UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Representação de narrativas e extração de suas unidades de informação para automatização de testes neuropsicológicos

Leandro Borges dos Santos

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Leandro Borges dos Santos

Representação de narrativas e extração de suas unidades de informação para automatização de testes neuropsicológicos

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Sandra Maria Aluísio

USP – São Carlos
Fevereiro de 2020

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S237r

Santos, Leandro Borges dos
Representação de narrativas e extração de suas unidades de informação para automatização de testes neuropsicológicos / Leandro Borges dos Santos; orientadora Sandra Maria Aluísio. -- São Carlos, 2020.
140 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2020.

1. Processamento de linguagem natural. 2. Automatização de exames médicos. 3. Classificação de narrativas. I. Aluísio, Sandra Maria, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Leandro Borges dos Santos

Narratives representation and extraction of their information units for automation of neuropsychological tests

Thesis submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Sandra Maria Aluísio

USP – São Carlos
February 2020

Dedico este trabalho:
Aos meus pais Gilda Alberton Spancerski e Zauri Borges dos Santos,
À minha companheira Sheena Mary John e
A todos que contribuíram diretamente ou indiretamente.

AGRADECIMENTOS

Primeiramente, agradeço a toda a minha família, em especial, à minha mãe, Gilda, que sempre me apoiou e me ajudou. Aos meus tios Ilda, Marcelo, Noeli e a minha avó Helena.

Um agradecimento especial à minha companheira, Sheena, que eu conheci durante a minha jornada, e foi paciente e me ajudou muito durante esse caminho.

Agradeço à minha orientadora Dra. Sandra Maria Aluísio por todos os ensinamentos, pela amizade, pela paciência, e apoio para realizar este trabalho.

Agradeço à Profa. Dra. Letícia Lessa Mansur (*in memoriam*), por todos os ensinamentos e pela contribuição para realização deste trabalho. *A senhora foi uma das pessoas mais queridas que eu conheci e sempre se preocupou comigo mesmo quando tinha assuntos mais importantes.*

Aos professores que conheci durante a minha jornada, em especial à Profa. Dra. Lilian Hübner por ter cedido os dados para essa pesquisa, e ao Prof. André Maleztke por ser meu primeiro orientador e ter me incentivado a seguir na área acadêmica.

Agradeço aos meus amigos do Núcleo Interinstitucional de Linguística Computacional (NILC-USP) pela companhia, amizade, discussões, e aprendizado. Em especial aos que estiveram mais tempo comigo: André, Erick, Edílson, Fernando, Lianet, Márcio, Marco, Nathan, Roney, e Vanessa. Muito obrigado pela contribuição na coleta dos dados e nas discussões sobre esta pesquisa Anderson, Cíntia, Gabriela, e Paolla. Aos amigos e colegas que conheci na UNIOESTE e em Foz do Iguaçu, Antonio, Gustavo, Paulo, Vinícius, e Wesley. Também, aos meus colegas de trabalho: Emílio, Danilo, Patrick, e Thiago por todo aprendizado.

Agradeço a todos os funcionários do ICMC, em especial ao Rogério por todo o seu trabalho desempenhado para atender as necessidades do NILC, e pelos momentos de descontração no café.

Agradeço também ao CNPq, pela bolsa de doutorado e mestrado, processos números 130100/2015-3, 155137/2015-8, e 153047/2016-0, e ao programa *Google Research Awards for Latin America*.

“Be fearful when others are greedy
and greedy when others are fearful”
(Warren Buffett)

RESUMO

SANTOS, L. B. **Representação de narrativas e extração de suas unidades de informação para automatização de testes neuropsicológicos**. 2020. 140 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

O aumento da expectativa de vida tem ocasionado um aumento nas taxas de doenças neurodegenarativas na população idosa. Entre os vários tipos de demências, a principal é a Doença de Alzheimer (DA), correspondendo a $50-75\%$ dos casos. Outra enfermidade que tem recebido atenção nos últimos anos é o Comprometimento Cognitivo Leve (CCL), sendo considerado uma condição pré-clínica da DA, sendo assim importante o seu diagnóstico precoce. Para a identificação de demências e outras doenças relacionadas, são utilizados testes que avaliam a função cognitiva e aspectos linguísticos. Alguns desses exames utilizam como subtestes o reconto de narrativas. Nessa avaliação, a narrativa é dividida em partes, chamadas de unidades de informação, podendo ser palavras ou orações. O escore final do teste representa a quantidade de unidades recordadas. Em geral, é atribuído um ponto para cada unidade. Entretanto, as principais dificuldades no uso de tarefas de reconto são a demanda de tempo e a subjetividade da análise humana. Assim, aplicação de métodos computacionais que automatizem a avaliação é bem-vinda tanto para a larga utilização da tarefa de reconto como para a manutenção da uniformidade na correção, em uma análise longitudinal, por exemplo. O objetivo deste projeto de doutorado, na área de Processamento de Línguas Naturais (PLN) aplicado à área médica, é a avaliação de métodos para automatizar o exame de reconto de narrativas em Português, utilizado na Bateria Arizona para Desordens de Comunicação em Demências (*ABCD*) e na Bateria de Avaliação da Linguagem no Envelhecimento (BALE). Neste trabalho, avaliamos um método de similaridade semântica que se destacou na Avaliação de Similaridade Semântica e Inferência Textual (ASSIN), e desenvolvemos um método baseado na similaridade de *word embeddings*. Transformamos o problema multirrótulo de identificação de elementos de uma narrativa recontada em problemas de classificação binária, e encontramos um ponto de corte para o valor de similaridade de cada unidade de informação. Visando uma triagem automática, esses elementos são usados como atributos para os algoritmos de classificação binária (idosos saudáveis *versus* idosos com comprometimentos cognitivos). Além desses atributos, utilizamos métricas linguísticas, e desenvolvemos um léxico com propriedades psicolinguísticas. Também, propusemos uma abordagem para enriquecer as redes de adjacências, permitindo extrair métricas das propriedades topológicas de redes complexas. Por fim, combinamos todos os atributos para identificar automaticamente em um cenário binário (idosos saudáveis *versus* idosos com comprometimentos cognitivos). Os métodos de identificação de unidades superaram os *baselines* em ambas as baterias clínicas avaliadas. Na classificação binária, os resultados foram semelhantes aos da anotação manual, demostrando a adequação dos métodos desenvolvidos. Em geral, os

resultados experimentais das métricas psicolinguísticas e de redes de adjacência enriquecidas ficaram acima de 50% de acurácia. Entretanto a combinação de todos os atributos investigados ou desenvolvidos não apresentou ganhos; acreditamos que a grande quantidade de atributos e o baixo número de exemplos causou esse resultado negativo.

Palavras-chave: Testes Neuropsicológicos, Reconto de Narrativas, Identificação de Unidades de Informação, Avaliação de Similaridade Semântica.

ABSTRACT

SANTOS, L. B. **Narratives representation and extraction of their information units for automation of neuropsychological tests**. 2020. 140 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

Increased life expectancy can be accompanied by neurodegenerative diseases. Among the various types of dementia, the main one is Alzheimer's Disease (AD), corresponding to 50-75% of cases. Another disease that has received increased attention over the last few years is Mild Cognitive Impairment (MCI), which is considered a preclinical stages of AD, and therefore important to diagnose early. Diagnosis of dementia and related syndromes are commonly based on the analysis of a patient's cognitive functions and linguistic aspects by applying neuropsychological batteries. Some of these batteries use a narrative retelling as a subtest, and are divided into chunks, called units of information, which can be words or phrases. The final score represents the number of units recalled. In general, one point being awarded for each unit recalled. However, the main difficulties of using narratives are the time required and the subjectivity of the manual analysis. Thus, the application of computational methods to automate the assessment is welcome both for the wide use of the task of retelling and to maintain assessment consistency, in a longitudinal analysis, for example. The purpose of this research project in Natural Language Processing (NLP) applied to the medical domain, is the evaluation of methods to automate specifically the retelling of narratives in Portuguese, using the Arizona Battery of Communication Disorders in Dementia (ABCD), and the *Bateria de Avaliação da Linguagem no Envelhecimento* (*BALE*). We evaluated the best ranked semantic similarity method in the *Avaliação de Similaridade Semântica e Inferência Textual* (ASSIN shared task), and we also developed a method based on the similarity of word embeddings. We transformed the multilabel problem of element identification of a narrative into binary classification problems, finding a cutoff point for the similarity value of each information unit. For automatic screening, these elements are then used as features for classification algorithms. In addition to these features, we used linguistic metrics and we also developed a lexicon with psycholinguistic properties. Moreover, we proposed an approach to enrich adjacency networks, allowing the extraction of metrics from topological properties of complex networks. Finally, we combined all of these features to automatically identify narratives in a binary classification task (healthy versus impaired elderly groups). The methods of units identification outperformed the baselines in both clinical batteries; for the binary classification task, the results were similar to manual annotation, demonstrating the adequacy of the methods. In general, the experimental results of the psycholinguistic metrics and enriched adjacency networks were above 50% accuracy. However, as combination of all features, investigated or developed, showed no gains, we believe that the large number of attributes and the low number of examples impacted this evaluation.

Keywords: Neuropsychological tests, Narrative retelling, Identification of information units, Semantic Textual Similarity.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABCD	*Arizona Battery for Communication Disorders of Dementia*
AID	*Alzheimer's Disease International*
BERT	*Bidirectional Encoder Representations from Transformers*
BoW	*Bag-of-Words*
brat	*brat rapid annotation tool*
BR	*Binary Relevance*
CBOW	*Continuous Bag-of-Words*
CDT	Teste do Desenho do Relógio
CFS	*Correlation-based Feature Selection*
DeepBonDD	*Deep neural approach to Boundary and Disfluency Detection*
ELMo	*Embeddings from Language Models*
GloVe	*Glove Vectors*
KNN	*K- Nearest Neighbor*
LP	*Label Powerset*
LSA	*Latent Semantic Analysis*
MEEM	Mini-exame do Estado Mental
MSE	*Mean Square Error*
RBF	*Radial Basis Function*
STS	*Semantic Textual Similarity*
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency-Inverse document Frequency*
TPE	*Tree-of-Parzen-Estimators*
ULMFiT	*Universal Language Model Fine-tuning*
WMS	*Wechsler Memory Scale*
AM	Aprendizado de Máquina
ASSIN	Avaliação de Similaridade Semântica e de INferência textual
CCL	Comprometimento Cognitivo Leve
DA	Doença de Alzheimer
FV	fluência verbal
IBGE	Instituto Brasileiro de Geografia e Estatística
INAF	Indicador de Alfabetismo Funcional

PB	Português Brasileiro
PE	Português Europeu
PLN	Processamento de Línguas Naturais

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Contexto e Motivação

O envelhecimento da população é uma tendência social conhecida em países desenvolvidos e que tem se tornado cada vez mais pronunciada também nos países em desenvolvimento. O Brasil, por exemplo, está mudando sua pirâmide etária, segundo os censos do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2000 e 2010[1].

No envelhecimento normal, ocorrem mudanças em habilidades cognitivas específicas, em especial na memória, na atenção e nas funções executivas, explicadas abaixo.

Com relação à memória[2], as mais afetadas são a de trabalho (curto prazo) e a episódica (ou autobiográfica) (longo prazo) que se referem, respectivamente, à realização de cálculos matemáticos e manipulação de informação, por exemplo, e a lembrar-se de uma história curta ou do que comeu no jantar do dia anterior ou ainda o que fez no último aniversário (CLEMENTE; RIBEIRO-FILHO, 2008). Com relação aos dois outros sistemas de memória de longo prazo que são a de procedimentos — mede o desempenho de habilidades e nos ajuda, por exemplo, a dirigir um carro ou andar de bicicleta — e a semântica, que nos ajuda a diferenciar conceitos ou objetos usados no dia a dia ou ainda conceitos científicos, dependendo da área de trabalho da pessoa, estas duas costumam estar relativamente preservadas no envelhecimento não patológico.

Quanto à atenção, o desempenho é melhor na atenção seletiva, mas na dividida, em que várias tarefas são realizadas simultaneamente, o rendimento é inferior ao de jovens (CLEMENTE; RIBEIRO-FILHO, 2008).

Finalmente, as funções executivas compreendem a seleção, análise, planejamento, organização de ações para alcançar objetivos, ou mesmo a formulação de conceitos. No envelhecimento

[1] <censo2010.ibge.gov.br/sinopse/webservice/frm_piramide.php>

[2] Embora não haja acordo quanto ao número de sistemas de memórias existentes, é comum autores citarem os quatro sistemas de memória descritos neste parágrafo.

normal ou patológico, as funções executivas tendem a estar prejudicadas, mas mudanças impostas pela idade não interferem na autonomia destas pessoas.

O aumento da expectativa de vida é um ganho inestimável para cada cidadão, porém trará sérios problemas financeiros e sociais, sobretudo na área da saúde, pois o envelhecimento pode ser acompanhado de doenças neurodegenerativas, que demandam novos recursos e estruturas médicas. Assim, cresce em importância as pesquisas em neurociência sobre detecção precoce e gestão de doenças que normalmente afetam pessoas com idade avançada, como as demências.

A Doença de Alzheimer (DA) é a demência mais prevalente no mundo. Ela gera déficits cognitivos sérios o bastante para interferir na vida diária de um indivíduo, que pioram com o passar do tempo, pois a doença destrói neurônios, causando perda de memória, habilidades de raciocínio, e de julgamento (SHEN *et al.*, 2014). O relatório mais recente da *Alzheimer's Disease International* (*AID*) aponta que existem 50 milhões de pessoas vivendo com demência no mundo; em 2050 serão 152 milhões de pessoas, e a cada três segundos uma pessoa desenvolve demência (Alzheimer's Disease International, 2019). Atualmente, o custo anual é estimado em 1 trilhão de dólares, e esse valor pode dobrar em 2030. Desse modo, devido aos seus custos sociais e econômicos, as demências são consideradas pela Organização Mundial de Saúde (OMS) como um desafio para as próximas décadas (WORTMANN, 2012).

Também cresce em importância o estudo de uma síndrome menos conhecida, com as primeiras referências datadas de 1999, chamada Comprometimento Cognitivo Leve (CCL). O CCL é definido como um declínio cognitivo maior do que o esperado para indivíduos de mesma idade e escolaridade, cuja interferência nas atividades do dia a dia é mínima, pois se manifesta somente em situações complexas e não pode ser considerado demência. O CCL pode acometer um ou múltiplos domínios. O tipo mais frequente e com maior taxa de conversão para Doença de Alzheimer (15% por ano, versus 1 a 2% da população em geral) é o que acomete a memória – o CCL amnéstico. A taxa de conversão se intensifica quando o acometimento de memória se associa a outros domínios (CLEMENTE; RIBEIRO-FILHO, 2008).

O principal desafio no gerenciamento clínico da DA vem do fato de que o início do processo neurodegenerativo pode se dar anos, às vezes décadas, antes que os efeitos cognitivos possam ser percebidos (SPERLING; KARLAWISH; JOHNSON, 2013). O melhor tratamento atualmente disponível para a demência consiste em retardar a progressão da doença quando da detecção de seus primeiros sinais. Apesar de não haver tratamentos que modifiquem a doença, o consenso na área é que, quando tratamentos desse gênero se tornarem disponíveis, será imperativo iniciar o tratamento muito antes que danos clinicamente significativos tenham ocorrido ao cérebro (JARROLD *et al.*, 2010). Assim, melhorar os mecanismos de diagnóstico precoce é fundamental para retardar o desenvolvimento da doença. Quanto ao CCL, o desafio é a identificação do subgrupo de indivíduos que progredirá para DA, para que também se possa dar tranquilidade ao grupo que não evoluirá para DA.

Prud'hommeaux e Roark (2012) reforçam que o diagnóstico definitivo do CCL requer

uma entrevista abrangente tanto com o paciente quanto com a família ou cuidador. Dado este grande esforço e pelo fato de que certos testes cognitivos não capturam a síndrome, o CCL passa despercebido nas avaliações, o que atrasa tanto o tratamento para retardar o seu avanço quanto para não permitir a sua evolução para casos moderados e severos de demência.

O diagnóstico das demências e síndromes relacionadas é feito com base na análise das funções cognitivas do paciente, via baterias neuropsicológicas, visando as funções que são mais afetadas como memória, orientação, linguagem e resolução de problemas. As baterias são usadas antes, durante e depois de tratamentos (ABREU; FORLENZA; BARROS, 2005).

O Mini-exame do Estado Mental (*MEEM*, *Mini Mental State Exam*) (FOLSTEIN; FOLSTEIN; MCHUGH, 1975) é o teste cognitivo mais utilizado para a avaliação de pacientes com demências no mundo, sendo usado como teste de rastreio e triagem. É composto por sete categorias de questões, cada uma com a finalidade de avaliar funções cognitivas específicas como orientação para tempo, orientação para local, registro de três palavras, atenção e cálculo, lembrança das três palavras, linguagem e capacidade construtiva visual. Segundo, Abreu, Forlenza e Barros (2005) o *MEEM* foi traduzido e adaptado em 1994 no Brasil. Em uma avaliação em larga escala, utilizou-se 530 indivíduos classificados segundo suas idades e escolaridades, usando-se os quatro níveis de escolaridade (analfabetos, baixa, média e alta). Constatou-se a importância da escolaridade no escore total do teste e, assim, os pontos de corte para o *MEEM* foram adaptados segundo o nível de escolaridade no Brasil. O estudo mostrou também que adaptar testes desenvolvidos para outras línguas requer uma avaliação em larga escala da população alvo, para que ajustes culturais, sociais e educacionais sejam realizados.

Inclusive baterias curtas têm sido propostas para dar conta do aumento do número de avaliações neuropsicológicas em hospitais e também da necessidade de se avaliar pacientes com níveis de letramento baixo (NITRINI *et al.*, 2007; HÜBNER *et al.*, 2019), dado que 30% dos brasileiros são considerados analfabetos funcionais, de acordo com dados do Indicador de Alfabetismo Funcional (INAF), relatório de 2018[3].

Em geral, as baterias incluem testes que avaliam a linguagem, como nomeação e a fluência verbal (FV). No teste de nomeação são apresentadas folhas de papeis com objetos desenhados, e o clínico solicita ao paciente para nomear o objeto. Enquanto que no teste de fluência verbal é encorajado ao paciente falar o maior número possível de palavras de uma categoria em 1 minuto, geralmente. Para a avaliação da memória visual-espacial, é frequentemente utilizado o teste do desenho do relógio (*CDT*, *Clock Drawing Test*) , que consiste em pedir para o participante desenhar um relógio de leitura analógica marcando uma determinada hora, e, embora não seja um indicador definitivo de DA ou CCL, pode indicar a severidade dos problemas.

As baterias neuropsicológicas também avaliam a memória em razão da alta prevalência da Doença de Alzheimer. Assim sendo, os demais aspectos cognitivos como linguagem, atenção,

[3] <https://www.ipm.org.br/relatorios>

funções executivas, praxias e aspectos visual-espaciais ficam em segundo plano, ou são analisadas sob o foco da memória. Pesquisas recentes têm reconhecido a heterogeneidade da própria DA e do CCL, o que amplia a relevância da análise de outras habilidades cognitivas, como a linguagem.

A avaliação da linguagem tem na produção discursiva, principalmente narrativas, uma alternativa interessante pelo fato de permitir a análise de microestruturas linguísticas (ANDRETTA; CANTAGALLO; MARINI, 2012) e componentes fonético-fonológicos, morfossintáticos, semântico-lexicais assim como macro estruturas semântico-pragmáticas. Por ser uma forma natural de comunicação, favorece a observação da funcionalidade do paciente na vida cotidiana. Além disso, fornece dados para a observação da interface linguagem-habilidades cognitivas como funções executivas (planejamento, organização, atualização de dados, monitoramento).

Publicações recentes têm mostrado que a produção de discurso é uma tarefa sensível para detectar efeitos do envelhecimento e diferenciar indivíduos com CCL e dementes (ROARK; MITCHELL; HOLLINGSHEAD, 2007; PRUD'HOMMEAUX; ROARK, 2012; CUNHA, 2015; ALUÍSIO; CUNHA; SCARTON, 2016; FRASER; MELTZER; RUDZICZ, 2016; SANTOS *et al.*, 2017). Nessa direção, o teste de reconto de narrativas tem sido um dos mais utilizados na avaliação linguístico-cognitiva. Dados obtidos a partir dessas narrativas permitem a avaliação do discurso, o que tem sido desvalorizado nas avaliações cognitivas que restringem a análise à memória verbal e pontuação do número de itens evocados.

Os testes de reconto de narrativas envolvem subtestes de reconto imediato e o reconto tardio. Estes dois subtestes utilizam uma história curta que é contada ao paciente e o teste pressupõe que se avalie o número de elementos da história que podem ser relembrados pelo paciente imediatamente após a contação da história e 30 minutos depois. Durante os dois testes há a aplicação de outros testes de natureza diferente.

Um desafio na escolha da avaliação neuropsicológica de indivíduos com DA e CCL, com a indicação do grau de comprometimento cognitivo, é o uso de uma bateria que consiga distinguir estes indivíduos. A Bateria Arizona para Desordens da Comunicação e Demência (*ABCD*, *Arizona Battery for Communication Disorders of Dementia*) (BAYLES; TOMOEDA, 1993) surge como opção, uma vez que é capaz de detectar a DA em estágio leve. Também é importante citar a Bateria de Avaliação da Linguagem no Envelhecimento (BALE) (HÜBNER *et al.*, 2019), que foi desenvolvida para a aplicação em pacientes com diferentes graus de escolaridade, incluindo analfabetos.

Enquanto uma análise qualitativa do discurso pode revelar o tipo da doença apresentada pelo paciente, uma análise quantitativa é capaz de revelar a intensidade do dano cerebral existente. A grande dificuldade de análises quantitativas de discurso é sua exigência de esforços: o processo de análise rigorosa e detalhada da produção oral (que é transcrita manualmente ou com ajuda de ferramentas semiautomáticas) ou escrita é bastante laborioso, o que dificulta sua adoção em larga escala. Por isso, análises computadorizadas se tornam uma solução de interesse.

Se um conjunto adequado de dados de sujeitos saudáveis e com quadros de demências estiver disponível, a área de pesquisa em Processamento de Línguas Naturais (PLN), aliada à de Aprendizado de Máquina (AM), é capaz de produzir modelos preditivos que respondam com precisão elevada se o sujeito é portador de uma síndrome/demência ou não (PEINTNER *et al.*, 2008; JARROLD *et al.*, 2010; FRASER; MELTZER; RUDZICZ, 2016). Mas, o mais importante talvez seja o efeito tranquilizador para muitas pessoas ao poderem saber que, após uma análise longitudinal de avaliações de seus discursos, não desenvolverão demências.

1.2 Definição do Problema

Existem vários exames para a identificação de demências, como a utilização de biomarcadores, ressonância magnética e neuroimagem molecular (MAPSTONE *et al.*, 2014; MCKHANN *et al.*, 2011). Entretanto, a definição de critérios de diagnóstico é conduzida principalmente pelos sintomas cognitivos apresentados pelos pacientes em testes padronizados e pelos prejuízos funcionais à vida diária (MCKHANN *et al.*, 2011).

Como a linguagem é uma das habilidades que mais fornece informações eficazes das funções cognitivas, torna-se importante sua utilização em exames médicos. Portanto, alguns testes utilizam o reconto de narrativas como uma ferramenta para auxiliar a identificar e quantificar o grau de demência. Como exemplos de testes temos: o subteste do Exame Wechsler de Memória Lógica, a Bateria Arizona para Desordens da Comunicação e Demência, Bateria de Avaliação da Linguagem no Envelhecimento, entre outros (WECHSLER, 1997; BAYLES; TOMOEDA, 1993; HÜBNER *et al.*, 2019).

O teste de reconto de narrativas permite mensurar a capacidade de memória de um paciente; para isso a narrativa é dividida em unidades de informação, as unidades podem ser palavras, sintagmas ou uma oração. Em geral, é atribuído um ponto para cada unidade recordada, e o escore final representa a quantidade de unidades recordadas.

As principais vantagens da utilização de narrativas é que elas mensuram as habilidades morfossintáticas, semânticas e pragmáticas (MEURIS; MAES; ZINK, 2014). Entretanto, as suas principais dificuldades são: (i) a demanda de tempo por ser uma tarefa de avaliação manual: (ii) a subjetividade do avaliador na checagem da presença das unidades de informação da narrativa no reconto feito pelo paciente, pois é necessário levar em conta as possíveis paráfrases. Assim, torna-se bem-vinda e importante a aplicação de métodos computacionais tanto para a automatização dessa tarefa, viabilizando sua aplicação em larga escala, como para a manutenção da uniformidade na correção.

Entretanto, há desafios computacionais também para a automatização do cálculo do escore por um sistema computacional. O sistema deverá resolver vários fenômenos que são comuns para essa tarefa, por exemplo: mudança da ordem de palavras da história original, uso de palavras similares às da história original, comentários que não estão relacionados com a história,

e disfluências que tornam a história recontada bastante diferente da original.

Na Figura 1 (a), apresentamos a história do teste do reconto da bateria *ABCD*, traduzida para o português. Ela possui 17 unidades de informação, com possíveis alternativas entre parênteses, sendo 17 a sua pontuação máxima. Em (b), apresentamos um reconto imediato de um paciente com CCL, com pontuação 12, pois a avaliação manual de seu reconto contabilizou 12 unidades relembradas. Neste reconto, há trechos com disfluências ((1) e (3)), duplicação de unidades de informação recontadas ((1) e (3); (5) e (6)) e comentários ((2), (10) a (13)).

Figura 1 – (a) Narrativa utilizada na *ABCD*, separada em unidades de informação; as nove unidades marcadas em negrito são as principais, o resto são detalhes.(b) Transcrição do reconto imediato de um paciente com CCL, segmentada manualmente.

(a) **Senhora (mulher)** (1) **// estava fazendo compras (na loja, foi às compras, foi ao mercado)** (2) // Sua carteira (seu porta-notas, sua moedeira) (3) **// carteira caiu (derrubou a carteira, perdeu a carteira, perdeu a bolsa)** (4) //da sua bolsa (da sua mochila, de sua pasta) (4) **// Ela não viu a carteira cair (ela não notou)** (5) // No caixa (quando ela foi pagar, guichê) (6) **// não tem como pagar (ela não tinha dinheiro, não tinha sua carteira)** (7)// Coloca as mercadorias de lado (coloca as mercadorias de volta) **// foi para sua casa (voltou para sua casa)** // Quando ela abriu a porta (quando ela chegou em casa, assim que ela entrou) **// telefone tocou (fone tocou, ela recebeu uma ligação)** // Pequena (jovem) // menina (garota) // lhe disse (falou, contou) **// ela achou a carteira (achou sua moedeira, achou o porta-notas) // Senhora aliviada (senhora estava feliz, senhora estava radiante, senhora estava agradecida)**
(b) (1) ahm uma senhora foi fazer compras no me foi no mercado. (2) não lembrava o local. (3) no me fazer compras. (4) e quando ela foi pagar a conta no caixa percebeu que estava sem a carteira. (5) aí ela foi deixou a mercadoria. (6) não levou a mercadoria. (7) voltou para casa. (8) chegando em casa toca o telefone. (9) era uma garotinha avisando ela que que tinha achado a carteira. (10) é isso. (11) tem mais coisa. (12) não cortei. (13) eu resumi o que eu ouvi.

Existem poucos trabalhos na literatura que tratam da automatização de identificação de unidades de informação em recontos de narrativas. Podemos dividi-los em: métodos de busca de palavras (PAKHOMOV *et al.*, 2010; FRASER; MELTZER; RUDZICZ, 2016), métodos de alinhamento (PRUD'HOMMEAUX; ROARK, 2015), e métodos de *clustering* (YANCHEVA; RUDZICZ, 2016; FRASER; FORS; KOKKINAKIS, 2019). Neste trabalho, é apresentado o resultado do método de busca de palavras. O método de *clustering* foi explorado mas não encontrou-se grupos coesos e o método de alinhamento não foi explorado, pois necessita de um conjunto de dados anotados no formato de alinhamento de palavras. Como algumas unidades de informação dos conjuntos de dados utilizados representam sintagmas, não seria adequado utilizar esse último método.

Para identificar automaticamente as unidades de informação foram desenvolvidos dois métodos que utilizam similaridade semântica. A razão desta escolha se deu pelo fato da similaridade semântica conseguir indicar um grau de semelhança entre dois textos curtos. Dessa forma, a similaridade é útil para identificar as variações na forma de recontar trechos da narrativa.

Durante o projeto de doutorado os recontos de narrativas foram segmentados manualmente em sentenças, e para cada sentença marcou-se as unidades de informação presentes. Como as sentenças podem possuir mais de uma unidade de informação, essa tarefa é chamada de classificação multirrótolo. Para contornar essa situação transformou-se o problema multirrótulo de identificação de elementos de uma narrativa recontada em problemas de classificação binária.

Para predizer as unidades de informação encontrou-se um ponto de corte do valor de similaridade para cada unidade de informação.

Após a identificação das unidades de informação, estas são combinadas com métricas que caracterizam as narrativas com o objetivo de auxiliar no diagnóstico de demências. Dentre dessas métricas, temos as linguísticas que extraem as informações lexicais e sintáticas das narrativas (FRASER *et al.*, 2014; ALUÍSIO; CUNHA; SCARTON, 2016).

Além dessas métricas, podemos representar um texto como uma rede complexa, em que os vértices representam as palavras e as arestas representam as relações de adjacência, e computar métricas que mensuram as propriedades topológicas da rede. A exploração dessas métricas para identificar automaticamente narrativas de pacientes com demência é inédita.

1.3 Objetivos, Lacunas e Hipótese

O objetivo geral do projeto, chamado de ANAA-Dementia (**A**nálise de **N**arrativas **A**utomatizada para **A**poiar o diagnóstico de demências), foi desenvolver métodos que utilizam a similaridade semântica textual para automatizar o reconto imediato e tardio de narrativas em português. Neste doutorado, foram utilizados o subteste de memória episódica da *ABCD* e o reconto de uma narrativa pertencente ao conjunto de tarefas da análise discursiva da BALE (HÜBNER *et al.*, 2019).

Os demais objetivos dessa tese foram: investigar métricas linguísticas e desenvolver métodos para caracterizar as narrativas em redes complexas para a identificação de pacientes com CCL ou DA, visando uma triagem automática de pacientes no cenário clínico.

O projeto ANAA-Dementia está contido num projeto de escopo maior que envolve a colaboração de outros alunos de mestrado e doutorado. No projeto maior, chamado de *Aging@Brazil*[4], são usadas ferramentas automáticas para auxiliar o estudo e a identificação de pacientes com CCL/DA. Uma parte dos dados foram coletados pela equipe coordenada pela Profa. Dra. Letícia Lessa Mansur da Faculdade de Medicina da Universidade de São Paulo. O projeto também conta com dados que foram coletados pela equipe coordenada pela Profa. Dra. Lilian Cristine Hübner da Escola de Humanidades da Pontifícia Universidade Católica do Rio Grande do Sul. No projeto de doutorado, utilizou-se os dados de ambos os centros de coleta.

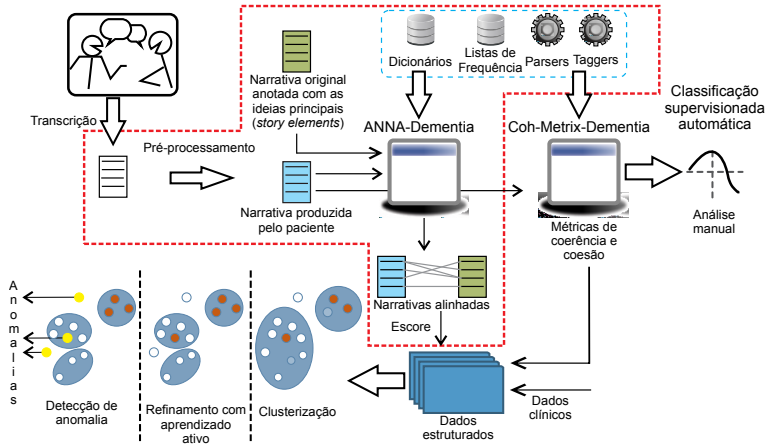
Na Figura 2 são ilustradas as relações de cada projeto; as narrativas coletadas podem ser analisadas pelo sistema *Deep neural approach to Boundary and Disfluency Detection* (*DeepBonDD*), desenvolvido durante o mestrado de Marcos Vinícius Treviso. O sistema é capaz de segmentar e remover as disfluências automaticamente, produzindo assim, uma narrativa adequada para o ambiente Coh-Metrix-Dementia, responsável por extrair métricas linguísticas, e que foi criado durante o mestrado de André Cunha, ambos foram orientados pela Profa. Sandra

[4] <http://www.nilc.icmc.usp.br/agingbrazil>

Maria Aluísio.

Finalmente, todos os projetos auxiliam o trabalho focado na identificação automática de pacientes que realizaram uma transição do estado normal para CCL amnéstico ou estado com DA, utilizando métodos de *ensemble* de agrupamento de dados e aprendizado ativo para a identificação de anomalias.

Figura 2 – Relações de cada um dos 4 subprojetos (*DeepBonDD*, ANAA-Dementia, Coh-Metrix-Dementia, e detecção de anomalia por meio de *ensemble* de agrupamento de dados), no escopo do projeto Aging@Brazil.



Fonte: Elaborada pelo autor.

O objetivo deste projeto está na intersecção de duas áreas de pesquisa: o PLN, com o estudo de métodos de similaridade semântica e caracterização de narrativas com métricas, e a neurociência, com os testes de avaliação neuropsicológica. Para alcançar o objetivo geral deste trabalho foram definidos os seguintes objetivos específicos de pesquisa:

- Desenvolver recursos para possibilitar a pesquisa da similaridade semântica textual em português;
- Investigar a utilização de similaridade semântica textual para identificar as unidades de informação de narrativas em português, no cenário de poucos dados de treinamento;
- Estender o conjunto de métricas desenvolvidos por Cunha *et al.* (2015), pois este não conta com nenhuma métrica que explora as propriedades psicolinguísticas das palavras;
- Investigar a representação de narrativas em redes complexas, dado que caracterização de textos em redes tem se mostrado promissora e não necessita de recursos caros de PLN como parser sintático ou de dependência ou dicionários, como é o caso de algumas métricas linguísticas.

As lacunas que este trabalho procurou atacar são tanto da área médica como do PLN:

- Necessidade de se automatizar exames de baterias neuropsicológicas para identificação de demências, dada a demanda crescente destes exames nos anos atuais;
- Falta de uniformidade na pontuação de exames que utilizam narrativas, dada a avaliação humana, que naturalmente insere tendências;
- Necessidade de conjuntos de dados com as unidades de informação anotadas;
- Escassez de conjuntos de dados para desenvolver métodos de similaridade semântica;
- Falta de um léxico com as propriedades psicolinguísticas das palavras para o Português Brasileiro;
- Necessidade de se representar textos curtos, como as narrativas de recontos, em redes complexas.

Dados o objetivo geral e os objetivos específicos, pode-se declarar a hipótese central deste projeto como: *métodos de similaridade semântica de texto podem viabilizar a automatização da análise de reconto de narrativas em exames neuropsicológicos, visando uma triagem automática com desempenho próximo da avaliação manual, para a língua portuguesa*.

1.4 Organização da Tese

Esta tese está organizada em cinco capítulos. No Capítulo 2, são apresentadas as informações sobre a DA e o CCL, e como essas doenças afetam o discurso; informações sobre a Bateria Arizona para Desordens da Comunicação em Demência e a Bateria de Avaliação da Linguagem no Envelhecimento que utilizam o reconto de narrativas. Também são apresentados conceitos sobre classificação monorrótulo e multirrótulo, e abordagens para caracterizar uma narrativa no processo de classificação, e Similaridade Semântica Textual e a Inferência Textual. No Capítulo 3, são apresentados os trabalhos relacionados às duas tarefas abordadas nesse projeto de doutorado: classificação de narrativas de pacientes com CCL e DA e a identificação de unidades de informação. Os recursos criados para o desenvolvimento do projeto de pesquisa e os métodos desenvolvidos, junto com seus resultados, são apresentados no Capítulo 4. Por fim, o Capítulo 5 traz as conclusões da tese.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

Este capítulo traz uma caracterização dos problemas de linguagem de pacientes com Doença de Alzheimer e com Comprometimento Cognitivo Leve (Seção 2.1). Na Seção 2.2 são apresentadas a Bateria Arizona para Desordens da Comunicação e Demência e a Bateria de Avaliação da Linguagem no Envelhecimento, que são de interesse especial neste trabalho. Também são apresentadas outras baterias usadas em testes neuropsicológicos que fazem uso do reconto de narrativas. Na Seção 2.3 são apresentados conceitos de classificação tradicional monorrótulo e classificação multirrótulo, que foi a escolhida para a tarefa de classificação de narrativas nesta tese. Na Seção 2.4 são abordadas técnicas para representar textos na tarefa de classificação de narrativas. Por fim, na Seção 2.5 são apresentados conceitos sobre Similaridade Semântica Textual e Inferência Textual, e também o conjunto de dados da Avaliação de Similaridade Semântica e de INferência textual (ASSIN).

2.1 Linguagem no Envelhecimento e na Demência

O envelhecimento acarreta algumas perdas de funcionalidades, como a capacidade motora, diminuição dos mecanismos de defesa natural do organismo e de adaptação ao ambiente e também a diminuição de funcionalidades cognitivas, como a linguagem. Alguns fatores podem afetar esse déficit funcional, entre eles, doenças frequentes no envelhecimento e fatores ambientais, sendo estes não determinantes para a perda da funcionalidade (GARCIA; MANSUR, 2006).

A linguagem tem um papel fundamental na vida das pessoas, possibilitando a comunicação e as demais atividades sociais. A sua modificação não ocorre de forma isolada, estando relacionada com as alterações na memória operacional, atenção e habilidade visual-espacial (FREITAS, 2010).

Nos idosos são notadas alterações dos padrões discursivos conforme o estímulo. Para

tarefas nas quais é exigido o reconto de narrativas ouvidas recentemente, são obtidos textos curtos e simples ao contrário das tarefas em que é necessária a produção de narrativas livres. Nestas últimas, os idosos tendem a elaborar textos mais longos, mas contendo um número maior de informações irrelevantes e com baixa coesão (GARCIA; MANSUR, 2006).

Nas próximas seções serão apresentadas as principais características da Doença de Alzheimer (DA) e do Comprometimento Cognitivo Leve (CCL).

2.1.1 Doença de Alzheimer

Em McKhann *et al.* (2011) são apresentados critérios para a identificação de pacientes com DA, dividindo-a em provável, possível e definida. Ela é caracterizada como provável quando preenche critérios de modo consistente; é possível quando há fatores incertos como história de doença vascular; e é definida quando comprovada por exame anatomo-patológico, realizado pós-mortem.

No Brasil, as recomendações para o diagnóstico foram elaboradas em 2011 pelos membros do Departamento de Neurologia Cognitiva e do Envelhecimento da Academia Brasileira de Neurologia (FROTA *et al.*, 2011). As recomendações incluem os critérios clínicos para o diagnóstico de demência (presença de sintomas cognitivos ou comportamentais que interferem no trabalho ou atividades usuais), declínio em níveis prévios de desempenho, não explicáveis por doenças psiquiátricas ou delirium (estado confusional agudo).

Entre os critérios estão o déficit cognitivo, podendo ser de duas formas: (i) a forma amnésica (quando predomina o comprometimento de memória, associado ao comprometimento em outro domínio cognitivo), e (ii) não-amnésica (deve haver outro domínio afetado) com alterações na linguagem (lembranças de palavras, por exemplo), aspectos visual-espaciais, agnosia (dificuldade de reconhecer objetos ou faces, e dificuldade de leitura relacionada a aspectos visual-espaciais), funções executivas (alteração do raciocínio, julgamento e solução de problemas).

O declínio cognitivo progressivo é confirmado com exames sucessivos e a positividade de biomarcadores. Também são utilizados exames de imagens para exclusão de outros diagnósticos.

Mesmo que a perda de memória seja a característica mais frequente, alterações na linguagem também podem aparecer nos estágios iniciais da DA. Uma das formas de se avaliar a linguagem é a produção de narrativas, sendo observado que estas narrativas apresentam sentenças simples e curtas, maior número de proposições irrelevantes, vocabulário pobre, ruptura no desenvolvimento do tema, maior número de erros ortográficos e menor nível de complexidade sintática (MANSUR *et al.*, 2005).

Abaixo são apresentados os aspectos da linguagem na DA, enfatizando a produção da linguagem, que é de interesse imediato neste projeto. Os níveis descritos abaixo envolvem o lexico-semântico, o sintático e o discursivo, pois a aplicação dos métodos de classificação

pressupõem a fala transcrita dos recontos dos pacientes [1].

Nível lexico-semântico

É consenso entre os pesquisadores que habilidades semânticas constituam o cerne das perdas da linguagem, causadas pelo processo degenerativo da Doença de Alzheimer. No que diz respeito à produção da linguagem, esses déficits têm sido estudados principalmente em tarefas de nomeação (por confrontação visual e por definição) e fluência verbal.

Quando convidados a emitir itens relacionados, durante um tempo restrito (um minuto), os indivíduos com DA produzem um menor número de palavras do que idosos saudáveis (VLIET *et al.*, 2003). Além disso, quando esse teste de fluência verbal é baseado em critérios semânticos torna-se sensível para discriminar idosos saudáveis e indivíduos com Doença de Alzheimer (CERHAN *et al.*, 2002; SALMON *et al.*, 2002).

A fluência verbal semântica, entre outras habilidades, depende da integridade da bagagem semântica, razão pela qual se supõe que o déficit de memória semântica na DA reflita uma degradação desse repertório (HENRY; CRAWFORD; PHILLIPS, 2004).

Do ponto de vista qualitativo, sabe-se que os portadores da DA produzem na tarefa de fluência verbal um menor número de *switches* (mudanças de critério de evocação de itens em determinado campo) e produzem clusteres (agrupamentos de itens de determinada categoria) menores quando comparados com idosos saudáveis.

Outra modalidade largamente utilizada para avaliação de memória semântica é a nomeação. Frequentemente testa-se a nomeação em testes de confrontação visual, sendo ainda utilizados, por exemplo, a definição de conceitos e a nomeação a partir da definição.

O sucesso no teste de nomeação está associado à preservação do conhecimento de atributos semânticos (GARRARD *et al.*, 2005). O empobrecimento da capacidade de definição (fornecimento de atributos semânticos) está associada à performance comprometida na nomeação. A perda semântica é gradual e no início do processo há vulnerabilidade dos conceitos distintivos sem distinção entre perdas nas diferentes categorias. A perda de atributos distintivos leva a falhas quando o portador de DA é solicitado a optar entre conceitos próximos.

Marques, Cappa e Sartori (2011) constataram que a relevância e o tipo de traço semântico (não sensorial) eram importantes para a representação conceitual e a recuperação lexical. Na nomeação a partir da definição, a relevância do traço semântico parece ser decisiva para o desempenho de idosos normais e pacientes com DA.

Uma questão interessante proposta nos estudos sobre nomeação em pacientes com DA é se existe vantagem na nomeação de verbos de ação quando comparada a nomeação de

[1] A descrição destes três níveis foi resumida da monografia de mestrado de André Cunha, que trabalhou com avaliação de textos de pacientes com DA e CCL, e foi orientado pela mesma pesquisadora que orientou esta pesquisa. Todas as referências da literatura sobre linguagem no envelhecimento saudável e com demências nos foi passada pela Profa. Dra. Letícia Mansur

substantivos. Essa questão fundamenta-se no fato de a DA acometer prioritariamente regiões posteriores do cérebro, poupando as redes anteriores frontais que dão suporte à nomeação de verbos. O estudo de Druks *et al.* (2006) mostrou que tanto os sujeitos controles quanto aqueles com DA tiveram mais dificuldades na nomeação de verbos do que na nomeação de substantivos.

O conhecimento semântico pode afetar outras habilidades de pacientes com DA, como por exemplo a memória de curta duração (PETERS *et al.*, 2009) e o uso da linguagem (ALTMANN; MCCLUNG, 2008).

Finalmente, as alterações em memória semântica acham-se comprometidas já em fase pré-clínica da DA (CUETOS; RODRÍGUEZ-FERREIRO; MENÉNDEZ, 2009), razão pela qual as pesquisas sobre habilidades léxico-semânticas merecem especial atenção dos pesquisadores.

Nível sintático

As descrições detalhadas sobre sintaxe na doença de Alzheimer sempre apontaram que algumas habilidades são relativamente preservadas, no início da apresentação da doença, como é o caso das relações verbo-sujeito e aspectos morfológicos. Com o progresso da doença, os portadores da doença tendem a simplificar sentenças e reduzir o conteúdo das proposições e a linguagem fica reduzida a sentenças curtas, familiares, repetitivas ou fragmentos, chegando ao mutismo (KEMPER; THOMPSON; MARQUIS, 2001).

A redução das habilidades sintáticas está relacionada à perda das bases semânticas da linguagem. É o que se observou no estudo seminal de Snowdon, Greiner e Markesbery (2000). Esses pesquisadores analisaram aspectos sintáticos indissociados dos semânticos na produção textual escrita de 93 religiosas, no contexto do "Estudo das freiras", um estudo longitudinal sobre DA. As religiosas idosas foram avaliadas do ponto de vista neuropsicológico, sendo que, para a linguagem, tomou-se como dado comparativo longitudinal o diário escrito por ocasião do ingresso no convento. O estudo neuropatológico realizado pos-mortem foi utilizado para comprovação de diagnóstico de DA, em 14 sujeitos. Os autores observaram que as religiosas cujo estudo pos-mortem confirmou o diagnóstico de DA já na juventude apresentavam traços indicativos da doença. Um desses indicativos era o que chamaram de "simplificação da sintaxe".

Nível discursivo

O estudo da produção de discurso na DA é recente. No discurso dos pacientes com DA nota-se o impacto de déficits cognitivos já no início da doença. Por essa razão, do ponto de vista de diagnóstico, o discurso torna-se interessante para observar aspectos microlinguísticos e sua interação com aspectos não linguísticos, por exemplo, seleção, planejamento, organização.

Os portadores de DA tornam-se repetitivos, esquecem o que ouviram ou leram, perdem o tópico. Ao longo do tempo, o discurso torna-se empobrecido, fragmentado, caracterizado por falta de coerência. Nota-se ainda tangencialidade e perseverações, i.e., repetições de ideias (HOOPER; BAYLES, 2007).

A produção de discurso de portadores de DA tem sido examinada a partir de estímulos visuais com cenas em prancha única ou sequências de pranchas, discursos de procedimento, e ainda em situação espontânea como relatos e diálogos em conversação.

Forbes-McKay e Venneri (2005) avaliaram o discurso de indivíduos idosos saudáveis e portadores de DA em pranchas classificadas como simples ou complexas, de acordo com o número de subtemas. Os autores verificaram os efeitos de idade e escolaridade no desempenho da tarefa. Além disso, o desempenho dos pacientes com DA esteve associado a outras habilidades de processamento semântico. Concluíram que a produção de discurso a partir de prancha complexa pode detectar alterações de linguagem na DA, já no início do quadro.

Carlomagno *et al.* (2005) investigaram fatores subjacentes à redução de conteúdo e falta de referência no discurso de pacientes com DA. As amostras de discurso dos portadores estudados foram colhidas a partir da descrição da clássica figura do "Roubo dos Biscoitos" (GOODGLASS; KAPLAN; BARRESI, 1983) e de uma tarefa de comunicação sensibilizada para observação de aspectos lexicais, elaboração de aspectos pragmático/conceituais da informação e efetividade no estabelecimento de referências. Nessa última tarefa, cada um dos participantes recebia figuras idênticas, porém em sequências diferentes. A solicitação era que reorganizassem as figuras, buscando alcançar a mesma sequência. Os autores valorizaram falhas na elaboração pragmático-conceitual como um dos fatores que se associaram à redução de informação e falta de referência na "fala vazia" dos pacientes com DA e ressaltaram a importância de se investigar o discurso por meio da situação sensibilizada, além da prancha única.

Lira *et al.* (2011) analisaram aspectos microlinguísticos da sequência de figuras "*The Dog Story*" (BOEUF, 1971) e constataram maior número de erros lexicais e menor índice de complexidade sintática numa amostra de 121 indivíduos portadores de DA. Esse índice representa a razão entre o número total de sentenças e os subtipos (subordinadas, coordenadas e orações reduzidas) produzidos pelo indivíduo. O discurso produzido era notadamente mais simples do que o da população de controle, com predomínio de sentenças coordenadas. Entre os erros lexicais, foram proeminentes as dificuldades de acesso lexical, repetição de palavras, uso de termos indefinidos, ao lado de um maior número de revisões e correções nos pacientes com DA. Os autores não puderam diferenciar os indivíduos controle dos portadores de DA em algumas medidas de interesse como dificuldade de acesso lexical, embora as demais medidas lexicais tenham se mostrado sensíveis, como a repetição de palavras, uso de termos indefinidos e revisões.

Ska e Duong (2005) estudaram simultaneamente diferentes níveis de representação nas narrativas de pacientes com DA, por meio do modelo de construção-integração (KINTSCH, 1988). As narrativas eram produzidas em duas situações: a partir de uma prancha única e pranchas em sequência. O objetivo do estudo era determinar níveis de representação discursivos comprometidos nos pacientes com DA, quando comparados a sujeitos normais. O modelo de construção-integração do discurso inclui quatro níveis de representação desde a superfície na qual se analisam componentes linguísticos do discurso (índice lexical, índice sintático e índice

referencial) até a organização dos esquemas narrativos abstratos. Os autores verificaram que a prancha única provocou maior número de dificuldades para gerar discursos, entre os pacientes. Além disso, constataram que embora todos os níveis estivessem comprometidos na DA, eles diferiram dos controles em três níveis: nível de superfície, o modelo de situação e a organização da estrutura narrativa.

Dificuldades como repetição de informação, também consideradas sintoma de "esvaziamento do discurso", que ocorre frequentemente na DA, aparecem de forma privilegiada em situações espontâneas, como entrevistas. Cook, Fay e Rockwood (2009) estudaram a fala de pacientes com DA produzida nessa situação. As ocorrências de repetição foram categorizadas por unidades de repetição (sons, palavras, afirmações, sintagmas, histórias), o tópico ou foco da repetição (por exemplo, retomada de evento passado), o intervalo da repetição (minutos, horas), e a constância da repetição dos episódios (diária, semanal). O tipo de repetição mais frequente foi sobre questões relacionadas a eventos prospectivos.

Ainda sobre aspectos semânticos em fala encadeada foi desenvolvida a investigação com pacientes cujo exame neuropatológico comprovou o diagnóstico de DA (AHMED *et al.*, 2013). O discurso de indivíduos sadios, com declínio cognitivo e em fase leve da doença, foi estudado por meio das medidas de "*idea density and efficiency*". A medida de *idea density* foi feita manualmente, a partir da definição do total de unidades semânticas dividido pelo total de palavras em uma amostra e a *idea efficiency* foi definida como o total de unidades semânticas divididas pela duração da fala em segundos. Além do fato de estudar pacientes com o status confirmado do ponto de vista neuropatológico, o estudo traz o interesse de usar a linguagem para estudar longitudinalmente a perda cognitiva desde a normalidade até a condição patológica da DA.

2.1.2 Comprometimento Cognitivo Leve

O comprometimento cognitivo leve é difícil de identificar. Geralmente são utilizados testes neuropsicológicos, por serem mais sensíveis, não existindo uma norma para o valor do ponto de corte. Essa dificuldade decorre pelo fato de ser uma situação entre o envelhecimento normal e a demência (FROTA *et al.*, 2011). Em Frota *et al.* (2011) são sugeridas as principais características utilizadas para identificação do CCL:

- Queixa de alteração cognitiva relatada pelo paciente;
- Evidência de comprometimento cognitivo em um ou mais dos seguintes domínios: memória, função executiva, linguagem e habilidades visual-espaciais;
- Preservação da independência funcional; e
- Não preenche critérios de demência.

O CCL tem sido descrito como uma condição pré-clínica da DA, sendo que um grande número de pacientes convertem para o quadro de demência. Mas como o funcionamento cognitivo pode se reverter para um estado normal e compatível com indivíduos da mesma faixa etária e nível de escolaridade o CCL é definido como uma condição heterogênea (FROTA *et al.*, 2011).

Recentemente, há evidências de que indivíduos com CCL têm mais risco para desenvolver DA, devido a comprometimentos em múltiplos domínios, incluindo a linguagem. Por essa razão é importante compreender a natureza do comprometimento de linguagem. São citados abaixo alguns trabalhos que trazem características da produção da linguagem, foco de interesse deste projeto.

Em Fleming e Harris (2008) foi realizado um estudo comparativo do discurso produzido por pacientes com CCL e idosos normais. Os autores observaram que o discurso produzido por pacientes com CCL contem um número menor de palavras, e as suas características se comparam com os estágios iniciais de DA. Enquanto que em Chapman *et al.* (2002) foram comparadas as habilidades de compreensão, memória e expressão de texto discursivo extenso. Os autores identificaram que a capacidade de fornecer informações detalhadas e realizar a síntese de ideias a partir das narrativas estavam comprometidas quando comparadas com as habilidade dos pacientes normais, sendo muito similar à de pacientes acometidos por DA. A partir desses trabalhos é possível notar a dificuldade da identificação de pacientes com CCL e DA em estágios iniciais.

Hodges *et al.* (1996) examinaram o desempenho de controles saudáveis e indivíduos com diversos graus de comprometimento de DA em tarefas de nomeação e geração de definições e reconheceram que a qualidade da definição produzia diferenças entre os grupos.

Os estudos sobre descrição (oral e escrita) de figuras simples e complexas realizado por Forbes-McKay e Venneri (2005) também distinguem indivíduos com DA em grau leve de indivíduos sadios.

Os testes de nomeação, geração de definição e produção de fala espontânea e descrição de figuras mostram que os indivíduos com CCL têm comprometimento semântico, embora nem sempre seja possível distinguir esse grupo dos idosos sadios. Porém, o fato de detectarem diferenças nos pacientes em estágio leve da DA, aguçou o interesse pela possibilidade de aplicação dos mesmos testes nos CCL.

Em resumo, para avaliar a linguagem de indivíduos com CCL é importante dispor de instrumentos sensíveis para detectar déficits sutis. Além disso, o monitoramento dessas dificuldades também carece de instrumentos acurados. A análise do discurso mostra-se interessante, pois abrange os diferentes componentes da linguagem, em uma perspectiva linguístico-cognitiva, cujo declínio é típico das condições mencionadas – envelhecimento saudável, Comprometimento Cognitivo Leve e Doença de Alzheimer. Porém, seu emprego somente será viável se dispusermos de instrumentos que permitam a organização de uma base de dados com número representativo

de informantes.

2.2 Testes Neuropsicológicos

Nesta seção, são apresentados três testes neuropsicológicos para identificação de demências que incluem a tarefa de reconto como subteste. Um desafio na escolha da avaliação neuropsicológica de indivíduos com DA e CCL, com a indicação do grau de comprometimento cognitivo, é o uso de uma bateria que consiga distinguir estes indivíduos.

2.2.1 Bateria Arizona para Desordens da Comunicação e Demência (ABCD)

A *ABCD*[2] foi desenvolvida por fonoaudiólogas americanas para fornecer uma medida funcional de linguagem dos pacientes residentes em instituições de longa permanência e medir a gravidade da demência (FREITAS, 2010). A *ABCD* fornece informações abrangentes sobre orientação, memória episódica, aspectos visual-espaciais e sobre processos de recepção e produção da linguagem oral e escrita. No que diz respeito à linguagem, avalia tarefas básicas, que estão presentes na rotina de pacientes com demência, como reproduzir uma breve história, nomear um objeto, repetir uma frase, definir um conceito, entre outras tarefas (FREITAS, 2010).

Em resumo, a *ABCD* é um conjunto de testes utilizados para identificar pacientes com demências e quantificar o grau da demência. Esse conjunto de testes fornece informações sobre cognição, orientação, memória e comunicação funcional. Para sua utilização são necessários testes preliminares identificando a presença de afasia grave, apraxia, agnosia ou analfabetismo, com o objetivo de excluir pacientes que possam comprometer o resultado da ABCD (FREITAS, 2010).

Especificamente, a *ABCD* é composta por 17 subtestes divididos em 5 domínios (ou construtos): Estado Mental, Memória Episódica, Expressão Linguística, Compreensão Linguística e Construção Visual-espacial, descritos abaixo.

- Estado Mental: é composto por um subteste de 13 perguntas sobre conhecimentos gerais e orientação espacial e temporal;
- Memória Episódica: composto por 5 subtestes, sendo eles reconto imediato e tardio de história, aprendizado de palavras com evocação livre e com pistas e aprendizado de palavras;
- Expressão Linguística: é formado pelos subtestes de descrição de objetos, nomeação por confrontação visual, geração de nomes e definição de conceitos;

[2] A tradução e adaptação para o português foi realizada por Danielle Rüegg, Isabel Maranhão de Carvalho, Letícia Lessa Mansur e Márcia Radanovic.

- Compreensão Linguística: fazem parte desse domínio os subtestes com questões comparativas, seguimento de ordens, repetição, compreensão de leitura de palavra e compreensão de leitura de sentença;

- Construção Visual-espacial: é composto pelos substeste de geração de desenho e de cópia de três figuras.

Neste trabalho de pesquisa, temos especial interesse nos avanços computacionais que foram devotados a um dos construtos - a Memória Episódica, pois nele há 2 de 5 subtestes que podem mostrar como avançar o estado da arte em métodos automáticos de alinhamento monolíngue de palavras: o reconto imediato de histórias e o reconto tardio de histórias. Estes dois subtestes envolvem uma história curta que é contada ao paciente e o teste pressupõe que se avalie o número de elementos da história que podem ser relembrados pelo paciente imediatamente após a contação da história e 30 minutos depois, sendo que durante os 2 testes há a aplicação de outros testes de natureza diferente. Os outros testes do construto da Memória Episódica devem ser avaliados manualmente, para que se complete a sua avaliação.

2.2.2 Bateria de Avaliação da Linguagem no Envelhecimento (BALE)

A BALE é uma das poucas baterias desenvolvidas por pesquisadores brasileiros (HÜBNER *et al.*, 2019). Foi elaborada para abordar algumas das deficiências de linguagem geralmente associadas ao CCL e à DA. Além disso, as tarefas foram desenvolvidas de modo a possibilitarem a administração junto a analfabetos e pessoas com menor escolaridade, amostras populacionais muito comuns no sistema público de saúde brasileiro.

A bateria é composta por dez tarefas de avaliação da linguagem que englobam desde o nível da palavra ao do texto, abrangendo tanto a produção quanto a compreensão, descritos abaixo.

- Hábitos de leitura e escrita dos pacientes: são levantados os hábitos de leitura e escrita dos pacientes;

- Compreensão de frases: é avaliada a compreensão escrita por meio da leitura de uma frase e posterior execução da ação solicitada na frase;

- Aprendizagem e recordação de figuras: Nesta subtarefa o participante deve aprender/memorizar e recordar 16 figuras;

- Reconto e compreensão de discurso oral: Nesta subtarefa o participante deve memorizar uma história contada oralmente e, em seguida, reproduzir a história também de forma oral. Na sequência, deve responder oralmente a questões de compreensão da história ouvida (História da Lúcia);

- Discurso oral: Nesta subtarefa o participante deve produzir oralmente uma história com viés cômico que tenha experienciado, presenciado, ou que tenham contado a ele. Além disso, deve produzir oralmente uma notícia (informativa) que tenha ouvido ou lido recentemente;
- Narrativa oral baseada em sequência de figuras: Nesta subtarefa o participante deve produzir oralmente uma narrativa com começo, meio e fim, tendo como base uma sequência de figuras relacionadas entre si, já apresentadas na sequência correta (História do Cachorro);
- Fluência verbal: Nesta subtarefa o participante deve produzir oralmente o maior número possível de nomes dentro da categoria semântica "animais" durante um minuto. Na segunda etapa devem ser produzidas palavras que comecem com o fonema /p/;
- Nomeação, designação e reconhecimento de figuras: essa subtarefa é dividida em três partes, sendo que na primeira parte, o participante deve nomear 60 itens, entre figuras animadas e inanimadas. Na segunda parte, é preciso identificar e designar dentre 10 figuras da mesma categoria o item solicitado oralmente pelo avaliador. Na terceira parte, o participante deve identificar dentre 10 figuras, os dois estímulos que não apareceram na primeira parte da tarefa;
- Associação semântica: Nesta subtarefa o participante deve estabelecer a relação semântica entre a figura principal apresentada e outra figura (alvo); e
- Conhecimento semântico: Nesta subtarefa são contadas metáforas e provérbios populares e o participante deve explicar os seus significados.

Neste trabalho estamos interessado na subtarefa de reconto e compreensão de texto de discurso oral, a História da Lúcia, que possui 24 unidades de informação.

2.2.3 Wechsler Memory Scale (WMS)

O teste *Wechsler Memory Scale* é utilizado para quantificar a memória visual de trabalho e a memória auditiva e visual de forma imediata e tardia. A sua versão atual *WMS*-IV é composta por 10 subtestes. O teste só pode ser aplicado por psicólogos, isto é, é realmente restrito ao uso deste profissional, o que inviabiliza a sua aplicação de forma computacional por uma equipe que não possua este profissional.

No subteste *Verbal Paired Associates* são apresentadas de forma oral pares de palavras para o paciente. Em seguida, o examinador lê a primeira palavra e pergunta para o paciente qual é palavra correspondente de cada par. Este subteste é utilizado para avaliar a memória auditiva.

Também faz parte desse conjunto o subteste *Logical Memory*, que é de especial interesse para essa pesquisa, sendo executado do seguinte modo (WECHSLER, 1997; HOELZLE; NELSON; SMITH, 2011; PRUD'HOMMEAUX, 2012): o sujeito sendo avaliado ouve uma breve história, mostrada na Figura 3 e, em seguida, deve recontar a história imediatamente após ouvi-la

(Memória Lógica I) e uma segunda vez após um intervalo de 30 minutos (Memória Lógica II). Para fins de pontuação, o texto é segmentado em 25 elementos, indicados na Figura 3 com barras. Durante o exame, o examinador observa quais elementos da história foram recuperados e relata uma pontuação resumida, que é simplesmente o número total de elementos relembrados da história. Os elementos podem ser parafraseados e eles não precisam estar em ordem. As diretrizes de pontuação descrevem substituições permitidas, como "*Ann*" para o elemento, "*Anna*".

A Figura 4 mostra um exemplo de reconto de Memória Lógica I, que inclui os elementos "*Anna*", "*employed*", "*Boston*", "*as a cook*", "*was robbed of*", "*she had four*", "*small children*", "*reported*", "*station*", "*touched by the woman's story*", "*took up a collection*", e "*for her*". P ara o reconto, a pontuação resumida foi de 12, de acordo com as diretrizes publicadas para pontuação manual. Observe a alteração da ordem de alguns dos elementos da história ("*worked*" aparece antes de "*Boston*"); o assunto fora de tópico ("*Is that right?*"); detalhamentos ("*so that she can feed the children*").

Figura 3 – Narrativa do exame *Wechsler Logical Memory*.

Anna / Thompson / of South / Boston / employed / as a cook / in a school / cafeteria / reported / at the police / station / that she had been held up / on State Street / the night before / and robbed of / fifty-six dollars. / She had four / small children / the rent was due / and they hadn't eaten / for two days. / The police / touched by the woman's story / took up a collection / for her.

Fonte: Prud'hommeaux e Roark (2011).

Figura 4 – Exemplo de um reconto de narrativa do *WLM*.

Ann Taylor worked in Boston as a cook. And she was robbed of sixty-seven dollars. Is that right? And she had four children and reported at the some kind of station. The fellow was sympathetic and made a collection for her so that she can feed the children.

Fonte: Prud'hommeaux e Roark (2011).

Com objetivo de avaliar a memória visual é utilizado o subteste *Visual Reproduction* em que são exibidas figuras geometricamente complexas. Após, é solicitado para o paciente selecionar a figura que tinha sido apresentada anteriormente a partir de um conjunto de desenhos (BRITO-MARQUES; CABRAL-FILHO; MIRANDA, 2009). Também é utilizado o subteste *Designs*, em que é mostrado para o paciente um tabuleiro de dimensão 4x4, e neste algumas posições contêm figuras. Após, é entregue para o paciente um tabuleiro em branco e cartões com as figuras, e então solicitado para o paciente selecionar e colocar as figuras nas posições corretas (WECHSLER, 1997).

Para a avaliar a memória de trabalho, é utilizado o subteste *Spatial Addition* que consiste em exibir para o paciente dois tabuleiros de tamanho 4x4. Nestes, algumas posições contêm

pontos azuis e/ou vermelhos. Após, é entregue um tabuleiro sem marcações e cartões com pontos brancos, azuis e vermelhos. Em seguida, é solicitado para o paciente colocar os pontos azuis no tabuleiro onde ele viu pontos azuis e branco nas posições onde ambos os tabuleiros continham pontos azuis. Também é utilizado o *Symbol Span*, sendo que neste subteste é mostrado para o paciente uma sequência de símbolos e, em seguida, o paciente deve selecionar os símbolos na ordem correta a partir de um conjunto de símbolos (WECHSLER, 1997).

2.3 Classificação Monorrótulo e Multirrótulo

O objetivo de um algoritmo de aprendizado de máquina é aprender a mapear uma entrada $\mathbf{x}_i$ para uma saída y_i, dado um conjunto de dados $\mathscr{D} = \{(\mathbf{x}_i, y_i)_i^N\}$, onde N é quantidade de pares de entrada e saída, $\mathbf{x}_i$ é um vetor D-dimensional formado por valores contínuos ou categóricos chamados de atributos (*features*), variáveis, ou covariáveis (MURPHY, 2012).

Na tarefa de classificação, a resposta y_i é um valor categórico. Caso a resposta seja um valor real, o problema é chamado de regressão. Se a reposta for ordinal, então chamamos o problema de ranqueamento (*ranking*) (MURPHY, 2012).

Tradicionalmente, os problemas de classificação são monorrótulo, ou seja, cada exemplo possui um único rótulo, podendo ser: i) binário quando existem apenas duas classes, ii) ou multiclasse, quando temos mais que duas classes. Porém, em algumas aplicações os exemplos naturalmente podem estar associados a mais de um rótulo. Por exemplo, um texto pode possuir mais que um tópico (MCCALLUM, 1999), ou podemos estar interessados em categorizar as sentenças de um resumo em unidades retóricas (DAYRELL *et al.*, 2012); músicas podem estar associadas a diferentes gêneros musicais (SANDEN; ZHANG, 2011; ORAMAS *et al.*, 2018); imagens podem conter diferentes objetos (BOUTELL *et al.*, 2004; WANG *et al.*, 2016). Nos casos elencados acima, o problema é chamado de multirrótulo, e o objetivo é mapear um exemplo $\mathbf{x}_i$ para um subconjunto de rótulos Y_i, onde $Y_i \subseteq L$ e $L = \{y_j : j = 1 \dots q\}$ é o conjunto dos q rótulos simples que participam dos multirrótulos $Y_i, i = 1 \dots N$ (TSOUMAKAS; KATAKIS; VLAHAVAS, 2009).

A seguir, são apresentados os principais métodos para a construção de classificadores multirrótulo.

2.3.1 Métodos Multirrótulos

A grande maioria dos algoritmos de aprendizado de máquina foram desenvolvidos para problemas binários ou multiclasse; estes não conseguem lidar naturalmente com problemas multirrótulos, sendo necessárias adaptações dos métodos já existentes ou a proposição de novos algoritmos. Esses métodos podem ser divididos em: Transformação de Problema e Adaptação de Algoritmos (TSOUMAKAS; KATAKIS; VLAHAVAS, 2009; ZHANG; ZHOU, 2013).

Na transformação de problema, são utilizadas diferentes formas de transformar um problema multirrótulo em um ou mais problemas monorrótulo. Já na adaptação de algoritmos são criados novos métodos para lidar diretamente com problemas multirrótulo.

2.3.1.1 Transformação de Problema

As abordagens de transformação de problemas alteram a forma como o rótulo é representado, decompondo o problema, e possibilitando a utilização de qualquer algoritmo de classificação. As técnicas mais famosas são: *Label Powerset* e o *Binary Relevance*. Para exemplificar os métodos de transformação de problemas, vamos utilizar como exemplo a Tabela 1, em que representamos um conjunto de dados $\mathscr{D}$ multirrótulo, com quatro exemplos rotulados com um ou mais rótulos pertencentes ao conjunto $L = \{y_1, y_2, y_3, y_4\}$.

Tabela 1 – Exemplo de conjunto de dados multirrótulo.

Exemplo	Atributos	Y
1	$\mathbf{x}_1$	$Y_1 = \{y_1, y_4\}$
2	$\mathbf{x}_2$	$Y_2 = \{y_3, y_4\}$
3	$\mathbf{x}_3$	$Y_3 = \{y_1\}$
4	$\mathbf{x}_4$	$Y_4 = \{y_2, y_3, y_4\}$

Fonte: Tsoumakas, Katakis e Vlahavas (2009).

O método *Label Powerset* (*LP*) transforma o problema multirrótulo em um problema multiclasse, em que cada novo monorrótulo é composto pela concatenação dos rótulos Y_i. Na Tabela 2, mostramos o resultado do método *LP* no conjunto de dados da Tabela 1. Por exemplo, na primeira linha o subconjunto de rótulos $Y_1 = \{y_1, y_4\}$ é transformado no monorrótulo $y_{1,4}$. Após a transformação, o nosso conjunto de dados utilizado como exemplo será composto pelas seguintes classes $L = \{y_{1,4}, y_{3,4}, y_1, y_{2,3,4}\}$.

Tabela 2 – Exemplo de conjunto de dados *Label Powerset*.

Exemplo	Y
1	$y_{1,4}$
2	$y_{3,4}$
3	y_1
4	$y_{2,3,4}$

Fonte: Tsoumakas, Katakis e Vlahavas (2009).

Esse método é simples, mas cada combinação de classes é substituída por um monorrótulo, podendo chegar até 2^q combinações. Dessa forma, podemos criar um número muito grande de classes, sendo que teremos poucos exemplos associados a combinações menos frequentes, tornando o processo de aprendizado mais complexo.

Tabela 3 – Exemplo de conjunto de dados *Binary Relevance*.

Exemplo	Y
1	y_1
2	$\neg y_1$
3	y_1
4	$\neg y_1$

Exemplo	Y
1	$\neg y_2$
2	$\neg y_2$
3	$\neg y_2$
4	y_2

Exemplo	Y
1	$\neg y_3$
2	y_3
3	$\neg y_3$
4	y_3

Exemplo	Y
1	y_4
2	y_4
3	$\neg y_4$
4	y_4

Fonte: Tsoumakas, Katakis e Vlahavas (2009).

Outro método simples e popular na literatura é o *Binary Relevance* (*BR*). Esse método adota a estratégia *one-vs-all* utilizada em classificação multiclasse, em que o conjunto de dados é transformado em problemas binários independentes.

Na Tabela 3 é apresentado a criação de quatro conjuntos de dados com a mesma quantidade de exemplos do conjunto original, mas para cada conjunto os exemplos são rotulados como positivo se eles estiverem associados a esse rótulo ou negativo caso não estejam associados.

Essa técnica possibilita a aplicação de qualquer algoritmo de aprendizado binário, mas tem como desvantagem a utilização de todos os exemplos do conjunto em q processos de aprendizado, e também não é considerada nenhuma relação de dependência entre os rótulos.

2.3.1.2 Métodos de Adaptação de Algoritmo

Os métodos de adaptação de algoritmos têm como objetivo o desenvolvimento de algoritmos para lidar diretamente com problemas multirrótulos sem a necessidade de transformar o problema em monorrótulo. A seguir, são comentadas brevemente algumas adaptações de técnicas populares para o cenário multirrótulo:

- Algoritmos baseados em árvores: em geral são realizadas adaptações no cálculo da entropia e alterações nas folhas para retornar o subconjunto de rótulos (CLARE; KING, 2001; KOUZANI; NASIREDING, 2009; AGRAWAL *et al.*, 2013)

- *K- Nearest Neighbor* (*KNN*): uma das adaptações mais famosas é o *ML-KNN* que combina as probabilidade a priori e posteriori dos rótulos (ZHANG; ZHOU, 2007). O primeiro passo do método é obter a frequência de cada classe no conjunto de treinamento, isto é, as probabilidades a priori. Na etapa de inferência são obtidos os K exemplos mais próximos de um exemplo não visto, em seguida as probabilidades a posteriori, e finalmente são combinadas as probabilidades a priori e posteriori para determinar as classes.

- *Support Vector Machine* (*SVM*): Elisseeff e Weston (2002) propuseram a criação de um hiperplano para cada classe e a função de custo foi adaptada para ranquear os rótulos. Recentemente, Chen *et al.* (2016) estenderam o *Twin Support Vector Machine* (KHEMCHANDANI; CHANDRA *et al.*, 2007) para o problema multirrótulo.

- Redes neurais: os primeiros trabalhos utilizavam uma arquitetura simples baseadas em *Perceptrons* e adaptações do algoritmo *back-propagation* para o cenário multirrótulo (ZHANG; ZHOU, 2006), mas nos últimos anos essas técnicas foram superadas por arquiteturas mais complexas que utilizam Redes Recorrentes para predizer os multirrótulos (WANG *et al.*, 2016; NAM *et al.*, 2017)

2.3.2 Medidas de Avaliação de Classificadores

Na avaliação de classificadores binários, podemos utilizar a matriz de confusão, a qual ilustra o número de predições corretas e incorretas em cada classe.

Na Tabela 4, é apresentada uma matriz de confusão para um problema de classificação binária, onde temos uma classe positiva representada como $C+$ e uma classe negativa representada como $C-$. Nessa tabela, Verdadeiro Positivo (VP) corresponde ao número de exemplos da classe $C+$ que foram classificados corretamente, e Verdadeiro Negativo (VN) o número de exemplos classificados corretamente pertencente a classe $C-$. Já o valor de Falso Negativo (FN) e Falso Positivo (FP) correspondem ao número de exemplos classificados incorretamente como $C+$ e o número de exemplos classificados como $C-$ incorretamente, respectivamente (HAN; PEI; KAMBER, 2011).

Tabela 4 – Matriz de confusão.

Classe	**Predita C+**	**Predita C−**
Verdadeira $C+$	VP	FN
Verdadeira $C-$	FP	VN

A partir da matriz de confusão podem ser obtidas diversas medidas para avaliar o modelo, como a precisão, acurácia, *recall*, $F1$-*score*. A seguir, são apresentadas algumas dessas medidas (HAN; PEI; KAMBER, 2011).

Acurácia é calculada pela soma dos valores da diagonal principal da matriz, dividida pela a soma dos valores de todos os elementos da matriz. O erro pode ser obtido pelo complemento da acurácia.

$$Acc = \frac{VP+VN}{VP+VN+FP+FN} \tag{2.1}$$

Precisão é a proporção de exemplos positivos classificados corretamente entre todos aqueles preditos como positivos.

$$Pr = \frac{VP}{VP+FP} \tag{2.2}$$

Recall representa uma medida de perfeição, sendo o seu valor a fração de exemplos classificados corretamente pelo numero de exemplos positivos. Também chamada de taxa de verdadeiros positivos.

$$Rec = \frac{VP}{VP+FN} \tag{2.3}$$

Uma outra forma usual de avaliar é combinar a precisão e o *recall* em uma única métrica, utilizando uma média harmônica. Essa medida é chamada de $F1$-*score* (conhecida também F-*score*) e F_β.

$$F1 = \frac{2 \times Pr \times Re}{Pr + Re} \tag{2.4}$$

$$F_\beta = (1+\beta^2) \times \frac{Pr \times Re}{\beta^2 \times Pr + Re} \tag{2.5}$$

onde β é valor não negativo, utilizado para ponderar a importância do *recall*, desse mdo, β representa quantas vezes o *recall* é mais importante que a precisão. No $F1$-*score recall* e precisão têm o mesmo peso.

Em tarefas multirrótulo com $L = \{y_1, y_2, \ldots, y_q\}$ é possível utilizar as medidas descritas anteriormente para avaliar os classificadores combinando com duas operações de médias: *macro-averaged* e *micro-averaged* (TSOUMAKAS; KATAKIS; VLAHAVAS, 2009).

Sendo B uma medida monorrótulo qualquer calculada com os números de VP, VN, FP, e FN, e sendo $TP_{y_i}, FP_{y_i}, VN_{y_i}$, e FP_{y_i} o número de verdadeiros positivos, falsos positivos, verdadeiros negativos, e falsos positivos para um rótulo y_i. As versões de *macro-averaged* e *micro-averaged* para essa medida B são definidas pelas seguintes equações:

$$B_{macro} = \frac{1}{q} \sum_{i=1}^{q} B(VP_{y_i}, FP_{y_i}, VN_{y_i}, FN_{y_i}) \tag{2.6}$$

$$B_{micro} = B(\sum_{i=1}^{q} VP_{y_i}, \sum_{i=1}^{q} FP_{y_i}, \sum_{i=1}^{q} VN_{y_i}, \sum_{i=1}^{q} FN_{y_i}) \tag{2.7}$$

Conceitualmente, a *macro-averaged* atribui um peso igual para todas as classes e a *micro-averaged* atribui um peso igual para cada exemplo (ZHANG; ZHOU, 2013). Para se obter uma medida B *macro-averaged*, é calculado o seu valor para cada rótulo individualmente e retornada a média sobre esses valores. Enquanto a *micro-averaged* soma os dividendos e divisores que compõem alguma métrica B para calcular um quociente de forma global. Assim, a abordagem *macro-averaged* é mais sensível ao desempenho de categorias raras, enquanto a abordagem *micro-averaged* é mais afetada pelas principais categorias (TANG; RAJAN; NARAYANAN, 2009).

Além dessas medidas, existem outras que avaliam o subconjunto de rótulos preditos e o subconjunto de rótulos verdadeiros do exemplo como um todo, e não utilizam operações de média com alguma medida binária. Umas dessas medidas é a *Hamming-Loss*, definida na

Equação 2.8.

$$Hamming\text{-}Loss = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \Delta Z_i|}{|L|} \tag{2.8}$$

onde Z_i é conjunto de rótulos preditos, Y_i é conjunto de rótulos verdadeiros, $|L|$ é quantidade de rótulos, N é quantidade de exemplos, e Δ representa a diferença simétrica entre dois conjuntos. Essa medida avalia a fração de rótulos faltantes ou preditos de forma incorreta no exemplo. No caso de uma classificação perfeita, a *Hamming-Loss* é zero (TSOUMAKAS; KATAKIS; VLAHAVAS, 2009; ZHANG; ZHOU, 2013).

Outra medida bastante utilizada na literatura é a *SubsetAccuracy*, definida na Equação 2.9. Essa medida é bem conservadora, pois o conjunto multirrótulo predito precisa ser exatamente igual ao conjunto multirrótulo verdadeiro, não levando em conta acertos parciais (TSOUMAKAS; KATAKIS; VLAHAVAS, 2009; ZHANG; ZHOU, 2013).

$$SubsetAccuracy = \frac{1}{N}\sum_{i=1}^{N} I(Y_i = Z_i) \tag{2.9}$$

onde I retorna 1 se o conjunto de rótulos for exatamente igual, e 0 caso contrário.

2.4 Abordagens para Representação de Textos

Nesta seção, são apresentadas algumas abordagens utilizadas para representar um texto na tarefa de classificação, em geral. Para uma melhor organização, dividimos as seções em 4 categorias: representações tradicionais (Seção 2.4.1); uso de métricas de complexidade e coerência textual para representação (Seção 2.4.2); representação de textos em redes complexas e sua caracterização topológica (Seção 2.4.3); e representações com modelos densos (Seção 2.4.4).

2.4.1 Representações Tradicionais

A forma mais tradicional de representar um texto na tarefa de classificação é o modelo de espaço vetorial (*vector-space model*). Usualmente, usa-se a palavra *termo* para denotar uma palavra ou uma sequência de palavras (*n-grams*). Nessa representação cada termo é representado como um atributo. Dado um conjunto $\mathscr{D}$ contendo N documentos (exemplos), em que M é a quantidade de termos presentes em $\mathscr{D}$, cada exemplo é representado como um vetor com M dimensões, $\mathbf{x}_i = (x_{i1}, \ldots, x_{iM})$, onde x_{ij} pode representar a presença ou ausência desse termo ou a ocorrência do termo no exemplo. Essa representação é conhecida como *Bag-of-Words* (*BoW*), sendo que a ordem de ocorrência de cada palavra no documento não é considerada (MANNING; MANNING; SCHÜTZE, 1999; MANNING; RAGHAVAN; SCHÜTZE, 2008).

Outra forma de ponderar um termo é pela frequência do termo pelo inverso do número de documentos em que o termo ocorre. Essa técnica é chamada de *Term Frequency-Inverse document Frequency* (*TF-IDF*).

O modelo de espaço vetorial apresenta como desvantagens a alta dimensionalidade causada pela grande número de termos, e alta esparsidade, pois grande parte dos atributos terão valores zero devido à baixa ocorrência dos termos. Além disso, não representa a ordem das palavras ou a semântica (MANNING; MANNING; SCHÜTZE, 1999; MANNING; RAGHAVAN; SCHÜTZE, 2008).

2.4.2 Métricas de Complexidade e Coerência Textual

Como a frequência das palavras não traz muitas informações, pode-se extrair características linguísticas dos textos para suprir essa carência. Sistemas computacionais foram desenvolvidos com objetivo de facilitar a extração de métricas para representar os textos, e entre esses sistemas, o mais famoso é o Coh-Metrix (MCNAMARA; LOUWERSE; GRAESSER, 2002; GRAESSER; MCNAMARA; KULIKOWICH, 2011), desenvolvido para a língua inglesa. Coh-Metrix pode ser usado para várias tarefas do Processamento de Línguas Naturais, por exemplo, para investigar a coesão de um dado texto e a coerência de sua representação mental. Para o português, temos o *Coh-Metrix-Port*, o qual foi criado via adaptações das métricas e recursos do Coh-Metrix para Português Brasileiro (SCARTON; ALUÍSIO, 2010; SCARTON; GASPERIN; ALUISIO, 2010), tendo sido desenvolvido como parte do projeto PorSimples (ALUISIO; GASPERIN, 2010). Inicialmente foi criado com 48 métricas validadas para a tarefa de classificação de textos infantis versus para adultos, e atualmente conta com 46 métricas em sua versão 3.0, tendo todas as métricas validadas por uma linguista[3] e uma documentação bastante completa[4].

Cunha (2015) desenvolveu o *Coh-Metrix-Dementia*, que é uma expansão do *Coh-Metrix-Port* para a tarefa de classificação de narrativas usadas em testes neuropsicológicos de idosos com envelhecimento normal (controles) e portadores de CCL e DA. O novo sistema possui 73 métricas, sendo 48 métricas reimplementadas da versão inicial do *Coh-Metrix-Port* e 25 métricas adicionadas, principalmente do nível sintático, dada a carência destas no conjunto inicial.

Após 10 anos do início do projeto PorSimples, em 2018, todas as métricas desenvolvidas para estudar a complexidade textual em projetos de alunos do NILC foram agrupadas em um único site, com cada uma testada e documentada. O conjunto das métricas se chama NILC-Metrix[5], agrupa as métricas de complexidade textual desenvolvidas no início do projeto PorSimples (ALUISIO; GASPERIN, 2010), algumas desenvolvidas no projeto Coh-Metrix-Dementia e outras no convênio Guten-ICMC.

[3] Agradecemos a atenção e trabalho cuidadoso de Magali Sanches Duran.
[4] <http://fw.nilc.icmc.usp.br:23380/cohmetrixport>
[5] <https://simpligo.sidle.al/>

Em geral, esses sistemas podem ser acomodados em *frameworks* teóricos (GRAESSER; MCNAMARA, 2011), que organizam as métricas em cinco níveis linguísticos: (1) palavras, (2) sintaxe, (3) base textual, (4) modelo situacional, (5) gênero e estrutura retórica. A seguir, são apresentadas algumas métricas que podem ser extraídas de cada nível.

Nesta tese são utilizadas apenas as métricas atreladas aos níveis de palavras, sintaxe e base textual. Para o Português Brasileiro, não foram desenvolvidas as métricas para os níveis de modelo situacional, e gênero e estrutura retórica.

2.4.2.1 Palavras

Pode-se analisar as palavras pela quantidade de categorias gramaticais ou morfossintáticas presentes nos textos, por exemplo, a quantidade de substantivos, verbos, preposições, palavras de conteúdo, pronomes, conectivos, entre outras.

A frequência de ocorrência das palavras fornece informações sobre a compreensão dos textos, dado que a utilização de palavras frequentes torna o texto mais simples de se entender, entretanto, o emprego de somente palavras com alta frequência pode indicar textos pobres; ao passo que uma única palavra com baixa frequência em uma sentença pode comprometer o entendimento dela. Para computar métricas de ocorrências de palavras é necessário coletar a frequência das palavras em grandes corpora (GRAESSER; MCNAMARA; KULIKOWICH, 2011).

Outros exemplos são as medidas psicolinguísticas das palavras, como idade de aquisição, concretude, imageabilidade, e familiaridade. Medidas como a concretude informam se o texto trata de objetos, pessoas, lugares ou coisas que podem ser experienciadas pelos sentidos; um texto com muitas palavras concretas é mais simples enquanto que se a maioria forem palavras abstratas pode ser complexo. Medidas como a idade de aquisição e familiaridade podem fornecer indicativos sobre a complexidade do vocabulário (GRAESSER; MCNAMARA, 2011). Já a imageabilidade é a facilidade e rapidez com que uma palavra evoca uma imagem mental e é correlacionada com concretude.

2.4.2.2 Sintaxe

Nessa categoria estão as métricas de complexidade das sentenças, por exemplo, sentenças curtas que seguem a estrutura sujeito-verbo-objeto são consideradas mais simples do que sentenças que possuem muitos modificadores por sintagma.

É possível extrair métricas como a quantidade de palavras antes do verbo principal ou a quantidade de modificadores por sintagma nominal. Outras métricas analisam a estrutura sintática como: a complexidade de Yngve (YNGVE, 1960) é uma métrica que utiliza a árvore sintática para medir o quanto a árvore está ramificando para a esquerda; a complexidade de Frazier (FRAZIER, 1985) é similar à de Yngve, mas Frazier propôs dividir a sentença em

trigramas para fazer o cálculo. Para calcular a complexidade da sentença, calcula-se a soma das pontuações das palavras em cada trigrama, usando o máximo dessas somas numa varredura da esquerda pra direita. A complexidade de um texto é a média da complexidade de Frazier para cada sentença; e a distância de dependência (ROARK; MITCHELL; HOLLINGSHEAD, 2007) utiliza uma árvore de dependências para realizar o cálculo; a cada relação de dependência está associada uma distância entre as palavras na superfície textual. Estudos da literatura mostram que essas distâncias entre palavras nas relações de dependência são diretamente proporcionais ao tempo de processamento em tarefas de compreensão de sentenças, sendo que grandes distâncias entre palavras relacionadas geram sobrecarga de memória.

Detalhes com ilustrações de como são computadas essas métricas podem ser encontradas em Cunha (2015).

2.4.2.3 Base Textual

A base textual se refere ao sentido das ideias explicitadas no texto ao invés da superfície textual. Nesse nível, um aspecto importante é a coesão referencial, que ocorre quando um nome, pronome, ou sintagma nominal se refere a outro constituinte na base textual.

Neste sentido, existem métricas que analisam a correferência, entre elas: sobreposição de palavras de conteúdo, sobreposição de nomes, sobreposição de argumentos, e sobreposição de radical (GRAESSER; MCNAMARA; KULIKOWICH, 2011).

Outro fator importante é a diversidade lexical, que se relaciona à coesão, pois o emprego de uma grande quantidade de palavras distintas no texto implica em uma maior quantidade de conceitos a serem integrados ao contexto discursivo. Uma métrica que explora essa relação é a razão tipo por *token* (TTR), que consiste no número de palavras distintas presente em um texto (os tipos) dividido pelo número total de palavras (os *tokens*).

2.4.2.4 Modelo Situacional

Esse nível envolve a construção de um modelo situacional (ou modelo mental) que representa o texto, estando além das palavras explícitas. Por exemplo, em narrativas o modelo situacional inclui personagens, objetos, cenário espacial, ações, eventos, processos, planos, pensamentos e emoções dos personagens (GRAESSER; MCNAMARA; KULIKOWICH, 2011). Em textos descritivos, o modelo situacional inclui os assuntos que estão sendo descritos (GRAESSER; MCNAMARA; KULIKOWICH, 2011). Zwaan e Radvansky (1998) propuseram cinco dimensões do modelo situacional de textos narrativos: causalidade, intencionalidade, tempo, espaço e protagonistas.

O *Coh-Metrix* possui métricas atreladas a casualidade e intencionalidade, como a incidência de verbos causais e verbos intencionais. A coesão temporal é calculada como a média de repetições de tempo e aspecto. Especificamente, as 8 métricas do modelo situacional implemen-

tadas no Coh-Metrix 3.0 (cf. <http://tool.cohmetrix.com/>) são: (1) incidência de verbos causais, (2) incidência de verbos intencionais, (3) proporção de partículas causais e verbos causais, (4) proporção de partículas intencionais e verbos intencionais, (5) sobreposição de verbos no modelo LSA, (6) sobreposição de verbos na Wordnet, (7) média de coesão temporal, repetição de tempo e aspecto, (8) incidência de verbos causais e partículas causais.

2.4.2.5 Gênero e Estrutura Retórica

No *framework* de Graesser, McNamara e Kulikowich (2011) os textos são categorizados de acordo com o gênero: narração, exposição, persuasão, ou descrição. Além do gênero global, um texto possui uma composição retórica que fornece uma organização funcional. Seções, parágrafos e sentenças têm funções discursivas que estão coerentemente vinculadas à macroorganização do texto. No *Coh-Metrix* foi desenvolvido um score que indica a *narratividade* do texto, para isso selecionou-se 53 métricas e aplicou-se a Análise de Componentes Principais (*PCA*, *Principal Component Analysis*) e uma das componentes foi definida como esse score (GRAESSER; MCNAMARA; KULIKOWICH, 2011).

2.4.3 Redes complexas para representação de textos e suas métricas

Além de utilizar a frequência de palavras ou métricas de complexidade e coerência textual, também é possível representar um texto em uma rede complexa. Essa representação vem sendo aplicada com sucesso em diferentes tarefas como: classificação de textos (ARRUDA; COSTA; AMANCIO, 2016), classificação de autoria (MARINHO; HIRST; AMANCIO, 2016), sumarização (ANTIQUEIRA *et al.*, 2009; AMANCIO *et al.*, 2012; TOHALINO; AMANCIO, 2018) e desambiguação de sentidos de palavras (SILVA; AMANCIO, 2012; CORRÊA; LOPES; AMANCIO, 2018).

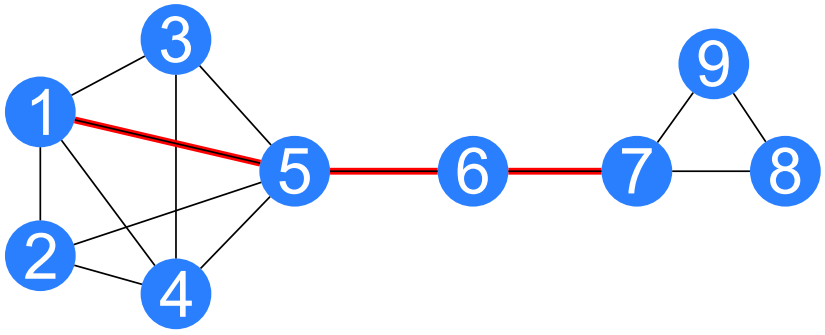
Matematicamente, uma rede é definida como um grafo $G = \{V, E\}$, formada por um conjunto de vértices $V = \{v_1, v_2, ..., v_n\}$, e um conjunto de arestas $E = \{e_1, e_2, ..., e_m\}$. Quando aplicadas ao estudo da linguagem, os vértices representam unidades linguísticas (palavras, morfemas, sentenças, parágrafos), e as arestas alguma relação entre as unidades, como: adjacência (ou co-ocorrência), sintática, e semântica (CANCHO; SOLÉ, 2001; CANCHO; SOLÉ; KÖHLER, 2004; SOLÉ *et al.*, 2010; CONG; LIU, 2014; AMANCIO, 2015a; AMANCIO, 2015b).

As redes de adjacência de palavras são uma representação amplamente usada (CANCHO; SOLÉ, 2001). Nesse modelo, os nós representam as palavras enquanto as arestas conectam as palavras adjacentes. A rede pode ser representada em uma matriz de adjacência A, onde os elementos A_{ij} são iguais a 1 sempre que houver uma aresta conectando os nós (palavras) i e j e, caso contrário, será igual a 0.

A partir das redes complexas, podemos obter diversas métricas que caracterizam a sua topologia. A seguir são apresentadas as métricas utilizadas na tese, e para exemplificar algumas

métricas será utilizado o grafo não direcionado da Figura 5. O grafo possui oito vértices e as arestas não possuem pesos.

Figura 5 – Exemplo de uma rede complexa.



Fonte: Elaborada pelo autor.

Grau É uma das medidas de redes complexas mais simples; o grau é o número de arestas associadas a um vértice. Dada uma matriz de adjacência A, o grau do vértice i é

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{2.10}$$

Por exemplo, na Figura 5 o nó 1 possui grau 4, enquanto que o nó 5 possui grau 5.

Centralidade de autovetor É uma medida que define a importância de um nó com base na soma ponderada da conectividade dos nós vizinhos. Mesmo que um nó i tenha um grau baixo, este pode estar conectado a vizinhos importantes e possuir uma centralidade de autovetor alta.

$$s_i = \lambda_i^{-1} \sum_j A_{ij} s_j \tag{2.11}$$

onde λ_i é o seu maior autovalor e s_j autovetor. No exemplo da Figura 5 o nó 6 possui uma centralidade de autovetor de $0,26$, enquanto que nó 7 possui uma centralidade de autovetor de $0,07$; ambos possuem a mesma quantidade de arestas, entretanto, o nó 6 está conectado ao nó 5 que possui várias conexões.

PageRank É uma medida que reflete a relevância de um nó com base em suas conexões com outros nós relevantes (BRIN; PAGE, 1998). É uma medida utilizada no algoritmo de ranqueamento de páginas da *web* do motor de busca do *Google*.

$$pr_i^{(t)} = \frac{(1-d)}{N} + d \sum_{j \to i} \frac{pr_j^{(t-1)}}{k_j^{out}} \tag{2.12}$$

onde pr_j^{t-1} é o *PageRank* do vértice j, k_j^{out} é o grau de saída do nó j, d é probabilidade de saltos aleatórios, N é número de nós na rede, e $j \to i$ denota a conexão de um vértice

j para o vértice i. O *PageRank* é calculado iterativamente até convergir a um *threshold* dado ou até um número de iterações executadas. Neste trabalho, foi adotado o valor de $d = 0,85$, o valor de convergência de $0,001$ e 1000 iterações.

Comprimento médio dos caminhos mínimos É a distância média entre um nó i e todos os outros nós da rede. Seja $dist(i,j)$ o comprimento de menor caminho entre o nó i e j, podemos obter o comprimento médio l_i pela equação:

$$l_i = \frac{1}{N-1} \sum_{j=1}^{N} dist(i,j) \tag{2.13}$$

Excentricidade Dado um nó i, são obtidos os caminhos mínimos para todos os nós presentes na rede, e a excentricidade é o caminho mínimo mais longo:

$$ecc_i = \max\{dist(i,j)\} \tag{2.14}$$

Na Figura 5, o caminho mínimo mais longo do nó 7 é representado em vermelho, e o seu valor de excentricidade é 3.

Betweenness É uma medida de centralidade que considera que um nó é relevante se for altamente acessado por caminhos mais curtos. A métrica *betweenness* de um nó i é definida como a fração dos caminhos mais curtos que passam pelo nó i.

$$b_i = \sum_{st} \frac{n_{st}^i}{g_{st}} \tag{2.15}$$

onde n_{st}^i é o número de caminhos mais curtos entre os nós s e t passando pelo nó i, g_{st}' é número de menores caminhos entre os nós s e t.

O maior valor de *betweenness* do exemplo da Figura 5 é o nó 5, pois os caminhos mínimos dos nós 1, 2, 3, 4 para os nós 6, 7, 8, e 9 passam pelo nó 5.

Grau médio dos vizinhos de um nó É a média do grau dos vizinhos conectados a um nó i.

$$k_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^{N} A_{ij} k_j \tag{2.16}$$

onde k_i e k_j representam o grau do nó i e j. O valor do grau médio dos vizinhos do nó 8 da rede apresentada na Figura 5 é $2,5$, o nó 8 esta conectado ao nó 9, que possui grau 2, e ao nó 7, que possui grau 3.

Coeficiente de aglomeração Mede a probabilidade de que dois vizinhos de um nó estejam conectados.

$$cc(i) = \frac{2e_i}{k_i(k_i-1)} \tag{2.17}$$

onde e_i representa o número de arestas entre os vizinhos do nó i e k_i é o grau do nó i. Se $k_i = 1$ então $cc(i) = 0$.

Podemos obter o coeficiente de aglomeração da rede inteira, que é a média dos coeficientes de aglomeração dos nós.

$$\langle cc \rangle = \frac{1}{N} \sum_{i=1}^{N} cc(i) \tag{2.18}$$

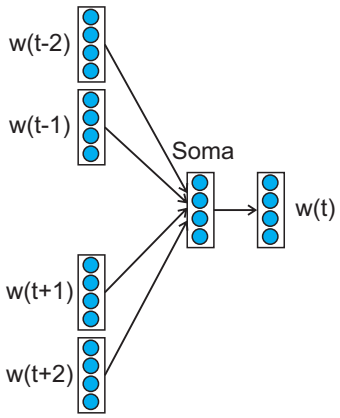
Diâmetro É obtido pelo maior caminho mínimo da rede, definido como $d = \max dist(i, j)$. O diâmetro da rede da Figura 5 é 4; o maior caminho mínimo é 1-5-6-7-8 ou 1-5-6-7-9.

2.4.4 Representações com Modelos Densos

Em geral, quando utilizamos métricas para caracterizar um texto, estamos fazendo uma engenharia de atributos (*feature engineering*), ou seja, utilizando o conhecimento sobre o problema para criar algum atributo que represente esse conhecimento. Uma abordagem alternativa à engenharia de atributos é fazer com que o próprio modelo de aprendizado de máquina crie uma representação dos dados; esse é o princípio do aprendizado de atributos (*feature learning*) ou aprendizado de representação (*representation learning*) (GOODFELLOW; BENGIO; COURVILLE, 2016; LECUN; BENGIO; HINTON, 2015). Nesse cenário, os exemplos são representados em um vetor de números reais e são aplicadas várias transformações não-lineares com o objetivo de aprender a modelar uma função. Quanto mais complexo o problema, mais transformações não-lineares serão necessárias (LECUN; BENGIO; HINTON, 2015).

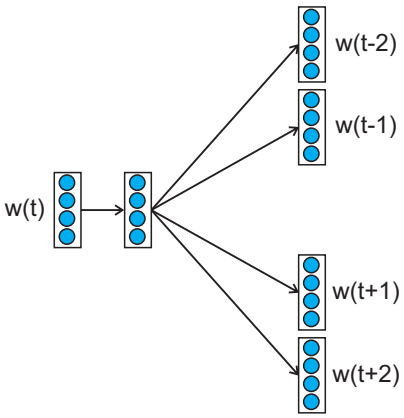
No processamento de línguas podemos representar um texto com uma soma dos vetores de cada palavra que compõe esse texto. Esses vetores são conhecidos como *word embeddings* e representam a semântica e a sintaxe de forma latente. Dentre os mais importantes e conhecidos temos o *Continuous Bag-of-Words* (*CBOW*) e o *Skip-gram* (MIKOLOV *et al.*, 2013). Ambos tentam predizer uma palavra em uma janela local, sendo que para isso é utilizada uma rede neural com uma única camada. No modelo *CBOW*, dada uma janela de palavras, o objetivo é prever a palavra do meio. Na Figura 6 é ilustrado esse modelo utilizando-se uma janela de tamanho 5.

Figura 6 – Modelo *CBOW*.



Fonte: Adaptada de Mikolov *et al.* (2013).

Figura 7 – Modelo *Skip-gram*.



Fonte: Adaptada de Mikolov *et al.* (2013).

Já no modelo *Skip-gram*, utilizando-se uma única palavra são preditas as palavras em torno dela, como pode ser visto na Figura 7.

Em Pennington, Socher e Manning (2014) é proposto o método *Glove Vectors* (*GloVe*), que explora estatísticas globais da coocorrência de palavras. Primeiramente, é construída uma matriz $\mathbf{M}$ de coocorrência de palavras em uma janela de tamanho n (geralmente n é igual 10); cada posição da matriz $\mathbf{M}_{ij}$ representa um peso da palavra i estar próxima da palavra j. A partir da matriz $\mathbf{M}$ são criados os vetores e estes devem obedecer a Equação 2.19.

$$\mathbf{w}_i\mathbf{w}_j + b_i + b_j = \mathbf{M}_{ij} \tag{2.19}$$

em que $\mathbf{w}_i$ é o vetor da palavra i e $\mathbf{w}_j$ é o vetor da palavra j; b_i e b_j são os termos de *bias*.

Os modelos descritos anteriormente são limitados por um vocabulário definido previamente, e caso uma palavra não esteja nesse vocabulário não é possível obter a sua representação. Bojanowski *et al.* (2017) propõem uma adaptação do *Skip-gram* para lidar com *n-grams* de caracteres, assim, a representação densa de alguma palavra é formada pela soma dos vetores de cada *n-gram*. No trabalho original, foi utilizado de 3-*grams* até 6-*grams*, e a própria palavra é incluída como um *n-gram* independente do seu tamanho. Esse método está disponível pela biblioteca *fastText* e conta com modelos pré-treinados para diversas línguas, incluindo o Português [6].

Diversos trabalhos da literatura têm explorado técnicas mais complexas como a combinação de várias camadas de Redes Recorrentes (HOCHREITER; SCHMIDHUBER, 1997; SUTSKEVER; VINYALS; LE, 2014) ou a utilização de redes com atenção (*Transformers*) (VASWANI *et al.*,) para representar um texto. Em geral, esses modelos são treinados como modelos de língua em que o objetivo é predizer a próxima palavra de uma sentença. Após o treinamento, é realizado um ajuste nos pesos para a tarefa final. Nessa categoria, temos o *Universal Language Model Fine-tuning* (*ULMFiT*) (HOWARD; RUDER, 2018) e o *Embeddings from Language Models* (*ELMo*) (PETERS *et al.*, 2018). Outro modelo que tem sido aplicado em diversas tarefas é o *Bidirectional Encoder Representations from Transformers* (*BERT*) (DEVLIN *et al.*, 2019). Esse modelo é treinado para predizer as palavras que estão faltando de uma sentença e também dadas duas sentenças se a segunda é a próxima sentença. Uma descrição detalhada deses métodos está além do escopo do presente doutorado.

2.5 Similaridade Semântica Textual e Inferência Textual

Nesta seção, são apresentadas duas tarefas importantes do Processamento de Línguas Naturais: a Similaridade Semântica Textual e a Inferência Textual, sendo ambas utilizadas em aplicações similares. Na Seção 2.5.3 é apresentada a construção do conjunto de dados para a primeira avaliação conjunta que contempla essas duas tarefas para o Português Brasileiro (PB)

[6] <https://github.com/facebookresearch/fastText>

e Português Europeu (PE), e na Seção 2.5.4 são apresentados os melhores métodos para cada tarefa da avaliação conjunta.

2.5.1 Similaridade Semântica Textual

Na tarefa de Similaridade Semântica (*STS*, *Semantic Textual Similarity*), dado um par de textos $(\mathbf{S}_i^1, \mathbf{S}_i^2)$ o objetivo é indicar o grau de similaridade entre eles. Estamos interessados em atribuir um valor y_i em alguma escala, geralmente de 0 a 5 ou 1 a 5 (AGIRRE *et al.*, 2012; AGIRRE *et al.*, 2015; MARELLI *et al.*, 2014; FONSECA *et al.*, 2016). Assim, essa gradação naturalmente captura as diferenças sutis de similaridade, como sentenças que possuem o mesmo significado (pontuação 5), possuem pequenas diferenças semânticas (pontuação 4), compartilham apenas alguns detalhes (pontuação 3), sentenças não relacionadas, mas versam sobre o mesmo assunto (pontuação 2), ou mesmo que não tem nada em comum (pontuação 1).

Os métodos de *STS* podem ser aplicados para diversas tarefas como *QA* (*Question Answering*, ou Perguntas e Respostas), sumarização, busca semântica, e outras (AGIRRE *et al.*, 2012; AGIRRE *et al.*, 2015; FONSECA *et al.*, 2016).

Na Tabela 5, são mostrados dois exemplos de pares de sentenças com valores de similaridade em que o valor de similaridade semântica mínimo é 1 e o valor máximo é 5[7]. Na primeira linha, os textos se referem ao produto *iPhone*, mas não são relacionados, e o segundo par de textos possuem o mesmo significado, isto é, são paráfrases. Cabe ressaltar que não existe uma definição unificada sobre *STS*, assim, cada conjunto de dados de *STS* pode possuir escalas de gradações diferentes e definições diferentes para cada gradação.

Por exemplo, a primeira tarefa de *STS* (AGIRRE *et al.*, 2012) possui uma gradação de 0 a 5, enquanto que Marelli *et al.* (2014) e Fonseca *et al.* (2016) utilizaram uma gradação de 1 a 5.

Tabela 5 – Exemplos para os valores de similaridade semântica.

Similaridade	$\mathbf{S}^1$	$\mathbf{S}^2$
2	As previsões da Apple confirmaram-se: os novos modelos iPhone 6S e 6S Plus bateram recorde de vendas.	A Square Trade testou os iPhones 6S e 6S Plus e não foi nada meiga.
5	Prometi ao estúdio que entregaria uma última trilogia para terminar a saga.	Prometi ao estudo que faria uma última trilogia para finalizar a saga.

[7] Esses exemplos foram retirados do conjunto de dados ASSIN (FONSECA *et al.*, 2016).

2.5.2 Inferência Textual

A tarefa de Inferência Textual (*RTE*, *Recognizing Textual Entailment*) pode ser definida como uma relação unidirecional entre dois textos, em que se uma pessoa ao ler um texto **T** conclui que uma hipótese **H** é verdadeira; diz-se que **T** implica (*entails*) **H**. Embora subjetiva, a definição é largamente aceita na comunidade de processamento de línguas naturais (DAGAN; GLICKMAN; MAGNINI, 2005; MARELLI *et al.*, 2014; FONSECA *et al.*, 2016). Com essa definição, é assumido que pessoas lendo o par $(\mathbf{T}, \mathbf{H})$ compartilham: (i) o conhecimento da linguagem em que os textos são formulados, e (ii) possuem o mesmo conhecimento prévio sobre o tema (DAGAN; GLICKMAN; MAGNINI, 2005). O par 1 exemplifica uma relação de inferência, em que 1a implica em 1b.

(1) a. Fontes hospitalares palestinas registraram que ao menos 15 palestinos ficaram feridos.

b. Ao menos 15 palestinos ficaram feridos, segundo fontes palestinas.

Lembrando que toda a informação de **H** que não seja conhecimento de mundo deve estar contida em **T**, no exemplo 2 o texto 2b possui informações sobre quais trechos estarão bloqueados, então esses pares não são considerados como inferência.

(2) a. O bloqueio do Rodoanel deve ocorrer até as 9h, segundo a Polícia Militar.

b. Segundo a PM, o bloqueio entre os quilômetros 7 e 16 do Rodoanel deve ocorrer até as 9h.

Um caso particular de inferência textual é a paráfrase e nesse caso ambos os textos são semanticamente equivalentes, sendo uma inferência bidirecional (exemplo 3).

(3) a. A CBF encomendou ao IBOPE uma pesquisa sobre a percepção do torcedor quanto ao Campeonato Brasileiro de 2015.

b. A CBF encomendou uma pesquisa ao Ibope sobre a avaliação dos torcedores sobre a edição do Brasileirão de 2015.

Outra relação abordada nas tarefas de *RTE* é a contradição, em que segmentos dos pares possuem informações conflitantes. No exemplo 4, é mostrado um exemplo de contradição, extraído e traduzido do *Stanford Natural Language Inference* (*SNLI*) (BOWMAN *et al.*, 2015)).

(4) a. Várias frutas estão disponíveis em um mercado ao ar livre no que parece ser a Índia.

b. As frutas estão pegando fogo.

2.5.3 ASSIN

Para o desenvolvimento de sistemas de similaridade é necessário um córpus anotado com os valores de similaridade semântica. Como esse recurso não existia para o português, o autor desta tese auxiliou fortemente o desenvolvimento do córpus da Avaliação de Similaridade Semântica e de Inferência Textual (ASSIN) (FONSECA *et al.*, 2016), possibilitando o uso de sistemas de similaridade semântica para o reconhecimento de unidades de informação, tarefa aqui explorada.

O ASSIN foi uma avaliação conjunta (*shared task*) apresentada no PROPOR 2016[8], consistindo em duas subtarefas: a similaridade semântica textual e a inferência textual. A avaliação criou o primeiro córpus anotado para as duas tarefas em português brasileiro e português europeu.

Para tarefa de *STS* adotamos uma gradação de 1 a 5, conforme o *Sentences Involving Compositional Knowldedge* (*SICK*) (MARELLI *et al.*, 2014), enquanto Agirre *et al.* (2012) utilizou 0 a 5. Abaixo é apresentada a definição de cada valor da escala, sendo que essas definições foram passadas para os anotadores.

1. As sentenças são completamente diferentes. É possível que elas falem do mesmo fato, mas isso não é visível examinando-as isoladamente, sem contexto.

2. As sentenças se referem a fatos diferentes e não são semelhantes entre si, mas são sobre o mesmo assunto (jogo de futebol, votações, variações cambiais, acidentes, lançamento de produtos).

3. As sentenças têm alguma semelhança entre si, e podem se referir ao mesmo fato ou não.

4. O conteúdo das sentenças é muito semelhante, mas uma (ou ambas) tem alguma informação exclusiva. A diferença pode ser mencionar uma data, local, quantidade diferente, ou mesmo um sujeito ou objeto diferente.

5. As sentenças têm praticamente o mesmo significado, possivelmente com uma diferença mínima (como um adjetivo que não altera a sua interpretação).

A Tabela 6 mostra exemplos de pares para cada valor da escala. É importante destacar que foi solicitado aos anotadores considerar apenas o conteúdo das sentenças em análise, e não os contextos possíveis nos quais elas poderiam aparecer. Na Tabela 6, o primeiro exemplo possui similaridade 1, embora seja possível que ambas as sentenças venham do mesmo texto e sejam fortemente relacionadas (o que é o caso nesse exemplo), a anotação não deve considerar essas suposições (FONSECA *et al.*, 2016).

[8] <http://nilc.icmc.usp.br/assin/>

Tabela 6 – Exemplos para os valores de similaridade semântica.

1	Mas esta é a primeira vez que um chefe da Igreja Católica usa a palavra em público.	A Alemanha reconheceu ontem pela primeira vez o genocídio armênio.
2	Como era esperado, o primeiro tempo foi marcado pelo equilíbrio.	No segundo tempo, o panorama da partida não mudou.
3	Houve pelo menos sete mortos, entre os quais um cidadão moçambicano, e 300 pessoas foram detidas.	Mais de 300 pessoas foram detidas por participar de atos de vandalismo.
4	A organização criminosa é formada por diversos empresários e por um deputado estadual.	Segundo a investigação, diversos empresários e um deputado estadual integram o grupo.
5	Outros 8.869 fizeram a quadra e ganharão R$ 356,43 cada um.	Na quadra 8.869 apostadores acertaram, o prêmio é de R$ 356,43 para cada.

Fonte: Fonseca *et al.* (2016).

Tabela 7 – Exemplos para as categorias de inferência textual.

Inferência	Como não houve acordo, a reunião será retomada nesta terça, a partir das 10h. As partes voltam a se reunir nesta terça, às 10h.
Paráfrase	Vou convocar um congresso extraordinário para me substituir enquanto presidente. Vou organizar um congresso extraordinário para se realizar a minha substituição como presidente.
Sem relação	As apostas podem ser feitas até as 19h (de Brasília). As apostas podem ser feitas em qualquer lotérica do país.

Fonte: Fonseca *et al.* (2016).

Quanto à inferência textual, o ASSIN possui três categorias: implicação, paráfrase e neutro. A relação de contradição não foi utilizada, pois é fenômeno raro no conjunto de dados; no *SICK* as contradições foram criadas adicionando ou removendo negações, e no *SNLI* foram obtidas manualmente.

Na Tabela 7 mostra um caso em que a primeira sentença implica a segunda; um caso de implicação bidrecional ou paráfrase; e um terceiro caso em que não há implicação (neutro).

Foram utilizados os agrupamentos de notícias fornecidos pelo *Google News* e aplicado o método *Latent Dirichlet Allocation* (LDA) (BLEI; NG; JORDAN, 2003) para obter uma representação densa das sentenças e selecionar os pares similares. Após a seleção, os pares foram revisados manualmente, e por fim anotados por quatro anotadores diferentes, de um conjunto de 36 anotadores.

Na seleção dos pares, utilizamos um valor mínimo e máximo de similaridade, s_{min} e s_{max},

e fixamos uma proporção mínima e máxima de tokens exclusivos para cada sentença, α_{min} e α_{max}.

Dado que o processo de seleção, revisão e anotação foi realizado em lotes, diferentes valores de parâmetros foram utilizados. Para os valores mínimo e máximo de similaridade usamos: s_{min} de $0,65$ e $0,6$ e s_{max} de $0,9$; o número de tokens encontrados em uma sentença, mas não em outra (sem contar *stopwords*): α_{min} foi fixado em $0,1$ e para o valor de α_{max} utilizou-se $0,7$ e $0,8$.

Um grupo de quatro anotadores revisou os pares coletados em um processo manual. Se um par continha uma sentença sem sentido, era descartado. Sentenças foram também editadas para correção de erros ortográficos e gramaticais, ou para alterar casos em que a presença de implicação era pouco clara.

Na anotação, os pares foram anotados por quatro pessoas, selecionadas aleatoriamente pelo sistema de anotação. Cada anotador seleciona um valor de similaridade de 1 a 5, e também uma das quatro opções para inferência: a primeira sentença implica a segunda; a segunda implica a primeira; paráfrase, ou nenhuma relação.

Caso um par não tivesse concordância de pelo menos três votos para a tarefa de inferência textual, esse par era descartado. Nosso entendimento foi que esses pares eram controversos e assim não seriam boas escolhas para serem incluídos no córpus final. Observa-se que os anotadores poderiam indicar implicação tanto da primeira para a segunda sentença como da segunda para a primeira, porém, no córpus final, invertemos a ordem dos pares necessários para que todos os casos de inferência fossem da primeira sentença para a segunda. O valor final de similaridade para cada par é a média das quatro pontuações.

A anotação foi realizada via uma interface *Web* construída especialmente para a tarefa. Os anotadores receberam treinamento para calibrar os conceitos das tarefas a serem realizadas, com ajuda de um conjunto de 18 pares exemplificando todos os fenômenos tratados. Em caso de dúvidas, perguntas poderiam ser enviadas via e-mail para a equipe de anotadores, o que permitia discutir casos muito difíceis de decidir, principalmente no começo da anotação.

As tabelas 8 e 9 mostram estatísticas sobre as anotações de similaridade e inferência, respectivamente. Pode-se ver que as pontuações de similaridade mais comuns estão no intervalo entre 2 e 3. Já quanto à inferência, percebe-se que a relação neutra é a classe majoritária, enquanto as paráfrases são uma porção pequena do córpus.

2.5.4 Avaliação Conjunta

O córpus foi dividido em seções de treinamento (com três mil pares de cada variante) e teste (com os dois mil restantes de cada). A metade brasileira do córpus de treinamento foi disponibilizada em 20 de novembro de 2015, e a metade portuguesa foi disponibilizada dois meses depois.

Tabela 8 – Estatísticas de similaridade do ASSIN.

Similaridade	**PB**	**PE**	**Total**
4,0 – 5,00	1.074	1.336	2.410
3,0 – 3,75	1.591	1.281	2.872
2,0 – 2,75	1.986	1.828	3.814
1,0 – 1,75	349	555	904
Média	3,05	3,05	3,05

Fonte: Fonseca *et al.* (2016).

Tabela 9 – Estatísticas de inferência do ASSIN.

Relação	**PB**	**PE**	**Total**
Sem relação	3.884	3.432	7.316
Implicação	870	1.210	2.080
Paráfrase	246	358	604

Fonte: Fonseca *et al.* (2016).

Os participantes receberam o conjunto de teste (sem os rótulos corretos dos pares) em 4 de março de 2016, e tiveram 8 dias para enviar aos organizadores os arquivos com as respostas produzidas por seus sistemas. Cada participante pôde enviar até três resultados.

As métricas usadas na avaliação das duas tarefas são consoantes com as usadas em avaliações conjuntas internacionais. Na tarefa de similaridade textual, foi usada a correlação de Pearson, tendo o erro quadrático médio (*MSE*, *mean square error*) como medida secundária. Idealmente, os sistemas devem ter a maior correlação possível e o menor *MSE* possível. Para a inferência, foi usada a medida F1, tendo a acurácia como medida secundária.

Foram usadas duas estratégias como *baseline* para o ASSIN: a primeira memoriza a média das similaridades do córpus de treino e a classe de inferência mais comum, e emite esses valores para todos os pares de teste. A segunda, um pouco mais sofisticada, consiste no treinamento de um classificador baseado em regressão logística e um regressor linear. Estes dois modelos são treinados com apenas dois atributos: a proporção de tokens exclusivos da primeira e da segunda sentença.

Para a tarefa de inferência textual, o melhor sistema foi o segundo método *baseline*, indicando que a presença de inferência no ASSIN é fortemente relacionada com a sobreposição lexical.

A equipe Solo Queue (HARTMANN, 2016), que obteve os melhores resultados na avaliação de similaridade semântica textual, desenvolveu um método baseado no valor da similaridade do cosseno de duas representações vetoriais, detalhadas abaixo.

Na primeira abordagem para obter o valor de similaridade, as palavras são convertidas

em uma representação vetorial densa, obtida pelo word2vec (MIKOLOV *et al.*, 2013), e então é calculada a média das *embeddings* que compõem cada par, e por fim é calculada a similaridade do cosseno.

Na segunda abordagem, são obtidos os *stems* das palavras de conteúdo e é realizada uma expansão do vocabulário; nessa etapa, para cada palavra de conteúdo são buscados os sinônimos no TEP (Thesaurus para o português do Brasil) (MAZIERO *et al.*, 2008). Essa expansão é restrita apenas para palavras que possuem até um sinônimo.

Em seguida, os pares são transformados em uma representação vetorial esparsa utilizando o *TF-IDF*. Por fim, os cossenos entre as duas representações (*TF-IDF* e *word2vec*) de cada par são dadas como entrada para um regressor linear que determina a similaridade do par.

CAPÍTULO

3

TRABALHOS RELACIONADOS

Neste capítulo, é apresentada uma revisão da literatura relacionada às tarefas abordadas no projeto de doutorado.

Na Seção 3.1 são apresentados os trabalhos relacionados à classificação de narrativas para identificação de pacientes com CCL e DA, principalmente, pois são os grupos de interesse desta pesquisa. Os trabalhos abordam classificações binárias (CCLs *versus* Saudáveis, DAs *versus* Saudáveis, DAs + CCLs versus saudáveis) ou multiclasse (CCLs, DAs, Saudáveis). Já na Seção 3.2, são abordados os trabalhos para identificação automática das unidades de informação, foco principal dos esforços da pesquisa.

3.1 Triagem Automática de Pacientes: Classificação de Narrativas de Exames Neuropsicológicos

3.1.1 Os trabalhos de Roark, Mitchell e Hollingshead (2007) e Roark et al. (2011)

A utilização de métricas sintáticas extraídas de forma automática é algo recente na literatura. Roark, Mitchell e Hollingshead (2007) analisaram narrativas de reconto do teste Memória Lógica de *Wechsler* de 18 sujeitos diagnosticados com CCL, e 29 sujeitos com envelhecimento saudável. Os autores extraíram: quantidade de palavras por cláusulas, proporção de número de nós pela quantidade de palavras, complexidade de Yngve (YNGVE, 1960), complexidade de Frazier (FRAZIER, 1985), entropia cruzada de *PoS-tagging*, nível de desenvolvimento ou nível-D (*Developmental level* ou *D-level*), que é uma escala com oito níveis, que se baseia no nível de desenvolvimento de sentenças complexas em crianças com desenvolvimento normal (ROSENBERG; ABBEDUTO, 1987; CHEUNG; KEMPER, 1992).

As métricas de quantidade de palavras por cláusulas, número de nós pela quantidade de

palavras, complexidade de Frazier, e entropia cruzada de *PoS-tagging* apresentaram diferenças estatísticas entre os grupos CCL *versus* saudáveis no reconto imediato. Enquanto que, para o reconto tardio, a métrica distância de dependência apresentou diferença estatística entre os grupos CCL *versus* saudáveis.

Roark *et al.* (2011) é uma extensão do trabalho apresentado em Roark, Mitchell e Hollingshead (2007) e é um dos primeiros trabalhos com objetivo de identificar automaticamente transcrições de narrativas de pacientes com CCL, separando-os dos indivíduos com envelhecimento saudável. Foram utilizadas narrativas do teste de Memória Lógica de *Wechsler*, com 37 indivíduos saudáveis, e 37 indivíduos com CCL.

Os autores combinaram diversos atributos com nove escores de exames obtidos manualmente.

Entre as métricas lexicais, selecionaram: número total de palavras, número total de sentenças, palavras por sentenças. Também extraíram nove métricas sintáticas: palavras por cláusulas, proporção do número de nós pela quantidade de palavras, distância de dependência, complexidade Yngve, complexidade Frazier, entropia cruzada de *PoS-tagging* de um modelo de língua treinado em um córpus de conversas telefônicas, e o mesmo modelo adaptado ao domínio pela técnica *maximum a posteriori* (BACCHIANI *et al.*, 2006), densidade de ideias e densidade de conteúdo. Além dessas, as seguintes métricas acústicas foram extraídas:

- **Pausas por reconto:** é o número total de pausas do reconto.
- **Tempo total de pausa:** é a duração total de todas as pausas em segundos.
- **Duração média das pausas:** é o tempo total das pausas dividido pela quantidade de pausas.
- **Taxa de pausas padronizadas:** é o número de palavras da narrativa dividido pela quantidade de pausas.
- **Tempo total de fonação:** é a quantidade de tempo, em segundos, que contém eventos de fala.
- **Tempo total de locução:** é a quantidade de tempo incluindo fala e pausas.
- **Taxa de Fonação:** é o tempo total de fonação dividido pelo tempo total de locução.
- **Taxa de fonação transformada:** é o arco-seno da raiz quadrada da taxa de fonação.
- **Tempo de fonação padronizado:** é a quantidade de palavras dividida pelo tempo total de fonação.
- **Taxa verbal:** é a quantidade de palavras dividido pela tempo total de locução.

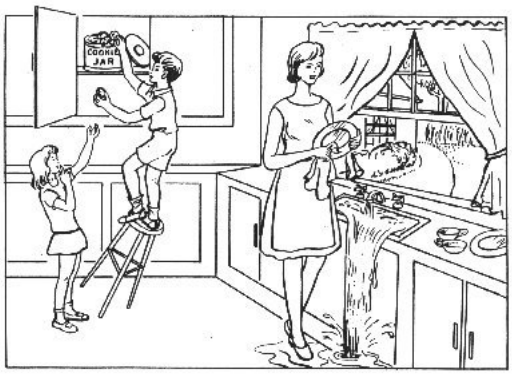
Na tarefa de classificação, os atributos foram normalizados de modo a terem valores no intervalo de 0 a 1. Aplicaram o algoritmo de classificação *SVM* com *kernel* polinomial de segunda ordem, *leave-pair-out cross-validation*, e a métrica *AUC* (área sob a curva, *Area Under the Curve*).

Obtiveram uma *AUC* de $0,86$ com a combinação da taxa de pausas padronizada, tempo total de fonação, taxa de fonação, taxa de fonação transformada, palavras por cláusulas, complexidade de Yngve, distância de dependência, entropia cruzada do modelo adaptado ao domínio, densidade de conteúdo e os nove escores de exames obtidos manualmente.

3.1.2 Os trabalhos que usaram o conjunto de dados DementiaBank

Nos últimos anos, diversos trabalhos têm explorado o *DementiaBank*, que é um dos maiores conjuntos de dados públicos longitudinais de fala espontânea de indivíduos com e sem demência (ORIMAYE; WONG; GOLDEN, 2014; FRASER; MELTZER; RUDZICZ, 2016; YANCHEVA; RUDZICZ, 2016; ORIMAYE *et al.*, 2017; FRASER; FORS; KOKKINAKIS, 2019). Os dados foram coletados entre 1983 e 1988 como parte do Programa de Pesquisa em Doença de Alzheimer da Universidade de Pittsburgh (BECKER *et al.*, 1994). Os pacientes foram solicitados a descrever a cena do Roubo do Biscoito (cf. Figura 8), que é uma subtarefa da *Boston Diagnostic Aphasia Examination* (*BDAE*) (GOODGLASS; KAPLAN; BARRESI, 1983).

Figura 8 – Cena do roubo do biscoito.



Fonte: Adaptada de Goodglass, Kaplan e Barresi (1983).

Orimaye, Wong e Golden (2014) utilizaram o *DementiaBank*, entretanto, agruparam todos os participantes com demências em um único grupo. Desse modo, o grupo de participantes com demência era formado por 239 participantes com Doença de Alzheimer provável, 21 com Doença de Alzheimer possível, 5 Demência Vascular 43 com CCL, 3 com problemas de memória. O grupo de controle continha 242 participantes saudáveis.

Para representar as narrativas, os autores extraíram nove atributos sintáticos com o parser de *Stanford*: quantidade de sentenças coordenadas e subordinadas, *reduced sentences* (sentenças subordinadas sem conjunção mas com formas verbais nominais), quantidade de predicados,

média da quantidade de predicados, distância de dependência, quantidade de relações únicas de dependência, média da quantidade de relações únicas de dependência na sentença, e quantidade de regras de produção exclusivas.

Também utilizaram 11 métricas lexicais: quantidade de *utterance* (elocução), palavras por *utterance*, quantidade de palavras funcionais, quantidade de palavras únicas, quantidade de palavras, quantidade de caracteres, quantidade de sentenças, quantidade de repetições, quantidade de revisões, quantidade de bigramas únicos, e quantidade de morfemas.

Na classificação utilizaram o *10-fold-cross-validation* em um conjunto de dados balanceado, e avaliaram os classificadores: SVM com *kernel Radial Basis Function* (*RBF*), *Naïve Bayes*, Árvore de Decisão, Redes Neurais, e Redes Bayesianas. Na Tabela 10 são mostrados os resultados de cada modelo em termos de Precisão, *Recall* e *F1*. O *SVM* possui o maior valor de *F1*, obtendo $0,74$.

Tabela 10 – Desempenho dos algoritmos de aprendizado de máquina na classificação de demência vs saudável no *DementiaBank*.

Classificadores	**Pr**	***Rec***	**F1**
SVM	0,75	**0,73**	**0,74**
Naïve Bayes	**0,79**	0,53	0,63
Árvore de Decisão	0,78	0,69	0,71
Redes Neurais	0,74	0,67	0,71
Redes Bayesianas	0,77	0,66	0,71

Fonte: Adaptada de Orimaye, Wong e Golden (2014).

Orimaye *et al.* (2017) estenderam o trabalho anterior adicionando atributos de bigramas e trigramas, e utilizaram 23 atributos sintáticos e léxicos: quantidade de sentenças coordenadas e subordinadas, *reduced sentences*(sentenças subordinadas sem conjunção mas com formas verbais nominais), quantidade de predicados, média da quantidade de predicados, distância de dependência, quantidade de relações únicas de dependência, média da quantidade de relações únicas de dependência na sentença, quantidade de regras de produção exclusivas, quantidade de *utterance*, palavras por *utterance*, quantidade de palavras funcionais, quantidade de palavras únicas, quantidade de palavras, quantidade de caracteres, quantidade de sentenças, quantidade de repetições, quantidade de revisões, quantidade de morfemas, quantidade de *utterance* incompletas, quantidade de substituições (quando o paciente utiliza uma palavra incorreta, após perceber e realiza uma correção), quantidade de palavras incompletas, e quantidade de palavras de preenchimento (essa marca do discurso é empregada pelo interlocutor para indicar hesitação ou manter o controle de uma conversa como, por exemplo, “ah”, “eh”, “hum”, “bom”, “então”, “digo”).

Em vez de agrupar pacientes com diferentes demências, os autores focaram na identificação de pacientes com Doença de Alzheimer. Assim, utilizaram narrativas de 99 participantes com Doença de Alzheimer provável, e 99 participantes com envelhecimento saudável.

Na avaliação, utilizaram o *leave-pair-out*, a métrica *AUC* e o algoritmo de classificação *SVM* com kernel *RBF*. Foram considerados três cenários na classificação: 23 atributos sintáticos e léxicos, seleção de 1000 atributos com o *Information Gain* para os *n-grams*, e seleção de atributos considerando os atributos sintáticos, léxicos e *n-grams*.

Na Tabela 11, são apresentados os resultados para os cenários. O melhor resultado é de $0,93$ de *AUC*, combinando os atributos sintáticos, lexicais e *n-grams* com a aplicação do *Information Gain*.

Tabela 11 – Desempenho dos algoritmos de aprendizado de máquina na classificação CCL vs saudável no *DementiaBank*.

Modelos	***AUC***
23-sintáticos-lexicais	0,80
top-1000-n-gram	0,91
top-1000-sintáticos-lexicais-n-gram	**0,93**

Fonte: Adaptada de Orimaye *et al.* (2017).

3.1.3 Os trabalhos para o Português do Brasil

Para o Português Brasileiro, Aluísio, Cunha e Scarton (2016) foi o primeiro trabalho a identificar pacientes com CCL e DA. Utilizaram 77 métricas linguísticas, como contagens básicas de classes morfossintáticas, complexidade sintática, densidade de ideias (CUNHA *et al.*, 2015) e coesão de texto por meio de *Latent Semantic Analysis* (*LSA*).

Os autores analisaram as transcrições de reconto de narrativas da história de Cinderela, produzidas por 60 sujeitos: 20 controles saudáveis, 20 pacientes com DA e 20 pacientes com CCL. Para cada sujeito, foram mostradas 22 cenas representando a história da Cinderela, na forma de um livro, e tiveram o tempo necessário para examinar o livro ilustrado da história. Após, o paciente foi solicitado a narrar a história com o máximo de detalhes possível. A narrativa foi gravada e depois transcrita manualmente. O tempo gasto foi registrado, mas não houve limite imposto.

Os autores exploraram os seguintes algoritmos de classificação: *Naïve Bayes* (*NB*), *Support Vector Machines* (*SVM*) com kernel linear, *Multilayer Perceptron* (*MLP*), Regressão Logística, JRip, J48, e *Random Forest*. Na avaliação utilizaram o *leave-one-out* e a medida $F1$.

Para avaliar a contribuição dos conjuntos de métricas, dividiram as métricas em quatro conjuntos: (i) 48 métricas derivadas do Coh-Metrix-Port (conjunto denominado CMP), (ii) 73 métricas (48 do Coh-Metrix-Port + 25 adicionadas, formando o conjunto CMP+Novas ou Coh-Metrix-Dementia), (iii) com apenas as 25 métricas novas (conjunto Novas), e (iv) com um subconjunto das métricas selecionadas pelo algoritmo *Correlation-based Feature Selection* (*CFS*). Avaliaram os métodos em dois problemas: multiclasse (participantes com envelhecimento

saudável, Doença de Alzheimer, e Comprometimento Cognitivo Leve) e binário (participantes com envelhecimento saudável e CCL). Como *baseline*, utilizaram a classe majoritária.

Os resultados obtidos por Aluísio, Cunha e Scarton (2016) para os diferentes cenários e algoritmos são apresentados na Tabela 12. No cenário multiclasse, o melhor desempenho foi obtido com as métricas selecionadas pelo *CFS*, utilizando o algoritmo *Naïve Bayes*, que apresentou 0,817 de medida *F1*. Já no cenário binário, o algoritmo J48 obteve 0,900 de *F1* para o cenário CMP+Novas, Novas e *CFS*.

Tabela 12 – Resultados dos métodos de classificação de narrativas da Cinderela para diferentes conjuntos de métricas.

Modelos	**Multiclasse**				**Binário**			
	CMP	**CMP+Novas**	**Novas**	***CFS***	**CMP**	**CMP+Novas**	**Novas**	***CFS***
NB	0,651	0,733	**0,767**	**0,817**	0,725	0,825	0,850	0,825
SVM	**0,669**	0,715	0,731	0,753	**0,775**	0,747	0,798	0,848
MLP	0,566	0,536	0,633	0,601	0,725	0,699	0,775	0,825
RL	0,616	0,701	0,718	0,750	0,749	0,747	0,697	0,749
JRip	0,500	0,699	0,750	0,732	0,697	0,875	0,800	0,775
J48	0,498	0,666	0,633	0,748	0,596	**0,900**	**0,900**	**0,900**
RF	0,635	**0,750**	0,733	0,752	0,750	0,799	0,850	0,850
Baseline	0,333	0,333	0,333	0,333	0,500	0,500	0,500	0,500

Fonte: Adaptada de Aluísio, Cunha e Scarton (2016).

3.2 Identificação Automática de Unidades de Informação em Recontos de Narrativas

Existem poucos trabalhos na literatura que tratam da automatização da identificação de unidades de informação em recontos de narrativas. Podemos dividi-los em: métodos de busca de palavras (PAKHOMOV *et al.*, 2010; FRASER; MELTZER; RUDZICZ, 2016), métodos de alinhamento (PRUD'HOMMEAUX; ROARK, 2015), e métodos de *clustering* (YANCHEVA; RUDZICZ, 2016; FRASER; FORS; KOKKINAKIS, 2019), detalhados nas próximas seções.

3.2.1 Métodos de Busca de Palavras

Pakhomov *et al.* (2010) compilaram uma lista com palavras e frases que representavam algum conceito da cena do Roubo do Biscoito, que é uma subtarefa da Bateria de Boston (*Boston Diagnostic Aphasia Examination — BDAE*) (GOODGLASS; KAPLAN; BARRESI, 1983). As narrativas foram divididas em *n-grams*, de 1 à 4, e para cada *n-gram* os autores realizaram uma busca na lista de palavras. Se o *n-gram* era encontrado, considerou-se que o paciente se relembrou dessa unidade de informação.

Os autores utilizaram 38 narrativas de idosos com Degeneração Lobar Frontotemporal, com o seguintes subtipos: Afasia Progressiva Primária, Demência Semântica, variante comportamental da Demência Frontotemporal, e Afasia Logopênica. Entretanto, não encontraram diferença estatisticamente significativa entre os três grupos usando a contagem de unidades de informação recordadas.

Fraser, Meltzer e Rudzicz (2016) utilizaram uma lista de palavras para cada possível unidade de informação. Para as unidades de informação que representam uma ação, os autores utilizaram o *parser* de *Stanford* para identificar o verbo e o sujeito, e analisaram se essa combinação estava na lista de palavras. As unidades de informação foram utilizadas como atributos binários em conjunto com as métricas: (i) de PoS, (ii) de complexidade sintática, (iii) psicolinguísticas, (iv) de diversidade lexical, (v) de constituintes gramaticais, (vi) de repetividade de informações, e (vii) acústicas, totalizando 370 atributos. No trabalho, o objetivo dos autores foi distinguir narrativas de pacientes com Doença de Alzheimer e envelhecimento saudável. Os autores utilizaram as narrativas do *DementiaBank* (BECKER *et al.*, 1994), com 233 narrativas de 97 participantes com envelhecimento saudável e 240 narrativas de 168 participantes com possível ou provável Doença de Alzheimer.

Para a classificação final, usaram o algoritmo de Regressão Logística, *10-fold-cross-validation*, e a métrica acurácia para avaliação, dado que a classificação era binária. O melhor resultado foi 0,819 de acurácia, utilizando 35 atributos selecionados com o método de Correlação de Pearson. Adicionalmente, aplicaram o *Principal Axis Factors* (PAF) com quatro fatores, sendo que um dos fatores possui uma correlação com os atributos de unidades de informação, demostrando que esses atributos podem ser úteis na tarefa de identificação de pacientes com algum tipo de comprometimento cognitivo.

3.2.2 Métodos de Alinhamento

Prud'hommeaux e Roark (2015) propuseram um método de alinhamento baseado em grafos, utilizando a técnica de passeios aleatórios (*Random Walks*) para automatizar o teste de reconto de narrativas do teste de Memória Lógica de *Wechsler*. Na abordagem proposta, cada palavra do reconto ou da narrativa original representa um nó do grafo e o alinhamento entre as palavras representa as arestas.

O método foi comparado com o alinhador *Berkeley*, sendo 72 pacientes com CCL, 163 com envelhecimento saudável, e 48 narrativas de pacientes inelegíveis, i.e., que não se enquadraram em algum critério e não podem fazer parte dos grupos CCL ou Saudáveis.

Os métodos de alinhamentos de palavras recebem um córpus paralelo sentencialmente alinhado e para a utilização desses métodos os autores consideraram as narrativas como sentenças. Para o treinamento, os autores compilaram três conjuntos de dados paralelos os quais são apresentados a seguir:

- **Conjunto** 1: contém as narrativas de 235 pacientes alinhadas sentencialmente com a narrativa de origem (72 CCL e 163 saudáveis) e mais 48 narrativas de pacientes inelegíveis. Como os pacientes dos grupos CCL e saudáveis produziram uma narrativa imediata e outra tardia, esse conjunto contém no total 518 narrativas, ou 518 linhas;

- **Conjunto** 2: contém a combinação entre as 518 narrativas, totalizando 268.324 narrativas (linhas);

- **Conjunto** 3: contém todas as palavras presentes na narrativa original e no reconto, alinhadas sentencialmente com ela mesma, totalizando 976 linhas.

A partir desses conjuntos foram construídos dois modelos para cada método de alinhamento. Os modelos chamados de *pequeno* foram treinados no Conjunto 1 e 3, e os modelos chamados de *grande* foram treinados no Conjunto 1 e 2 e 100 cópias do Conjunto 3.

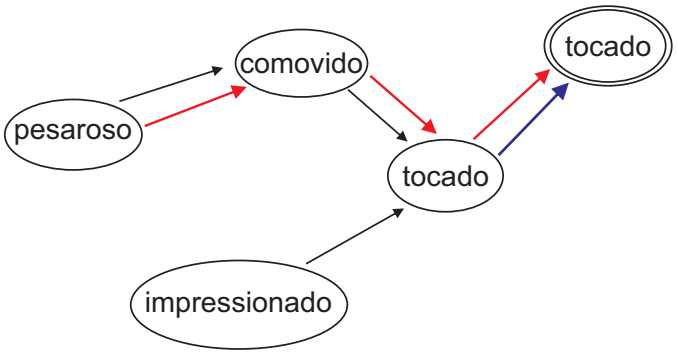
No método proposto, primeiramente, cada narrativa de reconto é alinhada com a narrativa original e as demais narrativas de reconto. Para obter os alinhamentos iniciais é utilizado o alinhador *Berkeley Aligner* (LIANG; TASKAR; KLEIN, 2006). A partir desses alinhamentos é construído um grafo, em que é verificado se o alinhamento possui uma probabilidade maior que 0.5. Neste caso, é adicionado um vértice entre essas palavras. Desse modo, podem existir dois tipos de alinhamentos: o alinhamento com uma palavra da narrativa fonte, e o alinhamento com uma palavra da narrativa de reconto. Dada uma palavra da narrativa de reconto, esta é definida como o vértice inicial da caminhada aleatória. A cada passo da caminhada é gerado um valor aleatório e, caso este seja maior que um λ, é realizada uma transição para uma palavra da narrativa original; caso contrário, a transição é realizada para palavra da narrativa de reconto. Quando a caminhada aleatória atingir uma palavra da narrativa fonte, é proposto um novo alinhamento entre a palavra inicial e a palavra fonte, e a caminhada é encerrada.

Na Figura 9 é demonstrado o funcionamento do método proposto, em que as setas em vermelho demonstram o passeio aleatório, a seta em azul é o alinhamento da palavra da narrativa de reconto com a palavra da narrativa original e as setas escuras são os alinhamentos entre as palavras da narrativa do reconto. Nesse exemplo, iniciando o passeio aleatório pela palavra *pesaroso* é realizada uma transição para a palavra *comovido*, a próxima transição é para a palavra *tocado*, e como esta palavra está alinhada com uma palavra fonte é criado um novo alinhamento entre as palavras *pesaroso* e *tocado*.

Para cada palavra presente nas narrativas de reconto são realizados mil passeios aleatórios. O novo alinhamento, entre a palavra do reconto e a palavra fonte, é definido pelo alinhamento mais frequente dos passeios aleatórios. Na Tabela 13 é apresentado o resultado da avaliação do método de alinhamento baseado em grafo com o *Berkeley Aligner*.

Após a obtenção dos alinhamentos, estes são utilizados como atributos para um classificador final; se alguma palavra da narrativa de reconto estiver alinhada com a narrativa original é

Figura 9 – Exemplo do algoritmo baseado em grafo (os índices das palavras e das narrativas foram omitidos).



Fonte: Adaptada de Prud'hommeaux e Roark (2015).

Tabela 13 – Comparação do desempenho entre o alinhador baseado em Grafo e o *Berkeley Aligner*.

Alinhadores	**Pre**	***Rec***	**F1**
Berkeley-Pequeno	0,721	0,796	0,756
Berkeley-Grande	0,786	0,805	0,795
Grafo-Pequeno	0,779	0,812	0,795
Grafo-Grande	0,855	0,796	0,810

Fonte: Adaptada de Prud'hommeaux e Roark (2015).

considerado que o paciente se recordou desse trecho.

Na tarefa de classificação final (Normal *versus* CCL), Prud'hommeaux e Roark (2015) exploram duas representações para cada paciente:

1. O *Summary score* é quantidade de unidades de informações recordas no reconto imediato e tardio;

2. *Element scores* em que cada unidade de informação representa um atributo. É marcado se o paciente se recordou ou não dessa unidade de informação no reconto imediado e no tardio.

Na Tabela 14 é apresentado a *AUC* (Área Sob a Curva *ROC*, *Area Under Curve ROC*) resultado do classificador *SVM* com *kernel radial basis function*, utilizando o *leave-pair-out*. É possível perceber que os resultados de classificação para todos os quatro modelos de alinhamento são muito próximos e não possuem uma diferença grande para o manual no *Element scores*.

3.2.3 Métodos de Clustering

Yancheva e Rudzicz (2016) e Fraser, Fors e Kokkinakis (2019) automatizaram a análise de unidades de informação aplicando algoritmos de agrupamento, em que os *clusters* são

Tabela 14 – Resultados da classificação de CCL vs saudáveis em narrativas do exame *WLM* utilizando o método de grafo para recuperar as unidades de informação.

Modelos	***Summary***	***Element***
Manual	0,733	0,813
Berkeley-Pequeno	0,733	0,779
Berkeley-Grande	0,751	0,792
Grafo-Pequeno	0,742	0,789
Grafo-Grande	0,748	0,786

Fonte: Adaptada de Prud'hommeaux e Roark (2015).

considerados como um indicador (*proxy*) para as unidades de informação e são utilizados para extrair atributos. Os detalhes de cada método são explicados a seguir.

Yancheva e Rudzicz (2016) avaliaram o método no *DementiaBank*. Neste conjunto de dados, os pacientes são solicitados a descrever a cena do Roubo do Biscoito. Os autores utilizaram 241 narrativas de 98 participantes com envelhecimento saudável e 255 narrativas participantes de 168 participantes com possível ou provável Doença de Alzheimer.

Os verbos e os substantivos das transcrições dos sujeitos são convertidos em uma representação densa com o *GloVe*; para cada grupo (participantes com envelhecimento saudável ou Doença Alzheimer) é aplicado o algoritmo *K-means* com a distância euclidiana e *k* igual à 10. A partir dos *clusters* são criados atributos baseados nas distâncias dos centroides:

- $\mathbf{C_i}$: Para cada *cluster-i* do grupo de Controle, é computada a distância média entre o centroide e todas as palavras atribuídas ao *cluster-i*;
- $\mathbf{D_i}$: Para cada *cluster-i* do grupo de Doença de Alzheimer, é computada a distância média entre o centroide e todas as palavras atribuídas a esse *cluster*;
- **Densidade da ideia**: É o número de *clusters* mencionados, dividido pela quantidade de palavras na narrativa. São consideradas menções se a distância da palavra para o centroide for menor que 3 desvios padrões; e
- **Eficiência da densidade ideia**: O número de *clusters* mencionados, dividido pelo tempo total da narrativa em segundos.

Na tarefa de classificação final, os autores optaram pelo classificador *Random Forest* e *10-fold-cross-validation*.

A abordagem proposta foi comparada com o resultado da classificação utilizando uma lista de palavras para recuperar as unidades de informação. Os autores também adicionaram atributos de métricas Linguísticas e Acústicas (L&A).

Na Tabela 15, são apresentados os resultados dos métodos. Os autores reportaram os resultados dos atributos extraídos de cada modelo de *cluster* e a combinação deles. O melhor

resultado é 0,80 de acurácia, obtido pelo modelo de *cluster* Controle + Demência em conjunto com as métricas linguísticas e acústicas. Observa-se que o *baseline* com lista de palavras de cada unidade de informação obteve 0,73% de acurácia.

Tabela 15 – Resultados da classificação de AD vs saudáveis no *DementiaBank* utilizando com métodos de *clustering* para recuperar as unidades de informação.

Modelos	***Acc***
Lista de palavras	0,73
L&A	0,76
Lista de palavras + L&A	0,80
Controle	0,74
Demência	0,74
Controle + Demência	0,74
Controle + L&A	0,79
Demência + L&A	0,77
Controle + Demência + L&A	0,80

Fonte: Adaptada de Yancheva e Rudzicz (2016).

Fraser, Fors e Kokkinakis (2019) substituíram o modelo *GloVe* pelo *FastText*, possibilitando inferir palavras que não estão presentes no vocabulário do modelo de *embeddings*. Os autores optaram pela distância do cosseno em vez da euclidiana.

Foram utilizados três conjuntos de dados: o *DementiaBank*, com 97 participantes saudáveis e 19 participantes com CCL; o *Gothenburg* (WALLIN *et al.*, 2016), que é composto por transcrições da descrição da cena do Roubo do Biscoito de pacientes suecos, com 36 participantes saudáveis e 31 com CCL; o *Karolinska* (CROMNOW; LANDBERG, 2009), em que é solicitado a 96 indivíduos com envelhecimento saudável que produzam uma descrição escrita da cena do Roubo do Biscoito em 5 minutos.

Os autores extraíram todos os verbos e os substantivos das transcrições, em seguida as palavras foram transformadas em uma representação densa com o *fastText*, e aplicaram o algoritmo *k-means* com três variações do parâmetro k, sendo: 10, 23, e k_{sil}, onde $k_{sil} \in \{2,3,\ldots,30\}$. O valor de k_{sil} é selecionado de forma automática pelo método da silhueta (KAUFMAN; ROUSSEEUW, 2009). Para cada configuração de k foram construídos modelos de agrupamento para o Inglês, Sueco, e uma versão multilíngue (Inglês e Sueco). Após a obtenção dos agrupamentos foram extraídos os seguintes atributos:

- $\mathbf{C_i}$: Para cada *cluster i* é encontrada a distância média do cosseno entre o centróide e todas as palavras atribuídas a esse *cluster*;
- $\mathbf{N_i}$: Para cada *cluster i* são descartadas as palavras com distância maior que 3 desvios padrões da distância média do centróide no conjunto de treinamento. Por fim, é contada a quantidade de palavras associadas a cada *cluster*;

- **P_i**: Para cada *cluster i* é obtido o valor de N_i e divido pelo número de palavras da narrativa;
- **Densidade da ideia**: O número de *clusters* mencionados, dividido pelo número total de palavras na narrativa;
- **Eficiência da densidade ideia**: O número de *clusters* mencionados, dividido pelo tempo total da narrativa em segundos;
- **Densidade informação**: O número de palavras que são atribuídas aos *clusters*, dividido pelo número total de palavras na narrativa;
- **Eficiência de informação**: O número de palavras atribuídas aos *clusters*, dividido pelo tempo total da narrativa em segundos;
- **Densidade de N+V**: é o número de verbos e substantivos dividido pelo número de palavras na narrativa; e
- **Eficiência de N+V**: é o número de verbos e substantivos dividido pelo pelo tempo total da narrativa em segundos.

Para a etapa de classificação, os autores utilizaram o *SVM* linear e *leave-one-out*. Avaliaram a acurácia no conjunto de dados do *DementiaBank* balanceado, e do *Gothenburg*. No treinamento dos modelos de agrupamentos foram adicionados o conjunto *Karolinska* e 78 participantes saudáveis restantes do *DementiaBank*. Para cada iteração do *leave-one-out* os autores executaram um *inner-cross-validation* para selecionar os parâmetros do *SVM*, e selecionaram o modelo de agrupamento a partir de 10 execuções.

Na Tabela 16 são apresentados os resultados dos métodos para os conjuntos *DementiaBank* e *Gothenburg*. No *DementiaBank* o melhor resultado foi o modelo multilíngue com k igual a 10, que obteve uma acurácia de $0,63$; para o modelo de agrupamento monolíngue a melhor acurácia foi de $0,47$ com k igual a 10 e k_{sil}. No *Gothenburg* o melhor resultado foi de $0,72$ com o modelo de multilíngue, e k igual a 23, enquanto que o melhor resultado do modelo monolíngue foi de $0,55$ com k igual a 10.

3.3 Considerações finais

Em geral, os trabalhos focados na identificação automática de pacientes com Doença de Alzheimer e Comprometimento Cognitivo Leve utilizam conjuntos de dados pequenos quando comparados com outras de tarefas de Aprendizado de Máquina. Além disso, existe uma falta de recursos computacionais adequados para esse cenário, por exemplo, todos os trabalhos abordados aqui utilizaram as narrativas transcritas manualmente devido a falta de sistemas automáticos de transcrição adequados.

Tabela 16 – Resultados da classificação de CCL vs saudáveis com métodos de *clustering*.

Treinamento	**k**	***Acc***	
		DementiaBank	***Gothenburg***
Inglês	10	0,47	-
Inglês	23	0,47	-
Inglês	2–30	0,47	-
Sueco	10	-	0,55
Sueco	23	-	0,40
Sueco	2–30	-	0,52
Inglês + Sueco	10	0,63	0,51
Inglês + Sueco	23	0,55	0,72
Inglês + Sueco	2–30	0,55	0,55

Fonte: Adaptada de Fraser, Fors e Kokkinakis (2019).

Para representar as narrativas no processo de classificação são utilizadas diversas métricas lexicais e sintáticas. Fraser, Meltzer e Rudzicz (2016) foram os únicos autores que exploram métricas psicolinguísticas; apenas Roark *et al.* (2011) e Fraser, Meltzer e Rudzicz (2016) extraíram métricas acústicas. Cabe ressaltar que nem sempre os áudios estão disponíveis e/ou as gravações possuem boa qualidade para serem utilizadas.

Além dessas métricas, alguns trabalhos extraem as unidades de informação. A abordagem mais simples para obter esses atributos é a busca por palavras, em que são desenvolvidas listas contendo as possíveis palavras para cada unidade de informação. Entretanto, por ser um processo manual sofre da subjetividade na criação das possíveis palavras para cada unidade de informação e exige tempo para construção e análise dessas listas.

O método de alinhamento proposto por Prud'hommeaux e Roark (2015) é capaz de inferir as unidades de informação presentes no reconto, mas a principal limitação é que o método produz alinhamentos um-para-um, e esse formato pode não ser o mais adequado para avaliações que contenham unidades de informação compostas por sintagmas nominais, sintagmas verbais ou mesmo uma oração. Apesar do método possuir uma versão não-supervisionada em que cada palavra da narrativa é considerada uma unidade de informação, ainda é necessário um conjunto de dados anotados no formato de alinhamentos para obter uma métrica de erro.

Enquanto que os métodos de *clustering* são abordagens interessantes, podem gerar agrupamentos que não necessariamente condizem com as unidades de informação definidas nas avaliações neuropsicológicas, e por serem métodos não-supervisionados os agrupamentos podem não ser coesos.

CAPÍTULO

4

PROJETO ANAA-DEMENTIA

Neste capítulo são apresentados os recursos e resultados do projeto ANAA-Dementia, que avalia a tarefa de similaridade semântica para automatizar o teste do reconto de baterias neuropsicológicas. Especificamente, são apresentados os dois córpus empregados neste trabalho (Seção 4.1), bem como a metodologia utilizada para compilá-los e uma análise de suas características. Na Seção 4.2, são apresentados o método de criação de um dicionário com as propriedades psicolinguísticas de palavras (imageabilidade, concretude, frequência subjetiva e idade de aquisição — *AoA*, *Age of Aquisition*), e também o próprio dicionário, que foi criado para suprir a carência desse recurso para o Português. Na Seção 4.3, são apresentados os métodos de identificação automática de unidades de informação desenvolvidos nesta pesquisa, bem como a avaliação desses métodos, usando os conjuntos de dados compilados para a pesquisa e as *baselines* propostas. Na Seção 4.4, é apresentada a exploração das métricas criadas nesta pesquisa para a tarefa de triagem automática de pacientes no cenário clínico, avaliando cada modelo separadamente e em conjunto.

4.1 Conjuntos de Dados Compilados

Utilizou-se dois conjuntos de dados de reconto. A Tabela 17 apresenta as estatísticas dos dois conjuntos de dados, que trazem uma média do tamanho de sentenças bem próxima entre CCLs e Controles na Bateria Arizona para Desordens de Comunicação e Demência (*ABCD*) (BAYLES; TOMOEDA, 1993) (diferença de 0,5), mas uma diferença maior dos grupos DA e CCL com o grupo de controle da Bateria de Avaliação da Linguagem no Envelhecimento (BALE) (HÜBNER *et al.*, 2019) (diferença de aproximadamente 1,6). O mesmo padrão se repete para a média das palavras por sentenças.

O primeiro conjunto de dados é formado por transcrições dos subtestes de reconto imediato e tardio da História da Carteira da *ABCD*. O teste do reconto foi aplicado em 23 idosos com CCL e 12 adultos com envelhecimento saudável, na Faculdade de Medicina da USP. O teste

Tabela 17 – Estatísticas dos conjuntos de dados.

Bateria	Grupo	Sujeitos	Número de narrativas	Média Sentenças (Desvio Padrão)	Média de palavras por sentença (Desvio Padrão)
ABCD	CCL	23	46	8,17 (1,92)	60,76 (17,39)
	Controle	12	24	7,67 (2,06)	58,96 (14,73)
BALE	DA	11	11	6,09 (2,63)	36,18 (17,10)
	CCL	5	5	6,00 (1,00)	36,40 (5,68)
	Controle	53	53	7,68 (2,67)	52,06 (19,18)

possui 17 unidades de informação, apresentadas na Figura 10, com possíveis alternativas entre parênteses, sendo assim a sua pontuação máxima é de 17 pontos (SANTOS *et al.*, 2019).

Figura 10 – Narrativa utilizada na *ABCD*, separada em unidades de informação; as nove unidades da macroestrutura são marcadas em negrito.

Senhora (mulher) // estava fazendo compras (na loja, foi às compras, foi ao mercado) // Sua carteira (seu porta-notas, sua moedeira) // **carteira caiu (derrubou a carteira, perdeu a carteira, perdeu a bolsa)** // da sua bolsa (da sua mochila, de sua pasta) // **Ela não viu a carteira cair (ela não notou)** // No caixa (quando ela foi pagar, guichê) // **não tem como pagar (ela não tinha dinheiro, não tinha sua carteira)** // Coloca as mercadorias de lado (coloca as mercadorias de volta) // **foi para sua casa (voltou para sua casa)** // Quando ela abriu a porta (quando ela chegou em casa, assim que ela entrou) // **telefone tocou (fone tocou, ela recebeu uma ligação)** // Pequena (jovem) // menina (garota) // lhe disse (falou, contou) // **ela achou a carteira (achou sua moedeira, achou o porta-notas)** // **Senhora aliviada (senhora estava feliz, senhora estava radiante, senhora estava agradecida)**

O segundo conjunto de dados é formado por transcrições da tarefa de reconto e compreensão de texto de uma história apresentada oralmente (História da Lúcia) da BALE, que possui originalmente 24 unidades de informação que foram reagrupadas neste trabalho (cf. Subseção 4.1.1), resultando em 21 unidades (Figura 11). O teste do reconto foi aplicado em 11 idosos com Alzheimer, 5 idosos com CCL e 53 adultos com envelhecimento saudável e estão em disponíveis em Hübner *et al.* (2019).

Figura 11 – Narrativa utilizada na BALE, separada em unidades de informação; as onze unidades da macroestrutura são marcadas em negrito.

Lúcia // mora // interior // do Paraná // Numa manhã de 2a feira // **ela saiu de casa** // **para buscar emprego (foi para uma uma entrevista, foi buscar trabalho)** // na capital do estado (em Curitiba) // **Foi para rodoviária** // **foi de carona (pegou carona)** // **com amigo Pedro (com Pedro)** // Estava chovendo // naquela manhã // O carro // passou (caiu) // por um buraco // **o pneu furou** // Pensou que ia perder (achou que ia perder) // o ônibus // **Pegou um táxi** // **conseguiu chegar chegou a tempo (chegou a tempo)**

Nas Figuras 10 e 11, as unidades da macroestrutura estão em negrito, seguindo o modelo de análise de (KINTSCH; DIJK, 1978) em que as unidades de informação do texto são organizadas de forma hierárquica; a macroestrutura correspondente às ideias principais e a microestrututura às ideias acessórias e detalhes. As duas baterias estão disponibilizadas publicamente[1].

[1] <https://github.com/nilc-nlp/DNLT-BP>

4.1.1 Metodologia Proposta para a Anotação das Unidades de Informação

Para cada conjunto de dados, o áudio do participante foi transcrito manualmente, seguindo os princípios do NURC/SP No 338 EF e 331 D^2 (PRETI, 2005) e segmentado manualmente em orações por um anotador experiente, usando conhecimento prosódico (pausas), sintático e semântico. Chamamos essas duas etapas de pré-processamento.

Na segmentação em orações, manteve-se as disfluências, uma vez que estas caracterizam fortemente os grupos clínicos, mas foram eliminadas as marcas de incompreensão de palavras/-segmentos, prolongamentos de vogais e consoantes, silabação, interrogação, pausas curtas e longas e comentários descritivos do transcritor. Primeiro segmentou-se as orações bem formadas sintaticamente e segmentaram-se também as orações coordenadas, pois estas formam ideias isoladamente. Palavras com ortografia incorreta não representam um problema para a tarefa de segmentação sentencial. Em seguida, delimitou-se as orações que são mal formadas sintática e/ou semanticamente.

Para criar os conjuntos de dados anotados com as unidades de informação sobre as unidades de interesse (orações anotadas no pré-processamento), utilizou-se o sistema de anotação *brat rapid annotation tool* (*brat*) (STENETORP *et al.*, 2012), realizando a anotação em duas fases. Na primeira fase, cada sentença da transcrição foi classificada de acordo com a lista de unidades de informação de cada bateria por um único anotador; na segunda fase, outro anotador revisou a anotação e os casos discordantes foram discutidos, visando a uma anotação concordante (cf. Figura 12).

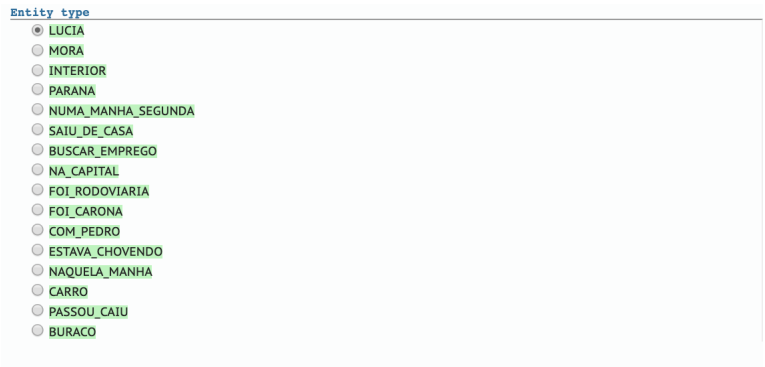
O reconto da *ABCD* foi mantido com as 17 unidades de informação originais, mas para as narrativas da BALE realizou-se algumas modificações nas unidades de informação (ora separando, ora juntando) para termos uma anotação manual uniforme, sem discrepâncias e possibilitar a aplicação de métodos automáticos. A partir dessas modificações, finalizamos com 21 unidades de informação (Figura 11) em vez das 24 unidades originais, com 14 delas sendo unidades macroestruturais. Dentre essas modificações, agrupou-se as unidades que eram precedidas pelo verbo "ir" como "foi para a rodoviária" e "foi de carona" e removeu-se a unidade de informação "foi". Também foram removidas as unidades que estavam repetidas (havia duas proposições "Lúcia" e duas proposições relacionadas com "a rodoviária", variando somente a preposição "para" e "até"), que dificultam a análise automática. Essas mudanças alteraram a pontuação máxima da narrativa de 24 para 21 pontos, com onze unidades macroestruturais. E se mostram uma limitação somente para anotação com repetições de trechos idênticos que usam diferentes categorizações na estrutura do texto. Mais especificamente, no caso da anotação de "Lúcia", na primeira vez é classificada como macroestrutura e na segunda lembrança anotada como unidade da microestrutura. Esse esquema pode, entretanto, ser anotado com indexação ("Lúcia1", "Lúcia2") ou com refraseamentos, como, por exemplo: "(foi) para a rodoviária", anotado como unidade macro e "(um táxi) até a rodoviária", anotado como micro, por ser um

Figura 12 – Exemplo da anotação das unidades de informação em uma narrativa com 12 orações e 19 unidades no *brat*.

Fonte: Elaborada pelo autor.

reforço somente. Entretanto, essa anotação com rótulos muito similares pode sobrecarregar o anotador, dado que a anotação usa uma lista de entidades descontextualizada (cf. Figura 13), levando a possíveis erros de anotação. Esta foi, então, a razão para alterar a pontuação de 24 para 21 pontos, para avaliar o sucesso (ou não) da anotação com menos pontos com vistas à identificação de semelhanças e diferenças entre os grupos de interesse.

Figura 13 – Exemplo do esquema de anotação com lista de entidades no *brat*.



Fonte: Elaborada pelo autor.

4.1.2 Caracterização da Anotação Manual

4.1.2.1 Descrição da Anotação Manual

Na Tabela 18, apresenta-se uma análise da quantidade das unidades de informação recordada por cada grupo da *ABCD*. A primeira coluna indica a unidade de informação; a segunda e a quarta colunas indicam o percentual que as unidades foram recordadas por cada grupo; a terceira e quinta colunas indicam a média e o desvio padrão da quantidade de vezes que a unidade foi recorda por cada grupo. Identificamos que os idosos do grupo de controle apresentaram uma porcentagem de unidades relembradas para componentes da microestrutura (KINTSCH; DIJK, 1978) da narrativa muito mais marcante (diferença maior que 5,8 pontos) do que os do grupo CCL para as unidades "Sua carteira", "Da sua bolsa" e "Pequena". Mas o que é interessante é a grande discrepância para a unidade da macroestrutura "Senhora ficou aliviada", que foi mais relembrada pelo grupo CCL (diferença de 38,4 pontos); o grupo CCL relembrou somente 1 elemento da microestrutura com diferença marcante (7,3 pontos) do que o grupo de controle ("Quando abriu a porta").

Tabela 18 – Porcentagem, média e desvio padrão das unidades de informação recordadas por cada grupo da ABCD. Unidades em negrito são unidades da macroestrutura.

	Controle		CCL	
Unidades de informação	**Unidades Recordadas %**	**Média (Desvio Padrão)**	**Unidades Recordadas %**	**Média (Desvio Padrão)**
Senhora	91,67	1,00 (0,42)	93,48	0,96 (0,29)
estava fazendo compras	91,67	1,25 (0,61)	97,83	1,13 (0,40)
Sua carteira	62,5	0,63 (0,49)	50	0,54 (0,59)
carteira caiu	58,33	0,58 (0,50)	47,83	0,52 (0,59)
da sua bolsa	16,67	0,17 (0,38)	10,87	0,11 (0,31)
Ela não viu	33,33	0,38 (0,58)	41,3	0,48 (0,66)
No caixa	83,33	0,96 (0,55)	82,61	0,85 (0,42)
não tem como pagar	83,33	0,88 (0,45)	86,96	1,00 (0,56)
Colocou de lado	75,00	0,75 (0,44)	78,26	0,78 (0,42)
foi para casa	91,67	0,96 (0,36)	89,13	0,89 (0,31)
Quando abriu a porta	66,67	0,67 (0,48)	73,91	0,74 (0,44)
telefone tocou	91,67	0,92 (0,28)	91,3	0,91 (0,28)
Pequena	70,83	0,75 (0,53)	52,17	0,52 (0,51)
menina	87,5	0,92 (0,41)	82,61	0,83 (0,38)
lhe disse	83,33	0,83 (0,38)	84,78	0,85 (0,36)
achou carteira	95,83	1,04 (0,36)	93,48	0,98 (0,33)
Senhora ficou aliviada	33,33	0,38 (0,58)	71,74	0,74 (0,49)

Na Tabela 18 é apresentada uma análise da quantidade das unidades de informação recordadas por cada grupo da BALE. A primeira coluna indica a unidade de informação; a segunda, quarta e sexta colunas indicam o percentual que as unidades foram recordadas por cada grupo; a terceira, quinta e oitava colunas indicam a média e o desvio padrão da quantidade de vezes que a unidade foi recorda por cada grupo. Diferentemente da *ABCD*, os idosos do grupo de controle apresentaram um número de unidades relembradas maior (com diferença marcante) do que os pacientes do grupo CCL para várias unidades de informação da microestrutura como "Paraná", "Numa manhã de segunda-feira", "na capital", "estava chovendo", "passou", "pensou

que ia perder". Os idosos do grupo Alzheimer apresentaram um número menor de unidade de informação recordadas quando comparados com grupo de indivíduos saudáveis ou mesmo com grupo com CCL. Essa diferença é mais acentuada nas seguintes unidades: "Paraná", "Numa manhã de segunda-feira", "na capital", "estava chovendo", "carro", "passou","buraco", "pensou que ia perder".

Em geral, os idosos do grupo CCL apresentaram uma taxa de recordação maior que o grupo com Alzheimer, exceto para as unidades de informação "Lúcia", "Foi para rodoviária" e "Pegou um táxi".

Tabela 19 – Porcentagem, média e desvio padrão das unidades de informação recordadas por cada grupo da BALE. Unidades em negrito são unidades da macroestrutura.

	Controle		CCL		Alzheimer	
Unidades de informação	**Unidades Recordadas %**	**Média (Desvio Padrão)**	**Unidades Recordadas %**	**Média (Desvio Padrão)**	**Unidades Recordadas %**	**Média (Desvio Padrão)**
Lucia	96,23	1,04(0,34)	80	0,8(0,45)	90,91	1,18(0,60)
mora	66,04	0,68(0,51)	20	0,2(0,45)	9,09	0,09(0,30)
interior	54,72	0,58(0,57)	20	0,2(0,45)	9,09	0,09(0,30)
Paraná	66,04	0,68(0,51)	20	0,20(0,45)	9,09	0,09(0,30)
Numa manhã de segunda-feira	13,21	0,13(0,34)	0	–	0	–
saiu de casa	5,66	0,06(0,23)	20	0,20(0,45)	0	–
buscar emprego	56,6	0,62(0,60)	20	0,20(0,45)	18,18	0,18(0,40)
na capital	13,21	0,13(0,34)	0	–	0	–
Foi para rodoviária	54,72	0,64(0,65)	20	0,20(0,45)	36,36	0,45(0,69)
foi de carona	54,72	0,58(0,57)	40	0,60(0,89)	27,27	0,27(0,47)
com o Pedro	43,4	0,45(0,54)	20	0,20(0,45)	0	–
Estava chovendo	28,3	0,38(0,69)	20	0,20(0,45)	9,09	0,09(0,30)
naquela manhã	3,77	0,04(0,19)	0	–	0	–
Carro	62,26	0,64(0,52)	60	0,80(0,84)	18,18	0,18(0,40)
passou	35,85	0,36(0,48)	20	0,20(0,45)	18,18	0,18(0,40)
buraco	43,4	0,43(0,50)	40	0,40(0,55)	27,27	0,27(0,47)
pneu furou	71,7	0,75(0,52)	60	0,80(0,84)	45,45	0,55(0,69)
Pensou que iria perder	49,06	0,49(0,50)	40	0,40(0,55)	18,18	0,18(0,40)
ônibus	30,19	0,30(0,46)	60	0,80(0,84)	27,27	0,27(0,47)
Pegou um táxi	64,15	0,72(0,60)	40	0,60(0,89)	45,45	0,45(0,52)
conseguiu chegar a tempo	56,6	0,58(0,53)	60	0,60(0,55)	9,09	0,09(0,30)

4.1.2.2 Análise Automática das Narrativas

Para descrever automaticamente os conjuntos de dados, utilizou-se métricas comumente utilizadas na tarefa de classificação automática de narrativas (ROARK *et al.*, 2011; ALUÍSIO; CUNHA; SCARTON, 2016; SANTOS *et al.*, 2017; FRASER; FORS; KOKKINAKIS, 2019) ou na análise de narrativas, para caracterização de métricas discriminativas para avaliação clínica (TOLEDO *et al.*, 2018).

As métricas selecionadas se dividem em: (i) contagens básicas (média de palavras por sentença, média de sentenças da narrativa, razão de substantivos por palavras do texto, razão de verbos por palavras do texto); (ii) métricas baseadas na análise sintática (distância de dependência, complexidade de Yngve (YNGVE, 1960), complexidade de Frazier (FRASER; FORS; KOKKINAKIS, 2019), quantidade média de orações por sentenças e a média dos tamanhos médios dos sintagmas nominais nas sentenças). Não realizou-se nenhum tratamento

para remover disfluências automaticamente porque visou-se a construção de um *dataset gold standard*, embora haja um sistema, chamado de DeepBonDD[2], que extrai automaticamente as disfluências (cf. Treviso e Aluísio (2018)). Este sistema remove as disfluências do tipo pausas preenchidas e marcadores do discurso com bastante precisão, embora os tipos de disfluências mais complexos (repetições e revisões) não tenham a mesma precisão.

Na Tabela 20, apresenta-se os resultados da aplicação das 9 métricas. Na *ABCD*, os valores das métricas são muito próximos para os dois grupos analisados. Utilizou-se o teste de *Mann-Whitney* com um intervalo de confiança de 95% e não encontrou-se diferença estatística significativa entre os grupos. Para a BALE, utilizou-se o teste estatístico *Kruskal-Wallis*, pois este conjunto contém três grupos, e o pós-teste de *Dunnett* com um intervalo de confiança de 95%. Encontrou-se resultados estatisticamente relevantes entre os idosos dos grupos Controle vs CCL e CLL vs Doença de Alzheimer para uma métrica do grupo morfossintáticas (**Razão de substantivos por palavras do texto**) com p-valor de 0,0192 e 0,0170, respectivamente, e entre os idosos do grupo de Controle e Doença de Alzheimer para a métrica sintática **Complexidade de Yngve** com p-valor de 0,0128.

Tabela 20 – O valores médios (desvio padrão) das métricas por cada grupo clínico.

Métricas	***ABCD***		**BALE**		
	Controle	**CCL**	**Controle**	**CCL**	**Alzheimer**
Complexidade de Yngve	1,78 (0,13)	1,78 (0,12)	1,82 (0,17)	1,72 (0,13)	1,64 (0,22)
Complexidade de Frazier	6,79 (0,48)	6,64 (0,40)	6,59 (0,52)	6,67 (0,26)	6,31 (0,48)
Distância de dependência	11,35 (3,34)	10,38 (3,11)	8,66 (3,25)	7,74 (1,33)	7,33 (2,35)
Número de sentenças	7,67 (2,06)	8,17 (1,92)	7,68 (2,67)	6,00 (1,00)	6,09 (2,63)
Média de Palavras por Sentença	7,77 (1,30)	7,41(1,41)	6,81 (1,58)	6,11 (0,73)	5,85 (1,46)
Quantidade média de orações por sentença	3,09 (0,60)	2,82 (0,60)	2,31 (0,86)	2,57 (0,60)	2,10 (1,09)
Média dos tamanhos médios dos sintagmas nominais nas sentenças	2,52 (0,59)	2,43 (0,77)	2,84 (0,93)	2,92 (0,96)	2,47 (0,69)
Razão de substantivos por palavras do texto	0,24 (0,03)	0,24 (0,03)	0,30 (0,06)	0,24 (0,04)	0,31 (0,05)
Razão de verbos por palavras do texto	0,29 (0,04)	0,29 (0,04)	0,23 (0,04)	0,26 (0,05)	0,23 (0,04)

4.2 Inferência de Propriedades Psicolinguísticas

Além da compilação e anotação dos conjuntos de dados usados na pesquisa, desenvolveu-se um dicionário com as seguintes propriedades das palavras: imageabilidade, concretude, frequência subjetiva e idade de aquisição (*AoA*, *Age of Aquisition*). Imageabilidade é a facilidade e rapidez com que uma palavra evoca uma imagem mental; concretude é o grau para o qual palavras se referem a objetos, pessoas, lugares ou coisas que podem ser experienciadas pelos sentidos, frequência subjetiva é a estimativa do número de vezes que uma palavra é usada em sua forma escrita ou falada, e *AoA* é a estimativa da idade em que uma palavra foi aprendida, por uma pessoa.

[2] <https://mtreviso.github.io/deepbond/>

Essas propriedades são conhecidas como propriedades psicolinguísticas subjetivas que não podem ser extraídas diretamente de algum córpus, como a frequência de palavras, sem um processo de inferência, pois dependem das experiências que os indivíduos têm relacionadas a elas. Dessa forma, é necessária a aplicação de questionários nos quais os indivíduos avaliam as propriedades das palavras em uma dada escala numérica. Devido aos seus custos inerentes, a medição de propriedades psicolinguísticas é realizada em conjuntos de dados de tamanho pequeno/limitado (CAMEIRAO; VICENTE, 2010; JANCZURA *et al.*, 2007; MARQUES, 2005; SOARES *et al.*, 2016).

As propriedades psicolinguísticas subjetivas das palavras vêm sendo utilizadas para construção de atributos em várias tarefas do PLN, como simplificação de texto (VAJJALA; MEURERS, 2014; PAETZOLD; SPECIA, 2016b), avaliação de inteligibilidade (GRAESSER; MCNAMARA; KULIKOWICH, 2011), e classificação de narrativas de reconto (FRASER; MELTZER; RUDZICZ, 2016).

Para o idioma inglês, o banco de dados mais conhecido desse tipo é o *The Medical Research Council* (*MRC*) *Psycholinguistic Database*[3], que contém 27 propriedades psicolinguísticas subjetivas para 150.837 palavras.

Para a Português Brasileiro (PB), antes do desenvolvimento da base descrita nesta seção, existia um banco de dados psicolinguístico chamado Léxico do Português Brasileiro (Lex-PorBR)[4], contendo 21 propriedades de 215.175 palavras, mas sem propriedades psicolinguísticas subjetivas, de interesse desta pesquisa de doutorado.

Para superar essa lacuna, inferiu-se automaticamente as propriedades psicolinguísticas de imaginabilidade, concretude, *AoA* e frequência subjetiva de um grande banco de dados de 26.874 palavras de PB, usando uma combinação de regressores. Este trabalho baseou-se fortemente nos resultados de (PAETZOLD; SPECIA, 2016b) que propuseram um método automático para ampliar o banco de dados *MRC*, para a língua inglesa.

Até onde sabemos, existem apenas dois estudos que propõem métodos de regressão para estimar automaticamente as propriedades psicolinguísticas ausentes no banco de dados *MRC* (FENG *et al.*, 2011; PAETZOLD; SPECIA, 2016b).

Feng *et al.* (2011) propuseram um método para prever a concretude de palavras usando regressão linear com os seguintes atributos: (i) 21 categorias lexicais da *WordNet* (FELLBAUM, 1998), sendo que para cada categoria é calculada a porcentagem de sentidos que a palavra tem nessa categoria; (ii) 37 dimensões de um modelo de Análise Semântica Latente (*LSA*), (iii) o log da frequência da palavra no *CELEX Database*[5], e (iv) a quantidade de caracteres da palavra. Obtiveram uma correlação de Pearson de 0,82 entre os escores estimados de concretude e o escore de concretude no conjunto de testes.

[3] <websites.psychology.uwa.edu.au/school/MRCDatabase/mrc2.html>
[4] <www.lexicodoportugues.com>
[5] <celex.mpi.nl>

Paetzold e Specia (2016b) inferiu as propriedades psicolinguísticas ausentes no banco de dados *MRC* por meio de uma regressão com o método *bootstrapping*. Os autores exploraram modelos de *word embeddings* e 15 atributos lexicais, incluindo o número de sentidos, sinônimos, hiperônimos e hipônimos das palavras na *WordNet*, e a distância mínima, máxima e média entre os sentidos da palavra na *WordNet* e em um tesauro. A correlação de Pearson entre a pontuação estimada e a pontuação inferida para familiaridade foi de 0,846; 0,862 para *AoA*; 0,823 para imageabilidade e 0,869 para concretude.

4.2.1 Criação de um Léxico com Propriedades Psicolinguísticas

Na Tabela 21 são apresentados os dicionários com propriedades psicolinguísticas subjetivas para o Português Europeu (PE) e o Português Brasileiro (PB) usados para a inferência de propriedades psicolinguísticas desta pesquisa.

Tabela 21 – Normas para o português focadas em propriedades psicolinguísticas subjetivas.

Trabalho	Participantes	Palavras	Propriedade	Variante	Escala
Soares *et al.* (2016)	2357	3789	concretude, imageabilidade, frequência subjetiva	PE	1-7
Cameirao e Vicente (2010)	685	1748	*AoA*	PE	1-9
Janczura *et al.* (2007)	719	909	concretude	PB	1-7
Marques *et al.* (2007)	110	834	*AoA*	PE	1-7
Marques (2005)	103	249	imageabilidade, concretude	PE	1-7

Fonte: Adaptada de Santos *et al.* (2017).

Para o Português Brasileiro, há apenas 909 palavras com valores de concretude (JANCZURA *et al.*, 2007). Portanto, incorporou-se os léxicos em PE ao conjunto em PB. Para tornar os dicionários em Português Europeu utilizáveis para a pesquisa, realizou-se ajustes nas listas de palavras. A grande parte dos ajustes ocorreu na ortografia, como por exemplo: a palavra "acção" foi convertida para "ação", "adotoução" foi convertida para "adoção". Outros ajustes se referiram a conceitos que as duas variantes do português lexicalizam de maneiras diferentes, como: "ficheiro"/"arquivo", "assassínio"/"assassinato," "apuramento"/"apuração". Por fim, algumas palavras foram descartadas, pois lexicalizam conceitos relacionados à fauna, flora e traços culinários nativos de Portugal.

Após concluir os ajustes lexicais, converteu-se a escala de 9 pontos de Cameirao e Vicente (2010) para 7 pontos. Por fim, 6 dicionários escolares que serviram de atributos indicativos de idades/séries escolares foram selecionados[6] e foram combinados, eliminando-se palavras duplicadas. Os dicionários são enumerados abaixo:

1. Dicionário Ilustrado 1500 Palavras, Douglas Tufano, Moderna, 1996, com 1500 palavras, com letras e adjetivos de estados. Especialmente desenvolvido para 1ª e 2ª séries;

[6] Usamos dicionários sugeridos pelo Programa Nacional do Livro Didático (PNLD) do Ministério da Educação (MEC).

2. O Aurélio Com A Turma da Mônica - Dicionário Infantil Ilustrado, Escolha de palavras pertencentes à noção de dicionário, composta por duas partes: uma parte temática e um dicionário elementar. Apresenta 19 temas na primeira parte, que visam criar, em diferentes dimensões e níveis de relações, um primeiro acesso ao mundo das palavras, de modo a defini-las ou explicá-las de maneira variada a partir da contextualização delas. 1435 palavras (removidas as duplicações de gênero e adjetivos de estados). Idade Recomendada: de 8 a 11 anos;

3. Dicionário Ilustrado de Português, Maria Tereza Camargo Biderman, 2010, com 5850 palavras, com adjetivos de estados;

4. Meu Primeiro Dicionário, Caldas Aulete, com a Turma do Cocoricó, Lexikon, 2009. Indicado para crianças de 6 a 8 anos (3 primeiras anos escolares); 1371 palavras (removidas as duplicações de gênero; não há adjetivos de estados);

5. Dicionário Escolar da Língua Portuguesa Ilustrado com a TURMA DO SITIO DO PICA-PAU AMARELO, Editora Globo, 2009. Este dicionário é dirigido para alunos das primeiras séries do ensino fundamental; com 7002 palavras, incluíndo os adjetivos de estados. Indicado para os 4 e 5 anos;

6. Minidicionário Contemporâneo da Língua Portuguesa, Caldas Aulete, Lexikon Editorial, 2009. Com 31.000 palavras e locuções, com as letras. Indicado para os 6 a 9 anos.

A Tabela 22 mostra o número de entradas obtidas para cada propriedade, entre parênteses.

Para inferir as propriedades psicolinguísticas, utilizou-se um *ensemble* de regressores com 10 atributos treinados nos dicionários descritos anteriormente. Os atributos foram agrupados em três categorias, e para cada uma foi treinado um regressor: (i) lexicais (1-8); (ii) *word embeddings* do *Skip-Gram* (9); e (iii) *word embeddings* do *GloVe* (10):

1. Logaritmo da frequência no SUBTLEX-pt-BR (TANG, 2012), que é uma base de frequência em PB com mais de 50 milhões de palavras extraídas de legendas de séries e filmes;

2. Logaritmo da quantidade de legendas que contém a palavra;

3. Logaritmo da frequência no SubIMDb-PT (PAETZOLD; SPECIA, 2016a), que é córpus de legendas de filmes e séries para crianças e família;

4. Logaritmo da frequência da parte escrita do Córpus Brasileiro (SARDINHA; FILHO; ALAMBERT, 2008), que é córpus com aproximadamente um bilhão de palavras de Português Brasileiro;

5. Logaritmo da frequência da parte falada do Córpus Brasileiro (SARDINHA; FILHO; ALAMBERT, 2008),

6. Logaritmo de frequência do córpus do repositório NILC-*Embeddings* (HARTMANN *et al.*, 2017), com 1,4 bilhões de *tokens* de gêneros de texto misto no Português Brasileiro;

7. Tamanho da palavra;

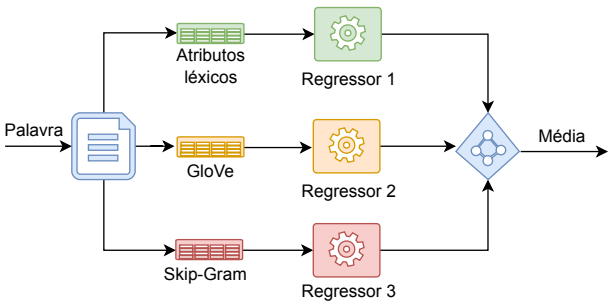
8. Criou-se um atributo categórico indicando qual é o primeiro dicionário que contém a entrada dessa palavra, sendo que cada dicionário é especifico para uma determinada série escolar. Assim, quanto mais complexa a palavra, maior será o valor desse atributo. O valor máximo do atributo é cinco (quando não é encontrada a palavra nos dicionários) e o valor mínimo é um (quando palavra é encontrada no primeiro dicionário);

9. *Word embeddings* do modelo *Skip-Gram*;

10. *Word embeddings* do modelo *GloVe*.

Utilizou-se um regressor linear de mínimos quadrados com regularização L2, também conhecido como Regressão *Ridge* (MURPHY, 2012). Selecionou-se esse método de regressão devido aos resultados promissores relatados por Paetzold e Specia (2016b). Foram treinados três regressores em diferentes espaços de atributos: lexicais, *word embeddings* do *Skip-Gram* e *word embeddings* do *GloVe*. O resultado final é obtido pela média dos três regressores. Na Figura 14 é apresentado o processo para inferir as propriedades. Essa abordagem possibilita em trabalhos futuros a exploração do melhor regressor para cada espaço de atributos.

Figura 14 – Arquitetura para inferir as propriedades psicolinguísticas das palavras.



Fonte: Adaptada de Santos *et al.* (2017).

Os experimentos foram avaliados com *20x5-fold-cross-validation*, e com as seguintes métricas de avaliação: o *Mean Square Error* (*MSE*), correlação Spearman (ρ), correlação e Pearson (r). Para os regressores que utilizam *words embeddings* avaliou-se modelos com 300, 600 e 1000 dimensões; os melhores resultados foram com 300 dimensões para ambos.

Na Tabela 22, são apresentados os resultados dos experimentos; os modelos com o menor *MSE* estão negrito. Aplicou-se o teste estatístico ANOVA com o pós-teste de Dunnett e com nível de significância de 0,05. Quase todos os melhores resultados apresentaram diferença estatisticamente significativa com os demais regressores, exceto para a propriedade *AoA*.

Para frequência subjetiva, o melhor resultado foi obtido pela combinação dos modelos Lexical, Skip-gram e GloVe. Para concretude e a imageabilidade os melhores resultados foram com a combinação dos modelos *Skip-Gram* e *GloVe*. Enquanto que para o *AoA*, o melhor resultado foi obtido pela combinação dos modelos Lexical e *GloVe*.

Tabela 22 – Resultado de *MSE*, correlação de Pearson, e correlação de Spearman dos regressores.

Regressores	**Concretude** (4088)			**Frequência subjetiva** (3735)			**Imageabilidade** (3735)			***AoA*** (2368)		
	MSE	*r*	ρ	*MSE*	*r*	ρ	MSE	*r*	ρ	*MSE*	*r*	ρ
Lexical	1,24	0,54	0,56	0,55	0,72	0,73	0,74	0,58	0,59	0,67	0,73	0,73
Skip-gram	0,52	0,84	0,84	0,58	0,70	0,71	0,46	0,77	0,77	0,81	0,66	0,66
GloVe	0,62	0,80	0,81	0,40	0,81	0,81	0,49	0,75	0,75	0,63	0,75	0,75
Lexical + *Skip-gram*	0,64	0,82	0,82	0,44	0,79	0,79	0,47	0,77	0,78	0,59	0,77	0,77
Lexical + *GloVe*	0,70	0,80	0,80	0,39	0,81	0,81	0,50	0,75	0,76	**0,54**	**0,79**	**0,79**
Skip-gram + *GloVe*	**0,49**	**0,85**	**0,85**	0,41	0,80	0,80	**0,42**	**0,79**	**0,79**	0,62	0,75	0,75
Lexical + *Skip-gram* + *GloVe*	0,55	0,85	0,84	**0,38**	**0,82**	**0,82**	0,43	0,79	0,78	0,54	0,79	0,79

Fonte: Adaptada de Santos *et al.* (2017).

Por fim, construiu-se um banco de dados de palavras simples, preenchido com os valores inferidos para as quatro propriedades psicolinguísticas. Para isso, explorou-se as entradas do Minidicionário Caldas Aulete (GEIGER, 2011) e sua respectiva primeira categoria gramatical. Selecionou-se apenas substantivos, verbos, adjetivos e advérbios. As palavras estrangeiras foram removidas. Em seguida, analisou-se a frequência de cada palavra no grande córpus de 1,4 bilhão de palavras que foi usado para treinar os modelos de *word embeddings*. Após uma análise manual, as palavras com menos de 8 ocorrências foram eliminadas, pois são muito incomuns. O léxico final está disponível no portal PortLex[7] e contém 26.874 palavras, sendo 15.204 substantivos, 4.305 verbos, 7.293 adjetivos e 72 advérbios com as informações das quatro propriedades psicolinguísticas inferidas, usando os melhores resultados com a menor quantidade de atributos (mostrados em negrito na Tabela 22).

4.3 Identificação Automática de Unidades de Informação em Recontos

4.3.1 *Exploração da Similaridade Semântica*

Métodos de similaridade semântica e inferência textual têm aplicações em diversas tarefas de PLN como: recuperação de informação, sistemas de perguntas-repostas, avaliação de sistemas de tradução, entre outras (AGIRRE *et al.*, 2012; AGIRRE *et al.*, 2015).

Uma das questões de pesquisa investigadas nesta tese foi a possibilidade de se utilizar sistemas de *STS* para identificar as unidades de informação recordadas e, até onde sabemos, essa abordagem é inédita.

[7] <http://nilc.icmc.usp.br/portlex/index.php/en/psycholinguistic>

Na Tabela 23, são apresentados exemplos de sentenças das narrativas de reconto e as sentenças da narrativa original (os dois primeiros são da História da da Carteira da *ABCD* e os dois últimos da História da Lúcia da BALE). Na primeira coluna, temos os valores de similaridade que foram obtidos pelo sistema *Solo Queue*[8] (HARTMANN, 2016), que obteve os melhores resultados na avaliação conjunta ASSIN, de 2016. Nesse exemplo, é possível perceber a viabilidade da exploração de *STS*. Nesta tese, chamamos esse método de STS. As duas primeiras linhas da Tabela 23 apresentam valores altos de similaridade semântica; esses valores indicam que as sentenças são muito semelhantes, mas apresentam algumas informações exclusivas, pois a pontuação máxima é 5. Enquanto os terceiro e quarto pares de sentenças apresentam valores de similaridade próximos de 3, sendo que esse valor indica que as sentenças possuem alguma similaridade e podem se referir ao mesmo fato, pela definição da tarefa de anotação do ASSIN.

Tabela 23 – Exemplos para os valores de similaridade semântica.

Similaridade	**Sentença do reconto**	**Sentença da narrativa original**
4,17	e ela ficou aliviada.	a senhora ficou muito aliviada.
4,55	uma senhora fazia as compras no mercado.	uma senhora fazia compras.
3,09	e ai foi pegou um táxi pra chegar com tempo.	então ela pegou um táxi até a rodoviária.
2,93	ela pegou carona.	ela foi para a rodoviária de carona com seu amigo.

Dado que o sistema de *STS* recebe como entrada dois textos curtos, para possibilitar a aplicação desse sistema na tarefa desta tese as narrativas originais de cada bateria foram sentenciadas e para cada sentença foram atribuídos seus respectivos rótulos de forma manual. Nas Tabelas 24 e 25 são apresentados os resultados dessa etapa. Assim como nas sentenças das narrativas dos pacientes (veja Seção 4.1), algumas sentenças possuem mais que um rótulo.

Tabela 24 – Sentenças da narrativa da BALE rotuladas com as unidades de informação.

Sentenças	**Rótulos**
Lúcia mora no interior do Paraná	LUCIA; MORA; INTERIOR; PARANA
numa manhã de segunda-feira	NUMA_MANHA_SEGUNDA
ela saiu de casa para mais uma entrevista de trabalho na capital do estado	SAIU_DE_CASA; BUSCAR_EMPREGO; NA_CAPITAL
ela foi para a rodoviária de carona com seu amigo Pedro	FOI_RODOVIARIA; COM_PEDRO
estava chovendo naquela manhã	ESTAVA_CHOVENDO; NAQUELA_MANHA
de repente o carro passou por um buraco	CARRO; PASSOU_CAIU; BURACO
e o pneu furou	PNEU_FUROU
Lúcia pensou que iria perder o ônibus	PENSOU_ACHOU_PERDER ONIBUS
então ela pegou um táxi até a rodoviária	PEGOU_TAXI
e conseguiu chegar a tempo	CONSEGUIU_CHEGAR_TEMPO

Como o córpus descrito na Seção 4.1.1 foi anotado no nível sentencial, cada sentença pode possuir mais do que uma unidade de informação. Na área de Aprendizado de Máquina, a

[8] O autor gentilmente nos forneceu o código fonte e os modelos utilizados no sistema.

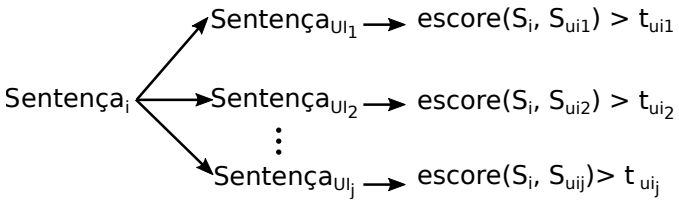
Tabela 25 – Sentenças da narrativa da *ABCD* rotuladas com as unidades de informação.

Sentenças	**Rótulos**
uma senhora fazia compras	SENHORA; ESTAVA_FAZENDO_COMPRAS
sua carteira caiu da bolsa	SUA_CARTEIRA; CARTEIRA_CAIU; DA_SUA_BOLSA
mas ela não viu	ELA_NAO_VIU_A_CARTEIRA_CAIR
quando ela foi ao caixa	NO_CAIXA
não tinha como pagar as compras	NAO_TEM_COMO_PAGAR
então ela colocou as compras de lado	COLOCA_AS_MERCADORIAS_DE_LADO
foi para casa	FOI_PARA_SUA_CASA
assim que ela abriu a porta da casa	QUANDO_ELA_ABRIU_A_PORTA
o telefone tocou	TELEFONE_TOCOU
uma menininha disse-lhe que tinha achado a carteira	PEQUENA; MENINA; LHE_DISSE ELA_ACHOU_A_CARTEIRA
a senhora ficou muito aliviada	SENHORA_ALIVIADA

identificação de mais de uma classe por exemplo é chamada de classificação multirrótulo. Neste trabalho, converteu-se o problema multirrótulo em vários problemas de classificação binária utilizando a abordagem *Binary Relevance*. Nessa abordagem são criados N conjuntos de dados com a mesma quantidade de exemplos, em que N é número de classes.

Na Figura 15 é apresentado o processo de identificação de unidades de informação. Para cada sentença da narrativa de reconto é calculada a similaridade com as sentenças da narrativa original, sendo que cada sentença contém uma unidade de informação (Tabelas 25 e 24). O valor de similaridade é comparado com um *threshold*, e caso seja maior considera-se que a unidade de informação está presente na sentença da narrativa de reconto. Assim, para cada sentença é criado um par $(\mathbf{S}_i^1, \mathbf{S}_{UI_j}^2)$, em que UI_j é a sentença que contém a unidade de informação j, e para cada par é obtido o valor de similaridade.

Figura 15 – Processo de identificação de unidades de informação.



Fonte: Elaborada pelo autor.

Na Tabela 26, são apresentados os valores de similaridade para a sentença "*uma senhora fazia as compras no mercado*", comparada com as sentenças que estão associadas a cada rótulo da narrativa original.

Na Figura 16, são mostrados os histogramas e a estimação de densidade por *kernel* para cada unidade de informação da *ABCD*. É possível perceber a separação para algumas unidades de

Tabela 26 – Valor de similaridade da sentença “uma senhora fazia as compras no mercado”.

Rótulo	**Sentença**	**Similaridade**
SENHORA	uma senhora fazia compras	4,55397
ESTAVA_FAZENDO_COMPRAS	uma senhora fazia compras	4,55397
SUA_CARTEIRA	sua carteira caiu da bolsa	1,2814
CARTEIRA_CAIU	sua carteira caiu da bolsa	1,2814
DA_SUA_BOLSA	sua carteira caiu da bolsa	1,2814
ELA_NAO_VIU_A_CARTEIRA_CAIR	mas ela não viu	1,24164
NO_CAIXA	quando ela foi ao caixa	1,20222
NAO_TEM_COMO_PAGAR	não tinha como pagar as compras	2,71347
COLOCA_AS_MERCADORIAS_DE_LADO	então ela colocou as compras de lado	2,48818
FOI_PARA_SUA_CASA	foi para casa	1,31341
QUANDO_ELA_ABRIU_A_PORTA	assim que ela abriu a porta da casa	1,46262
TELEFONE_TOCOU	o telefone tocou	1,08743
PEQUENA	uma menininha disse-lhe que tinha achado a carteira	1,08353
MENINA	uma menininha disse-lhe que tinha achado a carteira	1,08353
LHE_DISSE	uma menininha disse-lhe que tinha achado a carteira	1,08353
ELA_ACHOU_A_CARTEIRA	uma menininha disse-lhe que tinha achado a carteira	1,08353
SENHORA_ALIVIADA	a senhora ficou muito aliviada	2,52263

informação, como: SENHORA, ESTAVA_FAZENDO_COMPRAS, QUANDO_ELA_ABRIU_-A_PORTA. Entretanto, outras não apresentam uma separação clara, como ELA_NAO_VIU_A_-CARTEIRA_CAIR.

Dado o valor de similaridade semântica do par $(\mathbf{S}_i^1, \mathbf{S}_{UI_j}^2)$, é necessário definir um ponto de corte para transformar esse valor em uma resposta binária. Para encontrar o ponto de corte que maximizasse a medida $F1$ para a classe UI_j, aplicamos um otimizador Bayesiano com a técnica *Tree-of-Parzen-Estimators* (*TPE*) (BERGSTRA; YAMINS; COX, 2013), utilizando mil iterações para cada classe.

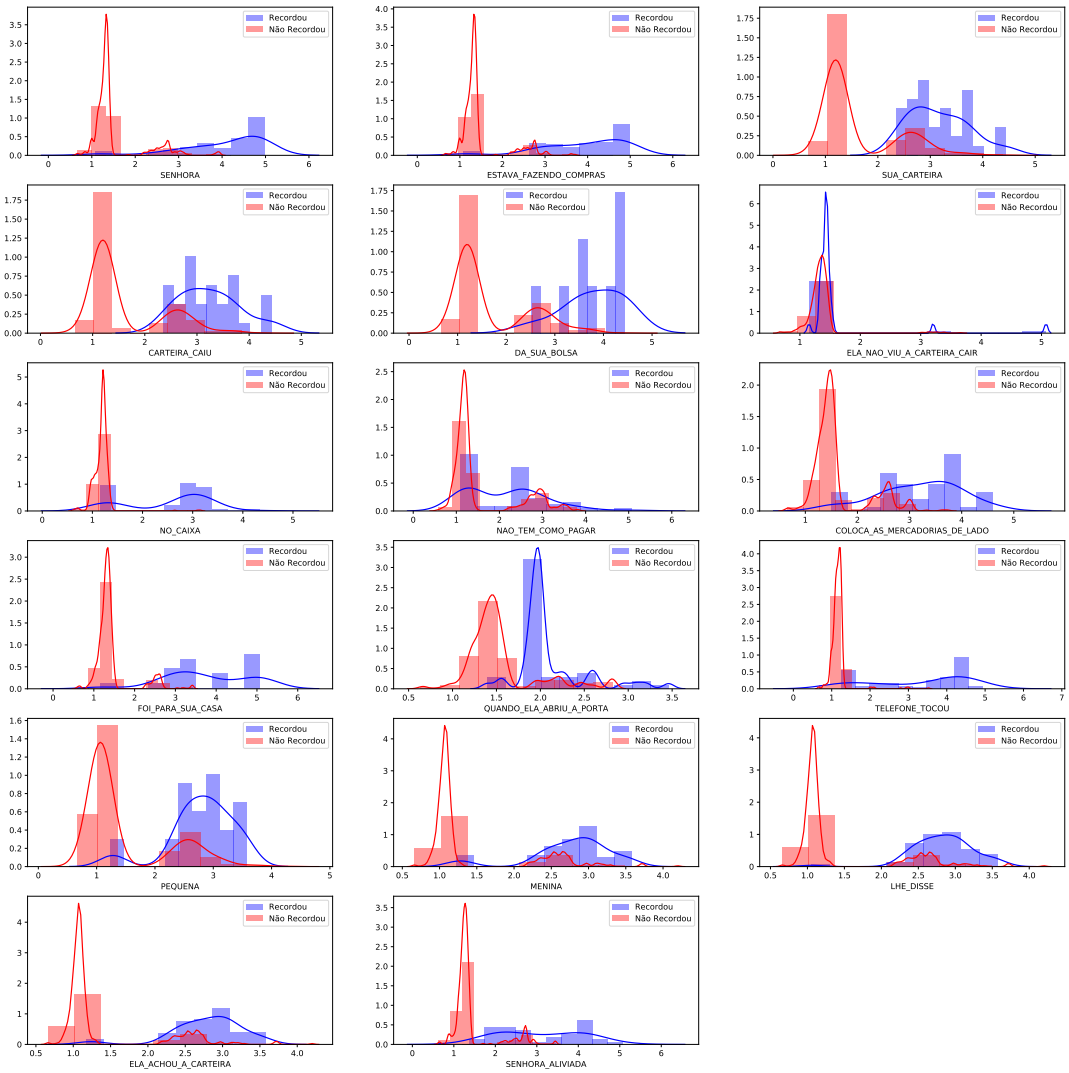
Os passos do segundo método avaliado e chamado aqui de *Chunking* são elencados abaixo:

1. Utilizou-se um *tagger* probabilístico (LÓPEZ; PARDO, 2015) que atribui a classe gramatical mais frequente do conjunto de dados para filtrar as palavras de conteúdo das sentenças dos recontos, de forma semelhante aos trabalhos de Yancheva e Rudzicz (2016) e Fraser, Fors e Kokkinakis (2019). Escolhemos esse *tagger*, pois as narrativas de reconto possuem ruídos que podem afetar o desempenho de *PoS taggers* treinados em córpus.

2. Em seguida, as palavras são convertidas para uma representação densa com o *FastText*.

3. Como o objetivo é identificar se a sentença contém uma respectiva unidade de informação ou não, para cada sentença cria-se um par $(\mathbf{S}_i^1, \mathbf{S}_{UI_j}^2)$, sendo que UI_j é a unidade de

Figura 16 – Histograma e distribuição acumulada para cada rótulo da ABCD.



informação j, e para cada par é obtido o valor de similaridade.

4. Dada uma sentença da narrativa de reconto, $\mathbf{S}_i^1$, esta é dividida em *n-grams*, variando de 1 a 3. Para cada n-gram, é calculada a média dos vetores que compõem esse *n-gram*.

5. É calculada a média dos vetores em $\mathbf{S}_{UI_j}^2$, e assim, obtém-se uma representação densa da sentença.

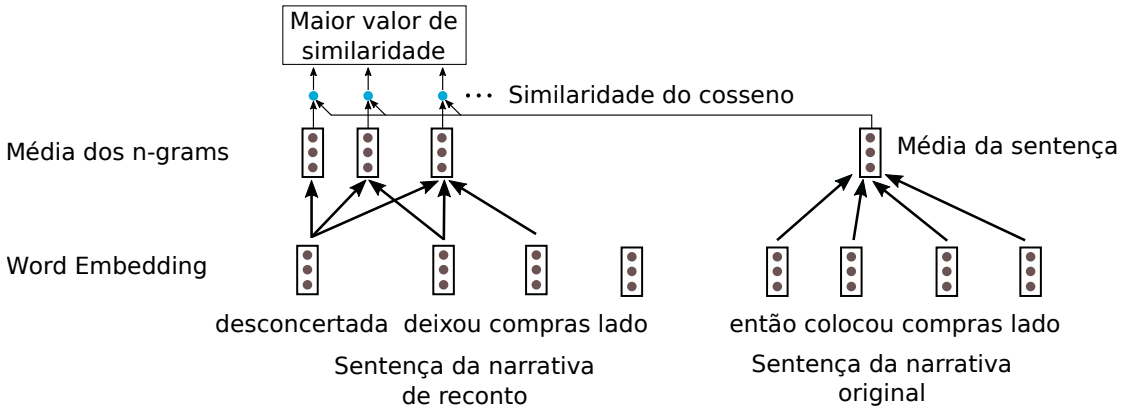
6. Por fim, calcula-se a similaridade do cosseno desses vetores e retorna-se o valor mais próximo de $\mathbf{S}_{UI_j}^2$.

7. Assim como no método anterior utilizou-se um otimizador Bayesiano com a técnica *TPE* para encontrar o ponto de corte.

Na Figura 17 é apresentado o funcionamento do método para obter a similaridade da sentença "*desconcertada ela deixou as compras de lado*" e a unidade de informação COLOCA_-AS_MERCADORIAS_DE_LADO que está associada a sentença "*então ela colocou as compras*

de lado". O primeiro passo do método é filtrar as palavras de conteúdo, assim, removemos as seguintes palavras funcionais de ambas as sentenças: "*ela*", "*as*" e "*de*". Na segunda etapa são obtidos os *embeddings* de cada palavra. Na etapa 4 é obtida a média dos seguintes *n-grams*: "*desconcertada*", "*desconcertada ela*", "*desconcertada ela deixou*" e assim por diante. Na etapa 5 é obtida a média da sentença da narrativa original. Na última etapa para obtenção da similaridade, é retornado o valor da similaridade do cosseno do *chunking* mais próximo de $\mathbf{S}^2_{UI_j}$. Nesse exemplo, o valor é 0,90 para o *chunking* "*deixou compras lado*".

Figura 17 – Método *Chunking* para a cálculo da similaridade.



Fonte: Elaborada pelo autor.

4.3.2 Baselines

Para comparar os métodos apresentados na Seção 4.3.1 na tarefa de identificação de unidades de informação, utilizou-se duas *baselines*.

A primeira, chamada de *Casamento exato*, utiliza uma lista de palavras para identificar as unidades de informação; essa abordagem também foi utilizada em outros trabalhos (PRUD'HOMMEAUX; ROARK, 2015; PAKHOMOV *et al.*, 2010; FRASER; MELTZER; RUDZICZ, 2016). A segunda utiliza a saída do sistema *baseline* de inferência textual do ASSIN. Nessa abordagem, chamada aqui de *Inferência*, consideramos que a sentença contém a unidade de informação se o sistema de inferência retornar os rótulos *Inferência* e *Paráfrase*.

4.3.3 Avaliação

Nesta seção, são apresentados os experimentos para identificação de unidades de informação. Os conjuntos foram separados em treinamento e teste, utilizando 70% para treinamento e 30% para teste, de forma estratificada para cada grupo.

Na *ABCD*, cada participante produz duas narrativas: uma imediatamente após ouvir a história e a outra após 30 minutos. Para não enviesar a avaliação, o par de narrativas está no

conjunto de treinamento ou no conjunto de teste. Na BALE, combinamos as narrativas dos idosos com CCL e Alzheimer em um único grupo, devido ao baixo número de idosos com CCL.

A Tabela 27 apresenta os resultados obtidos no conjunto de dados da *ABCD*. Por se tratar de um problema multirrótulo, reportamos a Precisão micro (Pr_{micro}), Precisão macro (Pr_{macro}), $F1$ micro ($F1_{micro}$), $F1$ macro ($F1_{macro}$), *SubsetAccuracy*, e o *HammingLoss*. A *baseline* Casamento exato obteve os valores mais baixos, já a *baseline* Inferência obteve os melhores resultados para precisão e o *HammingLoss*, mas foi superada pelo método de similaridade semântica *STS* nas outras medidas.

Tabela 27 – Resultados da identificação de unidades de informação na *ABCD*.

Método	**Pr$_{macro}$**		**Pr$_{micro}$**		**F1$_{macro}$**		**F1$_{micro}$**		***SubsetAccuracy***		***Hamming Loss***	
	Treino	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**
Casamento Exato	0,570	0,469	0,903	0,758	0,337	0,283	0,246	0,233	0,246	0,169	0,078	0,081
Inferência	0,858	**0,793**	0,891	**0,873**	0,552	0,531	0,500	0,478	0,348	0,384	0,062	**0,062**
Chunking	0,705	0,699	0,587	0,577	0,640	**0,624**	0,668	**0,656**	0,668	**0,395**	0,076	0,076
STS	0,651	0,569	0,595	0,552	0,672	0,598	0,670	0,552	0,670	0,273	0,072	0,081

A Tabela 28 mostra os resultados obtidos no conjunto de dados da BALE. As *baselines* Casamento exato e Inferência obtiveram os melhores resultados para precisão, enquanto o método de *STS* obteve os melhores resultados em $F1$ e *SubsetAccuracy*.

Tabela 28 – Resultados da identificação de unidades de informação na BALE.

Método	**Pr$_{macro}$**		**Pr$_{micro}$**		**F1$_{macro}$**		**F1$_{micro}$**		***SubsetAccuracy***		***Hamming Loss***	
	Treino	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**
Casamento Exato	0,680	**0,740**	0,830	**0,930**	0,510	0,540	0,440	0,460	0,430	0,300	0,040	**0,040**
Inferência	0,625	0,690	0,808	0,781	0,485	0,519	0,359	0,404	0,447	0,324	0,042	0,050
Chunking	0,620	0,580	0,510	0,480	0,600	0,570	0,630	0,560	0,460	0,340	0,060	0,070
STS	0,680	0,640	0,670	0,650	0,740	**0,700**	0,720	**0,650**	0,520	**0,430**	0,030	**0,040**

As *baselines Casamento Exato* e Inferência apresentaram resultados superiores nas métricas de Precisão, entretanto, nesse trabalho estamos interessados em identificar as unidades recordadas, assim, é importante acertar o rótulo completo e não parte dele. Desse modo, deu-se mais importância para as métricas *SubsetAccuracy* e a medida *F1* do que a Precisão Macro e Micro.

Como o objetivo final da pesquisa é criar um classificador para narrativas de testes neuropsicológicos de idosos saudáveis e idosos com comprometimento cognitivo (CCL e DA), para poder identificar os primeiros sinais de problemas cognitivos, na próxima seção, as unidades de informação extraídas automaticamente são utilizadas como atributos, com objetivo de identificar pacientes com CCL e Doença de Alzheimer.

Desse modo, é avaliado se os métodos com os maiores valores de *SubsetAccuracy* e e $F1$ conseguem obter bons resultados de classificação, visando uma triagem automática. Além disso, na Seção 4.4.4 as unidades de informação são combinadas com atributos de métricas linguísticas, e métricas de redes complexas, para avaliação do desempenho do classificador.

4.4 Classificação de Narrativas para Triagem Automática

Ao longo desta seção, são apresentados os resultados da classificação CCL *versus* Saudáveis e DA *versus* Saudáveis para diferentes conjuntos de atributos, alguns criados nesta pesquisa, outros apenas avaliados aqui. Em todas as avaliações, utilizamos os seguintes algoritmos de aprendizado de máquina: *SVM*, *Naïve Bayes*, Árvore de Decisão, *Gradient Boosting*, e *KNN*, implementados no *scikit-learn* versão 0.21.2, com os hiperparâmetros *default*. Algumas particularidades dos algoritmos são interessantes destacar:

Naïve Bayes É um dos algoritmos de aprendizado de máquina mais simples, pois assume que os atributos são independentes;

SVM É um algoritmo de classificação linear; sua função de otimização busca encontrar o hiperplano com margem máxima. Para esse algoritmo, utilizamos o *kernel* linear e o Radial Basis Function;

Árvore de Decisão É um algoritmo que recursivamente particiona o espaço de entrada, geralmente de forma binária, definindo um modelo local em cada região resultante do espaço de entrada. É possível visualizar o modelo final em forma de uma árvore, onde cada partição representa um nó;

Gradient Boosting Utiliza diversas árvores de decisão; cada árvore é treinada de forma sequencial para corrigir os erros da anterior;

KNN Pertence à categoria *lazy*, pois não necessita de uma fase de treinamento; para predizer um novo exemplo busca no conjunto de treinamento os k exemplos mais similares e retorna o rótulo mais frequente.

Para a tarefa de classificação binária, os conjuntos de dados foram balanceados. Para a *ABCD*, utilizamos 12 idosos por grupo (Controle e CCL), sendo que cada idoso produziu 2 narrativas. O conjunto de dados final possui 48 narrativas. Para a BALE os pacientes com CCL e Alzheimer foram agrupados em um único grupo (16 narrativas), e selecionamos de forma randômica as 16 narrativas do grupo de controle. O conjunto de dados final para a BALE possui 32 narrativas. Os resultados são reportados com o *10-fold-cross-validation* e a métrica de acurácia.

Como pré-processamento, os atributos que são valores reais foram normalizados, pois métodos como o *SVM* com *kernel RBF* assumem que os atributos possuem média 0 e desvio padrão 1. Esse processo também auxilia o algoritmo *KNN* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; MURPHY, 2012). Desse modo, utilizou-se a normalização *z-score*, apresentada a seguir:

$$\mathbf{X}' = \frac{\mathbf{X} - \mu}{\sigma} \tag{4.1}$$

onde μ é média de $\mathbf{X}$, σ é o desvio padrão de $\mathbf{X}$, lembrando que esses valores devem ser obtidos no conjunto de treinamento.

4.4.1 Classificação de Narrativas com Unidades de Informação

Nesta seção, são apresentados os resultados da classificação utilizando as unidades de informação. Como esses atributos são binários, não utilizou-se a normalização *z-score*. Para o treinamento dos métodos *STS* e *Chunking*, que necessitam de uma busca para encontrar os melhores *thresholds*, utilizou-se o *10-fold-cross-validation* para construir o novo conjunto de dados com as unidades de informação inferidas, e este foi utilizado na avaliação dos classificadores. Desse modo, não foram utilizadas as predições do conjunto de treinamento como atributos dos classificadores das narrativas.

Os resultados obtidos por todos os modelos na *ABCD*, em termos de acurácia, são apresentados na Tabela 29. Na segunda coluna da tabela, são apresentados os resultados para a anotação manual, que trouxe valores próximos para dois dos seis algoritmos de aprendizado (Árvores de Decisão e *Naïve Bayes*). Em geral, o classificador de Árvore de Decisão apresenta diferenças negativas maiores entre os valores da anotação manual e dos quatro modelos automáticos.

Para os dois métodos propostos para a identificação de unidades de informação (*STS* e *Chunking*), os melhores desempenhos para a *ABCD* foram do método *Chunking*. Já o método de Inferência apresentou, em geral, resultados melhores do que o *STS*.

Tabela 29 – Resultados na classificação utilizando as unidades de informação na *ABCD*.

Método	**Manual**	**Casamento Exato**	**Inferência**	***Chunking***	***STS***
Árvore de Decisão	**0,638**	0,475	0,475	0,525	**0,538**
Gradient Boosting	0,538	0,463	**0,625**	0,550	0,500
KNN	0,575	**0,513**	0,525	**0,663**	0,413
SVM-Linear	0,500	0,475	0,588	0,525	0,488
SVM-RBF	0,563	0,363	0,463	0,463	0,488
Naïve Bayes	0,625	0,425	0,588	0,638	0,525

Na Tabela 30, são mostrados os resultados dos modelos na BALE. Para a anotação manual, tivemos quatro empates no desempenho de classificadores. O método *Chunking* apresenta

diferenças negativas maiores entre os valores da anotação manual para todos os classificadores. Já o método *STS* apresenta as menores diferenças.

Em geral,a *baseline* Casamento exato superou os métodos automáticos propostos de identificação de unidades de informação.

Tabela 30 – Resultados na classificação utilizando as unidades de informação na BALE.

Método	**Manual**	**Casamento Exato**	**Inferência**	***Chunking***	***STS***
Árvore de Decisão	0,600	0,700	0,525	0,400	0,600
Gradient Boosting	**0,675**	0,725	0,450	0,400	0,575
KNN	0,650	0,575	0,475	0,525	0,625
SVM-Linear	**0,675**	0,625	0,600	0,550	0,625
SVM-RBF	**0,675**	0,525	**0,675**	**0,625**	**0,725**
NaiveBayes	**0,675**	**0,775**	0,550	0,550	0,700

Os métodos de identificação de unidades de informação desenvolvidos nesta pesquisa apresentaram desempenhos melhores que as *baselines* para a identificação de unidades de informação em termos da métrica *SubsetAccuracy*, sendo esta uma métrica bastante rigorosa na classificação multirrótulo, pois é necessário acertar todo o conjunto multirrótulo. Além disso, os métodos desenvolvidos trouxeram resultados para classificação próximos da anotação humana. Desse modo, se mostraram promissores em um processo visando uma triagem automática. Entretanto, é necessária uma avaliação em um conjunto maior de dados.

Além disso, nos experimentos realizados as unidades de informação auxiliaram mais na classificação final de pacientes com DA *versus* Controles saudáveis (caso da BALE). Esse comportamento é esperado, pois pacientes com Doença de Alzheimer possuem um comprometimento de memória maior do que pacientes com CCL.

Na próxima seção, são apresentados os resultados das métricas extraídas dos dicionários de propriedades psicolinguísticas, e a combinação delas com outras métricas linguísticas.

4.4.2 Classificação de Narrativas com Métricas Linguísticas

Todas as métricas linguísticas utilizadas no desenvolvimento desse trabalho foram extraídas automaticamente com o sistema NILC-Metrix[9]. O sistema conta com 189 métricas, cuja descrição está disponível no Anexo A. O sistema conta com 24 métricas que utilizam o dicionário inferido e descrito na Seção 4.2.1.

Para cada propriedade psicolinguística é calculada a média e o desvio padrão dos valores das palavras de conteúdo presentes no léxico. Também é calculada a proporção de palavras de conteúdo que estão nas seguintes faixas: 1 a 2,5; 2,5 a 4; 4 a 5,5; e 5,5 a 7. Na Tabela 31 são apresentadas as métricas psicolinguísticas utilizadas na avaliação dos experimentos.

[9] <https://simpligo.sidle.al/>

Tabela 31 – Métricas psicolinguísticas do NILC-Metrix.

	Métrica	Descrição
1	concretude_mean	Média dos valores de concretude das palavras de conteúdo do texto
2	concretude_std	Desvio padrão dos valores de concretude das palavras de conteúdo do texto
3	idade_aquisicao_mean	Média dos valores de idade de aquisição das palavras de conteúdo do texto
4	idade_aquisicao_std	Desvio padrão dos valores de idade de aquisição das palavras de conteúdo do texto
5	familiaridade_mean	Média dos valores de familiaridade das palavras de conteúdo do texto
6	familiaridade_std	Desvio padrão dos valores de familiaridade das palavras de conteúdo do texto
7	imageabilidade_mean	Média dos valores de imageabilidade das palavras de conteúdo do texto
8	imageabilidade_std	Desvio padrão dos valores de imageabilidade das palavras de conteúdo do texto
9	concretude_1_25_ratio	Proporção de palavras com valor de concretude entre 1 e 2,5
10	concretude_25_4_ratio	Proporção de palavras com valor de concretude entre 2,5 e 4
11	concretude_4_55_ratio	Proporção de palavras com valor de concretude entre 4 e 5,5
12	concretude_55_7_ratio	Proporção de palavras com valor de concretude entre 5,5 e 7
13	idade_aquisicao_1_25_ratio	Proporção de palavras com valor de idade de aquisição entre 1 e 2,5
14	idade_aquisicao_25_4_ratio	Proporção de palavras com valor de idade de aquisição entre 2,5 e 4
15	idade_aquisicao_4_55_ratio	Proporção de palavras com valor de idade de aquisição entre 4 e 5,5
16	idade_aquisicao_55_7_ratio	Proporção de palavras com valor de idade de aquisição entre 5,5 e 7
17	familiaridade_1_25_ratio	Proporção de palavras com valor de familiaridade entre 1 e 2,5
18	familiaridade_25_4_ratio	Proporção de palavras com valor de familiaridade entre 2,5 e 4
19	familiaridade_4_55_ratio	Proporção de palavras com valor de familiaridade entre 4 e 5,5
20	familiaridade_55_7_ratio	Proporção de palavras com valor de familiaridade entre 5,5 e 7
21	imageabilidade_1_25_ratio	Proporção de palavras com valor de imageabilidade entre 1 e 2,5
22	imageabilidade_25_4_ratio	Proporção de palavras com valor de imageabilidade entre 2,5 e 4
23	imageabilidade_4_55_ratio	Proporção de palavras com valor de imageabilidade entre 4 e 5,5
24	imageabilidade_55_7_ratio	Proporção de palavras com valor de imageabilidade entre 5,5 e 7

Para avaliação das métricas linguísticas, utilizou-se três cenários: (i) removeu-se as métricas psicolinguísticas do conjunto do NILC-Metrix — esse conjunto conta com 165 métricas (Base); (ii) avaliou-se a capacidade da classificação exclusivamente das métricas psicolinguísticas (Psicolinguístico); (iii) utilizou-se todas as métricas disponíveis do NILC-Metrix (Completo).

Na Tabela 32 são apresentados os resultados dos algoritmos de classificação para os três cenários. Para o conjunto Base quase todos os classificadores apresentaram resultados acima de $0,5$, com exceção do *Naïve Bayes*, sendo o melhor resultado o do *SVM*-Linear com $0,68$ de acurácia. Para as métricas psicolinguísticas, todos os classificadores apresentaram resultados acima de $0,5$, demostrando que essas métricas podem ajudar a identificar pacientes com CCL.

Entretanto, quando utilizou-se todas as métricas disponíveis no NILC-Metrix, quase todos os algoritmos apresentaram uma diferença negativa de acurácia quando comparado com o conjunto Base. Acreditamos que essa queda se deve ao fato de uma grande quantidade de atributos em uma baixa quantidade de exemplos. Esse problema é conhecido como maldição da dimensionalidade (*Curse of dimensionality*), sendo o *KNN* um dos algoritmos que mais sofrem com esse problema (MURPHY, 2012). Este algoritmo foi o que apresentou maior queda no valor da acurácia.

Na Tabela 33, são apresentados os resultados dos experimentos na BALE. O classificador *SVM-RBF* obteve o melhor resultado para o conjunto Base e para o conjunto Psicolinguístico. Em geral, o conjunto de dados com métricas psicolinguísticas apresentaram resultados melhores ou iguais ao conjunto Base. O conjunto Completo apresentou uma diferença negativa na acurácia

Tabela 32 – Resultados na classificação utilizando métricas linguísticas na *ABCD*.

Classificador	Base	Psicolinguístico	Completo
Árvore de Decisão	0,550	0,525	0,513
Gradient Boosting	0,600	0,525	0,575
KNN	0,550	0,525	0,388
SVM-Linear	**0,688**	0,600	0,575
SVM-RBF	0,600	0,563	**0,625**
Naïve Bayes	0,475	**0,675**	0,513

quando comparado com conjunto Base somente para os algoritmos de classificação Árvore de Decisão, *Gradient Boosting*, e o *SVM-RBF*.

Tabela 33 – Resultados na classificação utilizando métricas linguísticas na BALE.

Classificador	Base	Psicolinguístico	Completo
Árvore de Decisão	0,500	0,525	0,400
Gradient Boosting	0,600	0,625	0,425
KNN	0,575	0,625	0,575
SVM-Linear	0,600	0,550	0,575
SVM-RBF	**0,650**	**0,650**	0,600
Naïve Bayes	0,525	0,525	**0,625**

A seguir, é explorada a representação de narrativas em redes de adjacência, e a utilização de métricas que caracterizam a estrutura topológica das redes, visando uma triagem automática.

4.4.3 Classificação com Métricas de Redes Complexas

Antes de modelar textos em redes complexas, muitas vezes é necessário fazer um pré-processamento. O primeiro passo é a tokenização, em seguida, a remoção de *stopwords*. Para isso, foi usada a lista fornecida pela biblioteca NLTK[10] e sinais de pontuação, pois os dois elementos têm pouco significado semântico. Em geral, o processo de lematização é adotado na representação de textos em redes complexas, mas esse processo não foi aplicado. Essa decisão foi tomada pois trabalhos recentes mostraram que a lematização tem pouca ou nenhuma influência na modelagem de redes (MACHICAO *et al.*, 2018).

Recentemente, Amancio (2015b) demonstrou que textos curtos possuem menor capacidade de representação quando comparados com textos longos. Em geral, as narrativas de exames neuropsicológicos são muito curtas, tendo em torno de 20 a 60 palavras, dependendo do grupo em que se esta analisando.

Para possibilitar a aplicação de redes de adjacências em textos extremamente curtos, adaptou-se a abordagem de Perozzi *et al.* (2014) para induzir redes complexas. Nessa abordagem, as palavras são consideradas como vértices e para conectar duas palavras (dois vértices) foram

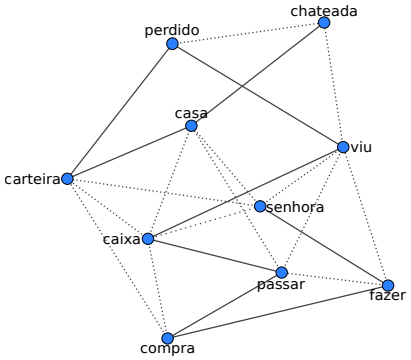
[10] <https://www.nltk.org/>

desenvolvidas duas estratégias: (1) adicionar arestas para os k-vizinhos mais próximos; (2) adicionar arestas quando a distância de similaridade é menor que um *threshold*.

Assim, explorou-se a similaridade do cosseno entre a representação vetorial das palavras para enriquecer a rede de adjacência (SANTOS *et al.*, 2017). No método proposto, é criado um novo vértice entre os pares que estão desconectados, caso a similaridade do cosseno seja maior que um *threshold*.

Na Figura 18 é apresentada uma rede de adjacência enriquecida. As linhas pontilhadas indicam as arestas que foram adicionadas para o seguinte trecho extraído de uma narrativa: "*uma senhora foi fazer compra. e quando ela foi passar na caixa ela viu que tinha perdido a carteira. ela foi para casa muito chateada.*". Podemos perceber que foram adicionadas arestas relacionadas à semântica das palavra como: *compra-carteira*, *carteira-caixa*, e *compra-caixa*. E também foram adicionadas arestas relacionadas a sintaxe: *fazer-passar*, *passar-viu*, e *viu-fazer*.

Figura 18 – Exemplo da rede de adjacência enriquecida para um trecho da narrativa.



Na Figura 19a é possível perceber a baixa capacidade de representação de padrões em textos curtos, pois somente uma palavra possui mais que uma aresta. Além dessa informação, pode-se extrair a quantidade de palavras presentes na narrativa. Na Figura 19b é apresentada a rede enriquecida para a mesma narrativa, sendo possível perceber uma estrutura topológica mais complexa.

Para extrair as características topológicas das redes complexas, utilizou-se as métricas apresentadas na Seção 2.4.3. Como algumas métricas apresentam informações locais, ou seja, sobre um nó, computou-se a média, o desvio padrão e a assimetria (obliquidade, *skewness* em inglês) para as seguintes métricas: grau, centralidade de autovetor, *PageRank*, *betweenness*, comprimento médio dos caminhos mínimos, excentricidade, grau médio dos vizinhos de um nó. Como o coeficiente de aglomeração e o diâmetro são métricas globais, utilizou-se seus valores. Desse modo, extraiu-se 23 métricas topológicas das redes complexas.

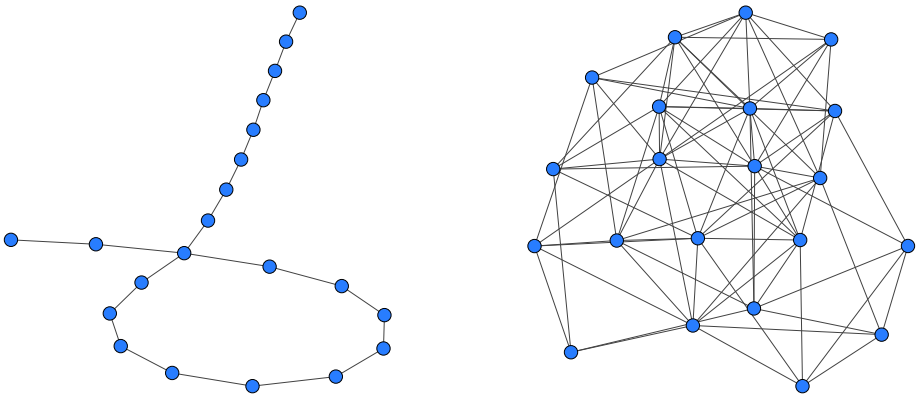
Para o enriquecimento da rede de adjacência utilizou-se o *FastText*, e variou-se o *threshold* entre $0,4$ a $0,9$.

Os resultados da classificação utilizando métricas de redes complexas na *ABCD* em termos da acurácia, são apresentados na Tabela 34. O enriquecimento da rede de adjacência

Figura 19 – Exemplos de rede de adjacência para uma narrativa de reconto.

(a) Rede de adjacência.

(b) Rede de adjacência enriquecida com *threshold* maior que 0,4.



trouxe ganhos de acurácia para todos os algoritmos, exceto para o *Gradient Boosting*. Os algoritmos Árvore de Decisão, *SVM*-Linear, e *Naïve Bayes* obtiveram os maiores valores de acurácia com o *threshold* de $0,5$, enquanto que o *KNN* e o *SVM-RBF* apresentaram os melhores valores com o *threshold* de $0,6$, sendo que o *SVM-RBF* apresentou o melhor valor de acurácia na tarefa com $0,663$.

Em geral, na *ABCD* os algoritmos de classificação apresentaram uma queda de acurácia quando utilizaram a rede de adjacência enriquecida com *threshold* de $0,4$. Possivelmente com esse *threshold* as redes se tornaram muito conectadas, apresentando poucos padrões úteis na classificação.

Tabela 34 – Resultados na classificação utilizando métricas topológicas de redes complexas na *ABCD*.

Classificador	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**Adjacência**
Árvore de Decisão	0,450	**0,575**	0,525	0,525	0,513	0,513	0,513
Gradient Boosting	0,413	0,438	0,550	0,488	**0,575**	**0,575**	**0,575**
KNN	0,300	0,500	**0,650**	0,575	0,638	0,638	0,638
SVM-Linear	0,575	**0,600**	0,475	0,525	0,525	0,525	0,525
SVM-RBF	0,375	0,550	**0,663**	0,638	0,550	0,550	0,550
Naïve Bayes	0,425	**0,588**	0,550	0,563	0,550	0,550	0,550

Na Tabela 35, são apresentados os resultados da classificação utilizando métricas de redes complexas na BALE. Nessa avaliação, a rede de adjacência apresentou resultados ruins; em geral, os valores de acurácia ficaram abaixo de $0,40$. Com a utilização da abordagem proposta de enriquecimento quase todos os algoritmos apresentaram resultados acima de $0,50$, exceto o algoritmo Árvore de Decisão.

Os maiores valores de acurácia são com a utilização do *threshold* de $0,4$ e $0,5$. O maior aumento foi obtido com o algoritmo *SVM*-Linear, em que o valor da acurácia aumentou em 2 vezes, passando de $0,375$ para $0,775$.

Tabela 35 – Resultados na classificação utilizando métricas topológicas de redes complexas na BALE.

Classificador	0,4	0,5	0,6	0,7	0,8	0,9	Adjacência
Árvore de Decisão	**0,400**	**0,400**	0,300	0,325	0,350	0,350	0,350
Gradient Boosting	**0,550**	0,300	0,225	0,500	0,275	0,350	0,350
KNN	0,525	**0,650**	0,425	0,400	0,425	0,350	0,350
SVM-Linear	**0,775**	0,450	0,550	0,350	0,400	0,375	0,375
SVM-RBF	**0,550**	0,450	0,300	0,350	0,350	0,375	0,375
Naïve Bayes	0,425	**0,525**	0,450	0,425	0,425	0,400	0,400

4.4.4 Classificação com Combinação dos Recursos

Por fim, combinou-se os atributos de métricas linguísticas, de identificação automática de unidades de informação com métodos *STS* e *Chunking*, e os atributos de redes complexas. Para o último conjunto, utilizou-se os atributos que geraram os maiores valores de acurácia. Na *ABCD* houve empate para o algoritmo *Gradient Boosting*, utilizando-se a rede de adjacência. Na BALE houve empate para o algoritmo Árvore de Decisão, utilizando-se a rede de adjacência enriquecida com *threshould* acima de 0,5.

Na Tabela 36 são apresentados os resultados da combinação dos atributos extraídos, em que Métricas é a combinação das métricas linguísticas e redes complexas, e *STS* e *Chunking* indicam o método que foi utilizado para extrair as unidades de informação de forma automática. Para facilitar a visualização são apresentados: (i) os valores de acurácia dos métodos de identificação de unidades de informação apresentados na Tabela 29, (ii) o resultado da utilização de todas as métricas linguísticas da Tabela 32, e (iii) o melhor resultado de cada algoritmo da Tabela 34.

A combinação de todos os atributos apresentou os melhores resultados somente para o algoritmo *Gradient Boosting* com método de *Chunking* e para o algoritmo *SVM*-Linear com o método *STS*. O método de *Chunking* apresentou os melhores resultados para os algoritmos *KNN* e *Naïve Bayes*, enquanto que os atributos de redes de adjacência enriquecidas apresentaram os melhores resultados para os algoritmos de Árvore de Decisão e *SVM-RBF*.

Tabela 36 – Resultados na classificação utilizando os diferentes conjunto de atributos e a combinação dos atributos na *ABCD*.

Classificador	*Chunking*	*STS*	Linguísticas	Redes	Métricas + *Chunking*	Métricas + *STS*
Árvore de Decisão	0,525	0,538	0,513	**0,575**	0,513	0,563
Gradient Boosting	0,550	0,500	0,575	0,575	**0,613**	0,575
KNN	**0,663**	0,413	0,388	0,650	0,563	0,488
SVM-Linear	0,525	0,575	0,575	0,600	0,600	**0,613**
SVM-RBF	0,463	0,488	0,625	**0,663**	0,650	0,625
Naïve Bayes	**0,638**	0,525	0,513	0,588	0,613	0,550

Na Tabela 37 são apresentados os resultados de cada conjunto de atributos e a sua combinação na BALE. Os métodos de identificação de unidades de informação e as métricas de redes de adjacência apresentaram os melhores resultados. Para essa bateria, as métricas linguísticas e a combinação de todos os atributos não apresentaram bons resultados.

Tabela 37 – Resultados na classificação utilizando os diferentes conjunto de atributos e a combinação dos atributos na BALE.

Classificador	***Chunking***	***STS***	**Linguísticas**	**Redes**	**Métricas + *Chunking***	**Métricas + *STS***
Árvore de Decisão	0,400	**0,600**	0,400	0,400	0,400	0,325
Gradient Boosting	0,400	**0,575**	0,425	0,550	0,450	0,375
KNN	0,525	0,625	0,575	**0,650**	0,525	0,550
SVM-Linear	0,550	0,625	0,575	**0,775**	0,550	0,575
SVM-RBF	0,625	**0,725**	0,600	0,550	0,600	0,625
Naïve Bayes	0,550	**0,700**	0,625	0,525	0,650	0,600

Os métodos de unidades de informação, e as métricas de redes complexas se mostraram úteis na identificação de pacientes com Comprometimento de Cognitivo Leve e com Doença de Alzheimer. O enriquecimento de redes de adjacência auxiliaram mais na classificação final de pacientes com DA *versus* Controles saudáveis (caso da BALE), enquanto que as métricas linguísticas não obtiveram resultados expressivos em certos cenários. Entretanto, acreditamos que uma combinação de certos atributos pode trazer resultados melhores, porém é necessário um conjunto de dados maior para uma validação extensiva da melhor combinação, e por fim uma investigação dos melhores parâmetros de cada algoritmo de classificação.

CAPÍTULO

5

CONCLUSÕES

Este capítulo retoma os objetivos deste trabalho, trazendo uma visão geral do trabalho desenvolvido, as suas contribuições, limitações e trabalhos futuros.

5.1 Objetivos e Contribuições

Este trabalho se debruçou sobre duas avaliações automáticas no cenário clínico: (i) avaliação da tarefa de similaridade semântica textual para a tarefa de identificação de unidades de informação em narrativas de recontos curtos, e (ii) classificação das narrativas acima em um cenário binário (grupos idosos saudáveis *versus* idosos com comprometimento cognitivo), via três conjuntos diferentes de atributos, desenvolvidas durante a pesquisa: métricas psicolinguísticas, métricas de redes complexas e as unidades de informação relembradas.

Para atingir o primeiro objetivo, propusemos uma anotação manual para a criação de dois conjuntos de dados com as unidades de informação identificadas. Após a anotação, cada conjunto de dados foi caracterizado com: (i) uma análise quantitativa das unidades de informação recordadas, e (ii) uma análise usando métricas linguísticas automáticas.

Como observado na revisão dos trabalhos da literatura, a grande dificuldade de utilizar listas de palavras para cada unidade de informação é a necessidade de um trabalho humano e subjetivo, pois nem sempre a lista possui todas as paráfrases/sinônimos possíveis. Métodos de *clustering* são úteis, pois conseguem criar automaticamente as unidades de informação, mas podem gerar unidades pouco representativas ou não relacionadas às unidades que um dado teste neuropsicológico avalia.

Em estudos envolvendo análise de narrativas clínicas, geralmente a quantidade de dados é limitada dado o alto custo de aquisição. E pelo fato das sentenças estarem categorizadas com várias unidades informação, caracterizando uma tarefa multirrótulo, o cenário tratado nesta tese é ainda mais desafiador. Atacamos essas dificuldades aproximando as tarefas de

similaridade semântica com a identificação automática de unidades de informação em narrativas. Avaliamos um método de similaridade semântica que se destacou na avaliação conjunta ASSIN, e transformamos o problema multirrótulo em problemas de classificação binária, encontrando um ponto de corte para o valor de similaridade de cada unidade de informação. Dessa forma, conseguimos superar ambas as *baselines* na identificação de unidades de informação para os dois conjuntos de dados. Após a identificação das unidades de informação, conseguimos utilizá-las combinadas com outros atributos. Na literatura, é comum a extração de diversas métricas léxicas e sintáticas, mas para o Português não existia nenhum trabalho que utilizava métricas de propriedades psicolinguísticas das palavras, como concretude, imageabilidade, idade de aquisição e frequência subjetiva, sendo essa última similar ao conceito de familiaridade. Essas propriedades são úteis no cenário clínico de identificação de narrativas de pacientes com CCL versus AD, pois ajudam a identificar se o discurso é composto predominante por palavras concretas, palavras adquiridas nos primeiros anos de vida, por exemplo. Para poder utilizar essas propriedades em métricas, induzimos automaticamente as propriedades para as palavras de um grande dicionário do português. Para isso, utilizamos a média de três regressores, sendo que cada um foi treinado em um conjunto específico de atributos: oito atributos lexicais, isto é, vindos de dicionários e de frequências de vários córpus, por exemplo; *word embeddings* do GloVe; e *word embeddings* do Skip-Gram.

Com relação ao segundo objetivo, isto é, a classificação das narrativas em um cenário binário (grupos idosos saudáveis *versus* idosos com comprometimento cognitivo), podemos dizer que:

1. Os atributos de propriedades psicolinguísticas se mostraram úteis na classificação de narrativas, para quase todos os algoritmos avaliados, e apresentaram resultados acima de 50% de acurácia para ambos os conjuntos de dados clínicos (*ABCD* e BALE).

2. As unidades de informação identificadas automaticamente apresentaram resultados próximos da anotação manual. Na *ABCD*, dentre as abordagens propostas, o método *Chunking* apresentou os melhores resultados, enquanto na BALE o método *STS* apresentou os melhores resultados.

3. A representação de narrativas de recontos em redes de adjacência se mostrou pouco efetiva na classificação, principalmente para o conjunto de dados clínicos da BALE, apresentando resultados abaixo de 40% de acurácia. Esse comportamento era esperado devido à estrutura linear da rede de adjacência, entretanto, após a aplicação do método de enriquecimento foi possível obter bons resultados para certos *thresholds*. Cabe também ressaltar que o melhor resultado na BALE foi obtido com a rede de adjacência enriquecida.

4. Em geral, a combinação de todos os atributos, investigados ou desenvolvidos neste projeto de doutorado apresentaram valores similares de acurácia quando comparados com os conjuntos de atributos de forma independente para os dados da *ABCD*, e menores para

o conjunto de dados da BALE. Acreditamos que a grande quantidade de atributos, 229 atributos na *ABCD* e 233 atributos na BALE, e o baixo número de exemplos causou esse resultado.

Os códigos utilizados no desenvolvimento do regressor para inferir as propriedades psicolinguísticas estão disponiveis em <https://github.com/lbsantos/psycholinguistic-regression>, e o léxico inferido se encontra em <http://nilc.icmc.usp.br/portlex/index.php/en/?option=com_content&view=article&layout=edit&id=23>. Os demais códigos e o conjunto de dados anotados estão disponíveis em <https://github.com/lbsantos/ANAA-Dementia>.

Artigos Publicados

Publicações diretamente relacionadas com o tema de pesquisa:

- Santos, L. B. D., & Aluísio, S. M. (2020). Identificação automática de unidades de informação em testes de reconto de narrativas usando métodos de similaridade semântica. Linguamática, 11(2), 47-63.
- Santos, L. B., Hübner, L. C., Smidarle, A. D., Mansur, L., & Aluísio, S. M. (2019). Anotação de unidades de informação em transcrições de fala na tarefa de reconto de narrativas em português. In Proceedings do XII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana.
- Treviso, M. V., Santos, L. B. D., Shulby, C., Hübner, L. C., Mansur, L. L., & Aluísio, S. M. (2018). Detecting mild cognitive impairment in narratives in Brazilian Portuguese: first steps towards a fully automated system. Letras de Hoje, 53(1), 48-58.
- Santos, L. B., Corrêa Júnior, E. A. C., Oliveira Júnior, O., Amancio, D., Mansur, L., & Aluísio, S. M. (2017). Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1284-1296). Association for Computational Linguistics.
- Santos, L. B., Duran, M. S., Hartmann, N. S., Candido, A., Paetzold, G. H., & Aluisio, S. M. (2017). A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In International Conference on Text, Speech, and Dialogue (pp. 281-289). Springer.
- Fonseca, E. R., dos Santos, L. B., Criscuolo, M., & Aluísio, S. M. (2016). Visão geral da avaliação de similaridade semântica e inferência textual. Linguamática, 8(2), 3-13.

Publicações resultantes de colaborações do aluno em áreas relacionadas ao tema de pesquisa:

- Hartmann, N., & Santos, L. B. (2018). NILC at CWI 2018: Exploring Feature Engineering and Feature Learning. In Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 335-340). Association for Computational Linguistics

- Toledo, C. M., Aluísio, S. M., dos Santos, L. B., Brucki, S. M. D., Trés, E. S., de Oliveira, M. O., & Mansur, L. L. (2018). Analysis of macrolinguistic aspects of narratives from individuals with Alzheimer's disease, mild cognitive impairment, and no cognitive impairment. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 10, 31-40.

- Corrêa Júnior, E. A., Marinho, V. Q., & Santos, L. B. (2017). NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 611-615). Association for Computational Linguistics.

- Corrêa, E. A., Marinho, V. Q., dos Santos, L. B., Bertaglia, T. F. C., Treviso, M. V., & Brum, H. B. (2017). PELESent: Cross-domain polarity classification using distant supervision. In Proceedings 6th Brazilian Conference on Intelligent Systems (pp. 49-54). IEEE.

- Hübner, L. C., Loureiro, F., Smidarle, A. D., Pedro, J. R., Garcia, V. R. M., Treviso, M. V.,Santos, L. B., Schilling, L. P., Mansur, L. L. & Aluísio, S. M. (2017). Automatically distinguishing Mild Cognitive Impairment, Alzheimer's Disease and education effect in healthy aging in narratives in Brazilian Portuguese. Stem-Spraak-en Taalpathologie, 22, 78-80.

5.2 Limitações

A principal limitação dessa pesquisa de doutorado é a quantidade de narrativas dos dois conjunto de dados utilizados no projeto, quando comparada com a quantidade de narrativas do *DementiaBank*, o principal conjunto de dados público de avaliações neuropsicológicas da língua Inglesa. Essa limitação é compreensível dado que existe um gasto para realizar o processo de triagem com o objetivo de selecionar os possíveis candidatos a um estudo. Após essa seleção são aplicados todos os testes e descartados os dados de pacientes que não atendem os critérios da pesquisa. Além disso, pode ser difícil encontrar pacientes com algum tipo de doença/síndrome na região em que o estudo está sendo conduzido.

Além disso, na BALE combinou-se as narrativas de pacientes com DA e CCL, devido ao baixo de número de sujeitos diagnosticados com CCL, o que não é uma solução ideal. Sabemos que os grupos DA e Controles são mais facilmente separáveis, portanto, acreditamos que essa combinação seja a razão pela qual os classificadores apresentaram resultados ligeiramente melhores na BALE quando comparados com a *ABCD*.

Em relação ao método de inferência das propriedades psicolinguísticas, utilizamos os valores das propriedades dos léxicos de Portugal para servirem de semente dos métodos de inferência. Assim, assumiu-se que a propriedades são equivalentes, ignorando as diferenças culturais. Também, não realizamos um estudo para identificar o melhor algoritmo para cada conjunto de atributos ou a melhor forma de combinar os regressores.

Quanto aos métodos de identificação de unidades de informação, não criou-se nenhum um modelo preditivo, apenas utilizamos a similaridade e encontrou-se um *threshold* no conjunto de treinamento para cada classe. Assim, seria interessante extrair mais atributos e utilizar um classificador, mas precisaríamos de um conjunto de dados maior para essa tarefa.

5.3 Trabalhos Futuros

Como trabalhos futuros, apontamos a identificação e a classificação de narrativas em um cenário somente com DAs *versus* Controles para a BALE; a coleta de dados desse conjunto está ainda em andamento.

O método desenvolvido para identificar as unidades de informação depende da qualidade do sistema de similaridade semântica. O sistema utilizado representa as sentenças com a média das *embeddings* das palavras (HARTMANN, 2016), mas nos últimos anos essa abordagem vem sendo superada por métodos mais complexos como *ELMo* (*Embeddings from Language Models*) (PETERS *et al.*, 2018) ou *BERT* (*Bidirectional Encoder Representations from Transformers*) (DEVLIN *et al.*, 2019). Como trabalhos futuros, pretende-se explorar esses modelos mais atuais, pois acreditamos que com sistemas melhores de similaridade semântica podemos obter métodos de identificação de unidades de informação também melhores. Além disso pretende-se realizar uma análise de erros dos métodos desenvolvidos e os novos métodos.

Propomos desenvolver um conjunto com propriedades psicolinguísticas com sementes anotadas manualmente e validar a correlação das propriedades obtidas com os léxicos de Português de Portugal, reavaliando o sistema de inferência.

REFERÊNCIAS

ABREU, I. D.; FORLENZA, O. V.; BARROS, H. L. de. Demência de alzheimer: correlação entre memória e autonomia. **Revista de Psiquiatria Clınica**, v. 32, p. 131–136, 2005. Citado na página 27.

AGIRRE, E.; BANEA, C.; CARDIE, C.; CER, D.; DIAB, M.; GONZALEZ-AGIRRE, A.; GUO, W.; LOPEZ-GAZPIO, I.; MARITXALAR, M.; MIHALCEA, R. *et al.* Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In: **Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)**. Denves, United States of America: Association for Computational Linguistics, 2015. p. 252–263. Citado nas páginas 60 e 92.

AGIRRE, E.; DIAB, M.; CER, D.; GONZALEZ-AGIRRE, A. Semeval-2012 task 6: A pilot on semantic textual similarity. In: **Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation**. Montréal, Canada: Association for Computational Linguistics, 2012. p. 385–393. Citado nas páginas 60, 62 e 92.

AGRAWAL, R.; GUPTA, A.; PRABHU, Y.; VARMA, M. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In: **Proceedings of the 22nd international conference on World Wide Web**. Rio de Janeiro, Brazil: Association for Computing Machinery, 2013. p. 13–24. Citado na página 48.

AHMED, S.; JAGER, C. A. de; HAIGH, A.-M.; GARRARD, P. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer;s disease. **Neuropsychology**, v. 27, n. 1, p. 79–85, 2013. ISSN 1931-1559. Disponível em: <http://www.biomedsearch.com/nih/Semantic-processing-in-connected-speech/23356598.html>. Citado na página 40.

ALTMANN, L. J.; MCCLUNG, J. S. Effects of semantic impairment on language use in Alzheimer's disease. **Semin Speech Lang**, v. 29, n. 1, p. 18–31, feb 2008. Citado na página 38.

ALUÍSIO, S.; CUNHA, A.; SCARTON, C. Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In: **Proceedings of 12th International Conference on Computational Processing of the Portuguese Language**. Tomar, Portugal: Springer International Publishing, 2016. p. 109–114. Citado nas páginas 28, 31, 71, 72 e 86.

ALUISIO, S.; GASPERIN, C. Fostering digital inclusion and accessibility: the Porsimples project for simplication of Portuguese texts. In: **Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas - Association for Computational Linguistics**. Los Angeles, United States of America: Association for Computational Linguistics, 2010. p. 46–53. Citado na página 52.

Alzheimer's Disease International. **World Alzheimer Report 2019: Attitudes to dementia**. London: Alzheimer's Disease Internationals London, 2019. Citado na página 26.

AMANCIO, D. R. A complex network approach to stylometry. **PloS one**, Public Library of Science, v. 10, n. 8, p. e0136076, 2015. Citado na página 55.

____. Probing the topological properties of complex networks modeling short written texts. **PloS one**, Public Library of Science, v. 10, n. 2, p. e0118394, 2015. Citado nas páginas 55 e 103.

AMANCIO, D. R.; NUNES, M. G. V.; JR., O. N. O.; COSTA, L. F. Extractive summarization using complex networks and syntactic dependency. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 391, n. 4, p. 1855–1864, 2012. Citado na página 55.

ANDREETTA, S.; CANTAGALLO, A.; MARINI, A. Narrative discourse in anomic aphasia. **Neuropsychologia**, Elsevier, v. 50, n. 8, p. 1787–1793, 2012. Citado na página 28.

ANTIQUEIRA, L.; JR., O. N. O.; COSTA, L. da F.; NUNES, M. das G. V. A complex network approach to text summarization. **Information Sciences**, v. 179, n. 5, p. 584 – 599, 2009. Citado na página 55.

ARRUDA, H. F. de; COSTA, L. F.; AMANCIO, D. R. Using complex networks for text classification: Discriminating informative and imaginative documents. **EPL (Europhysics Letters)**, IOP Publishing, v. 113, n. 2, p. 28007, 2016. Citado na página 55.

BACCHIANI, M.; RILEY, M.; ROARK, B.; SPROAT, R. Map adaptation of stochastic grammars. **Computer speech & language**, Elsevier, v. 20, n. 1, p. 41–68, 2006. Citado na página 68.

BAYLES, K.; TOMOEDA, C. **ABCD: Arizona Battery for Communication Disorders of Dementia**. Tucson, United States of America: Canyonlands Publishing, 1993. Citado nas páginas 28, 29 e 81.

BECKER, J. T.; BOILER, F.; LOPEZ, O. L.; SAXTON, J.; MCGONIGLE, K. L. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. **Archives of Neurology**, American Medical Association, v. 51, n. 6, p. 585–594, 1994. Citado nas páginas 69 e 73.

BERGSTRA, J.; YAMINS, D.; COX, D. D. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In: **Proceedings of the 12th Python in science conference**. Austin, United States of America: IOP Publishing Ltd, 2013. p. 13–20. Citado na página 95.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003. Citado na página 63.

BOEUF, C. **Raconte...: 55 historiettes en images**. Paris, France: Ecole, 1971. Citado na página 39.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. **Transactions of the Association for Computational Linguistics**, MIT Press, v. 5, p. 135–146, 2017. Citado na página 59.

BOUTELL, M. R.; LUO, J.; SHEN, X.; BROWN, C. M. Learning multi-label scene classification. **Pattern recognition**, Elsevier, v. 37, n. 9, p. 1757–1771, 2004. Citado na página 46.

BOWMAN, S. R.; ANGELI, G.; POTTS, C.; MANNING, C. D. A large annotated corpus for learning natural language inference. In: **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**. Lisbon, Portugal: Association for Computational Linguistics, 2015. p. 632–642. Citado na página 61.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. **Computer Networks and ISDN Systems**, v. 30, n. 1-7, p. 107–117, 1998. Citado na página 56.

BRITO-MARQUES, P. R. de; CABRAL-FILHO, J. E.; MIRANDA, R. M. Visual reproduction test in normal elderly: Influence of schooling and visual task complexity. **Dementia & Neuropsychologia**, v. 6, n. 2, p. 91–96, 2009. Citado na página 45.

CAMEIRAO, M. L.; VICENTE, S. G. Age-of-acquisition norms for a set of 1,749 portuguese words. **Behavior research methods**, Springer, v. 42, n. 2, p. 474–480, 2010. Citado nas páginas 88 e 89.

CANCHO, R. F. I.; SOLÉ, R. V. The small world of human language. **Proceedings of the Royal Society of London. Series B: Biological Sciences**, The Royal Society, v. 268, n. 1482, p. 2261–2265, 2001. Citado na página 55.

CANCHO, R. F. i; SOLÉ, R. V.; KÖHLER, R. Patterns in syntactic dependency networks. **Physical Review E**, American Physical Society, v. 69, n. 5, p. 051915, 2004. Citado na página 55.

CARLOMAGNO, S.; SANTORO, A.; MENDITTI, A.; PANDOLFI, M.; MARINI, A. Referential Communication in Alzheimer's Type Dementia. **Cortex**, v. 41, n. 4, p. 520–534, 2005. ISSN 0010-9452. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0010945208701928>. Citado na página 39.

CERHAN, J. H.; IVNIK, R. J.; SMITH, G. E.; TANGALOS, E. C.; PETERSEN, R. C.; BOEVE, B. F. Diagnostic Utility of Letter Fluency, Category Fluency, and Fluency Difference Scores in Alzheimer's Disease. **The Clinical Neuropsychologist**, v. 16, n. 1, p. 35–42, 2002. PMID: 11992224. Disponível em: <http://www.tandfonline.com/doi/abs/10.1076/clin.16.1.35.8326>. Citado na página 37.

CHAPMAN, S. B.; ZIENTZ, J.; WEINER, M.; ROSENBERG, R.; FRAWLEY, W.; BURNS, M. H. Discourse changes in early alzheimer disease, mild cognitive impairment, and normal aging. **Alzheimer Disease & Associated Disorders**, Lippincott Williams & Wilkins, v. 16, n. 3, p. 177–186, 2002. Citado na página 41.

CHEN, W.-J.; SHAO, Y.-H.; LI, C.-N.; DENG, N.-Y. Mltsvm: a novel twin support vector machine to multi-label learning. **Pattern Recognition**, Elsevier, v. 52, p. 61–74, 2016. Citado na página 48.

CHEUNG, H.; KEMPER, S. Competing complexity metrics and adults' production of complex sentences. **Applied Psycholinguistics**, Cambridge University Press, v. 13, n. 1, p. 53–76, 1992. Citado na página 67.

CLARE, A.; KING, R. D. Knowledge discovery in multi-label phenotype data. In: SPRINGER. **Proceedings of 5th European Conference on Principles of Data Mining and Knowledge Discovery**. Freiburg, Germany, 2001. p. 42–53. Citado na página 48.

CLEMENTE, R.; RIBEIRO-FILHO, S. Comprometimento cognitivo leve: aspectos conceituais, abordagem clínica e diagnóstica. **Revista HospClementeedro Ernesto**, v. 7, n. 1, 2008. Citado nas páginas 25 e 26.

CONG, J.; LIU, H. Approaching human language with complex networks. **Physics of life reviews**, Elsevier, v. 11, n. 4, p. 598–618, 2014. Citado na página 55.

COOK, C.; FAY, S.; ROCKWOOD, K. Verbal Repetition in People With Mild-to-Moderate Alzheimer Disease: A Descriptive Analysis From the VISTA Clinical Trial. **Alzheimer Disease & Associated Disorders**, v. 23, n. 2, 2009. ISSN 0893-0341. Disponível em: <http://journals.lww.com/alzheimerjournal/Fulltext/2009/04000/Verbal_Repetition_in_People_With_Mild_to_Moderate.8.aspx>. Citado na página 40.

CORRÊA, E. A.; LOPES, A. A.; AMANCIO, D. R. Word sense disambiguation: A complex network approach. **Information Sciences**, Elsevier, v. 442, p. 103–113, 2018. Citado na página 55.

CROMNOW, K.; LANDBERG, T. **Skriftliga beskrivningar av bilden Kakstölden. Insamling av referensvärden från friska försökspersoner**. Dissertação (Mestrado) — Division of Speech and Language Pathology, Karolinska institute, 2009. Citado na página 77.

CUETOS, F.; RODRÍGUEZ-FERREIRO, J.; MENÉNDEZ, M. Semantic markers in the diagnosis of neurodegenerative dementias. **Dementia and Geriatric Cognitive Disorders**, Hospital Álvarez Buylla, Mieres, Spain, v. 28, n. 3, p. 267–274, 2009. Citado na página 38.

CUNHA, A. L. V. d. **Coh-Metrix-Dementia: análise automática de distúrbios de linguagem nas demências utilizando Processamento de Línguas Naturais**. Dissertação (Mestrado) — Universidade de São Paulo, 2015. Citado nas páginas 28, 52 e 54.

CUNHA, A. L. V. D.; SOUSA, L. B. D.; MANSUR, L. L.; ALUÍSIO, S. M. Automatic proposition extraction from dependency trees: Helping early prediction of alzheimer's disease from narratives. In: **Proceedings of 28th International Symposium on Computer-Based Medical Systems**. São Carlos, Brazil: IEEE, 2015. p. 127–130. Citado nas páginas 32 e 71.

DAGAN, I.; GLICKMAN, O.; MAGNINI, B. The pascal recognising textual entailment challenge. In: **Machine Learning Challenges Workshop**. Southampton, United Kingdom: Springer, 2005. p. 177–190. Citado na página 61.

DAYRELL, C.; JR, A. C.; LIMA, G.; JR, D. M.; COPESTAKE, A.; FELTRIM, V.; TAGNIN, S.; ALUISIO, S. M. Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora. In: **Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)**. Istanbul, Turkey: European Language Resources Association (ELRA), 2012. p. 1604–1609. Citado na página 46.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, United States of America: Association for Computational Linguistics, 2019. p. 4171–4186. Citado nas páginas 59 e 113.

DRUKS, J.; MASTERSON, J.; KOPELMAN, M.; CLARE, L.; ROSE, A.; RAI, G. Is action naming better preserved (than object naming) in Alzheimer's disease and why should we ask? **Brain and Language**, v. 98, n. 3, p. 332–340, 2006. ISSN 0093-934X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0093934X06001167>. Citado na página 38.

ELISSEEFF, A.; WESTON, J. A kernel method for multi-labelled classification. In: **Proceedings of the 14th International Conference on Neural Information Processing Systems**. Vancouver, Canada: MIT Press, 2002. p. 681–687. Citado na página 48.

FELLBAUM, C. **Wordnet: An Electronic Lexical Database**. Cambridge, United States of America: MIT Press, 1998. Citado na página 88.

FENG, S.; CAI, Z.; CROSSLEY, S.; MCNAMARA, D. S. Simulating human ratings on word concreteness. In: **Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference**. Palm Beach, United States of America: Association for the Advancement of Artificial Intelligence, 2011. Citado na página 88.

FLEMING, V. B.; HARRIS, J. L. Complex discourse production in mild cognitive impairment: Detecting subtle changes. **Aphasiology**, Taylor & Francis, v. 22, n. 7-8, p. 729–740, 2008. Citado na página 41.

FOLSTEIN, M. F.; FOLSTEIN, S. E.; MCHUGH, P. R. "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. **Journal of psychiatric research**, Elsevier, v. 12, n. 3, p. 189–198, 1975. Citado na página 27.

FONSECA, E. R.; SANTOS, L. B. dos; CRISCUOLO, M.; ALUÍSIO, S. M. Visao geral da avaliaçao de similaridade semântica e inferência textual. **Linguamática**, v. 8, n. 2, p. 3–13, 2016. Citado nas páginas 60, 61, 62, 63 e 65.

FORBES-MCKAY, K.; VENNERI, A. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. **Neurological Sciences**, Springer-Verlag, v. 26, n. 4, p. 243–254, 2005. ISSN 1590-1874. Disponível em: <http://dx.doi.org/10.1007/s10072-005-0467-9>. Citado nas páginas 39 e 41.

FRASER, K. C.; FORS, K. L.; KOKKINAKIS, D. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. **Computer Speech & Language**, Elsevier, v. 53, p. 121–139, 2019. Citado nas páginas 30, 69, 72, 75, 77, 79, 86 e 95.

FRASER, K. C.; MELTZER, J. A.; GRAHAM, N. L.; LEONARD, C.; HIRST, G.; BLACK, S. E.; ROCHON, E. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. **Cortex**, Elsevier, v. 55, p. 43–60, 2014. Citado na página 31.

FRASER, K. C.; MELTZER, J. A.; RUDZICZ, F. Linguistic features identify alzheimer's disease in narrative speech. **Journal of Alzheimer's Disease**, IOS Press, v. 49, n. 2, p. 407–422, 2016. Citado nas páginas 28, 29, 30, 69, 72, 73, 79, 88 e 97.

FRAZIER, L. Syntactic complexity. In: DOWTY, D. R.; KARTTUNEN, L.; ZWICKY, A. M. (Ed.). **Natural language parsing: Psychological, computational, and theoretical perspectives**. Cambridge, United Kingdom: Cambridge University Press, 1985, (Studies in Natural Language Processing). Citado nas páginas 53 e 67.

FREITAS, M. I. D. **Habilidades linguísticas de pacientes com demência vascular: estudo comparativo com a doença de Alzheimer**. Tese (Doutorado) — Universidade de São Paulo, 2010. Citado nas páginas 35 e 42.

FROTA, N. A. F.; NITRINI, R.; DAMASCENO, B. P.; FORLENZA, O.; DIAS-TOSTA, E.; SILVA, A. B. d.; JUNIOR, E. H.; MAGALDI, R. M. Critérios para o diagnóstico de doença de alzheimer. **Dement. neuropsychol**, v. 5, n. supl 1, 2011. Citado nas páginas 36, 40 e 41.

GARCIA, F. H. A.; MANSUR, L. L. Habilidades funcionais de comunicação: idoso saudável. **Acta fisiátrica**, v. 13, n. 2, p. 87–89, 2006. Citado nas páginas 35 e 36.

GARRARD, P.; RALPH, M. A. L.; PATTERSON, K.; PRATT, K. H.; HODGES, J. R. Semantic feature knowledge and picture naming in dementia of Alzheimer's type: A new approach. **Brain and Language**, v. 93, n. 1, p. 79–94, 2005. ISSN 0093-934X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0093934X04002263>. Citado na página 37.

GEIGER, P. **Minidicionário Contemporâneo da Língua Portuguesa**. [S.l.]: Lexicon, 2011. Citado na página 92.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge, United States of America: MIT press, 2016. Citado na página 58.

GOODGLASS, H.; KAPLAN, E.; BARRESI, B. **The Assessment of Aphasia and Related Disorders**. Philadelphia, United States of America: Lippincott Williams & Wilkins, 1983. Citado nas páginas 39, 69 e 72.

GRAESSER, A. C.; MCNAMARA, D. S. Computational analyses of multilevel discourse comprehension. **Topics in cognitive science**, Wiley Online Library, v. 3, n. 2, p. 371–398, 2011. Citado na página 53.

GRAESSER, A. C.; MCNAMARA, D. S.; KULIKOWICH, J. M. Coh-metrix: Providing multilevel analyses of text characteristics. **Educational researcher**, Sage Publications Sage CA: Los Angeles, CA, v. 40, n. 5, p. 223–234, 2011. Citado nas páginas 52, 53, 54, 55 e 88.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. 3. ed. Waltham: Morgan Kaufmann, 2011. (Morgan Kaufmann Series in Data Management Systems). ISBN 9780123814807. Citado na página 49.

HARTMANN, N. S. Solo queue at assin: Combinando abordagens tradicionais e emergentes. **Linguamática**, v. 8, n. 2, p. 59–64, 2016. Citado nas páginas 65, 93 e 113.

HARTMANN, N. S.; FONSECA, E.; SHULBY, C.; TREVISO, M.; SILVA, J.; ALUÍSIO, S. M. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: **Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology**. Uberlandia, Brasil: Sociedade Brasileira de Computacão, 2017. p. 122–131. Citado na página 91.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition**. 2. ed. New York, United States of America: Springer New York, 2009. (Springer Series in Statistics). ISBN 9780387848587. Citado na página 100.

HENRY, J. D.; CRAWFORD, J. R.; PHILLIPS, L. H. Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. **Neuropsychologia**, v. 42, n. 9, p. 1212–1222, 2004. ISSN 0028-3932. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0028393204000296>. Citado na página 37.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997. Citado na página 59.

HODGES, J. R.; PATTERSON, K.; GRAHAM, N.; DAWSON, K. Naming and Knowing in Dementia of Alzheimer's Type. **Brain and Language**, v. 54, n. 2, p. 302–325, 1996. ISSN 0093-934X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0093934X96900772>. Citado na página 41.

HOELZLE, J. B.; NELSON, N. W.; SMITH, C. A. Comparison of wechsler memory scale–fourth edition (wms–iv) and third edition (wms–iii) dimensional structures: Improved ability to evaluate auditory and visual constructs. **Journal of clinical and experimental neuropsychology**, Taylor & Francis, v. 33, n. 3, p. 283–291, 2011. Citado na página 44.

HOOPER, T.; BAYLES, K. A. Management of neurogenic communication disorders associated with dementia. In: CHAPEY, R. (Ed.). **Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders**. 4. ed. Philadelphia: Wolters Kluwer, Lippincott Williams & Wilkins, 2007. p. 988–1008. Citado na página 38.

HOWARD, J.; RUDER, S. Universal language model fine-tuning for text classification. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 328–339. Citado na página 59.

HÜBNER, L. C.; LOUREIRO, F.; TESSARO, B.; SIQUEIRA, E. C. G.; JERÔNIMO, G. M.; SMIDARLE, A. Bale: Bateria de avaliação da linguagem no envelhecimento. In: ZIMMERMANN, N.; DELAERE, F.; FONSECA, R. P. (Ed.). **Tarefas de avaliação neuropsicológica para adultos: memória e linguagem**. 1. ed. Rio de Janeiro: Memnon, 2019. v. 3. Citado nas páginas 27, 28, 29, 31, 43, 81 e 82.

JANCZURA, G.; CASTILHO, G.; ROCHA, N.; ERVEN, T. van; HUANG, T. Normas de concretude para 909 palavras da língua portuguesa. **Psicologia: Teoria e Pesquisa**, scielo, p. 195–204, 2007. Citado nas páginas 88 e 89.

JARROLD, W. L.; PEINTNER, B.; YEH, E.; KRASNOW, R.; JAVITZ, H. S.; SWAN, G. E. Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic alzheimer's disease. In: YAO, Y.; SUN, R.; POGGIO, T.; LIU, J.; ZHONG, N.; HUANG, J. (Ed.). **Proceedings of International Conference on Brain Informatics (BI 2010)**. Arlington, United States of America: Springer Berlin Heidelberg, 2010. p. 299–307. Citado nas páginas 26 e 29.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data: an introduction to cluster analysis**. Hoboken: John Wiley & Sons, 2009. v. 344. Citado na página 77.

KEMPER, S.; THOMPSON, M.; MARQUIS, J. Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content. **Psychology and Aging**, American Psychological Association, v. 16, n. 4, p. 600–614, 2001. Citado na página 38.

KHEMCHANDANI, R.; CHANDRA, S. *et al.* Twin support vector machines for pattern classification. **IEEE Transactions on pattern analysis and machine intelligence**, Institute of Electrical and Electronics Engineers, v. 29, n. 5, p. 905–910, 2007. Citado na página 48.

KINTSCH, W. The role of knowledge in discourse comprehension: A construction-integration model. **Psychological Review**, v. 95, p. 163–182, 1988. Citado na página 39.

KINTSCH, W.; DIJK, T. A. van. Toward a model of text comprehension and production. **Psychological Review**, American Psychological Association, v. 85, n. 5, p. 363–394, 1978. Citado nas páginas 82 e 85.

KOUZANI, A. Z.; NASIREDING, G. Multilabel classification by bch code and random forests. **International journal of recent trends in engineering**, Academy Publisher, v. 2, n. 1, p. 113–116, 2009. Citado na página 48.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Nature Publishing Group, v. 521, n. 7553, p. 436, 2015. Citado na página 58.

LIANG, P.; TASKAR, B.; KLEIN, D. Alignment by agreement. In: **Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics**. New York City, United States of America: Association for Computational Linguistics, 2006. p. 104–111. Citado na página 74.

LIRA, J. O. de; ORTIZ, K. Z.; CAMPANHA, A. C.; BERTOLUCCI, P. H. F.; MINETT, T. S. C. Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. **International Psychogeriatrics**, v. 23, p. 404–412, 4 2011. ISSN 1741-203X. Disponível em: <http://journals.cambridge.org/article_S1041610210001092>. Citado na página 39.

LÓPEZ, R.; PARDO, T. A. Experiments on sentence boundary detection in user-generated web content. In: **Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics**. Cairo, Egypt: Springer, 2015. p. 227–237. Citado na página 95.

MACHICAO, J.; JR, E. A. C.; MIRANDA, G. H.; AMANCIO, D. R.; BRUNO, O. M. Authorship attribution based on life-like network automata. **PloS one**, Public Library of Science, v. 13, n. 3, p. e0193703, 2018. Citado na página 103.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. Cambridge, United Kingdom: Cambridge University Press, 2008. ISBN 9781139472104. Citado nas páginas 51 e 52.

MANNING, C. D.; MANNING, C. D.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Cambridge, United Kingdom: MIT press, 1999. Citado nas páginas 51 e 52.

MANSUR, L. L.; CARTHERY, M. T.; CARAMELLI, P.; NITRINI, R. Linguagem e cognição na doença de alzheimer. **Psicologia: reflexão e crítica**, SciELO Brasil, v. 18, n. 3, p. 300–307, 2005. Citado na página 36.

MAPSTONE, M.; CHEEMA, A. K.; FIANDACA, M. S.; ZHONG, X.; MHYRE, T. R.; MACARTHUR, L. H.; HALL, W. J.; FISHER, S. G.; PETERSON, D. R.; HALEY, J. M. *et al.* Plasma phospholipids identify antecedent memory impairment in older adults. **Nature medicine**, Nature Publishing Group, 2014. Citado na página 29.

MARELLI, M.; MENINI, S.; BARONI, M.; BENTIVOGLI, L.; BERNARDI, R.; ZAMPARELLI, R. A SICK cure for the evaluation of compositional distributional semantic models. In: **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)**. Reykjavik, Iceland: European Languages Resources Association (ELRA), 2014. p. 216–223. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf>. Citado nas páginas 60, 61 e 62.

MARINHO, V. Q.; HIRST, G.; AMANCIO, D. R. Authorship attribution via network motifs identification. In: **Proceedings og the 5th Brazilian Conference on Intelligent Systems (BRACIS)**. Recife, Brazil: IEEE, 2016. p. 355–360. Citado na página 55.

MARQUES, J. F. Normas de imagética e concreteza para substantivos comuns. **Laboratório de Psicologia**, Instituto Superior de Psicologia Aplicada, v. 3, p. 65–75, 2005. Citado nas páginas 88 e 89.

MARQUES, J. F.; CAPPA, S. F.; SARTORI, G. Naming from definition, semantic relevance and feature type: the effects of aging and Alzheimer's disease. **Neuropsychology**, v. 25, n. 1, p. 105–113, 2011. Citado na página 37.

MARQUES, J. F.; FONSECA, F. L.; MORAIS, S.; PINTO, I. A. Estimated age of acquisition norms for 834 portuguese nouns and their relation with other psycholinguistic variables. **Behavior Research Methods**, Springer, p. 439–444, 2007. Citado na página 89.

MAZIERO, E. G.; PARDO, T. A.; FELIPPO, A. D.; SILVA, B. C. Dias-da. A base de dados lexical e a interface web do tep 2.0-thesaurus eletrônico para o português do brasil. In: **Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web**. Vila Velha, Brazil: Association for Computing Machinery, 2008. p. 390–392. Citado na página 66.

MCCALLUM, A. Multi-label text classification with a mixture model trained by em. In: **Proceedings of the AAAI workshop on Text Learning**. [S.l.: s.n.], 1999. p. 1–7. Citado na página 46.

MCKHANN, G. M.; KNOPMAN, D. S.; CHERTKOW, H.; HYMAN, B. T.; JR, C. R. J.; KAWAS, C. H.; KLUNK, W. E.; KOROSHETZ, W. J.; MANLY, J. J.; MAYEUX, R. *et al.* The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. **Alzheimer's & Dementia**, Elsevier, v. 7, n. 3, p. 263–269, 2011. Citado nas páginas 29 e 36.

MCNAMARA, D. S.; LOUWERSE, M. M.; GRAESSER, A. C. **Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension**. Memphis, United States of America, 2002. Citado na página 52.

MEURIS, K.; MAES, B.; ZINK, I. Evaluation of language and communication skills in adult key word signing users with intellectual disability: Advantages of a narrative task. **Research in Developmental Disabilities**, v. 35, n. 10, p. 2585 – 2601, 2014. ISSN 0891-4222. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0891422214002649>. Citado na página 29.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient Estimation of Word Representations in Vector Space. In: **Proceedings of the International Conference on Learning Representations Workshop**. [S.l.: s.n.], 2013. Citado nas páginas 58 e 66.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. Cambridge, United Kingdom: MIT Press, 2012. v. 4. Citado nas páginas 46, 91, 100 e 102.

NAM, J.; MENCÍA, E. L.; KIM, H. J.; FÜRNKRANZ, J. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In: CURRAN ASSOCIATES INC. **Proceedings of the 31st International Conference on Neural Information Processing Systems**. Long Beach, United States of America, 2017. p. 5413–5423. Citado na página 49.

NITRINI, R.; CARAMELLI, P.; PORTO, C. S.; CHARCHAT-FICHMAN, H.; FORMIGONI, A. P.; CARTHERY-GOULART, M. T.; OTERO, C.; PRANDINI, J. C. Brief cognitive battery in the diagnosis of mild alzheimer's disease in subjects with medium and high levels of education. **Dementia Neuropsychol**, v. 1, p. 32–36, 2007. Citado na página 27.

ORAMAS, S.; BARBIERI, F.; NIETO, O.; SERRA, X. Multimodal deep learning for music genre classification. **Transactions of the International Society for Music Information Retrieval**, Ubiquity Press, v. 1, n. 1, p. 4–21, 2018. Citado na página 46.

ORIMAYE, S. O.; WONG, J.; GOLDEN, K. J. Learning predictive linguistic features for alzheimer's disease and related dementias using verbal utterances. In: **Proceedings of the 1st Workshop on Computational Linguistics and Clinical Psychology (CLPsych)**. Association for Computational Linguistics, 2014. p. 78–87. Disponível em: <www.aclweb.org/anthology/W/W14/W14-3210>. Citado nas páginas 69 e 70.

ORIMAYE, S. O.; WONG, J. S.; GOLDEN, K. J.; WONG, C. P.; SOYIRI, I. N. Predicting probable alzheimer's disease using linguistic deficits and biomarkers. **BMC bioinformatics**, BioMed Central, v. 18, n. 1, p. 34, 2017. Citado nas páginas 69, 70 e 71.

PAETZOLD, G.; SPECIA, L. Collecting and exploring everyday language for predicting psycholinguistic properties of words. In: **Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers**. Osaka, Japan: The COLING 2016 Organizing Committee, 2016. p. 1669–1679. Citado na página 90.

____. Inferring psycholinguistic properties of words. In: **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. San Diego, California: Association for Computational Linguistics, 2016. p. 435–440. Citado nas páginas 88, 89 e 91.

PAKHOMOV, S. V.; SMITH, G. E.; CHACON, D.; FELICIANO, Y.; GRAFF-RADFORD, N.; CASELLI, R.; KNOPMAN, D. S. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. **Cognitive and Behavioral Neurology**, NIH Public Access, v. 23, n. 3, p. 165, 2010. Citado nas páginas 30, 72 e 97.

PEINTNER, B.; JARROLD, W.; VERGYRI, D.; RICHEY, C.; TEMPINI, M. L. G.; OGAR, J. Learning diagnostic models using speech and language measures. In: **Proceedings of 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society**. Vancouver, Canada: IEEE, 2008. p. 4648–4651. Citado na página 29.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Citado na página 59.

PEROZZI, B.; AL-RFOU, R.; KULKARNI, V.; SKIENA, S. Inducing language networks from continuous space word representations. In: **Proceedings of the 5th Workshop on Complex Networks CompleNet 2014**. Bologna, Italy: Springer, 2014. p. 261–273. Citado na página 103.

PETERS, F.; MAJERUS, S.; BAERDEMAEKER, J. D.; SALMON, E.; COLLETTE, F. Impaired semantic knowledge underlies the reduced verbal short-term storage capacity in Alzheimer's disease. **Neuropsychologia**, v. 47, n. 14, p. 3067–3073, 2009. ISSN 0028-3932. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0028393209002942>. Citado na página 38.

PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTLEMOYER, L. Deep contextualized word representations. In: **Proceedings of NAACL-HLT**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 2227–2237. Citado nas páginas 59 e 113.

PRETI, D. (organizer). **O discurso oral culto**. 3. ed. São Paulo: Associação Editorial Humanitas, 2005. Projetos Paralelos. V.2. Citado na página 83.

PRUD'HOMMEAUX, E.; ROARK, B. Graph-based alignment of narratives for automated neurological assessment. In: **BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing**. Montréal, Canada: Association for Computational Linguistics, 2012. p. 1–10. Citado nas páginas 26 e 28.

PRUD'HOMMEAUX, E.; ROARK, B. Graph-based word alignment for clinical language evaluation. **Computational Linguistics**, MIT Press, v. 41, n. 4, p. 549–578, 2015. Citado nas páginas 30, 72, 73, 75, 76, 79 e 97.

PRUD'HOMMEAUX, E. T. **Alignment of Narrative Retellings for Automated Neuropsychological Assessment**. Tese (Doutorado) — Oregon Health & Science University, 2012. Citado na página 44.

PRUD'HOMMEAUX, E. T.; ROARK, B. Extraction of narrative recall patterns for neuropsychological assessment. In: **Proceedings of 12th Annual Conference of the International Speech Communication Association**. Florence, Italy: International Speech Communication Association, 2011. p. 3021–3024. Citado na página 45.

ROARK, B.; MITCHELL, M.; HOLLINGSHEAD, K. Syntactic complexity measures for detecting mild cognitive impairment. In: **Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing**. Prague, Czech Republic: Association for Computational Linguistics, 2007. p. 1–8. Citado nas páginas 28, 54, 67 e 68.

ROARK, B.; MITCHELL, M.; HOSOM, J.-P.; HOLLINGSHEAD, K.; KAYE, J. Spoken language derived measures for detecting mild cognitive impairment. **IEEE transactions on audio, speech, and language processing**, IEEE, v. 19, n. 7, p. 2081–2090, 2011. Citado nas páginas 67, 68, 79 e 86.

ROSENBERG, S.; ABBEDUTO, L. Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults 1. **Applied Psycholinguistics**, Cambridge University Press, v. 8, n. 1, p. 19–32, 1987. Citado na página 67.

SALMON, D. P.; THOMAS, R. G.; PAY, M. M.; BOOTH, A.; HOFSTETTER, C. R.; THAL, L. J.; KATZMAN, R. Alzheimer's disease can be accurately diagnosed in very mildly impaired individuals. **Neurology**, v. 59, n. 7, p. 1022–1028, 2002. Disponível em: <http://www.neurology.org/content/59/7/1022.abstract>. Citado na página 37.

SANDEN, C.; ZHANG, J. Z. Enhancing multi-label music genre classification through ensemble techniques. In: **Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval**. Beijing, China: Association for Computing Machinery, 2011. p. 705–714. Citado na página 46.

SANTOS, L. B.; Corrêa Júnior, E. A.; JÚNIOR, O. O.; AMANCIO, D.; MANSUR, L.; ALUÍSIO, S. Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 1284–1296. Disponível em: <https://www.aclweb.org/anthology/P17-1118>. Citado nas páginas 28, 86 e 104.

SANTOS, L. B.; HüBNER, L. C.; SMIDARLE, A. D.; MANSUR, L.; ALUÍSIO, S. M. Anotação de unidades de informação em transcrições de fala na tarefa de reconto de narrativas em português. In: **Proceedings do XII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Salvador, Brasil: [s.n.], 2019. Citado na página 82.

SANTOS, L. B. dos; DURAN, M. S.; HARTMANN, N. S.; CANDIDO, A.; PAETZOLD, G. H.; ALUISIO, S. M. A lightweight regression method to infer psycholinguistic properties for brazilian portuguese. In: **Proceedings of the 20th International Conference on Text, Speech and Dialogue**. Prague, Czech Republic: Springer, 2017. p. 281–289. Citado nas páginas 89, 91 e 92.

SARDINHA, T. B.; FILHO, J. M.; ALAMBERT, E. Corpus brasileiro. **Comunicaçao ao VII Encontro de Lingüıstica de Corpus**, 2008. Citado na página 90.

SCARTON, C.; GASPERIN, C.; ALUISIO, S. M. Revisiting the readability assessment of texts in portuguese. In: **Ibero-American Conference on Artificial Intelligence**. Bahía Blanca, Argentina: Springer, 2010. p. 306–315. Citado na página 52.

SCARTON, C. E.; ALUÍSIO, S. M. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. **Linguamática**, v. 2, n. 1, p. 45–61, 2010. Citado na página 52.

SHEN, D.; WEE, C.-Y.; ZHANG, D.; ZHOU, L.; YAP, P.-T. Machine learning techniques for ad/mci diagnosis and prognosis. In: **Machine Learning in Healthcare Informatics**. [S.l.]: Springer, 2014. p. 147–179. Citado na página 26.

SILVA, T. C.; AMANCIO, D. R. Word sense disambiguation via high order of learning in complex networks. **EPL (Europhysics Letters)**, v. 98, n. 5, p. 58001, 2012. Citado na página 55.

SKA, B.; DUONG, A. Communication, discours et démence. **Psychol NeuroPsychiatr Vieil**, v. 3, n. 2, p. 125–133, 2005. Citado na página 39.

SNOWDON, D. A.; GREINER, L. H.; MARKESBERY, W. R. Linguistic Ability in Early Life and the Neuropathology of Alzheimer's Disease and Cerebrovascular Disease: Findings from the Nun Study. **Annals of the New York Academy of Sciences**, Blackwell Publishing Ltd, v. 903, n. 1, p. 34–38, 2000. ISSN 1749-6632. Disponível em: <http://dx.doi.org/10.1111/j.1749-6632.2000.tb06347.x>. Citado na página 38.

SOARES, A. P.; COSTA, A. S.; J, M.; COMESANA, M. H. M. The minho word pool: Norms for imageability, concreteness, and subjective frequency for 3,800 portuguese words. **Behavior Research Methods**, 2016. Citado nas páginas 88 e 89.

SOLÉ, R. V.; COROMINAS-MURTRA, B.; VALVERDE, S.; STEELS, L. Language networks: Their structure, function, and evolution. **Complexity**, Wiley Online Library, v. 15, n. 6, p. 20–26, 2010. Citado na página 55.

SPERLING, R. A.; KARLAWISH, J.; JOHNSON, K. A. Preclinical alzheimer disease—the challenges ahead. **Nature Reviews Neurology**, Nature Publishing Group, v. 9, n. 1, p. 54–58, 2013. Citado na página 26.

STENETORP, P.; PYYSALO, S.; TOPIć, G.; OHTA, T.; ANANIADOU, S.; TSUJII, J. brat: a web-based tool for NLP-assisted text annotation. In: **Proceedings of the Demonstrations Session at EACL 2012**. Avignon, France: Association for Computational Linguistics, 2012. Citado na página 83.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. In: **Proceedings of the 27th International Conference on Neural Information Processing Systems**. Montréal, Canada: MIT Press, 2014. p. 3104–3112. Citado na página 59.

TANG, K. A 61 million word corpus of brazilian portuguese film subtitles as a resource for linguistic research. **UCL Working Papers in Linguistics**, v. 24, p. 208–214, 2012. Citado na página 90.

TANG, L.; RAJAN, S.; NARAYANAN, V. K. Large scale multi-label classification via metalabeler. In: **Proceedings of the 18th international conference on World wide web**. Madrid, Spain: Association for Computing Machinery, 2009. p. 211–220. Citado na página 50.

TOHALINO, J. V.; AMANCIO, D. R. Extractive multi-document summarization using multilayer networks. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 503, p. 526–539, 2018. Citado na página 55.

TOLEDO, C. M.; ALUÍSIO, S. M.; SANTOS, L. B. dos; BRUCKI, S. M. D.; TRÉS, E. S.; OLIVEIRA, M. O. de; MANSUR, L. L. Analysis of macrolinguistic aspects of narratives from individuals with alzheimer's disease, mild cognitive impairment, and no cognitive impairment. **Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring**, Elsevier, v. 10, p. 31–40, 2018. Citado na página 86.

TREVISO, M. V.; ALUíSIO, S. M. Sentence segmentation and disfluency detection in narrative transcripts from neuropsychological tests. In: **Proceedings of the 13th International Conference on the Computational Processing of the Portuguese Language (PROPOR)**. Canela, Brasil: Springer International Publishing, 2018. p. 409–418. Citado na página 87.

TSOUMAKAS, G.; KATAKIS, I.; VLAHAVAS, I. Mining multi-label data. In: **Data mining and knowledge discovery handbook**. 2. ed. New York, United States of America: Springer, 2009. p. 667–685. Citado nas páginas 46, 47, 48, 50 e 51.

VAJJALA, S.; MEURERS, D. Assessing the relative reading level of sentence pairs for text simplification. In: **Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics**. Gothenburg, Sweden: Association for Computational Linguistics, 2014. p. 288–297. Citado na página 88.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. In: CURRAN ASSOCIATES INC. **Proceedings of the 31st International Conference on Neural Information Processing Systems**. Long Beach, United States of America. p. 5998–6008. Citado na página 59.

VLIET, E. C.-V.; MANLY, J.; TANG, M.-X.; MARDER, K.; BELL, K.; STERN, Y. The neuropsychological profiles of mild Alzheimer's disease and questionable dementia as compared to age-related cognitive decline. **Journal of the International Neuropsychological Society**, v. 9, p. 720–732, 7 2003. ISSN 1469-7661. Disponível em: <http://journals.cambridge.org/article_S1355617703950053>. Citado na página 37.

WALLIN, A.; NORDLUND, A.; JONSSON, M.; LIND, K.; EDMAN, Å.; GÖTHLIN, M.; STÅLHAMMAR, J.; ECKERSTRÖM, M.; KERN, S.; BÖRJESSON-HANSON, A. *et al.* The gothenburg mci study: design and distribution of alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. **Journal of Cerebral Blood Flow & Metabolism**, SAGE Publications Sage UK: London, England, v. 36, n. 1, p. 114–131, 2016. Citado na página 77.

WANG, J.; YANG, Y.; MAO, J.; HUANG, Z.; HUANG, C.; XU, W. Cnn-rnn: A unified framework for multi-label image classification. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. Las Vegas, United States of America: IEEE, 2016. p. 2285–2294. Citado nas páginas 46 e 49.

WECHSLER, D. **Wechsler Memory Scale - Third Edition.** [S.l.]: The Psychological Corporation, San Antonio, TX., 1997. Citado nas páginas 29, 44, 45 e 46.

WORTMANN, M. Dementia: a global health priority-highlights from an adi and world health organization report. **Alzheimer's research & therapy**, BioMed Central, v. 4, n. 5, p. 40, 2012. Citado na página 26.

YANCHEVA, M.; RUDZICZ, F. Vector-space topic models for detecting alzheimer's disease. In: **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 2337–2346. Citado nas páginas 30, 69, 72, 75, 76, 77 e 95.

YNGVE, V. H. A model and an hypothesis for language structure. **Proceedings of the American philosophical society**, JSTOR, v. 104, n. 5, p. 444–466, 1960. Citado nas páginas 53, 67 e 86.

ZHANG, M.-L.; ZHOU, Z.-H. Multilabel neural networks with applications to functional genomics and text categorization. **IEEE transactions on Knowledge and Data Engineering**, Institute of Electrical and Electronics Engineers, v. 18, n. 10, p. 1338–1351, 2006. Citado na página 49.

____. Ml-knn: A lazy learning approach to multi-label learning. **Pattern recognition**, Elsevier, v. 40, n. 7, p. 2038–2048, 2007. Citado na página 48.

____. A review on multi-label learning algorithms. **IEEE transactions on knowledge and data engineering**, Institute of Electrical and Electronics Engineers, v. 26, n. 8, p. 1819–1837, 2013. Citado nas páginas 46, 50 e 51.

ZWAAN, R. A.; RADVANSKY, G. A. Situation models in language comprehension and memory. **Psychological bulletin**, American Psychological Association, v. 123, n. 2, p. 162, 1998. Citado na página 54.

ANEXO

A

MÉTRICAS DISPONÍVEIS NO NILC-METRIX

Nome da métrica	Descrição
relative_pronouns_diversity_ratio	Proporção de types de pronomes relativos em relação à quantidade de tokens de pronomes relativos no texto
hard_conjunctions_ratio	Proporção de conjunções difíceis em relação à quantidade de palavras do texto
easy_conjunctions_ratio	Proporção de conjunções fáceis em relação à quantidade de palavras do texto
simple_word_ratio	Proporção de palavras de conteúdo simples em relação a todas palavras de conteúdo do texto
noun_ratio	Proporção de substantivos em relação à quantidade de palavras do texto
min_cw_freq	Média das frequências das palavras de conteúdo mais raras das sentenças do texto
dalechall_adapted	Fórmula Dale Chall adaptada
adjunct_per_clause	Quantidade média de adjuntos adverbiais por oração do texto
content_words_ambiguity	Média de sentidos por palavra de conteúdo do texto
adverbs_ambiguity	Proporção de sentidos dos advérbios do texto em relação à quantidade de advérbios do texto
clauses_per_sentence	Quantidade média de orações por sentença
relative_clauses	Proporção de orações relativas em relação à quantidade de orações do texto

content_word_standard_deviation	Desvio padrão das proporções entre as palavras de conteúdo e a quantidade de palavras das sentenças
content_word_max	Proporção máxima de palavras de conteúdo em relação à quantidade de palavras das sentenças
content_word_min	Proporção Mínima de palavras de conteúdo por quantidade de palavras nas sentenças
indicative_pluperfect_ratio	Proporção de Verbos no Pretérito Mais que Perfeito do Indicativo em relação à quantidade de verbos flexionados no texto
honore	Estatística de Horoné
punctuation_diversity	Proporção de types de pontuações em relação à quantidade de tokens de pontuações no texto
mean_noun_phrase	Média dos tamanhos médios dos sintagmas nominais nas sentenças
min_noun_phrase	Mínimo entre os tamanhos de sintagmas nominais do texto
max_noun_phrase	Máximo entre os tamanhos de sintagmas nominais do texto
std_noun_phrase	Desvio-padrão do tamanho dos sintagmas nominais do texto
sentences	Quantidade de Sentenças no texto
infinite_subordinate_clauses	Proporção de orações subordinadas reduzidas pela quantidade de orações do texto
words_before_main_verb	Quantidade Média de palavras antes dos verbos principais das orações principais das sentenças
sentence_length_min	Quantidade Mínima de palavras por sentença
sentence_length_max	Quantidade Máxima de palavras por sentença
verbal_time_moods_diversity	Quantidade de diferentes tempos-modos verbais que ocorrem no texto
named_entity_ratio_sentence	Média das proporções de Nomes Próprios em relação à quantidade de palavras das Sentenças
apposition_per_clause	Quantidade média de apostos por oração do texto

negation_ratio	Proporção de palavras que denotam negação em relação à quantidade de palavras do texto
punctuation_ratio	Proporção de sinais de pontuação em relação à quantidade de palavras do texto.
oblique_pronouns_ratio	Proporção de pronomes oblíquos em relação a todos os pronomes do texto
preposition_diversity	Proporção de types de preposições em relação à quantidade de tokens de preposições no texto
third_person_possessive_pronouns	Proporção de pronomes possessivos nas terceiras pessoas em relação à quantidade de pronomes possessivos do texto
inflected_verbs	Proporção de verbos flexionados em relação a todos os verbos do texto
non-inflected_verbs	Proporção de verbos no gerúndio, particípio ou infinitivo em relação a todos os verbos do texto
adj_arg_ovl	Quantidade média de referentes que se repetem nos pares de sentenças adjacentes do texto
arg_ovl	Quantidade média de referentes que se repetem nos pares de sentenças do texto
adj_stem_ovl	Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças adjacentes do texto.
stem_ovl	Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças do texto
adj_cw_ovl	Quantidade média de palavras de conteúdo que se repetem nos pares de sentenças adjacentes do texto
indicative_imperfect_ratio	Proporção de Verbos no Pretérito Imperfeito do Indicativo em relação à quantidade de verbos flexionados no texto
indicative_condition_ratio	Proporção de Verbos no Futuro do Pretérito do Indicativo em relação à quantidade de verbos flexionados do texto
indicative_future_ratio	Proporção de Verbos no Futuro do Presente do Indicativo em relação à quantidade de verbos flexionados do texto
indicative_present_ratio	Proporção de Verbos no Presente do Indicativo em relação à quantidade de verbos flexionados no texto
subjunctive_present_ratio	Proporção de Verbos no Presente do Subjuntivo em relação à quantidade de verbos flexionados no texto

subjunctive_imperfect_ratio	Proporção de Verbos no Pretérito Imperfeito do Subjuntivo em relação à quantidade de verbos flexionados no texto
subjunctive_future_ratio	Proporção de Verbos no Futuro do Subjuntivo em relação à quantidade de verbos flexionados no texto
passive_ratio	Proporção de orações na voz passiva analítica em relação à quantidade de orações do texto
subordinate_clauses	Proporção de orações subordinadas pela quantidade de orações do texto
coordinate_conjunctions_per_clauses	Proporção de conjunções coordenativas em relação a todas as orações do texto
ratio_coordinate_conjunctions	Proporção de conjunções coordenativas em relação a todas as conjunções do texto
ratio_subordinate_conjunctions	Proporção de conjunções subordinativas em relação a todas conjunções do texto
sentences_with_zero_clause	Proporção de sentenças sem verbos em relação a todas as sentenças do texto
sentences_with_one_clause	Proporção de sentenças com 1 oração em relação a todas as sentenças do texto
sentences_with_two_clauses	Proporção de sentenças com 2 orações em relação a todas as sentenças do texto
sentences_with_three_clauses	Proporção de sentenças com 3 orações em relação a todas as sentenças do texto
sentences_with_four_clauses	Proporção de sentenças com 4 orações em relação a todas as sentenças do texto
sentences_with_five_clauses	Proporção de sentenças com 5 orações em relação a todas as sentenças do texto
sentences_with_six_clauses	Proporção de sentenças com 6 orações em relação a todas as sentenças do texto
sentences_with_seven_more_clauses	Proporção de sentenças com 7 ou mais orações em relação a todas as sentenças do texto
nouns_min	Proporção mínima de substantivos em relação à quantidade de palavras das sentenças

nouns_standard_deviation	Desvio padrão das proporções entre substantivos e a quantidade de palavras das sentenças
nouns_max	Proporção máxima de substantivos em relação à quantidade de palavras das sentenças
gunning_fox	Índice Gunning Fox
adjectives_min	Proporção mínima de adjetivos em relação à quantidade de palavras das sentenças
adjectives_standard_deviation	Desvio padrão das proporções entre adjetivos e a quantidade de palavras das sentenças
adjectives_max	Proporção máxima de adjetivos em relação à quantidade de palavras das sentenças
adverbs_min	Proporção mínima de advérbios em relação à quantidade de palavras das sentenças
adverbs	Proporção de Advérbios em relação à quantidade de palavras do texto
adverbs_standard_deviation	Desvio padrão das proporções entre advérbios e a quantidade de palavras das sentenças
adverbs_max	Proporção máxima de advérbios em relação à quantidade de palavras das sentenças
verbs_min	Proporção mínima de verbos em relação à quantidade de palavras das sentenças
verbs_standard_deviation	Desvio padrão das proporções entre verbos e a quantidade de palavras das sentenças
verbs_max	Proporção máxima de verbos por palavras em relação à quantidade de palavras das sentenças
noun_diversity	Proporção de types de substantivos em relação à quantidade de tokens de substantivos no texto
adjective_diversity_ratio	Proporção de types de adjetivos em relação à quantidade de tokens de adjetivos no texto
adjacent_refs	Média das proporções de candidatos a referentes na sentença anterior em relação aos pronomes pessoais do caso reto nas sentenças
adverbs_diversity_ratio	Proporção de types de advérbios em relação à quantidade de tokens de advérbios no texto

aux_plus_PCP_per_sentence	Proporção de verbos auxiliares seguidos de particípio em relação à quantidade de sentenças do texto
pronoun_diversity	Proporção de types de pronomes em relação à quantidade de tokens de pronomes no texto
content_word_diversity	Proporção de types de palavras de conteúdo em relação à quantidade de tokens de palavras de conteúdo no texto
verb_diversity	Proporção de types de verbos em relação à quantidade de tokens de verbos no texto
pronouns_min	Proporção mínima de pronomes em relação à quantidade de palavras das sentenças
personal_pronouns	Proporção de Pronomes Pessoais em relação à quantidade de palavras do texto
pronouns_standard_deviation	Desvio padrão das proporções entre pronomes e a quantidade de palavras das sentenças
pronouns_max	Proporção máxima de pronomes em relação à quantidade de palavras das sentenças
function_word_diversity	Proporção de types de palavras funcionais em relação à quantidade de tokens de palavras funcionais no texto
subtitles	Proporção de Subtítulos em relação à quantidade de sentenças do texto
sentence_length_standard_deviation	Desvio Padrão da quantidade de palavras por sentença
non_svo_ratio	Proporção de orações que não estão no formato SVO (sujeito-verbo-objeto) em relação a todas orações do texto
medium_long_sentence_ratio	Proporção de Sentenças Longas em relação a todas as sentenças do texto
long_sentence_ratio	Proporção de Sentenças Muito Longas em relação a todas as sentenças do texto
short_sentence_ratio	Proporção de Sentenças Curtas em relação a todas as sentenças do texto
medium_short_sentence_ratio	Proporção de Sentenças Médias em relação a todas as sentenças do texto
relative_pronouns_ratio	Proporção de Pronomes Relativos em relação à quantidade de pronomes do texto

temporal_adjunct_ratio	Proporção de adjuntos adverbiais de tempo em relação a todos os adjuntos adverbiais do texto
anaphoric_refs	Média das proporções de candidatos a referentes nas 5 sentenças anteriores em relação aos pronomes anafóricos das sentenças
coreference_pronoum_ratio	Média de candidatos a referente, na sentença anterior, por pronome anafórico do caso reto
demonstrative_pronoun_ratio	Média de candidatos a referente, na sentença anterior, por pronome demonstrativo anafórico
postponed_subject_ratio	Proporção de sujeitos pospostos em relação a todos os sujeitos do texto
adverbs_before_main_verb_ratio	Proporção de orações com advérbio antes do verbo principal em relação à quantidade de orações do texto
indefinite_pronouns_diversity	Proporção de types de pronomes indefinidos em relação à quantidade de tokens de pronomes indefinidos no texto
indefinite_pronoun_ratio	Proporção de pronomes indefinidos em relação a todos os pronomes do texto
if_ratio	Proporção do operador lógico SE em relação à quantidade de palavras do texto
or_ratio	Proporção do operador lógico OU em relação à quantidade de palavras do texto
and_ratio	Proporção do operador lógico E em relação à quantidade de palavras do texto
logic_operators	Proporção de Operadores Lógicos em relação à quantidade de palavras do texto
positive_words	Proporção de palavras de polaridade positiva em relação a todas palavras do texto
negative_words	Proporção de palavras de polaridade negativa em relação a todas palavras do texto
hypernyms_verbs	Quantidade Média de Hiperônimos por verbo nas sentenças
cw_freq	Média das frequências absolutas das palavras de conteúdo do texto

conn_ratio	Proporção de Conectivos em relação à quantidade de palavras do texto
add_pos_conn_ratio	Proporção de conectivos aditivos positivos em relação à quantidade de palavras do texto
add_neg_conn_ratio	Proporção de conectivos aditivos negativos em relação à quantidade de palavras do texto
tmp_pos_conn_ratio	Proporção de conectivos temporais positivosem relação à quantidade de palavras do texto
tmp_neg_conn_ratio	Proporção de conectivos temporais negativos em relação à quantidade de palavras do texto
cau_pos_conn_ratio	Proporção de conectivos causais positivos em relação à quantidade de palavras do texto
cau_neg_conn_ratio	Proporção de conectivos causais negativos em relação à quantidade de palavras do texto
log_pos_conn_ratio	Proporção de Conectivos Lógicos Positivos em relação à quantidade de palavras do texto
log_neg_conn_ratio	Proporção de Conectivos Lógicos Negativos em relação à quantidade de palavras do texto
function_words	Proporção de Palavras Funcionais em relação à quantidade de palavras do texto
content_words	Proporção de palavras de conteúdo em relação à quantidade de palavras do texto
pronoun_ratio	Proporção de pronomes em relação à quantidade de palavras do texto
adjective_ratio	Proporção de Adjetivos em relação à quantidade de palavras do texto
words	Quantidade de Palavras no texto
paragraphs	Quantidade de Parágrafos no texto
verbs	Proporção de Verbos em relação à quantidade de palavras do texto
syllables_per_content_word	Quantidade média de sílabas por palavra no texto
words_per_sentence	Média de Palavras por Sentença

indicative_preterite_perfect_ratio	Proporção de Verbos no Pretérito Perfeito Simples do Indicativo em relação à quantidade de verbos flexionados no texto
gerund_verbs	Proporção de verbos no gerúndio em relação a todos os verbos do texto
infinitive_verbs	Proporção de verbos no infinitivo em relação a todos os verbos do texto
participle_verbs	Proporção de verbos no particípio em relação a todos os verbos do texto
abstract_nouns_ratio	Proporção de substantivos abstratos em relação à quantidade de palavras do texto
sentences_per_paragraph	Quantidade média de sentenças por parágrafo no texto
prepositions_per_sentence	Quantidade Média de preposições por sentença no texto
prepositions_per_clause	Proporção de preposições em relação à quantidade de orações no texto
first_person_possessive_pronouns	Proporção de pronomes possessivos nas primeiras pessoas em relação à quantidade de pronomes possessivos do texto
second_person_possessive_pronouns	Proporção de pronomes possessivos nas segundas pessoas em relação à quantidade de pronomes possessivos do texto
first_person_pronouns	Proporção de pronomes pessoais nas primeiras pessoas em relação à quantidade de pronomes pessoais do texto
second_person_pronouns	Proporção de pronomes pessoais nas segundas pessoas em relação à quantidade de pronomes pessoais do texto
third_person_pronouns	Proporção de pronomes pessoais nas terceiras pessoas em relação à quantidade de pronomes pessoais do texto
dialog_pronoun_ratio	Proporção de pronomes pessoais que indicam uma conversa com o leitor em relação à quantidade de pronomes pessoais do texto
content_density	Proporção de palavras de conteúdo em relação à quantidade de palavras funcionais do texto
ttr	Proporção de types (despreza repetições de palavras) em relação à quantidade de tokens (computa repetições de palavras) no texto

verbs_ambiguity	Proporção de sentidos dos verbos do texto em relação à quantidade de verbos do texto
adjectives_ambiguity	Proporção de sentidos dos adjetivos do texto em relação à quantidade de adjetivos do texto
flesch	Índice Flesch
brunet	Índice de Brunet
nouns_ambiguity	Proporção de sentidos dos substantivos do texto em relação à quantidade de substantivos do texto
named_entity_ratio_text	Proporção de Nomes Próprios em relação à quantidade de palavras do Texto
idade_aquisicao_1_25_ratio	Proporção de palavras com valor de idade de aquisição entre 1 e 2,5 em relação a todas as palavras de conteúdo do texto
idade_aquisicao_25_4_ratio	Proporção de palavras com valor de idade de aquisição entre 2,5 e 4 em relação a todas as palavras de conteúdo do texto
idade_aquisicao_4_55_ratio	Proporção de palavras com valor de idade de aquisição entre 4 e 5,5 em relação a todas as palavras de conteúdo do texto
idade_aquisicao_55_7_ratio	Proporção de palavras com valor de idade de aquisição entre 5,5 e 7 em relação a todas as palavras de conteúdo do texto
familiaridade_1_25_ratio	Proporção de palavras com valor de familiaridade entre 1 e 2,5 em relação a todas as palavras de conteúdo do texto
familiaridade_25_4_ratio	Proporção de palavras com valor de familiaridade entre 2,5 e 4 em relação a todas as palavras de conteúdo do texto
familiaridade_4_55_ratio	Proporção de palavras com valor de familiaridade entre 4 e 5,5 em relação a todas as palavras de conteúdo do texto
familiaridade_55_7_ratio	Proporção de palavras com valor de familiaridade entre 5,5 e 7 em relação a todas as palavras de conteúdo do texto
imageabilidade_1_25_ratio	Proporção de palavras com valor de imageabilidade entre 1 e 2,5 em relação a todas as palavras de conteúdo do texto
imageabilidade_25_4_ratio	Proporção de palavras com valor de imageabilidade entre 2,5 e 4 em relação a todas as palavras de conteúdo do texto

imageabilidade_4_55_ratio	Proporção de palavras com valor de imageabilidade entre 4 e 5,5 em relação a todas as palavras de conteúdo do texto
imageabilidade_55_7_ratio	Proporção de palavras com valor de imageabilidade entre 5,5 e 7 em relação a todas as palavras de conteúdo do texto
concretude_1_25_ratio	Proporção de palavras com valor de concretude entre 1 e 2,5 em relação a todas as palavras de conteúdo do texto
concretude_25_4_ratio	Proporção de palavras com valor de concretude entre 2,5 e 4 em relação a todas as palavras de conteúdo do texto
concretude_4_55_ratio	Proporção de palavras com valor de concretude entre 4 e 5,5 em relação a todas as palavras de conteúdo do texto
concretude_55_7_ratio	Proporção de palavras com valor de concretude entre 5,5 e 7 em relação a todas as palavras de conteúdo do texto
idade_aquisicao_mean	Média dos valores de idade de aquisição das palavras de conteúdo do texto
concretude_mean	Média dos valores de concretude das palavras de conteúdo do texto
concretude_std	Desvio padrão dos valores de concretude das palavras de conteúdo do texto
idade_aquisicao_std	Desvio padrão dos valores de idade de aquisição das palavras de conteúdo do texto
familiaridade_mean	Média dos valores de familiaridade das palavras de conteúdo do texto
familiaridade_std	Desvio padrão dos valores de familiaridade das palavras de conteúdo do texto
imageabilidade_mean	Média dos valores de imageabilidade das palavras de conteúdo do texto
imageabilidade_std	Desvio padrão dos valores de imageabilidade das palavras de conteúdo do texto
yngve	Fórmula de Complexidade Sintática de Yngve
frazier	Fórmula de Complexidade Sintática de Frazier
dep_distance adj_mean	Distância na árvore de dependências Média de similaridade entre pares de sentenças adjacentes no texto

adj_std	Desvio padrão de similaridade entre pares de sentenças adjacentes no texto
all_mean	Média de similaridade entre todos os pares de sentenças no texto
all_std	Desvio padrão de similaridade entre todos os pares de sentenças no texto
paragraph_mean	Média entre parágrafos adjacentes
paragraph_std	Desvio padrão entre parágrafos adjacentes
givenness_mean	Média de *givenness* das sentenças, sendo que *givenness* de uma sentença é similaridade LSA entre a sentença e todo o texto que a precede
givenness_std	Desvio padrão de *givenness* das sentenças
span_mean	Média do *span* das sentenças. O *span* de uma sentença é o cosseno do vetor da sentença pela projeção das sentenças anteriores
span_std	Desvio padrão de *span* das sentenças


ICMC
USP
SÃO CARLOS