narrativas mais bem desenvolvidas nesse recorte — em direção aos com doença de Alzheimer, inclusive tendo o grupo com CCL como grupo intermediário dessa variação.

Assim, como visto nos gráficos de acurácia na Seção 4.3, quando processados os dados dos grupos de dois em dois, há uma maior facilidade em se distinguir DA do grupo-controle, os dois extremos nessa nossa “régua”; em seguida, com um pouco mais de dificuldade, distinguem-se os dados entre CCL e grupo-controle, uma vez que neste grupo clínico os reflexos linguísticos de origem neurodegenerativa começam a se manifestar na língua, dando menos pistas para a classificação automática do modelo. Por fim, o maior desafio é a diferenciação entre DA e CCL, uma vez que ambos apresentam problemas linguísticos iniciados semelhantemente.

Os gráficos de perda tendem a seguir um mesmo padrão: o treinamento se comporta de maneira linearmente decrescente; enquanto isso, a validação mostra menos consistência, uma vez que acompanha o movimento do treinamento, em queda. Entretanto, o movimento se reverte e passa a exibir um crescente na perda. Isso reflete a produção de *overfitting* nos dados de treinamento, conforme passam as épocas e a validação deixa de acompanhá-lo, aumentando a perda. Uma exceção a esse movimento são os casos em que o modelo processa os grupos DA e CCL juntos, em que a perda é alta boa parte do tempo e tende ao pior desempenho (o gráfico em que isso é mais evidente é o da Figura 4.13b), desenhando pouca melhora ao longo do tempo.

Já os resultados com base no modelo DistilBERT na Seção 4.5 sugerem que ele é capaz de aprender de maneira consistente a partir dos dados fornecidos, podendo diferenciar os pares de grupos de análise, especialmente quando se utiliza texto anotado. A escolha da combinação de texto também influencia o desempenho, com a combinação de DA com grupo-controle apresentando

os melhores resultados em média. A consistência nos resultados ao longo das execuções é encorajadora, indicando uma robustez geral do modelo. Essas observações fornecem *insights* valiosos para aprimoramentos futuros, ajustes na parametrização do treinamento do modelo e, principalmente, a implementação com mais dados.

Os resultados da rede BiLSTM são os menos robustos e não apresentam resultados que se podem afirmar como bons (afinal, as acurácias são bastante baixas) nem necessariamente consistentes para refutar nossas hipóteses. Diversos fatores podem vir a melhorar a performance no futuro, mesmo cumulativamente: diferente implementação da rede, diferente parametrização da rede escolhida, inclusão de mais dados.

4.7 Resultados das análises de fenômenos linguísticos

Com base na literatura de referência sobre fenômenos linguísticos relacionados a hesitações e apresentada no capítulo anterior, observamos nos nossos *corpora* os seguintes elementos descritos por Marcuschi (1999) e Marcuschi (2003): pausas preenchidas (às quais nos referiremos como *preenchedores*), repetições hesitativas (*repetições completas*), gaguejamentos e falsos inícios (estes dois últimos, aqui, incorporados ao que chamaremos de *repetições incompletas*). Os *preenchedores* foram anotados manualmente, enquanto as *repetições completas* e *incompletas* foram anotadas automaticamente com base em regras ao nível de caractere. Os resultados para as análises de cada um dos três agrupamentos que propomos são apresentados nas subseções a seguir, além de um quarto agrupamento que une as duas formas de repetições, cuja motivação será apresentada na sua respectiva subseção.

4.7.1 Preenchedores

Na Tabela 4.11 estão os resultados acerca da frequência com que preenchedores tendem a aparecer nas narrativas dos indivíduos de cada grupo. Essa métrica é a razão entre o número de preenchedores empregados na narrativa de um indivíduo e o número total de palavras (*tokens*) dessa narrativa. Calculamos a

média dessa métrica por grupo de indivíduos que se apresenta na tabela referida junto do erro padrão e também, graficamente, na Figura 4.15.

Grupo	Proporção média	Erro padrão
DA	0.0365	0.0048
Controle	0.0303	0.0023
CCL	0.0327	0.0034

Tabela 4.11: Médias e erros-padrão das razões entre o número de preenchedores e o número total de tokens por narrativa para cada grupo.

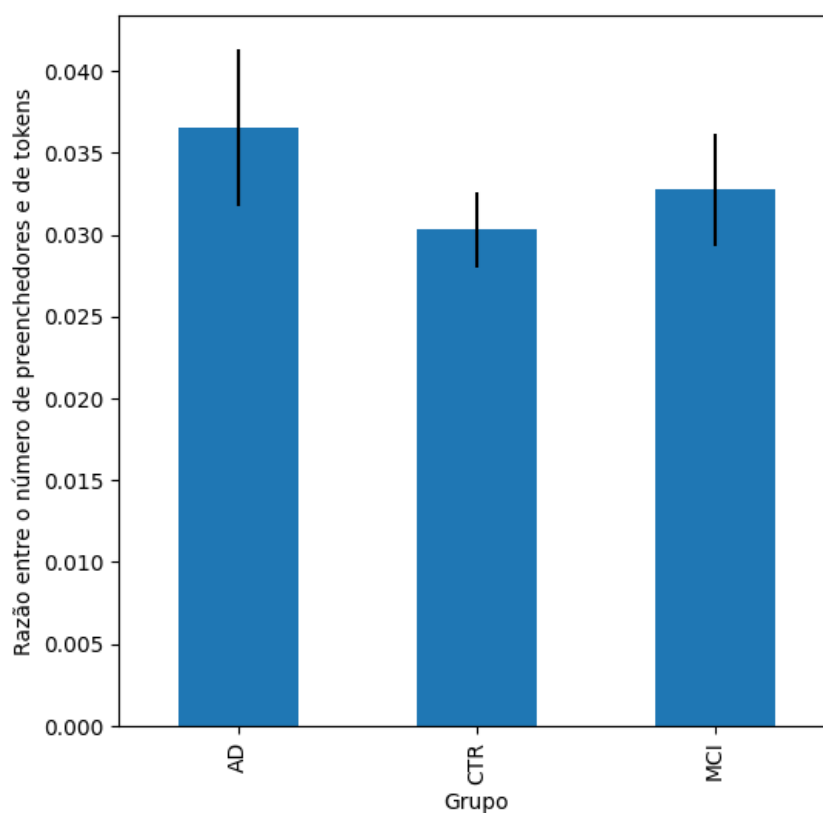


Figura 4.15: Comparação entre médias e erros-padrão das razões do número preenchedores sobre o número total de tokens por narrativa para cada grupo.

Encarando esses dados comparativamente entre si (o objetivo com eles não é considerá-los isoladamente a princípio), temos reafirmado o *continuum* que vimos apontado pela literatura no qual as manifestações de alterações linguísticas aumentam na direção da DA e são mais brandas, mas já notáveis, no CCL, enquanto indivíduos do grupo-controle apresentam a menor recorrência entre os fenômenos de hesitação. Assim sendo, vemos os dados apontando maior incidência de preenchedores em DA, menor em CCL e menor que ambos no

grupo-controle. Paralelo a isso, temos o erro padrão indicando a maior variabilidade dessa métrica em DA e grupo-controle a menor, com CCL figurando entre os outros dois grupos.

Submetendo os dados de média de cada indivíduo ao teste não paramétrico U de Mann-Whitney, confirma-se a possibilidade de se averiguar a distinção gradativa entre os grupos, uma vez que o valor- p obtido para a comparação dos grupos, dois a dois, foi menor que 0.001. Assim, rejeitando a hipótese nula, isto é, a ausência de correlação entre a progressão da manifestação de preenchedores na narrativa no *continuum* estabelecido entre os grupos. Os resultados gerados foram os presentes na Tabela 4.12.

Grupos (dois a dois)	Mann-Whitney	valor- p
DA e controle	17956.0	< 0.001
DA e CCL	4096.0	< 0.001
CCL e controle	17956.0	< 0.001

Tabela 4.12: Resultados do teste U de Mann-Whitney e valor- p ao se submeter as proporções de preenchedores sobre tokens na narrativa de cada indivíduo a cada dois grupos.

4.7.2 Repetições completas

Usando metodologia similar à análise de incidência de preenchedores, nesta subseção, apresentamos, por meio de médias entre os grupos, a verificação comparativa da incidência de tokens repetidos por completo, seja uma palavra inteira, seja casos em que uma palavra incompleta foi repetida da mesma maneira.

Na Tabela 4.13 e na Figura 4.16, apresentam-se os resultados dessas médias acompanhados do erro padrão. Dessa vez, o grupo que fica à frente na métrica é o CCL, apresentando a maior incidência média de repetições entre os três grupos. Ainda assim, o grupo-controle é o menos produtivo com relação a palavras repetidas. Vale notar também como os demais grupos apresentaram, cada um, cerca de 2,5 vezes mais variabilidade em relação ao grupo-controle ao se comparar o erro padrão entre eles. Isso aponta uma maior tendência, nos grupos clínicos, de produzir repetições completas em relação a indivíduos

saudáveis, além de haver maior variabilidade entre os indivíduos de ambos os grupos.

Grupo	Proporção média	Erro padrão
DA	0.0135	0.0026
Controle	0.0092	0.0010
CCL	0.0185	0.0027

Tabela 4.13: Médias e erros-padrão das razões entre o número de repetições completas sequenciais e o número total de tokens por narrativa para cada grupo.

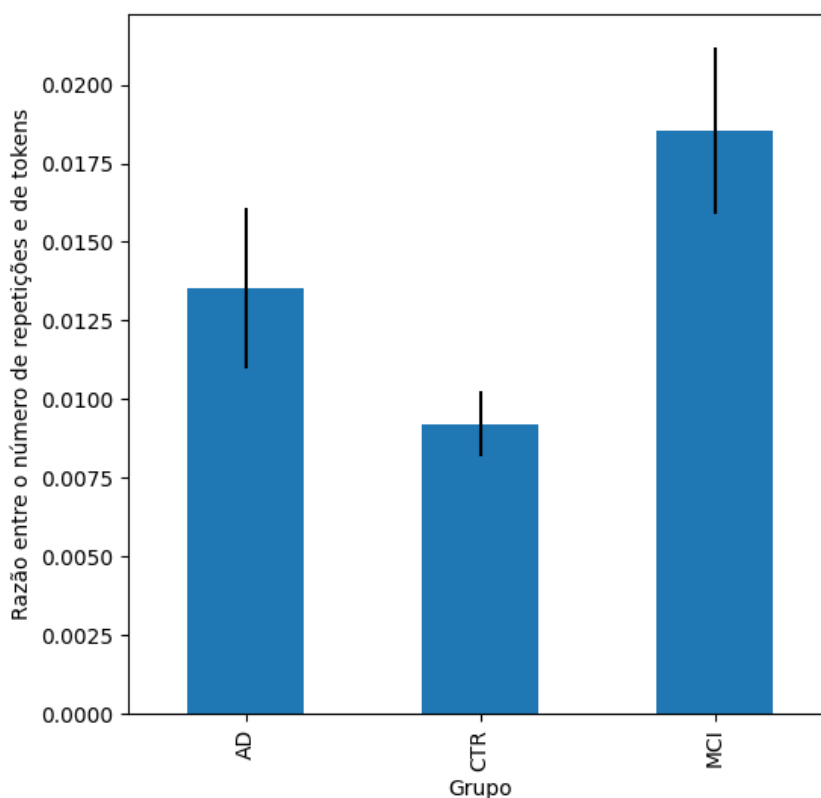


Figura 4.16: Comparação entre médias e erros-padrão das razões do número repetições completas sobre o número total de tokens por narrativa para cada grupo.

Na Tabela 4.14, a partir do teste estatístico de Mann-Whitney, indica-se uma diferença estatisticamente significativa na recorrência de alterações linguísticas entre cada par de grupo, com valores-p inferiores a 0.001. Essa possibilidade de relevância estatística já havia sido antecipada em inspeção visual pela baixa concorrência entre os erros-padrão de cada grupo na Figura 4.16.

Grupos (dois a dois)	Mann-Whitney	valor- <i>p</i>
DA e controle	17956.0	< 0.001
DA e CCL	4096.0	< 0.001
CCL e controle	17956.0	< 0.001

Tabela 4.14: Resultados do teste U de Mann-Whitney e valor-*p* ao se submeter as proporções de preenchedores sobre tokens na narrativa de cada indivíduo a cada dois grupos.

4.7.3 Repetições incompletas

Quanto às repetições incompletas, isto é, ocorrências de gaguejamentos e falsos inícios, ao se aplicar os testes que vimos aplicando para os outros dois fenômenos, encontramos uma maior distinção do grupo DA em relação aos demais, apresentando uma proporção de incidência do fenômeno em relação aos tokens cerca de 1,8 vezes maior, como presente na Tabela 4.15. O erro padrão desse grupo também é o mais expressivo apresentando maior variabilidade entre os indivíduos.

Dessa vez, no entanto, o CCL apresentou tanto a proporção média quanto o erro padrão bastante semelhantes ao controle se comparados ao DA; o que é graficamente perceptível pela Figura 4.17. Apesar disso, o grupo-controle continua apresentando o menor erro padrão, ou seja, menor variabilidade de incidência entre um indivíduo e o outro no mesmo grupo, ainda assim, tendendo a manter a viabilidade da interpretação da incidência do fenômeno como proporcional aos grupos de análise.

Grupo	Proporção média	Erro padrão
DA	0.0033	0.0012
Controle	0.0018	0.0004
CCL	0.0018	0.0006

Tabela 4.15: Médias e erros-padrão das razões entre o número de repetições incompletas sequenciais e o número total de tokens por narrativa para cada grupo.

Inspecionando métricas descritivas dos dados a fim de entender se havia algum aspecto dos dados que estivesse causando esse resultado menos marcado, notamos uma frequência alta de narrativas em que não havia nenhuma manifestação de repetições incompletas, contando 115 no total, enquanto as narrativas em que pelo menos uma repetição incompleta ocorreu foram apenas 51 — essas

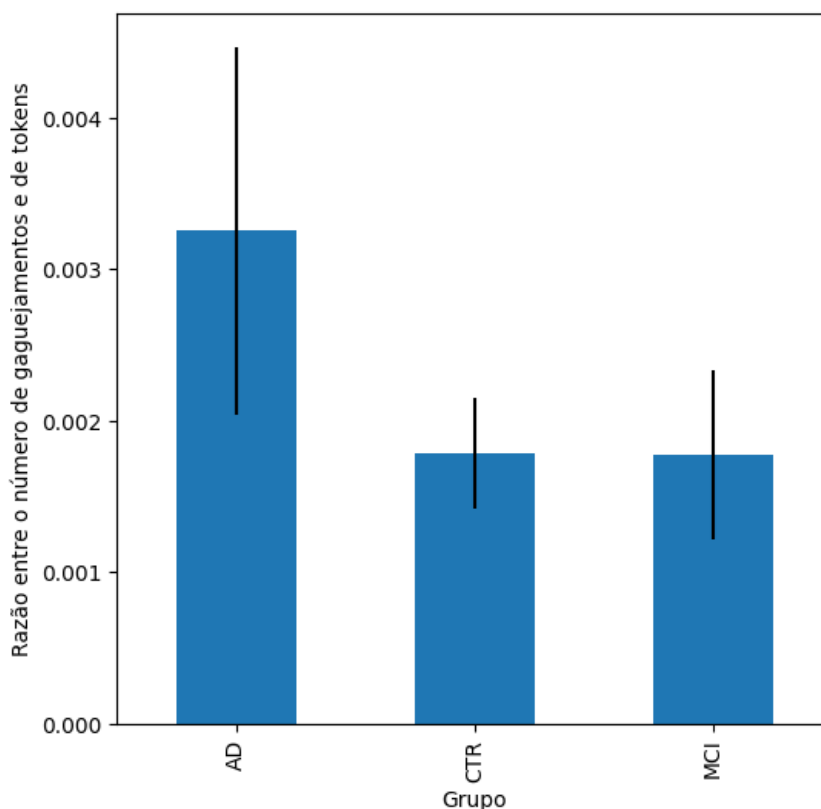


Figura 4.17: Comparação entre médias e erros-padrão das razões do número repetições incompletas sobre o número total de tokens por narrativa para cada grupo.

quantidades estão descritas para cada grupo na Tabela 4.16. Com isso em vista, optamos por executar um teste de qui-quadrado para avaliar a associação ou independência entre as duas variáveis: narrativas em que aparecem uma ou mais repetições incompletas em oposição às que não apresentam nenhuma.

Grupo	Narrativas sem rep. inc.	Narrativas com rep. inc.
DA	19	13
Controle	75	27
CCL	21	11

Tabela 4.16: Quantidade de narrativas em que não apareceram repetições incompletas e de narrativas em que apareceram uma ou mais repetições incompletas em cada grupo.

Os valores-p associados a cada estatística de qui-quadrado são maiores do que o nível de significância comumente adotado de 0.05, como constam na Tabela 4.17. Portanto, não podendo, a princípio, tomar a estatística dessa natureza de fenômeno linguístico como suficiente para rejeitar a hipótese nula de que não há diferença significativa entre as distribuições.

Grupo	χ^2	Valor-p
DA e controle	1.7036	0.1918
CCL e controle	0.4106	0.5217
DA e CCL	0.0667	0.7963

Tabela 4.17: Teste de qui-quadrado para repetições incompletas, considerando a divisão em cada grupo de narrativas sem presença do fenômeno em oposição àquelas em que ele ocorre uma ou mais vezes.

Por outro lado, dois fatores são passíveis de não nos ter garantido uma boa confiabilidade de resultados: primeiramente, a quantidade de dados gerais — houvessem mais narrativas, poderíamos ter resultados mais robustos para confirmar ou refutar nossa hipótese com maior assertividade —; e, principalmente, o provável fato de a metodologia de transcrição dos áudios dos pesquisadores que a executaram não ser orientada à anotação minuciosa de expressões como palavras incompletas, possibilitando até mesmo a inconsistência de um mesmo transcritor ou transcritores diferentes. Portanto, podendo ter, de fato, havido mais casos em que os indivíduos manifestaram esse fenômeno linguístico e que não foram registrados, dada a possibilidade de a concepção do protocolo não ter antevisto esse tipo de dado consistentemente.

4.7.4 Repetições completas e incompletas

Revisitando os dados e análises de incidência de repetições completas e incompletas, optamos por unir ambos os fenômenos de realização linguística semelhante num último agrupamento de análise, sobretudo devido à baixa incidência registrada nas transcrições das repetições incompletas, assim, poderemos ter uma forma de averiguar se os dados se complementam ou se os poucos dados destas atrapalham a maior direcionalidade na interpretação daquela. Os resultados da proporção média e erro padrão por grupo são apresentados na Tabela 4.18.

Como se pode notar pelos números e pela Figura 4.18, a tendência de resultados foi bastante semelhante à dos números somente com repetições completas. A ordem de incidência entre os grupos volta a ter o CCL como o mais produtivo dos fenômenos seguido por DA e controle como o menos produtivo; o que era de se esperar dada a maior quantidade de dados nas repetições completas, por-

Grupo	Proporção média	Erro padrão
DA	0.0168	0.0032
Controle	0.0110	0.0012
CCL	0.0203	0.0028

Tabela 4.18: Médias e erros-padrão das razões entre o número de repetições totais (completas e incompletas) sequenciais e o número total de tokens por narrativa para cada grupo.

tanto, influenciando a diferença entre um grupo e outro ser proporcionalmente semelhante.

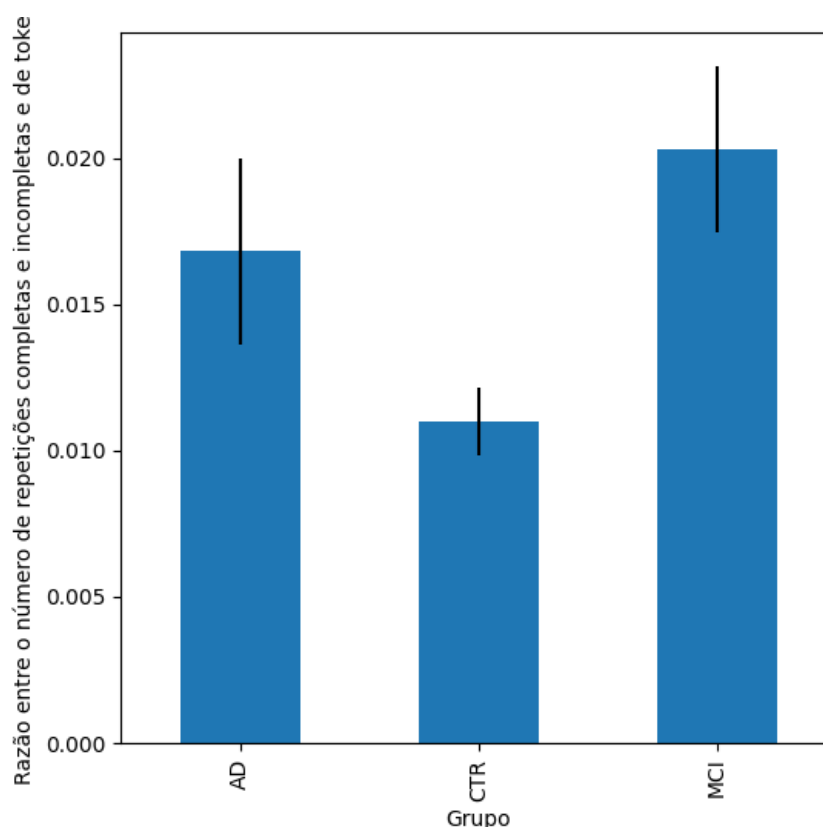


Figura 4.18: Comparação entre médias e erros-padrão das razões do número repetições totais (completas e incompletas) sobre o número total de tokens por narrativa para cada grupo.

Quanto à variabilidade entre os indivíduos de cada grupo, em comparação com o que obtivemos na análise das repetições completas, CCL aumenta um pouco o erro padrão (de 0,0026 para 0,0028) e DA aumenta proporcionalmente mais (de 0,0025 para 0,0032), o que faz com que DA volte a ser o grupo com maior variabilidade entre os três (como acontece com os preenchedores). Isso ajuda a fundamentar nossa sugestão de que há a necessidade de realizar um

teste com mais dados e/ou com uma garantia de que uma transcrição mais robusta com relação à anotação de tokens incompletos.

Em linha com o comentário de esses resultados se assemelharem ao caso dos dados a partir das repetições completas, aqui, notamos que a diferenciação entre CCL e grupo-controle voltar a ser estatisticamente relevante como visto no valor- p do teste de qui-quadrado na Tabela 4.19; diferenciação que nota visualmente pela distância entre ambos os grupos na Figura 4.18.

Grupo	χ^2	Valor- p
DA e controle	0.8257	0.3635
CCL e controle	3.8613	0.0494
DA e CCL	0.4391	0.5076

Tabela 4.19: Teste de qui-quadrado para repetições completas e incompletas.

4.8 Discussão de análises de fenômenos linguísticos

Ao longo da seção, discutimos a análise de fenômenos linguísticos específicos de hesitação baseados na hipótese de que a presença ou frequência deles se daria em um *crescendo* na seguinte ordem de coortes analisadas: grupo-controle, CCL e DA. A seguir, apresentam-se esquematicamente nossas descobertas para cada fenômeno e comentários sobre possíveis tendências da ocorrência desses fenômenos como um facilitador para a identificação de algum distúrbio neurodegenerativo em desenvolvimento ou em curso intensificado. Essas impressões não substituem em nenhuma hipótese a necessidade de um diagnóstico clínico, mas podem ser um recurso para auxiliar um indivíduo a ser conduzido à clínica, caso surja a suspeita de algum desses distúrbios.

- Preenchedores:
 - Os resultados indicam uma maior incidência de preenchedores em indivíduos com DA, seguido por CCL e controle.
 - A análise estatística (teste U de Mann-Whitney) confirma diferenças significativas entre os grupos, reforçando a progressão no desenvolvimento de distúrbios linguísticos observada na literatura.

- Repetições completas:
 - O grupo CCL apresentou a maior incidência média de repetições completas, seguido por DA e controle.
 - Testes estatísticos confirmaram diferenças significativas entre os grupos, destacando a relevância desse fenômeno como um marcador potencial.
- Repetições incompletas:
 - O grupo DA mostrou uma incidência maior desses fenômenos em comparação com CCL e controle.
 - CCL e controle praticamente não se distinguiram, a não ser pelo erro padrão maior para o primeiro.
 - Uma análise de qui-quadrado para narrativas com ou sem repetições incompletas indicou resultados não significativos, sugerindo a necessidade de uma análise mais aprofundada ou, talvez, uma possível revisão na metodologia de transcrição.
- Repetições completas e incompletas:
 - Ao combinar repetições completas e incompletas, os resultados mostraram uma tendência semelhante àquela observada apenas com repetições completas, provavelmente devido à maior quantidade de dados, em relação aos dados isolados de cada tipo de repetição.
 - A diferenciação entre o grupo CCL e o controle tornou-se estatisticamente relevante, indicando a importância de considerar ambos os tipos de repetições na análise.
- Considerações gerais:
 - A quantidade de dados e o objetivo direcionado a produções linguísticas na metodologia de transcrição diferentes das que analisamos são limitações importantes. Com essa maior robustez no *corpus*, em futuras pesquisas, os resultados menos expressivos e que aqui se mostram, em geral, como tendências, podem ser ratificados.

- Os resultados reforçam a utilidade dos fenômenos linguísticos analisados como indicadores potenciais de alterações cognitivas, até mesmo em estágios iniciais.

Em conclusão, as análises dos fenômenos linguísticos fornecem pistas valiosas sobre as diferenças entre os grupos estudados, indicando a relevância desses marcadores na identificação clínica de alterações cognitivas. No entanto, fazemos ressalvas quanto à robustez dos resultados, enfatizando a necessidade de abordagens mais aprofundadas e aprimoramento metodológico em pesquisas futuras, sobretudo na transcrição de dados. Esses achados contribuem para a compreensão da linguagem em contextos clínicos e podem ter implicações e implementações significativas no diagnóstico precoce e no acompanhamento de condições como Alzheimer e Comprometimento Cognitivo Leve.

Conclusão

Neste capítulo, apresentamos o que acreditamos ser nossas potenciais contribuições para o entendimento do impacto da DA e do CCL na produção linguística. Começamos por retomar o Capítulo 3, em que foram detalhados os procedimentos de análise dos dados, que incluíram a análise e processamento de narrativas dos três grupos de análise (DA, CCL e grupo-controle), bem como a aplicação de classificadores textuais de diferentes naturezas. Recapitulamos os resultados, apresentados no Capítulo 4, tanto das análises dos classificadores automáticos quanto das manifestações de fenômenos linguísticos relacionados à hesitação, que indicam alterações na produção linguística associadas a tais condições neurológicas.

A seguir, relatamos as contribuições mais imediatas que são fruto direto da nossa análise, além da relevância da metodologia de análise que trouxemos. No entanto, limitações foram reconhecidas, como o tamanho reduzido dos *corpora* e o desbalanceamento dos dados, impactando a robustez de alguns resultados dos modelos — os quais usualmente são treinados em quantidade vasta de dados. Por fim, trazemos uma série de propostas que podem gerar novas pesquisas com direções advindas das nossas contribuições, como a coleta recorrente de *corpora* especializados, a exploração de modelos e/ou parametrizações alternativas e a aplicação, análoga aos nossos métodos, de análises em dados acústicos. Essas sugestões visam a aprimorar a compreensão das alterações linguísticas associadas a condições cognitivas e contribuir para o desenvolvimento de ferramentas diagnósticas mais eficazes.

5.1 Breve sumarização dos resultados

No Capítulo 3, apresentamos os procedimentos adotados para análise dos dados utilizados na pesquisa. Os dados estão divididos em três grupos de indivíduos: um grupo com diagnóstico de Doença de Alzheimer (DA), um grupo com diagnóstico de Comprometimento Cognitivo Leve (CCL) e um grupo-controle, composto por indivíduos sem diagnóstico de DA ou CCL. Os participantes foram convidados a produzir narrativas a partir de estímulos visuais e essas narrativas foram analisadas com base em diferentes critérios, como a média de palavras por sentença, a riqueza vocabular e a taxa de *hapax legomena*. Para fins de Processamento de Linguagem Natural, executamos classificadores textuais de quatro diferentes naturezas (bayesiano, redes uni- e bidirecionais e um modelo *transformer*). Além disso, foram identificadas manifestações de fenômenos de hesitação nas narrativas, como pausas e diferentes tipos de repetições (repetições hesitativas, falsos inícios e gaguejamentos).

No Capítulo 4, apresentamos os resultados obtidos a partir da análise dos dados coletados de acordo com as abordagens supramencionadas. Os resultados indicam que os grupos DA e CCL apresentaram desempenho inferior em relação ao grupo controle em diversos critérios de análise, como a média de palavras por sentença e a riqueza vocabular. Além disso, apresentamos uma análise do desempenho dos modelos utilizados na pesquisa: os já mencionados modelo bayesiano, rede neural de propagação para frente, rede BiLSTM (bidirecional) e o modelo *transformers*. Por fim, foram identificadas diferenças significativas na ocorrência de fenômenos de hesitação entre os grupos.

Com base nesses achados, é possível concluir que a Doença de Alzheimer e o Comprometimento Cognitivo Leve afetam significativamente a produção linguística dos indivíduos, resultando em narrativas com menor complexidade e maior ocorrência de fenômenos de hesitação. Além disso, os modelos utilizados na pesquisa apresentaram desempenho variado na classificação das narrativas, indicando a necessidade de aprimoramento desses modelos. Em relação à literatura, destacamos a importância de estudos que continuem a buscar compreender como as condições neurológicas afetam a linguagem e como é possível

classificar essas narrativas, contribuindo para o desenvolvimento de estratégias de diagnóstico e intervenção mais eficazes e ágeis.

5.2 Contribuições principais

Nossa investigação apresenta resultados que sustentam padrões identificados na literatura, conforme evidenciado por estudos anteriores, como os de Abrisqueta-Gomez et al. (2004) e Nitrini, Caramelli et al. (2005). Esses autores destacam a crescente incidência de problemas linguísticos à medida que as fases de desenvolvimento da DA (Frota et al. 2011) progridem. Esses problemas linguísticos inicialmente se manifestam de maneira sutil na fase pré-clínica e se tornam mais proeminentes à medida que passa a haver uma fase de demência instaurada.

Além dos achados empíricos, outra contribuição de nossa pesquisa é relativa à metodologia. Estabelecemos o uso, com base em treinamento, de quatro diferentes classificadores textuais que avaliam se um dado texto transcrito pode ou não ter sido proferido por uma pessoa com DA ou CCL. Dos quatro classificadores testados, um deles (a rede BiLSTM) não obteve resultados conclusivos com base na implementação selecionada; entretanto, os outros três (bayesiano, rede de propagação para frente e DistilBERT) apresentaram um desempenho satisfatório para a classificação. A limitação percebida nos resultados da BiLSTM não está, portanto, ligada à análise dos dados, mas ao próprio uso dessa rede, com suas características próprias, no tratamento dos dados. Isso se deve, mais provavelmente, ao tamanho do conjunto de dados, característica à qual as redes recorrentes (como a BiLSTM) são mais sensíveis do que os classificadores mais simples.

Os melhores resultados orbitam em torno da distinção entre DA e grupo-controle — justamente os que deveriam de fato serem os mais distinguíveis em termos de intensidade da recorrência de alterações linguísticas atípicas. No caso do classificador bayesiano, esses grupos apresentaram as melhores métricas (acurácia acima de 75%) tanto no *corpus* balanceado, com cobertura e medida F1, como no desbalanceado, com acurácia e precisão (Tabela 4.2), patamar que nenhuma das outras combinações alcançou.

Na rede de propagação para frente, para cada uma das três diferentes configurações propostas, o par DA–Controle sempre atinge as marcas mais altas com a validação comparativamente aos demais confrontos (Figuras 4.3 a 4.13). Com a última iniciativa de classificador abordada, não foi diferente: na Tabela 4.10, vemos como o modelo DistilBERT lida melhor com a distinção entre DA e grupo-controle, apresentando acurácia de até 0.78 (inclusive com o menor desvio padrão), enquanto os outros pares apresentaram acurácia entre 0.69 e 0.72.

Dando um passo em paralelo ao da classificação automática, observamos também a produção de fenômenos linguísticos de hesitação nos grupos em foco. Identificamos diferenças significativas na ocorrência de preenchedores, além de repetições completas e incompletas entre os grupos, indicando uma progressão nos distúrbios linguísticos conforme documentado na literatura. No caso do primeiro e do último fenômeno, DA esteve à frente da produção; e as repetições completas foram lideradas pelo CCL. É relevante notar como os grupos-controle se mostram como os menos produtivos nesses processos de hesitação.

A constante distinção do grupo-controle em relação aos grupos clínicos nos ajuda a corroborar nossa hipótese: se conseguirmos classificar não só a DA mas também o CCL em oposição aos indivíduos-controle, poderemos, utilizar os métodos que apresentamos neste trabalho, ao lado de outras ferramentas já usuais, para auxiliar no diagnóstico precoce de condições neurodegenerativas como essas.

Essa abordagem metodológica proporciona uma base sólida para análise, promovendo a replicabilidade das investigações subsequentes e viabilizando a reprodutibilidade dos resultados ao se submeter mais dados de natureza semelhante à daqueles com que trabalhamos.

5.3 Posição dos achados frente à literatura

Ressaltamos aqui alguns apontamentos que absorvemos da literatura e que inicialmente incitaram nossa motivação para conduzir determinados testes

e que podemos chegar a resultados que corroboram o que outros autores já haviam proposto.

Karlekar, Niu e Bansal (2018) apontam como indivíduos com DA tendem a empregar mais preenchedores (como “uh” e “um”) que indivíduos saudáveis. Na nossa análise de elementos de hesitação, notamos justamente que preenchedores são mais bem frequentes no discurso desse grupo, seguido por CCL e, em menor incidência, pelo grupo-controle (Figura 4.15)

É consenso entre diversos autores, como Steiner et al. (2017), Abrisqueta-Gomez et al. (2004), Frota et al. (2011) e Nitrini, Brucki et al. (2021), que se deve enfatizar a necessidade de uma detecção precoce de um estágio de desenvolvimento DA. Em nossos resultados, pudemos atestar alguns casos em que, mesmo com menos distinção do que DA, o grupo CCL já se diferenciava do grupo-controle suficientemente, como, por exemplo, no registro do uso de preenchedores (Figura 4.15). Houve, também, o caso das repetições completas em que o CCL, na realidade, foi o mais incidente de todos os grupos, mais que dobrando a métrica em relação ao grupo-controle (Figura 4.16). Porém, mesmo nos casos de classificação em que o par entre grupo-controle e CCL teve o menor desempenho, a acurácia não foi baixa, atingindo 0.7 (Tabela 4.10).

Através dos resultados do classificador que mais bem se ajustou aos dados (DistilBERT), observamos uma clareza crescente na distinção na produção linguístico-discursiva que segue um contínuo entre os grupos na ordem Controle–CCL–DA. Em outras palavras, o classificador teve maior dificuldade, em primeiro lugar, em distinguir pessoas sem comprometimento cognitivo daquelas com comprometimento leve, e, em seguida, distinguir entre pessoas com diferentes graus de comprometimento (CCL e DA). Confirmando esse resultado, a distinção que se mostra mais fácil é entre pessoas com grau mais elevado de comprometimento (DA) e as do grupo controle.

Essas observações sugerem uma maior dificuldade de diferenciação diagnóstica nos estágios iniciais de comprometimento, quando as manifestações do CCL começam a surgir no discurso. No entanto, é preciso que se acrescente a ressalva de que as diferenças encontradas entre os grupos são pequenas, em parte em função da quantidade e da qualidade das transcrições disponíveis.

Acima de tudo, é necessário aprofundar a discussão com os especialistas em diagnóstico das áreas de saúde.

5.4 Limites desta pesquisa

Uma das questões que desafiavam a consistência das ferramentas e consequente reprodutibilidade de resultados é a composição dos *corpora*, uma vez que frequentemente eles contam com uma quantidade de dados pequena. Por vezes, o número de sujeitos gira em torno de uma centena de indivíduos ou menos, como ocorre com os DNLT-BP e que também é relatada na literatura, como no caso do *corpus* espanhol de Peraita e Grasso (2010). Sabemos que bons modelos de classificação, particularmente os de aprendizado de máquina, costumam precisar de uma quantidade mais extensa de dados para garantir o bom ajuste aos dados.

O não balanceamento dos *corpora*, tanto em número de participantes por grupo quanto em quantidade de informação produzida por cada participante, também pode prejudicar resultados, dependendo do modelo empregado, enviesando-os e minimizando recursos para decisões.

A quantidade de dados tanto é um fator sensível que observamos, durante nosso movimento de balancear os dados reduzindo a quantidade de sentenças do grupo-controle, que, em vez de melhorar a distinção entre grupos, o que aconteceu foi a piora das medidas de acurácia e aumento da perda na rede de propagação para frente.

Além disso, os participantes podem não constituir um recorte sociodemográfico que represente fidedignamente a realidade de determinada localidade ou de grupos mais abrangentes, o que tende a gerar resultados com algum grau de enviesamento, prejudicando, assim, a confiabilidade das aplicações. O fato também é apontado por Hernández-Domínguez et al. (2016, p. 13). Não dispomos de informações suficientes a respeito dessas variáveis.

Por fim, um último ponto ainda relacionado à confecção dos *corpora* que utilizamos é a metodologia de transcrição. O objetivo com que os transcritores devem ter executado sua tarefa não tinha como foco principal a produção

linguística fidedigna, algo como uma transcrição próxima à produção fonológica. Podemos notar algumas inconsistências na transcrição que não causam nenhum ruído numa eventual análise do conteúdo que os indivíduos estavam narrando, porém, no caso do nosso processamento automático, é importante, por exemplo, uma consistência na escrita de um determinado token de um mesmo preenchedor ou na transcrição assídua de palavras incompletas.

5.5 Pesquisas futuras

Para futuros estudos e aplicabilidades que nosso trabalho possa vir a suscitar, destacamos alguns pontos de sugestão:

- A coleta de *corpora* especializados numa mesma tarefa (como as tarefas de narração de que fizemos uso) para reproduzir as análises. Na possibilidade de compilar um *corpus* especificamente para o tópico, a instrução para se transcrever toda e qualquer manifestação linguística completa ou não é importante, pois, como vimos, realizações hesitativas reincidentes podem trazer indícios de que pode se tratar de um caso como CCL e DA.
- Do ponto de vista técnico dos modelos que executamos, haverá sempre mais formas implementá-los, reparametrizá-los, treiná-los, adicionar bases dados. Portanto, além de as nossas quatro propostas poderem ser cada vez mais aperfeiçoadas, outras redes e modelos, que lidem bem com sequências e a tarefa de classificação, podem ser testados com os dados que usamos.
- Nesta pesquisa, trabalhamos com análises comparativas, sempre estabelecendo uma comparação entre o desempenho de um modelo frente a outro, bem como entre um grupo de análise frente a outro. Uma tentativa que pode ser factível é o estabelecimento de limiares indicativos (Karlekar, Niu e Bansal 2018) para a classificação não comparativa de um grupo ou mesmo estágio de desenvolvimento da DA — para que essa métrica seja confiável, novamente, reforça-se uma grande quantidade de dados especializados como crucial para treinamento.

- Aqui, trabalhamos com o texto transcrito; no entanto, paralelamente, com o avanço das ferramentas de análise automática de áudio, uma metodologia semelhante poderia ser aplicada a esse canal a fim de verificar se os resultados obtidos também corroboram as tendências da literatura.
- Como devolutiva para a sociedade, a metodologia por trás de nossa pesquisa pode compor algum tipo de aplicação que conte com transcrição automática (incluindo palavras incompletas) para ajudar indivíduos, pacientes ou profissionais da saúde numa espécie de triagem como uma nova ferramenta para ajudar na decisão do encaminhamento para um diagnóstico especializado.

Referências

- Abrisqueta-Gomez, Jacqueline et al. (set. de 2004). “A longitudinal study of a neuropsychological rehabilitation program in Alzheimer’s disease”. Em: *Arquivos de Neuro-Psiquiatria* 62.3b, pp. 778–783. ISSN: 0004-282X. DOI: [10.1590/S0004-282X2004000500007](https://doi.org/10.1590/S0004-282X2004000500007). URL: <https://doi.org/10.1590/S0004-282X2004000500007>.
- Beltrami, Daniela et al. (2018). “Speech Analysis by Natural Language Processing Techniques: A Possible Tool for Very Early Detection of Cognitive Decline?” Em: *Frontiers in Aging Neuroscience* 10. ISSN: 1663-4365. DOI: [10.3389/fnagi.2018.00369](https://doi.org/10.3389/fnagi.2018.00369). URL: <https://www.frontiersin.org/article/10.3389/fnagi.2018.00369>.
- Boyé, Maïté, Thi Tran e Natalia Grabar (set. de 2014). “NLP-Oriented Contrastive Study of Linguistic Productions of Alzheimer’s and Control People”. Em: pp. 412–424. ISBN: 978-3-319-10887-2. DOI: [10.1007/978-3-319-10888-9_41](https://doi.org/10.1007/978-3-319-10888-9_41).
- Casanova, Edresson et al. (mai. de 2020). “Evaluating Sentence Segmentation in Different Datasets of Neuropsychological Language Tests in Brazilian Portuguese”. English. Em: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 2605–2614. ISBN: 979-10-95546-34-4. URL: <https://aclanthology.org/2020.lrec-1.317>.
- Cui, Zhiyong et al. (2020). “Stacked Bidirectional and Unidirectional LSTM Recurrent Neural Network for Forecasting Network-wide Traffic State with Missing Values”. Em: *CoRR* abs/2005.11627. arXiv: [2005.11627](https://arxiv.org/abs/2005.11627). URL: <https://arxiv.org/abs/2005.11627>.
- Devlin, Jacob et al. (2018). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. Em: *CoRR* abs/1810.04805. arXiv: [1810.04805](https://arxiv.org/abs/1810.04805). URL: <http://arxiv.org/abs/1810.04805>.

- Dreisbach, Caitlin et al. (2019). “A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data”. Em: *International Journal of Medical Informatics* 125, pp. 37–46. ISSN: 1386-5056. DOI: <https://doi.org/10.1016/j.ijmedinf.2019.02.008>. URL: <https://www.sciencedirect.com/science/article/pii/S1386505618313789>.
- Ferreira, M. e M. Lopes (2019). *Para conhecer: Linguística Computacional*. Editora Contexto. ISBN: 9788552001522.
- Frota, Norberto Anízio Ferreira et al. (jul. de 2011). “Criteria for the diagnosis of Alzheimer’s disease: Recommendations of the Scientific Department of Cognitive Neurology and Aging of the Brazilian Academy of Neurology”. Em: *Dementia & Neuropsychologia* 5.3, pp. 146–152. ISSN: 1980-5764. DOI: [10.1590/S1980-57642011DN05030002](https://doi.org/10.1590/S1980-57642011DN05030002). URL: <https://doi.org/10.1590/S1980-57642011DN05030002>.
- Hernández-Domínguez, Laura et al. (ago. de 2016). “Detection of Alzheimer’s disease based on automatic analysis of common objects descriptions”. Em: *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*. Berlin: Association for Computational Linguistics, pp. 10–15. DOI: [10.18653/v1/W16-1902](https://doi.org/10.18653/v1/W16-1902). URL: <https://aclanthology.org/W16-1902>.
- Karlekar, Sweta, Tong Niu e Mohit Bansal (jun. de 2018). “Detecting Linguistic Characteristics of Alzheimer’s Dementia by Interpreting Neural Models”. Em: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 701–707. DOI: [10.18653/v1/N18-2110](https://doi.org/10.18653/v1/N18-2110). URL: <https://aclanthology.org/N18-2110>.
- Koleck, Theresa A et al. (fev. de 2019). “Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review”. Em: *Journal of the American Medical Informatics Association* 26.4, pp. 364–379. ISSN: 1527-974X. DOI: [10.1093/jamia/ocy173](https://doi.org/10.1093/jamia/ocy173). eprint: <https://academic.oup.com/jamia/article-pdf/26/4/364/34151341/ocy173.pdf>. URL: <https://doi.org/10.1093/jamia/ocy173>.
- Kumar, Sayantan et al. (jul. de 2021). “Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review”. Em: *JAMIA Open* 4.3. ISSN: 2574-2531. DOI: [10.1093/jamiaopen/ooab052](https://doi.org/10.1093/jamiaopen/ooab052). URL: <http://dx.doi.org/10.1093/jamiaopen/ooab052>.

- Marcuschi, Luiz Antônio (1999). “Gramática do Português Falado: Volume VII: Novos Estudos”. Em: ed. por Maria Helena de Moura Neves. São Paulo: Humanitas/FFLCH/USP; Campinas: Editora da Unicamp. Cap. A hesitação, pp. 159–194.
- (2003). *Análise da Conversação*. Ática.
- Negash, Selam et al. (2007). “Effects of ApoE genotype and mild cognitive impairment on implicit learning”. Em: *Neurobiology of Aging* 28.6, pp. 885–893. ISSN: 0197-4580. DOI: <https://doi.org/10.1016/j.neurobiolaging.2006.04.004>. URL: <https://www.sciencedirect.com/science/article/pii/S0197458006001187>.
- Nitrini, Ricardo, Sonia Maria Dozzi Brucki et al. (abr. de 2021). “The Figure Memory Test: diagnosis of memory impairment in populations with heterogeneous educational background”. Em: *Dementia & Neuropsychologia* 15.2, pp. 173–185. ISSN: 1980-5764. DOI: [10.1590/1980-57642021dn15-020004](https://doi.org/10.1590/1980-57642021dn15-020004). URL: <https://doi.org/10.1590/1980-57642021dn15-020004>.
- Nitrini, Ricardo, Paulo Caramelli et al. (set. de 2005). “Diagnóstico de doença de Alzheimer no Brasil: avaliação cognitiva e funcional. Recomendações do Departamento Científico de Neurologia Cognitiva e do Envelhecimento da Academia Brasileira de Neurologia”. Em: *Arquivos de Neuro-Psiquiatria* 63.3a, pp. 720–727. ISSN: 0004-282X. DOI: [10.1590/S0004-282X2005000400034](https://doi.org/10.1590/S0004-282X2005000400034). URL: <https://doi.org/10.1590/S0004-282X2005000400034>.
- Patestas, Maria A. e Leslie P. Gartner (2006). *A Textbook of Neuroanatomy*. Blackwell Publishing.
- Peraita, Herminia e Lina Grasso (jan. de 2010). “Corpus lingüístico de definiciones de categorías semánticas de sujetos ancianos sanos y con la enfermedad de Alzheimer: Una investigación transcultural hispano-argentina”. Em: *Documentos de trabajo (Fundación BBVA)*, N.º 3, 2010.
- Santos, Leandro et al. (jul. de 2017). “Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts”. Em: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1284–1296. DOI: [10.18653/v1/P17-1118](https://aclanthology.org/P17-1118). URL: <https://aclanthology.org/P17-1118>.
- Savonenko, Alena V. et al. (2015). “Neurobiology of Brain Disorders: Biological Basis of Neurological and Psychiatric Disorders”. Em: ed. por Michael J. Zigmond, Joseph T. Coyle e Lewis P. Rowland. Elsevier. Cap. Alzheimer Disease, pp. 321–338.

- Steiner, Ana Beatriz Quintes et al. (jul. de 2017). “Mild cognitive impairment and progression to dementia of Alzheimer’s disease”. Em: *Revista da Associação Médica Brasileira* 63.7, pp. 651–655. ISSN: 0104-4230. DOI: [10.1590/1806-9282.63.07.651](https://doi.org/10.1590/1806-9282.63.07.651). URL: <https://doi.org/10.1590/1806-9282.63.07.651>.
- Teixeira, Camila Vieira Ligo et al. (2012). “Non-pharmacological interventions on cognitive functions in older people with mild cognitive impairment (MCI)”. Em: *Archives of Gerontology and Geriatrics* 54.1, pp. 175–180. ISSN: 0167-4943. DOI: <https://doi.org/10.1016/j.archger.2011.02.014>. URL: <https://www.sciencedirect.com/science/article/pii/S0167494311000409>.
- Toledo, Cíntia Matsuda et al. (2018). “Analysis of macrolinguistic aspects of narratives from individuals with Alzheimer’s disease, mild cognitive impairment, and no cognitive impairment”. Em: *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring* 10.1, pp. 31–40. DOI: <https://doi.org/10.1016/j.dadm.2017.08.005>. eprint: <https://alz-journals.onlinelibrary.wiley.com/doi/pdf/10.1016/j.dadm.2017.08.005>. URL: <https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1016/j.dadm.2017.08.005>.
- Vincze, Veronika et al. (ago. de 2016). “Detecting Mild Cognitive Impairment by Exploiting Linguistic Information from Transcripts”. Em: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 181–187. DOI: [10.18653/v1/P16-2030](https://doi.org/10.18653/v1/P16-2030). URL: <https://aclanthology.org/P16-2030>.
- Wu, Stephen et al. (dez. de 2019). “Deep learning in clinical natural language processing: a methodical review”. Em: *Journal of the American Medical Informatics Association* 27.3, pp. 457–470. ISSN: 1527-974X. DOI: [10.1093/jamia/ocz200](https://doi.org/10.1093/jamia/ocz200). eprint: <https://academic.oup.com/jamia/article-pdf/27/3/457/34152802/ocz200.pdf>. URL: <https://doi.org/10.1093/jamia/ocz200>.