
Coh-Matrix-Dementia: análise automática de
distúrbios de linguagem nas demências utilizando
Processamento de Línguas Naturais

Andre Luiz Verucci da Cunha

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Andre Luiz Verucci da Cunha

Coh-Metrix-Dementia: análise automática de distúrbios de linguagem nas demências utilizando Processamento de Línguas Naturais

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre 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
Dezembro de 2015**

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

C972c Cunha, Andre Luiz Verucci da
Coh-Metrix-Dementia: análise automática de
distúrbios de linguagem nas demências utilizando
Processamento de Línguas Naturais / Andre Luiz
Verucci da Cunha; orientadora Sandra Maria Aluísio. -
- São Carlos, 2015.
153 p.

Dissertação (Mestrado - Programa de Pós-Graduação
em Ciências de Computação e Matemática
Computacional) -- Instituto de Ciências Matemáticas
e de Computação, Universidade de São Paulo, 2015.

1. Doença de Alzheimer. 2. Comprometimento
Cognitivo Leve. 3. classificação textual. 4.
métricas textuais. I. Aluísio, Sandra Maria, orient.
II. Título.

Andre Luiz Verucci da Cunha

**Coh-Matrix-Dementia: automatic analysis of language
impairment in dementia using Natural Language Processing**

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Sandra Maria Aluísio

**USP – São Carlos
December 2015**

*Este trabalho é dedicado a todas as pessoas que,
mesmo no fim de suas vidas, puderem se beneficiar dele.*

AGRADECIMENTOS

Agradeço à minha orientadora, Prof.^a Dr.^a Sandra Aluísio, pela dedicação intensa e incessante ao meu trabalho, mesmo diante das etapas mais complexas do meu trajeto no mestrado. Agradeço também à minha co-orientadora, Prof.^a Dr.^a Letícia Mansur, por seu apoio inestimável durante a realização deste trabalho.

Agradeço aos meus amigos do Núcleo Interinstitucional de Linguística Computacional (NILC-USP) – cujos nomes não listo aqui para não correr o risco de esquecer alguém –, pela companhia e amizade excelentes, que tornaram minha passagem pelo laboratório inesquecível. Agradeço especialmente ao Leandro, que sempre garantiu que o bom humor estivesse presente na minha rotina, e ao Gustavo e à Marina, que me apoiaram e me levantaram mais do que eu jamais poderia retribuir.

Agradeço aos meus amigos “virtuais” (longes do corpo, mas não do coração) que conheci durante meu mestrado (Ana, Artur, Pablo, Wallker, e diversos outros), que surgiram na minha vida sem que eu nem me desse conta, se fizeram parte dela, e ajudaram a aliviar minha solidão quando não havia ninguém por perto.

Agradeço aos amigos que conheci no Canadá (Aldea, Jed, MaryAnn, Muyu, e outros), por terem feito da minha visita uma das melhores experiências que já vivi.

Agradeço à minha família, que sempre esteve ao meu lado durante o mestrado, mesmo em momentos dos mais turbulentos. Agradeço (com não menos importância) aos meus amigos de longa data (André, Carlos, Isabela, Jonas, Larissa, Marcelo, Ricardo, e tantos outros), por tornarem minha vida muito mais interessante e feliz.

Agradeço à CAPES e à FAPESP (processo 2013/16182-0) pelo apoio financeiro durante a execução deste projeto de mestrado.

“Science is the poetry of reality.”
(Richard Dawkins)

RESUMO

CUNHA, A. L. V.. **Coh-Metrix-Dementia: análise automática de distúrbios de linguagem nas demências utilizando Processamento de Línguas Naturais**. 2015. 153 f. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

(*Contexto*) Segundo a Organização Mundial da Saúde, as demências são um problema de custo social elevado, cujo manejo é um desafio para as próximas décadas. Demências comuns incluem a Doença de Alzheimer (DA), bastante conhecida. Outra síndrome menos conhecida, o Comprometimento Cognitivo Leve (CCL), é relevante por ser o estágio inicial clinicamente definido da DA. Embora o CCL não seja tão conhecido do público, pessoas com um tipo especial dessa síndrome, o CCL amnésico, evoluem para a DA a uma taxa bastante maior que a da população em geral. O diagnóstico das demências e síndromes relacionadas é feito com base na análise de aspectos linguísticos e cognitivos do paciente. Testes clássicos incluem testes de fluência, nomeação, e repetição. Entretanto, pesquisas recentes têm reconhecido cada vez mais a importância da análise da produção discursiva, especialmente de narrativas, como uma alternativa mais adequada, principalmente para a detecção do CCL. (*Lacuna*) 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 é bastante laborioso, o que dificulta sua adoção em larga escala. Nesse cenário, análises computadorizadas despontam como uma solução de interesse. Ferramentas de análise automática de discurso com vistas ao diagnóstico de demências de linguagem já existem para o inglês, mas nenhum trabalho nesse sentido foi feito para o português até o presente momento. (*Objetivo*) Este projeto visa criar um ambiente unificado, intitulado Coh-Metrix-Dementia, que se valerá de recursos e ferramentas de Processamento de Línguas Naturais (PLN) e de Aprendizado de Máquina para possibilitar a análise e o reconhecimento automatizados de demências, com foco inicial na DA e no CCL. (*Hipótese*) Tendo como base o ambiente Coh-Metrix adaptado para o português do Brasil, denominado Coh-Metrix-Port, e incluindo a adaptação para o português e inserção de vinte e cinco novas métricas para calcular a complexidade sintática, a densidade de ideias, e a coerência textual, via semântica latente, é possível classificar narrativas de sujeitos normais, com DA, e com CCL, em uma abordagem de aprendizado de máquina, com precisão comparável a dos testes clássicos. (*Conclusão*) Nos resultados experimentais, foi possível separar os pacientes entre controles, CCL, e DA com medida F de 81,7%, e separar controles e CCL com medida F de 90%. Os resultados indicam que o uso das métricas da ferramenta Coh-Metrix-Dementia é bastante promissor como recurso na detecção precoce de declínio nas habilidades de linguagem.

Palavras-chave: Doença de Alzheimer, Comprometimento Cognitivo Leve, classificação textual, métricas textuais.

ABSTRACT

CUNHA, A. L. V.. **Coh-Metrix-Dementia: análise automática de distúrbios de linguagem nas demências utilizando Processamento de Línguas Naturais**. 2015. 153 f. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

(Background) According to the World Health Organization, dementia is a costly social issue, whose management will be a challenge on the coming decades. One common form of dementia is Alzheimer's Disease (AD). Another less known syndrome, Mild Cognitive Impairment (MCI), is relevant for being the initial clinically defined stage of AD. Even though MCI is less known by the public, patients with a particular variant of this syndrome, Amestic MCI, evolve to AD in a considerably larger proportion than that of the general population. The diagnosis of dementia and related syndromes is based on the analysis of linguistic and cognitive aspects. Classical exams include fluency, naming, and repetition tests. However, recent research has been recognizing the importance of discourse analysis, specially narrative-based, as a more suitable alternative, specially for MCI detection. *(Gap)* While qualitative discourse analyses can determine the nature of the patient's disease, quantitative analyses can reveal the extent of the existing brain damage. The greatest challenge in quantitative discourse analyses is that a rigorous and thorough evaluation of oral production is very labor-intensive, which hinders its large-scale adoption. In this scenario, computerized analyses become of increasing interest. Automated discourse analysis tools aiming at the diagnosis of language-impairing dementias already exist for the English language, but no such work has been made for Brazilian Portuguese so far. *(Goal)* This project aims to create a unified environment, entitled Coh-Metrix-Dementia, that will make use of Natural Language Processing and Machine Learning resources and tools to enable automated dementia analysis and classification, initially focusing on AD and MCI. *(Hypothesis)* Basing our work on Coh-Metrix-Port, the Brazilian Portuguese adaption of Coh-Metrix, and including the adaptation and inclusion of twenty-five new metrics for measuring syntactical complexity, idea density, and text cohesion through latent semantics, it is possible to classify narratives of healthy, AD, and MCI patients, in a machine learning approach, with a precision comparable to classical tests. *(Conclusion)* In our experiments, it was possible to separate patients in controls, DA, and CCL with 81.7% F-measure, and separate controls and CCL with 90% F-measure. These results indicate that Coh-Metrix-Dementia is a very promising resource in the early detection of language impairment.

Key-words: Alzheimer's Disease, Mild Cognitive Impairment, text classification, textual metrics.

LISTA DE ILUSTRAÇÕES

Figura 1 – Tela do Coh-Metrix exibindo métricas de um texto de exemplo.	55
Figura 2 – Saída do Coh-Metrix-Port para um texto de exemplo.	60
Figura 3 – Exemplo de parte da tela de saída do AIC.	62
Figura 4 – Tela de saída do LIWC, com o dicionário para o inglês, versão de 2007, mostrando a análise de uma música.	63
Figura 5 – Exemplo de árvore sintática para cálculo da complexidade de Yngve (extraído de Pakhomov <i>et al.</i> (2011)).	66
Figura 6 – Exemplo de árvore sintática para cálculo da complexidade de Frazier (extraído de Pakhomov <i>et al.</i> (2011)).	67
Figura 7 – Exemplo da Figura 6 com os ramos individuais das palavras.	67
Figura 8 – Exemplo de estrutura de dependências.	68
Figura 9 – Exemplo de uso do CPIDR.	74
Figura 10 – Exemplo de uso do CPIDR no modo fala.	75
Figura 11 – Transcrição no formato CHAT da história da Cinderela (extraído de (MACWHINNEY <i>et al.</i> , 2010)).	80
Figura 12 – Arquitetura geral do Coh-Metrix-Dementia.	92
Figura 13 – Interface web do Coh-Metrix-Dementia: tela inicial.	93
Figura 14 – Interface web do Coh-Metrix-Dementia: lista dos textos do usuário.	93
Figura 15 – Interface web do Coh-Metrix-Dementia: tela para submissão de um novo texto.	94
Figura 16 – Interface web do Coh-Metrix-Dementia: visão geral de um texto analisado.	94
Figura 17 – Fluxo de trabalho para criação da ferramenta IDD3.	103
Figura 18 – Proposições geradas pelo IDD3 para uma sentença de exemplo.	103
Figura 19 – Conjuntos de regras para processamento da sentença de exemplo da Figura 18.	104
Figura 20 – Correlação, em termos de contagem de proposições, entre Chand <i>et al.</i> , IDD3, e CPIDR para 91 sentenças de exemplo (CUNHA <i>et al.</i> , 2015).	115
Figura 21 – Cenas da história da Cinderela apresentadas aos sujeitos.	120
Figura 22 – Exemplos de transcrição bruta e adaptada, para um sujeito com DA leve.	121
Figura 23 – Gráfico de dispersão do conjunto de dados normalizado. Dados de controles em vermelho, CCL em verde, DA em azul.	127
Figura 24 – Árvore de decisão gerada pelo algoritmo J48 para três classes (cenário 1) utilizando-se os atributos selecionados pelo CFS.	128
Figura 25 – Árvore de decisão gerada pelo algoritmo J48 para duas classes (cenário 2) utilizando-se os atributos selecionados pelo CFS.	130

LISTA DE TABELAS

Tabela 1 – Métricas do Coh-Metrix 3.0 (extraído de < http://141.225.42.101/CohMetrixHome/documentation_indices.html >).	49
Tabela 3 – Métricas do Coh-Metrix-Port.	56
Tabela 2 – Desempenho dos classificadores com características extraídas pelo Coh-Metrix-Port.	61
Tabela 4 – Exemplos de proposições e densidade de ideias (extraídos de Chand <i>et al.</i> (2010)).	73
Tabela 5 – Métricas de LSA do Coh-Metrix 3.0.	78
Tabela 6 – Sumário dos resultados de acurácia de classificação de Thomas <i>et al.</i> (2005).	82
Tabela 7 – Sumário dos testes de significância de Bryant <i>et al.</i> (2013) (X sinaliza significância).	86
Tabela 8 – Resultados (aproximados) obtidos por Peintner <i>et al.</i> (2008).	87
Tabela 9 – Resultados obtidos por Jarrold <i>et al.</i> (2010).	87
Tabela 10 – Acurácia (%) dos três classificadores de Fraser <i>et al.</i> (2012).	89
Tabela 11 – Métricas de disfluência acrescentadas ao Coh-Metrix-Dementia.	96
Tabela 12 – Córpus utilizados para geração do espaço LSA.	96
Tabela 13 – Medida F (%) dos métodos para o primeiro conjunto de experimentos (3-8 versus 9+).	111
Tabela 14 – Medida F (%) dos métodos com classes extremas e intermediárias.	112
Tabela 15 – Medida F (%) dos métodos de classificação com remoção de classes intermediárias.	112
Tabela 16 – Medida F (%) dos métodos de classificação e seleção de atributos com remoção de classes intermediárias.	113
Tabela 17 – Estatísticas do córpus de avaliação (CUNHA <i>et al.</i> , 2015).	116
Tabela 18 – Medida F total para as anotações (CUNHA <i>et al.</i> , 2015).	116
Tabela 19 – Medida F média para as anotações (CUNHA <i>et al.</i> , 2015).	117
Tabela 20 – Valores de correlação para a contagem de proposições entre o IDD3 e a anotação manual, no córpus de avaliação (CUNHA <i>et al.</i> , 2015).	117
Tabela 21 – Critérios de inclusão e exclusão dos sujeitos (TOLEDO, 2015).	119
Tabela 22 – Estatísticas do córpus utilizado na avaliação.	122
Tabela 23 – Medida F (%) dos métodos de classificação para diferentes conjuntos de métricas, com três classes (cenário 1).	123
Tabela 24 – Atributos selecionados pelo CFS para três classes (cenário 1).	124

Tabela 25 – Média e desvio padrão (normalizados) para as métricas selecionadas pelo CFS para três classes (cenário 1).	124
Tabela 26 – Matriz de confusão para o algoritmo Naive Bayes com três classes (cenário 1).	128
Tabela 27 – Medida F (%) dos métodos de classificação para diferentes conjuntos de métricas, com duas classes (cenário 2).	129
Tabela 28 – Atributos selecionados pelo CFS para duas classes (cenário 2).	129
Tabela 29 – Matriz de confusão para o algoritmo J48 para duas classes (cenário 2). . . .	130
Tabela 30 – p -valor para teste de Wilcoxon pareado unilateral, no cenário 1, Coh-Metrix-Dementia <i>versus</i> Coh-Metrix-Port.	131
Tabela 31 – p -valor para teste de Wilcoxon pareado unilateral, no cenário 2, Coh-Metrix-Dementia <i>versus</i> Coh-Metrix-Port.	131

SUMÁRIO

1	INTRODUÇÃO	21
2	ENVELHECIMENTO E LINGUAGEM	27
2.1	O envelhecimento normal	27
2.2	A linguagem nos indivíduos acometidos por Doença de Alzheimer	33
2.3	A linguagem nos indivíduos acometidos por Comprometimento Cognitivo Leve	40
2.4	Considerações finais	42
3	ANÁLISE AUTOMATIZADA DE LINGUAGEM	45
3.1	Ferramentas Automáticas	45
3.1.1	<i>O Coh-Metrix e o Coh-Metrix-Port</i>	46
3.1.2	<i>A ferramenta Análise de Inteligibilidade de Córpus (AIC)</i>	60
3.1.3	<i>O Linguistic Inquiry and Word Count (LIWC)</i>	62
3.2	Medidas Automatizadas	64
3.2.1	<i>Medidas de diversidade lexical</i>	64
3.2.2	<i>Medidas de complexidade sintática</i>	65
3.2.2.1	<i>Complexidade de Yngve</i>	65
3.2.2.2	<i>Complexidade de Frazier</i>	66
3.2.2.3	<i>Distância de dependência</i>	67
3.2.2.4	<i>Nível de desenvolvimento (nível D)</i>	68
3.2.2.5	<i>Entropia cruzada</i>	69
3.2.2.6	<i>Considerações finais</i>	70
3.2.3	<i>Medidas de densidade semântica</i>	71
3.2.3.1	<i>A Análise de Densidade de Ideias (AID)</i>	71
3.2.3.2	<i>A densidade de CSU's</i>	74
3.2.4	<i>Medidas de semântica latente</i>	75
4	TRABALHOS RELACIONADOS	79
4.1	Abordagem lexical	79
4.2	Abordagem baseada em complexidade sintática	82
4.3	Abordagem baseada em densidade de ideias	85
4.4	Abordagem baseada em traços semânticos e categorias morfossintáticas	86

4.5	Abordagem em vários níveis	88
5	COH-METRIX-DEMENTIA: IMPLEMENTAÇÃO	91
5.1	Arquitetura do Coh-Metrix-Dementia	91
5.2	As Métricas Herdadas do Coh-Metrix-Port	92
5.3	As Novas Métricas	95
5.3.1	<i>Métricas de disfluência</i>	95
5.3.2	<i>Métricas de Análise de Semântica Latente</i>	96
5.3.3	<i>Diversidade lexical e complexidade sintática</i>	99
5.3.4	<i>Densidade semântica</i>	102
6	COH-METRIX-DEMENTIA: AVALIAÇÃO	107
6.1	Experimento Piloto	107
6.2	Das Ferramentas Desenvolvidas	114
6.2.1	<i>Extração de Proposições</i>	114
6.3	Da Classificação Textual	118
6.3.1	<i>Cópus de pesquisa</i>	118
6.3.2	<i>Treinamento dos Classificadores</i>	122
7	CONCLUSÕES	133
7.1	Conclusões e Contribuições	133
7.2	Limitações e Trabalhos Futuros	134
	REFERÊNCIAS	137

INTRODUÇÃO

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 IBGE de 2000¹ e 2010². Segundo o censo de 2000, havia no país cerca de 19 mil homens e 37 mil mulheres entre 95 e 99 anos de idade; em 2010, já eram 32 mil homens e 67 mil mulheres. Para a faixa de 90 a 94 anos, havia 65 mil homens e 115 mil mulheres em 2000, enquanto em 2010 esses números subiram para 115 mil homens e 212 mil mulheres. Na faixa de 85 a 89 anos, o aumento nestes dez anos foi de 0,1% para 0,2% para homens, e de 0,2% para 0,3% para mulheres. A pirâmide se afina na base e se alarga no topo, e a estimativa para o país é de que, em 2025, existam 31,8 milhões de idosos com sessenta anos de idade ou mais (Jacob Filho, 2000).

O envelhecimento, da maturidade à senescência, traz mudanças, que têm como principal característica a redução de reservas cognitivas e funcionais. Independentemente de doenças, notam-se mudanças primárias, que resultam da passagem do tempo e que podem ser aceleradas ou retardadas de acordo com o estilo de vida e outros fatores, mas são geralmente evidentes na quarta ou quinta década de vida, graduais, e inexoráveis (SKA; JOANETTE, 2006).

Se, por um lado, a redução da reserva funcional torna o indivíduo vulnerável ao desenvolvimento de doenças, a manutenção da funcionalidade, em contraposição, constitui importante fator para a estabilidade da saúde em boas condições. Habilidades de linguagem refletem a saúde cognitiva do idoso e compõem importantes pilares para a manutenção da funcionalidade, do envelhecimento saudável (CARVALHO; MANSUR, 2008).

O aumento da expectativa de vida está associado aos avanços nas áreas de saúde e de saneamento básico, e reflete seus efeitos sobre a taxa de mortalidade (PASCHOAL, 1996). No entanto, esse aumento poderá trazer sérios problemas em termos de recursos financeiros e sociais,

¹ <<http://www.ibge.gov.br/home/estatistica/populacao/perfilidoso/perfidosos2000.pdf>>

² <http://censo2010.ibge.gov.br/sinopse/webservice/frm_piramide.php>

sobretudo na área da saúde, pela repercussão nos diversos níveis assistenciais e pela demanda por novos recursos e estruturas. Citem-se aqui as doenças crônico-degenerativas, como a Doença de Alzheimer (DA), que acompanham a maior sobrevivência, com sequelas e complicações (FRIED, 2000; PASCHOAL, 1996). Dessa forma, crescem em importância as pesquisas sobre detecção e gestão de doenças que normalmente afetam pessoas com idade avançada, como as **demências**.

As demências são o resultado de uma desordem neurodegenerativa progressiva e irreversível, que se manifesta de diversas formas. O principal desafio no gerenciamento das demências 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 as demências consiste em retardar a progressão da doença quando da detecção de seus primeiros sinais. Apesar de não haver tratamentos que modificam 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 modificar o desenvolvimento da doença.

Diversos biomarcadores vêm sendo disponibilizados como recursos na detecção das demências (RIVEROL; LÓPEZ, 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). A linguagem é uma das fontes mais eficazes de informações sobre as funções cognitivas. Mudanças na linguagem são frequentemente observadas em pacientes com demência, sendo normalmente as primeiras a serem notadas por eles e por seus familiares.

Algumas formas de demência são caracterizadas por profundo declínio das habilidades de linguagem, deixando outras habilidades cognitivas praticamente intactas no curso da doença. Tais demências são, geralmente, denominadas Afasias Progressivas Primárias (APP), dentre as quais identificam-se três subtipos principais (FRASER *et al.*, 2012; GORNO-TEMPINI *et al.*, 2011). A Afasia Progressiva Primária Não-fluente ou Agramatical (APP-g) caracteriza-se pela fala esforçada, hesitante e disfluente, com dificuldades de resgate verbal. A Afasia Progressiva Primária Semântica (APP-s) cursa com perda de conhecimento semântico, não só com déficits de acesso ao léxico, mas também com deterioração do conhecimento dos conceitos subjacentes às palavras.

Tanto a APP-g quanto a APP-s são consideradas subtipos de Demência Frontotemporal (DFT), uma desordem neurodegenerativa distinta da DA. Uma terceira forma de APP, chamada Afasia Progressiva Primária Logopênica (APPL), é, em vez disso, relacionada à DA na sua patologia subjacente (ROHRER; ROSSOR; WARREN, 2012), e inclui sintomas de linguagem característicos, como erros fonológicos, repetição pobre de palavras (mas não de frases ou sentenças), e uma redução global da taxa de produção do discurso (HENRY; GORNO-TEMPINI, 2010). Esses sintomas estão presentes em uma porção significativa de pacientes com DA,

especialmente nos casos de início precoce (LAM *et al.*, 2013). Além disso, muitos pacientes com DA apresentam outros problemas de linguagem, incluindo degradação de memória semântica semelhante à observada em APP-s, anomia³, e maior nível de déficits no discurso que impedem a conversação, como mudanças frequentes de assunto (TALER; PHILLIPS, 2008).

Embora a primeira descrição feita por Alois Alzheimer tenha sido a de um quadro com predomínio de alterações de linguagem, até algumas décadas aceitou-se que a DA era uma demência que comprometia predominantemente a memória. Entretanto, a constatação da heterogeneidade de manifestação dos quadros mudou a visão da doença. Atualmente, aceita-se que a DA acomete outras habilidades cognitivas além da memória, e que o comprometimento da linguagem ocorre já ao início da doença, principalmente nos casos de início precoce.

A variabilidade no comprometimento da linguagem é alta na população de pacientes com DA, pois os sintomas de linguagem são devidos à degeneração das redes cerebrais no córtex. A DA é definida pela diminuição da memória episódica, predominantemente associada a danos no hipocampo e no lobo temporal mesial. Essa forma característica de dano cerebral está presente na maioria dos pacientes com DA.

Outro quadro de interesse relacionado às demências é a síndrome conhecida como *Comprometimento Cognitivo Leve (CCL)*, que é definida como um declínio cognitivo maior do que o esperado para indivíduos de mesma idade e escolaridade, mas sem grande interferência nas atividades do dia a dia, e que é conhecido por ser o estágio inicial clinicamente definido da DA. Embora o CCL não seja tão conhecido do público em geral, pessoas com um tipo especial desta síndrome, o CCL amnésico (CCL-A), evoluem para a DA a uma taxa de 15% por ano, versus 1% a 2% da população em geral.

A linguagem é uma fonte importante de informações que podem revelar a região e a intensidade do dano, o que faz das análises linguísticas aliadas importantes na busca por diagnósticos precoces e tratamentos mais oportunos e efetivos (JARROLD *et al.*, 2010). Enquanto uma análise *qualitativa* da competência linguística de um paciente pode informar o tipo da demência que o afeta, uma análise *quantitativa* revela a gravidade do dano cortical existente. Análises qualitativas podem ser feitas manualmente com esforço razoável, mas análises quantitativas manuais são extremamente custosas e demoradas, o que limita sua aplicabilidade prática.

Segundo Fraser *et al.* (2012), o declínio nas habilidades linguísticas pode ser de difícil quantificação, por meio de testes padronizados, durante os estágios iniciais de desordens neurodegenerativas. Até recentemente, a maior parte da investigação sistemática da produção oral de pacientes com APP focava-se na produção de palavras isoladas (nomeação, leitura, repetição); entretanto, tem se tornado disponível uma pequena literatura concentrada em examinar a produção discursiva conexa (FRASER *et al.*, 2012).

³ Dificuldade em encontrar palavras. O paciente não consegue se lembrar da palavra que deseja utilizar, empregando pausas preenchidas e pedindo auxílio ao interlocutor para manter a comunicação enquanto tenta se lembrar.

O discurso vem sendo reconhecido como um componente discriminativo e essencial na interpretação de avaliações de linguagem (TOGHER, 2001). Tem-se pesquisado uma rica variedade de medidas e tipos de discurso, e reconhecido que este é uma forma natural de comunicação que pode fornecer informações importantes sobre macro e microestruturas linguísticas (ANDREETTA; CANTAGALLO; MARINI, 2012). Além disso, o discurso é capaz de informar como estão se integrando as habilidades linguísticas e cognitivas do paciente, como as de seleção, organização e planejamento (WILLS; CAPILOUTO; WRIGHT, 2012; CANNIZZARO; COELHO, 2012).

No contexto clínico, para avaliações de linguagem com vistas ao diagnóstico de lesões cerebrais, são considerados vários tipos de discurso: conversação, procedimento, narração, relato, entre outros. O discurso narrativo⁴ induzido por figuras é útil para pesquisas, pois elicia a fala de forma padronizada e permite a comparação entre indivíduos e grupos (COOPER, 1990).

Segundo Chand *et al.* (2010), testes padronizados de linguagem, como o de fluência verbal, frequentemente são capazes de rastrear déficits cognitivos e contribuir para o monitoramento da evolução do quadro, ainda em fase precoce; entretanto, a avaliação provida por eles das habilidades expressivas da linguagem é limitada. Esses autores asseveram que tais medidas nem sempre oferecem informações de diagnóstico e prognóstico úteis; isso porque elas apenas avaliam um aspecto restrito do comportamento linguístico, não sendo capazes de levantar informações sobre a linguagem tal como é usada para comunicação nem sobre a complexa interação entre conhecimentos linguísticos e semânticos necessária para tanto. Os autores concluem apontando a análise de discurso narrativo como método promissor na detecção precoce de mudanças no processamento de conhecimento semântico.

Até o presente momento, o principal fator limitante na análise quantitativa desse tipo de discurso é sua exigência de esforços: o processo sistemático de transcrição e análise de discurso é bastante laborioso, dificultando sua adoção em larga escala. Ainda assim, tem havido progresso, e a comunidade médica está começando a compreender as características da produção discursiva em algumas variantes de demências (FRASER *et al.*, 2012). De qualquer maneira, essas avaliações se beneficiariam grandemente da rapidez e sistematicidade das análises computadorizadas.

Recentemente, Fraser *et al.* (2012) realizaram um estudo comparativo dessa natureza entre três grupos de pessoas: pacientes com APP-s, pacientes com APP-g, e controles saudáveis. Ferramentas de Processamento de Línguas Naturais (PLN) foram utilizadas para extrair características léxicas e sintáticas de fragmentos de fala narrativa transcrita produzidos por esses grupos. Com base em métodos de Aprendizado de Máquina (AM), que são capazes de construir *classificadores* que identificam o grupo a que um indivíduo pertence, foi possível identificar corretamente pacientes com APP *versus* controles saudáveis com 100% de precisão e classificar

⁴ A maior parte dos estudos assume que o discurso descritivo é um componente do narrativo, considerado o gênero primordial.

APP-g vs APP-s com 80% de precisão.

Peintner *et al.* (2008) realizaram um trabalho semelhante, em que empregaram classificadores de vários paradigmas na separação de quatro grupos: controles saudáveis e pacientes com três subtipos de demência frontotemporal (DFT comportamental, APP-g, e APP-s). Para tanto, utilizaram características fonológicas extraídas de gravações em áudio dos pacientes, bem como características lexicais e semânticas extraídas da gravação transcrita. Ao final, classificadores treinados com essas características apresentaram, em geral, desempenho bastante acima do *baseline* estabelecido (dado pela classe majoritária).

O estudo de Jarrold *et al.* (2010) tratou a identificação de três distúrbios cerebrais: DA, CCL, e depressão clínica. Utilizando métricas textuais semelhantes às de Peintner *et al.* (2008) (características morfosintáticas, características extraídas do LIWC e Densidade de Ideias (CHAND *et al.*, 2010)), os classificadores conseguiram separar pacientes pré-DA de controles saudáveis com 73% de precisão e separar pacientes com alto comprometimento cognitivo dos controles com 83% de precisão.

O trabalho de Thomas *et al.* (2005) analisou transcrições de fala espontânea de pacientes com DA, e empregou técnicas de análise automatizada em nível lexical para detectar e quantificar a demência dos pacientes. Foram testados diversos cenários de classificação automática, com precisão para separação entre sujeitos com grau alto e baixo de demência em torno de 70%.

Os trabalhos de Roark *et al.* (2007), Roark, Mitchell e Hollingshead (2007) e Roark *et al.* (2011) analisaram gravações em áudio de exames neuropsicológicos feitos em sujeitos saudáveis e em sujeitos com CCL. Os autores empregaram métricas fonológicas e de análise de complexidade em nível sintático, extraídas automaticamente, e testaram a presença de diferença estatisticamente significativa entre os grupos com base nessas medidas, concluindo que seu uso combinado é bastante eficiente na detecção do CCL.

O estudo de Bryant *et al.* (2013) analisou uma medida denominada Densidade de Ideias (descrita na seção 3.2.3.1) no discurso de sujeitos normais e de sujeitos com afasia resultante de lesão vascular. Foram empregadas ferramentas de análise automatizada para extrair essa densidade, além de outras medidas, de transcrições de entrevistas com os participantes. No trabalho, algumas medidas, entre elas a densidade de ideias, não apenas se mostraram eficientes em separar sujeitos saudáveis de afásicos, como também apresentaram correlação com a gravidade da afasia.

Apesar de obterem sucesso, todos esses trabalhos se restringem à língua inglesa. Para o português do Brasil, no nosso conhecimento, nenhuma análise dessa natureza existe. Além disso, alguns trabalhos apresentam limitações, como por exemplo não relatarem quais características foram decisivas na classificação dos sujeitos, como apontado por Fraser *et al.* (2012). É essa ausência de ferramentas de análise automatizada de discurso narrativo para a língua portuguesa voltadas à identificação automática de demências que motiva este trabalho.

O objetivo desta pesquisa de mestrado é criar uma ferramenta de auxílio à tomada de decisão, intitulada Coh-Metrix-Dementia, capaz de instrumentalizar pesquisadores que lidam com o diagnóstico da DA e CCL. O desenvolvimento do Coh-Metrix-Dementia possui dois objetivos principais:

1. Contribuir para a detecção precoce automática de alterações linguístico-cognitivas, bem como para a identificação de características linguísticas distintivas no diagnóstico da DA e do CCL amnésico.
2. Fornecer à equipe de saúde a classe a que um indivíduo pertence (sadio, DA, ou CCL), com base em algoritmos de classificação da área de aprendizado de máquina, treinados com uso de atributos da área de PLN, para auxílio ao diagnóstico final.

Para alcançar tais objetivos, devemos responder às seguintes questões de pesquisa:

1. Existem características, que possam ser obtidas a partir de ferramentas de PLN disponíveis para o português, capazes de (i) distinguir indivíduos saudáveis de indivíduos com DA e CCL, e também (ii) distinguir indivíduos saudáveis dos com CCL, a fim de detectar precocemente o início da deterioração, com precisão aceitável para dar suporte à decisão médica? Se sim, quais são elas?
2. Qual a contribuição de métricas de complexidade sintática, densidade de ideias e coerência textual, via semântica latente, aqui propostas para serem adaptadas ao português, para as duas classificações propostas?
3. Quais, dentre os métodos de classificação considerados neste trabalho, desempenham melhor a tarefa de prever a classe de um indivíduo (sadio, DA, ou CCL) com base nas características linguísticas de seu texto?

Os capítulos que se seguem desenvolvem melhor as questões e objetivos apresentados acima, e organizam-se da seguinte maneira: no Capítulo 2, são apresentadas informações sobre o envelhecimento, tanto o natural quanto o patológico, na DA e no CCL. No Capítulo 3, são apresentadas várias medidas automatizadas, principalmente para a língua inglesa, usadas na avaliação de condições clínicas de pacientes a partir de textos transcritos ou escritos; é feita ainda uma revisão da literatura de trabalhos que automatizaram a avaliação de doenças neurológicas como a DA, o CCL, e as APPs. No Capítulo 4, são apresentados trabalhos que realizaram análises semelhantes às realizadas neste trabalho de mestrado. No Capítulo 5, são apresentadas as etapas realizadas para atingir os objetivos do trabalho; é apresentada a arquitetura geral do sistema, bem como a metodologia empregada na implementação das métricas. No Capítulo 6, são apresentados resultados de avaliações intrínsecas e extrínsecas realizadas sobre o ambiente Coh-Metrix-Dementia. No Capítulo 7, são apresentadas as conclusões do trabalho, e as sugestões para trabalhos futuros.

ENVELHECIMENTO E LINGUAGEM

No processo de envelhecimento, fatores sociais, biológicos, cognitivos, e afetivos respondem pelas modificações da linguagem e da cognição. Do ponto de vista neurobiológico, ocorrem manifestações anatômicas (diminuição do volume do cérebro), neurofisiológicas (diminuição do número e dimensões dos neurônios, e diminuição da eficiência do contato sináptico), e neuroquímicas (diminuição da concentração de neurotransmissores, especialmente dopamina) (SKA; JOANETTE, 2006). Frequentemente são relatadas alterações de atenção, memória, e habilidades visual-espaciais, ao lado de relativa preservação da linguagem.

Nesta Seção, são apresentados os fundamentos sobre cognição e linguagem de indivíduos normais e em condições patológicas. A revisão abaixo exposta foi organizada segundo a relevância temática, à qual se subordina a apresentação cronológica. A base de dados PubMed¹ foi a principal fonte de consulta para obtenção dos textos.

2.1 O envelhecimento normal

Um dos principais problemas no envelhecimento é a definição de “normal.” Do ponto de vista cognitivo, podem ser incluídos nessa categoria indivíduos com queixas subjetivas de memória, ou aqueles que, embora funcionais, apresentam leves alterações no desempenho de tarefas complexas do cotidiano. Há quem defenda a ideia de que um requisito para a inclusão no grupo “normal” é a estabilidade de desempenho cognitivo (SALTHOUSE, 1991), enquanto outros defendem a comparação com idosos da mesma faixa etária, e outros ainda a comparação com a população jovem (CROOK *et al.*, 1986).

Não existe um padrão que caracterize o envelhecimento: há idosos que não apresentam perdas cognitivas e de linguagem, há os que se comportam de modo similar aos jovens, e há outros que sofrem declínios em suas habilidades. A heterogeneidade se aplica ainda ao ritmo

¹ <<http://www.ncbi.nlm.nih.gov/pubmed>>

de perda e às habilidades afetadas. A diversidade de expressão do envelhecimento pode ser explicada por vários fatores, como estilo de vida, nível educacional, saúde, e personalidade, que modulam a plasticidade e a capacidade de compensações (SKA; JOANETTE, 2006).

As descrições da linguagem de idosos ampliaram-se consideravelmente nas últimas décadas. Porém, ainda há aspectos a serem investigados, entre eles o discurso. Dado que a heterogeneidade é característica do envelhecimento, os resultados das pesquisas nem sempre são consensuais. Quando comparada ao declínio de outras habilidades cognitivas, como a memória episódica e habilidades visual-espaciais, a linguagem mostra-se relativamente preservada e resistente ao processo de envelhecimento (PARK; BISCHOF, 2013). Ainda assim, a linguagem também é alvo de queixas frequentes de idosos, entre elas a dificuldade para o resgate de palavras e a compreensão em ambientes ruidosos (GARCIA; MANSUR, 2006).

Estilos prolixo ou simplificado podem ocorrer como efeito de tarefas. Nas tarefas que requerem o relato de histórias recentemente ouvidas, os idosos tendem a compor narrativas em estilo simplificado. Esse comportamento contrasta com a produção de histórias livres, nas quais os idosos tendem a ser mais prolixos e menos eficientes, com maior número de informações irrelevantes e menor número de marcadores de coesão.

Na linguagem espontânea, os idosos tendem a ser prolixos ou sucintos, e produzem erros mais frequentes (quando comparados aos jovens), embora sejam capazes de monitorá-los e corrigi-los. É o que se constata em estudos realizados no exterior (WILLIAMS *et al.*, 2010) e no país (MANSUR *et al.*, 2005). Numa investigação de caráter transversal longitudinal (ANDRADE; MARTINS, 2010), foram examinados 128 idosos brasileiros, incluindo nonagenários. Os autores constataram uma “tendência a decréscimo na taxa de fala e aumento da taxa de rupturas, ao longo das décadas avaliadas,” sem significância estatística, o que revela possibilidades compensatórias desse grupo de indivíduos saudáveis.

Recepção e compreensão

A admissão de dificuldades para recepção e compreensão da linguagem parece constituir ponto de acordo entre os autores (PINHEIRO; DESGUALDO, 2004; WINGFIELD; GROSSMAN, 2006; PICHORA-FULLER; LEVITT, 2012). Os próprios idosos frequentemente se queixam de dificuldades para entender a linguagem oral, principalmente em ambiente ruidoso e em situações em que a comunicação ocorre em grupos. Citam-se, sob os rótulos recepção e compreensão, inabilidades para o processamento da linguagem em níveis distintos, dos fonemas ao discurso.

Sabe-se que os idosos apresentam alterações periféricas e centrais de processamento da informação. Frequentemente, idosos queixam-se de dificuldades de compreensão em ambientes ruidosos ou com reverberação. Essas dificuldades coexistem com outras, de natureza periférica ou neurosensorial. Porém, em muitos casos, os déficits no processamento da fala são

desproporcionais à perda auditiva para tons puros (PINHEIRO; DESGUALDO, 2004).

A queixa dos idosos para *entender* sobrepuja a de dificuldades para *ouvir*. Além de fatores auditivos centrais, respondem por essas dificuldades fatores de outras naturezas: aspectos linguísticos, tais como a complexidade sintático-semântica do material a ser tratado; mecanismos relacionados ao tratamento do significado – como realização de inferências; fatores extralinguísticos, relacionados ao caráter temporal do processamento; e fatores de natureza cognitiva, como demanda atencional, de funções executivas, e de memória operacional e de longa duração.

Estudos recentes mostram que esses fatores interagem nas complexas atividades da vida cotidiana, de tal forma que mecanismos auditivos e cognitivos contribuem para a eficiência do processo de compreensão (ANDERSON *et al.*, 2013). O interessante é que o peso de participação desses sistemas difere, e que o status de audição periférica contribui menos do que experiências de vida (engajamento intelectual e treino musical), status cognitivo, e habilidades de processamento auditivo central, esses dois últimos respondendo pelo maior peso no complexo de interações requerido para ouvir em ambiente ruidoso.

O método de estudo com neuro-imagem funcional tem sido importante para entender os mecanismos de compreensão auditiva (SAUR *et al.*, 2010). Esses estudos têm mostrado que os modelos de processamento da linguagem falada são interessantes, porém insuficientes para que se entenda a audição de idosos, especialmente em determinadas condições de maior demanda, como em ambiente ruidoso.

Produção da linguagem oral nos idosos sadios

Em contraste com o conhecimento acumulado sobre dificuldades de percepção e de processamento no idoso, pouco se sabe sobre o impacto de mudanças da idade na *produção* da linguagem.

a. Nível fonético-fonológico

Os estudos a respeito da produção de linguagem não constatarem, no idoso, alterações no nível fonológico. No entanto, no que tange às características fonéticas, apontam-se imprecisões associadas a outras características, como respiração audível, tremor na voz, e alterações no timbre vocal e no ritmo de emissão, que conferem à produção oral do idoso características próximas às da fala de pacientes com disartria². A percepção de diferenças na voz (presbifonia) parece ser a mais evidente (KENDALL, 2007). Com a idade, a voz das idosas tende a se tornar mais grave, enquanto a dos idosos de sexo masculino tende a se agudizar; esse dado leva à indiferenciação da identidade sexual pela voz, o que traz importantes consequências para a interação comunicativa.

² Alteração da fala que resulta do comprometimento em suas bases: respiratória, fonatória, articulatória, ressoadora, e prosódica. A precisão da produção da linguagem oral depende de habilidades motoras relacionadas à força, direção, e duração do movimento na articulação da fala.

São citadas ainda hesitações, repetições de sons, fenômenos de disfluência (ANDRADE; MARTINS, 2010), e uso frequente de preenchedores (“uh”, “aham”), quando comparados com jovens (HORTON; SPIELER, 2007). A respeito das disfluências do idoso, Manning e Shirkey (1981) conceberam um modelo que distingue disfluências *motoras* e *formativas*. As primeiras compreenderiam aquelas rupturas intravocábulos (inclusas aí as repetições) e as sílabas acompanhadas de aumento de tónus global da emissão, enquanto as segundas abrangeriam as pausas intervocábulos e as repetições de segmentos. Para os autores, a repetição de sílabas é considerada disfluência do tipo motor, e está presente na produção oral de indivíduos disfluentes jovens e adultos. Já a repetição de palavras é classificada entre as disfluências formativas. No modelo, a previsão é de que ocorra diminuição de disfluências motoras e aumento das disfluências formativas (pausas – preenchidas ou não–, repetições de palavras e de segmentos) à medida que o indivíduo envelhece. Para eles, o aumento de disfluências formativas relacionadas à idade evidenciaria hesitações devidas a falhas psicolinguísticas (de acesso lexical, de formulação sintática, dentre outras).

b. Nível lexical

Dificuldades no nível lexical são queixas frequentes entre os idosos, e objeto de constante investigação. A tendência é que se aceite uma progressiva deterioração da informação semântica, desencadeada a partir da perda de atributos periféricos ao estímulo (processamento *bottom-up*). Esses atributos periféricos referem-se ao tratamento semântico circunstancial. Conforme observam Caramelli, Mansur e Nitrini (1998), no início do processo, a dificuldade de acesso à informação ocorre em paralelo à preservação de conhecimentos de traços essenciais dos itens.

Estudos mostram que os idosos são mais lentos do que os jovens quando devem decidir se um segmento é ou não uma palavra, porém melhoram seu desempenho, alcançando o mesmo tempo de reação, quando a tarefa é realizada com pré-ativações de significado. Para os autores, esses achados podem indicar que os idosos necessitam de maior tempo para realizar decisões lexicais do que os jovens sob condições que exigem maior esforço, porém mantém os mecanismos cognitivos que possibilitam as facilitações (como é o caso da pré-ativação).

Os estudos mais recentes com testes de nomeação apontam na direção de dificuldades maiores entre os idosos (MACKAY *et al.*, 2002; ALBERT *et al.*, 2009). Uma revisão de Verhaegen e Poncelet (2013) mostra que, já aos 50 anos, há declínio nessa habilidade. Por outro lado, diferenças entre verbos e substantivos têm sido menos exploradas entre idosos saudáveis, e MacKay *et al.* (2002) não encontraram diferenças na nomeação dessas classes de palavras.

Aceita-se igualmente que não existem alterações quanto a atividades de processamento passivo como designar itens a partir de estímulos auditivos (“mostre-me o sofá”); no entanto, ocorrem dificuldades quando a tarefa recruta o léxico de modo ativo, e em situações de maior demanda, com estímulos competitivos, como é o caso de determinadas situações de resgate ativo de informações específicas durante uma conversação, ou quando a conversação ocorre em

ambientes ruidosos que competem em recrutamento atencional. Essas dificuldades são chamadas *fenômenos de ponta de língua* (FACAL-MAYO *et al.*, 2006).

O fenômeno de ponta de língua está associado a falhas em recuperar palavras, especialmente nomes próprios de pessoas e lugares (FACAL-MAYO *et al.*, 2006). Esse fenômeno pode ocorrer em indivíduos jovens, mas é muito mais frequente entre os idosos e na fala espontânea. Na vigência do fenômeno de ponta de língua, os indivíduos são capazes de recordar várias características do item alvo, como o número de sílabas e o significado, porém falham no acesso à forma fonológica. Decorrido algum tempo, frequentemente podem recordar espontaneamente o item (FACAL-MAYO *et al.*, 2006).

A fluência verbal é outro teste frequentemente utilizado para estudar habilidades léxico-semânticas no envelhecimento normal e patológico. O efeito da idade nas provas de fluência verbal semântica na população brasileira sadia do ponto de vista cognitivo foi analisado por Brucki *et al.* (1997), Brucki e Rocha (2004), Steiner *et al.* (2008), Fischman *et al.* (2009), Yassuda *et al.* (2009), Silva *et al.* (2011), Amaral-Carvalho e Caramelli (2012), e Soares *et al.* (2012). Os autores constataram que tanto a idade quanto a escolaridade influenciaram os resultados. Os estudos mais recentes têm encontrado que o efeito da idade é mais proeminente nas provas de fluência verbal pelo critério semântico, enquanto o efeito da escolaridade se sobressai nas provas em que a geração de itens é determinada pelo critério fonêmico.

c. Nível sintático

A respeito de possível declínio de desempenho no nível sintático, Kemper, Herman e Lian (2003) estudaram a produção de sentenças de diferentes níveis de complexidade em grupos de jovens e idosos. Foram definidos sete níveis de complexidade, de acordo com o número e o tipo de construções (coordenadas ou subordinadas), termos encaixados, e complementos. O nível 7, por exemplo, incluiu sentenças com múltiplas formas de encaixes e subordinação. As respostas dos idosos foram similares às dos jovens em relação ao emprego de verbos transitivos e intransitivos. Por outro lado, os jovens produziram sentenças mais complexas do que os adultos idosos. Estes últimos também cometeram maior número de erros, e foram mais lentos nas respostas.

O estudo das freiras (KEMPER *et al.*, 2001) também mostra declínio da capacidade sintática, mais especificamente dificuldades em lidar com sentenças complexas no processo de envelhecimento. Nele os autores utilizaram medidas computadorizadas, como a densidade de ideias, para caracterizar os efeitos da idade. A mesma autora, em outra investigação (KEMPER; THOMPSON; MARQUIS, 2001), dirigiu-se à produção de sentenças em situação controlada, com e sem uma tarefa motora competitiva. Os autores observaram as fases de planejamento e produção em relação à extensão dos sintagmas nominais, assim como o tipo de verbo. Concluíram que o planejamento da sentença era mais custoso do que a produção, e que o custo do planejamento aumentou quando os participantes deveriam incorporar um sintagma nominal

longo na sua sentença.

A ideia de que habilidades sintáticas ficam estáveis no envelhecimento não é consensual. [Nippold, Cramond e Hayward-Mayhew \(2013\)](#) estudaram o discurso de idosos em comparação com outras faixas etárias, nas situações de conversação e em situação de conflito, e concluíram que não houve diferença na complexidade sintática exibida nas respostas dos grupos. A diversidade de resultados parece sinalizar a necessidade de estudos adicionais sobre o assunto. Além disso, não foram encontrados estudos brasileiros sobre o tema.

d. Nível discursivo

Na descrição do discurso de idosos saudáveis, é frequente que sejam identificados inúmeros segmentos considerados fora do tópico, e esses eventos são, em geral, associados a dificuldades cognitivas do tipo atencionais. [Wills, Capilouto e Wright \(2012\)](#), que estudaram idosos sadios com idades entre 40 e 80 anos, concluíram que a idade não foi fator determinante na ocorrência de eventos fora do tópico, embora tenham detectado déficits dos sujeitos mais idosos nos testes de atenção.

[Marini et al. \(2005\)](#) analisaram narrativas de 69 adultos dos pontos de vista micro e macrolinguístico. Foram encontradas diferenças atribuídas à idade em relação a parafasias semânticas³, paragramatismos⁴, complexidade sintática, grau de coerência global e local (como erros do tipo referência ambígua), e no nível de informatividade veiculado nas histórias. Para os autores, o abrupto declínio no grupo mais idoso sinaliza perda de habilidades relacionada ao envelhecimento.

Os efeitos de idade no relato de histórias foi notado no estudo de [Saling, Laroo e Saling \(2012\)](#). Os autores constataram que, ao contrário dos jovens, que tendiam a tornar as histórias cada vez mais sintéticas quando a enunciação ocorria em intervalos repetidos, os idosos mantinham o estilo menos conciso e prolixo nos vários relatos.

[Wright et al. \(2014\)](#) estudaram o efeito da idade e de processos cognitivos na coerência global, medida em diferentes tarefas de discurso. Os participantes produziram amostras de discurso e realizaram testes cognitivos de atenção e memória. Foram encontradas diferenças de coerência global nos grupos somente no relato de histórias. A influência de medidas cognitivas na manutenção da coerência global foi notada nos diferentes grupos etários, e positiva no caso dos mais idosos para o relato de histórias e discursos de procedimento.

[Fergadiotis, Wright e Capilouto \(2011\)](#) estudaram aspectos microlinguísticos, examinando o efeito do tipo de discurso na diversidade lexical. Os autores testaram quatro tipos de discurso: procedimento, eventos, contação de histórias, e relatos. Os textos foram produzidos por 86 adultos e idosos cognitivamente sadios. Os idosos estavam na faixa de 70 a 89 anos. As

³ Trocas lexicais que ocorrem em um mesmo campo semântico. Por exemplo, a troca de “mesa” por “cadeira.”

⁴ Alterações sintáticas que se caracterizam por violações de regras gramaticais da língua. Por exemplo, a atribuição de flexão de gênero ao advérbio: “simplo,” ao invés de “simples.”

amostras de discurso foram analisadas para estimar a diversidade lexical. Os resultados indicaram que a diversidade lexical é um dos índices influenciados pelo tipo de discurso e pela idade.

A pesquisa realizada por Toledo (2011) examinou aspectos micro e macrolinguísticos de 200 indivíduos brasileiros, sadios, com idade mínima de 30 anos e escolaridade mínima de 3 anos. Duas figuras estimularam a produção do discurso, cada uma retratando uma cena diferente: uma figura simples e uma complexa. Elegeram-se para análise os seguintes parâmetros: *extensão do discurso* (número total de palavras e expressões fáticas); *dificuldades no resgate lexical* (pausas maiores do que dois segundos, pausas preenchidas, erros semânticos e fonológicos, e repetições imediatas de palavras e sentenças); *conteúdo da descrição* (emissões irrelevantes e vagas, dificuldades de interpretação visual, e quantidade de informação), *habilidades sintáticas* (extensão e complexidade das sentenças); e o *tempo de descrição*. A autora verificou significativa influência da escolaridade e da idade nos discursos. O efeito da idade foi verificado no número de palavras, na repetição imediata de palavras, no emprego de termos indefinidos, na complexidade frasal, e no aumento do tempo de produção do discurso.

Ferguson *et al.* (2013) também avaliaram a influência da idade e escolaridade, porém na linguagem escrita. Os autores analisaram a informatividade do discurso de uma larga coorte de 19.512 participantes, e encontraram pequena mas significativa diminuição da densidade de proposições na faixa etária acima de 78 anos, e concluíram que essa medida mantém-se relativamente estável no envelhecimento.

2.2 A linguagem nos indivíduos acometidos por Doença de Alzheimer

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

As alterações são detectadas por questionários, anamnese, e exames objetivos. Os comprometimentos cognitivos ou comportamentais afetam no mínimo dois dos seguintes domínios:

- **Memória:** capacidade para adquirir ou evocar informações recentes. Os sintomas de quadro demencial incluem repetição das mesmas perguntas ou assuntos, e esquecimento de eventos e compromissos ou do lugar onde guardou seus pertences.
- **Funções executivas:** raciocínio, realização de tarefas complexas, e julgamento. Os sintomas de quadro demencial incluem compreensão pobre de situações de risco, e redução

da capacidade de cuidar das finanças, tomar decisões, e planejar atividades complexas ou sequenciais.

- **Habilidades visual-espaciais:** capacidade de reconhecer faces ou objetos comuns, encontrar objetos no campo visual, manusear utensílios, e vestir-se. No quadro demencial, ocorrem alterações dessas habilidades, não explicáveis por deficiência visual ou motora.
- **Linguagem:** expressão, compreensão, leitura, e escrita. Os sintomas do quadro demencial incluem dificuldade para encontrar e/ou compreender palavras, e erros ao falar e escrever, com trocas de palavras ou fonemas, não explicáveis por déficit sensorial ou motor.
- **Personalidade ou comportamento:** os sintomas do quadro demencial incluem alterações do humor (labilidade, flutuações incharacterísticas), agitação, apatia, desinteresse, isolamento social, perda de empatia, desinibição, e comportamentos obsessivos, compulsivos, ou socialmente inaceitáveis.

A DA é classificada como *provável* (quando preenche critérios de modo consistente), *possível* (quando há fatores incertos, como história de doença vascular), e *definida* (comprovada por exame anatomo-patológico, realizado pós-mortem). Os déficits cognitivos iniciais e mais proeminentes podem se apresentar na forma amnésica, quando predomina o comprometimento de memória, associado ao comprometimento em outro domínio cognitivo, ou não-amnésica (devendo haver outro domínio afetado), com alterações em linguagem (lembranças de palavras), aspectos visual-espaciais e funções executivas (alteração do raciocínio, julgamento, e solução de problemas), e agnosia⁵.

Assim, embora a perda de memória seja o traço cognitivo mais frequente na DA, as alterações de linguagem podem aparecer em fase precoce da doença. De maneira geral, parece existir acordo a respeito da diversidade de manifestações na doença e, embora as alterações de memória sejam prevalentes, há subgrupos de pacientes que apresentam significativos comprometimentos de linguagem ou visual-espaciais (MENDEZ *et al.*, 2012). Isso acontece principalmente nas manifestações precoces, abaixo de 65 anos de idade.

Na apresentação clássica da doença, em sua forma tardia, o predomínio é de alterações de memória (tanto de curta como de longa duração), ao lado de alterações de aspectos léxico-semânticos, com relativa preservação dos fonológico-sintáticos até os estágios mais avançados (TRAYKOV *et al.*, 2007).

Muito do que se sabe sobre a linguagem dos pacientes com DA provém de estudos que utilizam baterias de testes cognitivos. Mais recentemente, têm surgido estudos de linguagem em situações funcionais. Apresenta-se, a seguir, uma breve revisão da linguagem na DA, enfatizando a produção da linguagem, interesse da investigação deste trabalho.

⁵ Dificuldade de reconhecer objetos ou faces, e dificuldade de leitura relacionada a aspectos visual-espaciais.

Recepção e compreensão

Para os sujeitos com DA, a possibilidade de compreensão da linguagem oral não está relacionada apenas à complexidade da tarefa. Pesquisadores detectaram vários níveis de comprometimento, inclusive dificuldades em compreensão gramatical e de prosódia afetiva, já ao início da doença (TALER; PHILLIPS, 2008).

Deve-se notar que muitas avaliações de compreensão de linguagem representam problemas para os pacientes com DA, pois envolvem solicitações complexas de realização da resposta, e não da tarefa em si (é o que os autores chamam de dificuldades pós-interpretativas). Em outras palavras, os sujeitos compreenderam a questão, porém sentem dificuldade para organizar a resposta motora, por exemplo. O estudo de Grossman e Rhee (2001) controlou a complexidade na elaboração das respostas, no sentido de minimizar essa variável. Em sua pesquisa, os sujeitos eram convidados a identificar incongruências sintáticas. Os autores levantaram a hipótese de que a dificuldade de compreensão de sentenças gramaticalmente complexas na DA está relacionada à lentidão de processamento. Essa lentidão restringe a compreensão de sentenças cuja apreensão da estrutura exige a detecção da construção “em tempo” e, por outro lado, limita a inibição de interpretações canônicas (por exemplo, em que o sujeito é sempre o primeiro elemento).

As pesquisas sobre processamento de informação frequentemente valorizam o fator complexidade e a quantidade de dados a serem processados, o que compreensivelmente está afetado no caso de déficits de memória que ocorrem na apresentação típica da doença. É o caso do estudo de Creamer e Schmitter-Edgecombe (2010) sobre compreensão de narrativas. Os autores utilizaram um método em que os portadores de DA deveriam ler narrativas com pausas entre cada sentença, nas quais eram solicitados a “pensar em voz alta” sobre o que haviam compreendido. Seu objetivo era verificar não apenas a capacidade de compreender inferências, mas também o efeito de comprometimentos de memória no processamento. Sua conclusão foi que ambos os fatores estão associados às dificuldades, e que a memória (operacional) interfere na habilidade de integrar eventos por meio do uso de inferências e de criar a coerência global base para a compreensão das narrativas.

Produção oral

a. Níveis fonético-fonológico e sintático

Os primeiros estudos sobre linguagem na DA indicavam que esses pacientes não apresentavam dificuldades no aspecto fonético-fonológico da produção até estágios muito avançados da doença, o que os incluía na categoria de fluentes (BAYLES; TOMOEDA; TROSSET, 1992). Recentemente, essa ideia tem sido revista. Em primeiro lugar, tende-se a diferenciar, do ponto de vista cognitivo, os indivíduos com DA cuja apresentação é pré-senil daqueles em que a apresentação ocorre após os 65 anos de idade. Os indivíduos com apresentação pré-senil apresentam déficits proeminentes de linguagem, principalmente aqueles associados à memória operacional,

como os fonético-fonológicos (KALPOUZOS *et al.*, 2005). Além disso, aspectos relacionados à produção da linguagem e a praxias bucofaciais e de fala⁶ tem sido minuciosamente avaliados, como na investigação de Cera *et al.* (2013). Nesse estudo, foram encontradas alterações do tipo apraxia de fala em pacientes com apresentação tardia da DA, o que contraria a visão corrente da literatura.

Peters *et al.* (2009) constataram maior número de erros fonológicos nos sujeitos portadores de DA de apresentação tardia em tarefa de recordação imediata de sequências de palavras referentes a itens com alta e baixa imageabilidade⁷. Os autores discutem a possibilidade de a deterioração do conhecimento semântico interferir na memória verbal de curta duração, já que o efeito de imageabilidade manifestou-se de forma proeminente entre os pacientes com DA. Nesse grupo, houve acentuado decréscimo na memória de curta duração quando foram apresentadas palavras de baixa imageabilidade, enquanto a recordação de palavras de alta imageabilidade manteve-se preservada. Além disso, os pacientes com DA apresentaram erros fonológicos em proporção anormal na situação de baixa imageabilidade. A partir desses achados, os autores concluíram que o conhecimento semântico pode responder pelo comprometimento da memória de curta duração observada na DA.

As descrições detalhadas sobre sintaxe na doença de Alzheimer sempre apontaram que algumas habilidades estão relativamente preservadas ao 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 a 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 seminal estudo 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

⁶ **Praxias:** termo que define a capacidade de planejar movimentos não verbais (praxias bucofaciais), como por exemplo realizar sequências de atos como “pigarrear, estirar a língua, e estalar os lábios,” ou movimentos relacionados à fala (palavras ou sequências de sílabas sem significado).

⁷ **Imageabilidade** diz respeito à possibilidade de representação figurativa de um item. Itens prototípicos de uma categoria semântica são mais facilmente representados em imagens.

indicadores da doença. Um desses indicadores era o que chamaram de “simplificação da sintaxe.”

b. Nível lexical

É consenso entre os pesquisadores que habilidades semânticas constituem o cerne das perdas da linguagem, causadas pelo processo degenerativo da DA. 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 de fluência verbal. Quando convidados a emitir itens relacionados durante um tempo restrito (um minuto), os indivíduos com DA produzem 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 DA (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 doença produzem, na tarefa de fluência verbal, menor número de *switches* (mudanças de critério de evocação de itens em determinado campo) e produzem *clusters* (agrupamentos de itens de determinada categoria) menores quando comparados a 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 associado ao desempenho comprometido 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 de 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 à nomeação de 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.

Vale notar ainda que 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, cabe pontuar que 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.

c. Nível discursivo

O estudo da produção de discurso na DA é recente, escasso, e prevalecem investigações sobre aspectos fonológicos, sintáticos, e semânticos. Na data em que a revisão aqui apresentada foi realizada, na base de dados PubMed, a partir dos descritores “discourse and Alzheimer’s disease”, é possível recuperar 38 artigos, dos quais 21 dizem respeito ao tema pesquisado.

Produzir discursos é uma atividade complexa, com regras em diversos níveis: formais, estruturais, semânticas, e pragmáticas. No discurso dos pacientes com DA, nota-se o impacto de déficits cognitivos já ao 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, e organização). Os portadores de DA tornam-se repetitivos, esquecem o que ouviram ou leram, e perdem o tópico. Ao longo do tempo, o discurso torna-se empobrecido e fragmentado, caracterizado por falta de coerência. Nota-se ainda tangencialidade e perseverações (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 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, 2001) 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 à 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 reduzidas) produzidos pelo indivíduo. O discurso produzido era notavelmente mais simples do que o da população controle, com predomínio de sentenças coordenadas. Entre os erros lexicais, foram proeminentes as dificuldades de acesso lexical, as repetições de palavras, e o uso de termos indefinidos, ao lado de 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, o uso de termos indefinidos, e as revisões.

Ska e Duong (2005) estudaram simultaneamente diferentes níveis de representação nas narrativas de pacientes com DA, por meio de um modelo de construção-integração (KINTSCH, 1988). As narrativas eram produzidas em duas situações: a partir de uma prancha única, e de 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: no nível de superfície, no modelo de situação, e na organização da estrutura narrativa.

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

2.3 A linguagem nos indivíduos acometidos por Comprometimento Cognitivo Leve

O comprometimento cognitivo leve é uma entidade clínica dificilmente definida por suas próprias características. Com frequência, é identificado como uma situação intermediária entre o envelhecimento saudável e a condição demencial: o indivíduo (ou seus acompanhantes) reconhecem diferenças mínimas do ponto de vista cognitivo.

Petersen *et al.* (1999) estabeleceram critérios formais para o diagnóstico de CCL: i) queixa subjetiva de perda de memória; ii) perdas objetivas de habilidades; iii) preservação global de funções cognitivas; iv) atividades de vida diária preservadas; v) o indivíduo não preenche critérios para o diagnóstico de demência. O Consenso de Estocolmo propôs revisão do critério de Petersen (WINBLAD *et al.*, 2004): ii) o indivíduo não é nem normal nem demente; ii) há evidência de deterioração cognitiva, notada em medidas objetivas de declínio ou relatos de queixa feita pelo indivíduo ou informante, em conjunção com os déficits notados por meios objetivos; iii) as atividades de vida cotidiana estão preservadas, e as funções instrumentais complexas estão intactas ou minimamente comprometidas.

O CCL tem sido descrito como condição pré-clínica da DA e, de fato, um número considerável de indivíduos com CCL convertem o quadro para demência. Essa é uma das razões pelas quais tradicionalmente predominaram os estudos sobre memória nessa população. Recentemente, sabe-se que existe evidência de que indivíduos com CCL devido a comprometimentos em múltiplos domínios, incluindo-se aí a linguagem, têm mais risco de desenvolver DA. Por essa razão, é importante compreender a natureza do comprometimento de linguagem.

Aspectos léxico-semânticos

Medidas de processamento semântico, tais como fluência verbal e nomeação, estão incluídas na maioria dos estudos sobre CCL. Essa inclusão é justificada pela utilidade diagnóstica do teste no acompanhamento da DA e de idosos saudáveis. Déficits na fluência verbal são considerados preditivos de desenvolvimento de demência (DA e outros tipos), e estão presentes no CCL. Há maior concordância sobre o comprometimento da fluência verbal eliciada por critérios semânticos, e, recentemente, novas formas de avaliação sensibilizadas têm sido propostas. É o caso do estudo de Steiner (2012), no qual a fluência verbal baseada na geração de verbos de ação foi capaz de discriminar indivíduos sadios e com CCL.

Testes de nomeação, como o *Boston Naming Test* (KAPLAN; GOODGLASS; WEINTRAUB, 2001), também são utilizados para detectar comprometimentos e diagnosticar CCL. Instrumentos mais sensíveis do que os testes tradicionais têm sido cogitados, como o *Graded Naming Test* (GNT) (MCKENNA; WARRINGTON, 1983), que contém estímulos graduados por decréscimo em familiaridade, e os testes de nomeação de nomes próprios.

Os resultados do emprego de testes de fluência e nomeação para auxiliar o diagnóstico de CCL são controversos, e a utilização desses instrumentos carece de investigação adicional (TALER; PHILLIPS, 2008). Estudos isolados com outros testes, como o *Pyramids and Palm Trees* (HOWARD; PATTERSON; COMPANY, 1992), e baterias abrangentes, como o *Montreal Cognitive Assessment* (NASREDDINE *et al.*, 2005), para detectar CCL têm sido desenvolvidos. A descrição da prancha do Roubo dos Biscoitos também foi aplicada a indivíduos com CCL. Os autores Bschor, Kühl e Reischies (2001) obtiveram a distinção entre DA e CCL, mas não entre controles saudáveis e CCL.

Em resumo, os resultados de testes e baterias de linguagem mostram que os indivíduos com CCL frequentemente apresentam alterações semânticas. Assim sendo, é interessante que se desenvolvam instrumentos para detectar os déficits sutis, de modo a ampliar a sensibilidade do prognóstico de conversão para quadros demenciais.

Outros estudos com tarefas e testes não padronizados, como processamento de palavras isoladas, categorização, violação de regras, entre outros, mostraram-se promissores na diferenciação entre controles sadios e CCL. De particular interesse é a menção a outros estudos com verbos, como por exemplo relações semânticas entre verbos (GROSSMAN *et al.*, 2003).

Em relação à produção da linguagem, foco de interesse deste trabalho, 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. Na fala espontânea, sabe-se que os indivíduos com DA apresentam problemas semânticos, ao lado da relativa preservação sintática. Não existem, em nosso conhecimento, estudos com fala espontânea em CCL. 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 e indivíduos sadios.

Os testes de nomeação, geração de definição, e produção de fala espontânea e descrição de figuras mostram que os indivíduos com CCL tem 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 nos CCL.

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 numa perspectiva linguístico-cognitiva, cujo declínio é típico das condições mencionadas – envelhecimento saudável, CCL, e DA. Porém, seu emprego somente será viável se dispusermos de instrumentos que permitam a organização de uma base de dados com número representativo de informantes.

2.4 Considerações finais

O exame do cenário acima exposto nos permite destacar as seguintes lacunas no conhecimento de idosos saudáveis, com alterações cognitivas, e com quadros demenciais, que justificam o presente estudo.

1. Envelhecimento normal

As disfluências presentes na produção oral dos idosos refletem vários fenômenos, ainda não totalmente elucidados. As alterações lexicais presentes no idoso, tais como fenômenos de ponta de língua, necessitam ser entendidas do ponto de vista linguístico-cognitivo (natureza, etiologia, e concomitância de fatores cognitivos). Essas dificuldades ocorrem principalmente em fala espontânea e necessitam de larga base de dados representativa do fenômeno.

Empobrecimento sintático foi detectado em idosos saudáveis, falantes do Inglês, porém essa visão não é consensual. Na literatura, desconhece-se o desempenho de idosos brasileiros nesse aspecto, e a influência de fatores relacionados ao Português Brasileiro é caracterizado por amplas marcas morfológicas e redundância.

Estudos do discurso de idosos saudáveis brasileiros são escassos. Nos estudos com falantes do inglês, os autores constatam a influência de fatores relacionados ao tipo de tarefa proposta (conversação, descrição de prancha única ou em sequência, tarefas sensibilizadas para observação de uso de referenciação). Há necessidade de comprovação e obtenção de dados consistentes sobre essa interação de fatores. Há poucos dados sobre o impacto de escolaridade na produção do discurso.

Finalmente, cabe enfatizar que o estudo de idosos saudáveis é importante para constituir padrões de referência ao diagnóstico de detecção e evolução de alterações cognitivas, como os CCLs e a DA, desafios do mundo moderno. A diversidade e heterogeneidade do processo de envelhecimento exige que seu estudo seja baseado em métodos confiáveis e capazes de abranger amostras amplas e representativas.

2. Envelhecimento patológico: a Doença de Alzheimer

Desconhecemos a exata interação de fatores relacionados ao envelhecimento normal e patológico. Em outras palavras, não se sabe se as manifestações da DA seriam uma expressão exacerbada do envelhecimento normal, ou como a linguagem poderia contribuir para responder a essa questão.

Dado que a DA manifesta-se de forma heterogênea, há necessidade de estudos de linguagem em seus vários aspectos (fenômenos de disfluência e aspectos léxico-semânticos, sintáticos, e discursivos) nos sub-grupos de portadores de DA. Se, por um lado, o estudo sobre idosos saudáveis contribui para o entendimento da DA, por outro lado, o estudo da DA também pode auxiliar a entender o envelhecimento saudável, mostrando aspectos da linguagem

(semânticos, possivelmente) mais sensíveis ao processo demencial.

3. *Comprometimento Cognitivo Leve*

O CCL é uma área cinza entre demência e não-demência. Um dos maiores investimentos dos estudos cognitivos tem se dirigido na busca de traços que diferenciem o idoso saudável daquele que se queixa de alterações ou apresenta desempenho influenciado negativamente por situações complexas.

O rótulo de “comprometimento cognitivo” tem o benefício de reconhecer que existe comprometimento cognitivo, embora não seja possível saber a gravidade ou a causa subjacente. Existem poucos estudos sobre linguagem nesse grupo portador de CCL.

Um aspecto importante é a taxa de conversão de indivíduos com alteração cognitiva para quadros demenciais. Os fatores de risco para essa conversão ainda não estão claramente determinados. Investigações que levem em conta a heterogeneidade, no continuum da fronteira entre normalidade e doença, com métodos sensíveis são necessárias.

Embora tenham decorrido quase 20 anos do emprego de ferramentas computadorizadas para análise de conteúdo semântico da linguagem (SNOWDON *et al.*, 1996), verifica-se que o uso desse tipo de recurso ainda é pouco frequente, mesmo em estudos de língua inglesa. Não dispomos no Português Brasileiro, até o momento, de tais recursos, o que limita o estudo da linguagem nos idosos sadios e em condições patológicas.

ANÁLISE AUTOMATIZADA DE LINGUAGEM

Este Capítulo descreve, na Seção 3.1, várias medidas automatizadas, principalmente para a língua inglesa, usadas na avaliação de condições clínicas de pacientes a partir de textos transcritos ou escritos. Elas estão organizadas em quatro grupos: medidas de diversidade léxical, de complexidade sintática, de densidade semântica, e de coerência textual, via semântica latente.

Na Seção 3.2, é feita uma revisão da literatura de trabalhos que automatizaram a avaliação de doenças neurológicas como a DA, o CCL, e as APPs. Os nove trabalhos aqui analisados estão organizados com relação ao nível da língua com que as avaliações são realizadas; assim, são apresentados trabalhos que tratam: i) da avaliação de riqueza lexical; ii) da complexidade sintática; iii) da densidade de ideias; iv) do uso de métricas derivadas de avaliação de processos psicológicos, com ajuda do software LIWC; e v), de uma análise varrendo todos os níveis linguísticos.

3.1 Ferramentas Automáticas

Nesta Seção, são apresentadas ferramentas já implementadas que dão base ao Coh-Metrix-Dementia. Na Seção 3.1.1, é apresentado o Coh-Metrix e sua adaptação ao português brasileiro, o Coh-Metrix-Port, no qual este trabalho fortemente se baseia; na Seção 3.1.2, é apresentada a ferramenta AIC (Análise de Inteligibilidade de Córpus), desenvolvida durante o projeto PorSimples; na Seção 3.1.3, é descrita a ferramenta LIWC e seu dicionário, que foi traduzido para o português. Estas duas últimas ferramentas foram utilizadas no experimento piloto deste trabalho, descrito na Seção 6.1, e algumas das métricas do Coh-Metrix-Dementia se baseiam conceitualmente em métricas suportadas por elas.

3.1.1 O Coh-Metrix e o Coh-Metrix-Port

O Coh-Metrix¹ (GRAESSER *et al.*, 2004; MCNAMARA; LOUWERSE; GRAESSER, 2002) é uma ferramenta, desenvolvida para a língua inglesa, que extrai de um texto características que influenciam em sua coesão, em sua coerência, e em sua facilidade (ou dificuldade) de leitura (SCARTON; ALUÍSIO, 2010; GRAESSER *et al.*, 2004).

Os autores do Coh-Metrix propõem uma distinção entre coesão e coerência em um texto: a coesão é uma característica do texto, enquanto a coerência é uma característica da representação mental do conteúdo do texto estabelecida pelo leitor (GRAESSER *et al.*, 2004).

Palavras, expressões ou sentenças que guiam o leitor no processo de estabelecimento mental de uma representação consistente do conteúdo do texto são consideradas construções coesivas. Esses dispositivos linguísticos permitem ao leitor delimitar as ideias presentes no texto, conectá-las por meio das relações propostas e inseri-las em contextos de ordem mais alta, como tópicos e sub-tópicos.

A coerência está relacionada à representação mental que o leitor cria no transcorrer da leitura e é grandemente influenciada por sua bagagem cognitiva, ou seja, pelo conhecimento de mundo que ele traz, por suas habilidades de interpretação e raciocínio e pelos construtos coesivos do texto explícito.

Graesser *et al.* (2004) e McNamara *et al.* (1996) mencionam que a relação entre a coesão e a coerência é grandemente influenciada pelo conhecimento de mundo do leitor. Leitores com baixo conhecimento precisam de conectivos e explicações claras e presentes no texto, enquanto leitores com maior cultura se beneficiam de lapsos de coesão, uma vez que conseguem preenchê-los por meio de inferências baseadas em conhecimento prévio. Concluíram também que textos extremamente claros, explícitos, além de serem monótonos não contribuem com o desenvolvimento de habilidades de inferência, que são necessárias para uma melhoria das habilidades de leitura.

A ferramenta Coh-Metrix unifica a saída de diversas outras ferramentas de PLN e pode ser usada em diversos cenários de **análise e classificação** de textos. Os autores do Coh-Metrix coletaram e avaliaram centenas de métricas ao longo do tempo, que medem características do texto relacionadas a palavras, sentenças e à conexão entre sentenças. O Coh-Metrix está alinhado a diversos *frameworks* teóricos que objetivam identificar as representações, estruturas, estratégias e processos que ocorrem em diferentes níveis da linguagem e do discurso. Em particular, tais *frameworks* em geral englobam 5 níveis, descritos abaixo (GRAESSER; MCNAMARA; KULIKOWICH, 2011).

1. Palavras

Segundo Graesser, McNamara e Kulikowich (2011), o conhecimento de vocabulário

¹ <<http://cohmetrix.memphis.edu/cohmetrixpr/index.html>>

possui um impacto substancial no tempo de leitura e compreensão Crianças e adolescentes em idade escolar são expostos a textos cada vez mais complexos ao longo do tempo, e os leitores com representações lexicais de alta qualidade são aqueles que possuem associações ricas entre fonologia, ortografia, morfologia, e estrutura sintática de palavras. Dessa forma, é importante analisar as palavras sob a ótica de múltiplas características que sejam relevantes para o desenvolvimento da leitura e a construção do significado (GRAESSER; MCNAMARA; KULIKOWICH, 2011).

Um exemplo de característica é a **categoria gramatical ou morfossintática** das palavras. Pronomes, por exemplo, são mecanismos importantes de coesão da base textual e do modelo situacional, mas a resolução de seus antecedentes pode ser mais simples ou mais complexa, dependendo de quantos aparecem no texto e do gênero e número de seus antecedentes, impactando a facilidade de leitura. A **frequência de ocorrência** das palavras também influencia na compreensão; se o leitor está familiarizado com as palavras do texto, este flui naturalmente, ao passo que uma única palavra incomum em uma sentença pode comprometer o entendimento dela toda (GRAESSER; MCNAMARA; KULIKOWICH, 2011).

Outros exemplos são **medidas psicológicas** das palavras, como idade de aquisição, concretude, imageabilidade, e familiaridade. Medidas como a concretude informam se o texto trata de objetos e ideias concretos ou de construções abstratas, sendo estas últimas, em geral, mais difíceis de serem compreendidas; além disso, podem ser usadas para detectar um linguajar vago, impreciso, típico de alguns quadros demenciais.

São utilizadas também medidas de **conteúdo semântico**. Nesta última categoria, os substantivos são classificados como humano, animado, concreto, abstrato, entre outras; esse conhecimento pode ser usado para medir, por exemplo, a *polissemia* das palavras, que é seu número de significados básicos. Os verbos são classificados entre *intencionais* (produto da ação consciente e objetivada de um agente animado, como fazer um bolo) ou *causais* (ações que independem da vontade do sujeito, como chover), distinção que gera impacto no modelo situacional (GRAESSER; MCNAMARA; KULIKOWICH, 2011).

2. Sintaxe

Teorias sintáticas atribuem palavras a categorias gramaticais ou morfossintáticas (e.g., nomes, verbos, adjetivos), prescrevem como palavras se agrupam em sintagmas (nominais, verbais, preposicionais), e permitem levantar uma estrutura em árvore sintática para as sentenças. Algumas sentenças são curtas – seguindo, por exemplo, a estrutura sujeito-verbo-objeto –, possuem poucas orações subordinadas, e usam a voz ativa em vez da passiva. Sentenças assim são mais próximas da expressão oral, e tendem a ser mais fáceis de processar (GRAESSER; MCNAMARA; KULIKOWICH, 2011). Outras, em contrapartida, contém sintagmas com grande número de modificadores, colocam muitas palavras antes do verbo principal da oração principal (“enterram” o verbo, o que impõe demanda extra à memória operacional do leitor (GRAESSER

et al., 2006)), e possuem palavras que carregam sentido lógico (“e”, “ou”, “portanto”). Textos como os encontrados na imprensa estão mais próximos deste estilo. Como exemplo dessa estrutura complexa, veja a seguinte pergunta (GRAESSER; MCNAMARA; KULIKOWICH, 2011), extraída do *American Community Survey* de 1999:

At any time during the last 12 months, were you or any member of your household enrolled in or receiving benefits from free or reduced-price meals at school through the Federal School Lunch program or the Federal School Breakfast program?

O Coh-Matrix é capaz ainda de medir a frequência de aparição de voz passiva (que é mais difícil de ser processada que a voz ativa), e a similaridade sintática entre sentenças do texto, uma vez que o paralelismo sintático facilita a compreensão (GRAESSER; MCNAMARA; KULIKOWICH, 2011).

3. Base textual

Graesser, McNamara e Kulikowich (2011) definem a base textual como se referindo às ideias explicitadas no texto: o significado, ao invés da superfície textual. A **coesão referencial** é um aspecto importante à inteligibilidade do texto, e ocorre quando um nome, pronome, ou sintagma nominal se refere a outro constituinte na base textual. Por exemplo, em uma sentença como “A obra *Crime e Castigo*, de Fyodor Dostoyevsky, foi publicada em 1866, quando o escritor tinha 45 anos”, “o escritor” se refere a “Fyodor Dostoyevsky.” Um *gap* de coesão referencial, que ocorre quando uma sentença não se conecta bem às anteriores, apresentando referentes ambíguos ou difíceis de serem resolvidos, pode aumentar o tempo de leitura e prejudicar a compreensão do texto (GRAESSER; MCNAMARA; KULIKOWICH, 2011). O Coh-Matrix analisa diversos tipos de correferência, como **sobreposição de palavras de conteúdo**, **sobreposição de nomes**, **sobreposição de argumentos**, e **sobreposição de radical**.

Um outro fator importante, também analisado pelo Coh-Matrix, é a **diversidade lexical**, que se relaciona à coesão na medida em que um número grande de palavras distintas empregado no texto implica em um grande número de conceitos que devem ser assimilados e integrados ao contexto discursivo (GRAESSER; MCNAMARA; KULIKOWICH, 2011). Uma das medidas mais conhecidas da diversidade lexical é a relação **tipo por token**, que consiste no número de palavras distintas presente em um texto (os tipos) dividido pelo número total de palavras (os *tokens*).

4. Modelo situacional

Ainda seguindo-se as definições de Graesser, McNamara e Kulikowich (2011), o modelo situacional é o conteúdo do assunto sendo tratado, ou o mundo narrativo que o texto está

descrevendo. Em textos narrativos, isso inclui personagens, objetos, cenário espacial, ações, eventos, processos, planos, pensamentos e emoções dos personagens, entre outros detalhes. Zwaan e Radvansky (1998) propuseram cinco dimensões do modelo situacional de textos narrativos: causalidade, intencionalidade, tempo, espaço e protagonistas. Uma quebra na coesão textual ocorre quando há uma ruptura, uma descontinuidade em uma ou mais dessas dimensões do modelo conceitual. Quando uma quebra desse gênero ocorre, põe-se uma demanda inferencial extra sobre o leitor, demanda esta que gera aumento do tempo de compreensão do texto. Analogamente, o uso de conectivos temporais, causais, entre outros, ajuda o leitor a estabelecer uma figura coerente do modelo situacional. Vale lembrar o que já foi exposto: que leitores com maior conhecimento de mundo podem se beneficiar de lacunas coesivas, caso em que a ausência explícita de conectivos pode melhorar a compreensão ao estimular inferências (GRAESSER; MCNAMARA; KULIKOWICH, 2011).

5. Gênero e estrutura retórica

O gênero textual se refere à categoria do texto, que pode ser narração, exposição, persuasão, ou descrição (GRAESSER; MCNAMARA; KULIKOWICH, 2011). Essas categorias podem ser divididas ainda em sub-categorias, formando uma taxonomia, sendo o texto narrativo substancialmente mais fácil de ler, compreender, e relembrar que o texto informacional. Treinar os leitores para distinguir o gênero e as estruturas globais do texto os ajuda a melhorar a compreensão. O Coh-Metrix analisa a extensão à qual o texto pode ser classificado como narrativo ao invés de informativo, por meio de uma medida única, quantitativa, e contínua denominada **narratividade** (GRAESSER; MCNAMARA; KULIKOWICH, 2011).

A Figura 1 mostra um exemplo de tela do Coh-Metrix, com o resultado da análise de um texto de exemplo. As métricas do Coh-Metrix estão apresentadas na Tabela 1. São 108 métricas no total, agrupadas em 11 categorias, além de meta-dados acerca do texto.

Tabela 1 – Métricas do Coh-Metrix 3.0 (extraído de <http://141.225.42.101/CohMetrixHome/documentation_indices.html>).

Title	Title
Genre	Genre
Source	Source
UserCode	UserCode
LSASpace	LSASpace
Date	Date
Time	Time

Continua...

Tabela 1 – Continuação

	Label in Version 3.x	Label in Version 2.x	Description
Descriptive			
1	DESPC	READNP	Paragraph count, number of paragraphs
2	DESSC	READNS	Sentence count, number of sentences
3	DESWC	READNW	Word count, number of words
4	DESPL	READAPL	Paragraph length, number of sentences, mean
5	DESPLd	n/a	Paragraph length, number of sentences, standard deviation
6	DESSL	READASL	Sentence length, number of words, mean
7	DESSLd	n/a	Sentence length, number of words, standard deviation
8	DESWL _{sy}	READASW	Word length, number of syllables, mean
9	DESWL _{syd}	n/a	Word length, number of syllables, standard deviation
10	DESWL _{lt}	n/a	Word length, number of letters, mean
11	DESWL _{ltd}	n/a	Word length, number of letters, standard deviation
Text Easability Principal Component Scores			
12	PCNAR _z	n/a	Text Easability PC Narrativity, z score
13	PCNAR _p	n/a	Text Easability PC Narrativity, percentile
14	PCSYN _z	n/a	Text Easability PC Syntactic simplicity, z score
15	PCSYN _p	n/a	Text Easability PC Syntactic simplicity, percentile
16	PCCNC _z	n/a	Text Easability PC Word concreteness, z score
17	PCCNC _p	n/a	Text Easability PC Word concreteness, percentile
18	PCREF _z	n/a	Text Easability PC Referential cohesion, z score
19	PCREF _p	n/a	Text Easability PC Referential cohesion, percentile
20	PCDC _z	n/a	Text Easability PC Deep cohesion, z score
21	PCDC _p	n/a	Text Easability PC Deep cohesion, percentile
22	PCVERB _z	n/a	Text Easability PC Verb cohesion, z score
23	PCVERB _p	n/a	Text Easability PC Verb cohesion, percentile

Continua...

Tabela 1 – Continuação

24	PCCONNz	n/a	Text Easability PC Connectivity, z score
25	PCCONNp	n/a	Text Easability PC Connectivity, percentile
26	PCTEMPz	n/a	Text Easability PC Temporality, z score
27	PCTEMPp	n/a	Text Easability PC Temporality, percentile
Referential Cohesion			
28	CRFNO1	CRFBN1um	Noun overlap, adjacent sentences, binary, mean
29	CRFAO1	CRFBA1um	Argument overlap, adjacent sentences, binary, mean
30	CRFSO1	CRFBS1um	Stem overlap, adjacent sentences, binary, mean
31	CRFNOa	CRFBNaum	Noun overlap, all sentences, binary, mean
32	CRFAOa	CRFBAAum	Argument overlap, all sentences, binary, mean
33	CRFSOa	CRFBSSaum	Stem overlap, all sentences, binary, mean
34	CRFCWO1	CRFPC1um	Content word overlap, adjacent sentences, proportional, mean
35	CRFCWO1d	n/a	Content word overlap, adjacent sentences, proportional, standard deviation
36	CRFCWOa	CRFPCAum	Content word overlap, all sentences, proportional, mean
37	CRFCWOad	n/a	Content word overlap, all sentences, proportional, standard deviation
38	CRFANP1	CREFP1u	Anaphor overlap, adjacent sentences
39	CRFANPa	CREFPau	Anaphor overlap, all sentences
LSA			
40	LSASS1	LSAassa	LSA overlap, adjacent sentences, mean
41	LSASS1d	LSAassd	LSA overlap, adjacent sentences, standard deviation
42	LSASSp	LSApssa	LSA overlap, all sentences in paragraph, mean
43	LSASSpd	LSApssd	LSA overlap, all sentences in paragraph, standard deviation
44	LSAPP1	LSAppa	LSA overlap, adjacent paragraphs, mean
45	LSAPP1d	LSAppd	LSA overlap, adjacent paragraphs, standard deviation

Continua...

Tabela 1 – Continuação

46	LSAGN	LSAGN	LSA given/new, sentences, mean
47	LSAGNd	n/a	LSA given/new, sentences, standard deviation
Lexical Diversity			
48	LDTTRc	TYPTOKc	Lexical diversity, type-token ratio, content word lemmas
49	LDTTRa	n/a	Lexical diversity, type-token ratio, all words
50	LDMTLDa	LEXDIVTD	Lexical diversity, MTLT, all words
51	LDVOCDa	LEXDIVVD	Lexical diversity, VOC, all words
Connectives			
52	CNCAIi	CONi	All connectives incidence
53	CNCCaus	CONCAUSi	Causal connectives incidence
54	CNCLogic	CONLOGi	Logical connectives incidence
55	CNCADC	CONADVCONi	Adversative and contrastive connectives incidence
56	CNCTemp	CONTEMPi	Temporal connectives incidence
57	CNCTempx	CONTEMPEXi	Expanded temporal connectives incidence
58	CNCAdd	CONADDi	Additive connectives incidence
59	CNCPos	n/a	Positive connectives incidence
60	CNCNeg	n/a	Negative connectives incidence
Situation Model			
61	SMCAUSv	CAUSV	Causal verb incidence
62	SMCAUSvp	CAUSVP	Causal verbs and causal particles incidence
63	SMINTEp	INTEi	Intentional verbs incidence
64	SMCAUSr	CAUSC	Ratio of casual particles to causal verbs
65	SMINTEr	INTEC	Ratio of intentional particles to intentional verbs
66	SMCAUSlsa	CAUSLSA	LSA verb overlap
67	SMCAUSwn	CAUSWN	WordNet verb overlap
68	SMTEMP	TEMPta	Temporal cohesion, tense and aspect repetition, mean
Syntactic Complexity			
69	SYNLE	SYNLE	Left embeddedness, words before main verb, mean
70	SYNNP	SYNNP	Number of modifiers per noun phrase, mean
71	SYNMEDpos	MEDwtm	Minimal Edit Distance, part of speech

Continua...

Tabela 1 – *Continuação*

72	SYNMEDwrd	MEDawm	Minimal Edit Distance, all words
73	SYNMEDlem	MEDalm	Minimal Edit Distance, lemmas
74	SYNSTRUTa	STRUTa	Sentence syntax similarity, adjacent sentences, mean.
75	SYNSTRUTt	STRUTt	Sentence syntax similarity, all combinations, across paragraphs, mean
Syntactic Pattern Density			
76	DRNP	n/a	Noun phrase density, incidence
77	DRVP	n/a	Verb phrase density, incidence
78	DRAP	n/a	Adverbial phrase density, incidence
79	DRPP	n/a	Preposition phrase density, incidence
80	DRPVAL	AGLSPSVi	Agentless passive voice density, incidence
81	DRNEG	DENNEGi	Negation density, incidence
82	DRGERUND	GERUNDi	Gerund density, incidence
83	DRINF	INFi	Infinitive density, incidence
Word Information			
84	WRDNOUN	NOUNi	Noun incidence
85	WRDVERB	VERBi	Verb incidence
86	WRDADJ	ADJi	Adjective incidence
87	WRDADV	ADVi	Adverb incidence
88	WRDPRO	DENPRPi	Pronoun incidence
89	WRDPRP1s	n/a	First person singular pronoun incidence
90	WRDPRP1p	n/a	First person plural pronoun incidence
91	WRDPRP2	PRO2i	Second person pronoun incidence
92	WRDPRP3s	n/a	Third person singular pronoun incidence
93	WRDPRP3p	n/a	Third person plural pronoun incidence
94	WRDFRQc	FRCLacwm	CELEX word frequency for content words, mean
95	WRDFRQa	FRCLaewm	CELEX Log frequency for all words, mean
96	WRDFRQmc	FRCLmcsm	CELEX Log minimum frequency for content words, mean
97	WRDAOAc	WRDAacwm	Age of acquisition for content words, mean
98	WRDFAMc	WRDFacwm	Familiarity for content words, mean
99	WRDCNCc	WRDCacwm	Concreteness for content words, mean
100	WRDIMGc	WRDIacwm	Imagability for content words, mean

Continua...

Tabela 1 – Continuação

101	WRDMEAc	WRDMacwm	Meaningfulness, Colorado norms, content words, mean
102	WRDPOLc	POLm	Polysemy for content words, mean
103	WRDHYPn	HYNOUNaw	Hypernymy for nouns, mean
104	WRDHYPv	HYVERBaw	Hypernymy for verbs, mean
105	WRDHYPnv	HYPm	Hypernymy for nouns and verbs, mean
Readability			
106	RDFRE	READFRE	Flesch Reading Ease
107	RDFKGL	READFKGL	Flesch-Kincaid Grade Level
108	RDL2	L2	Coh-Metrix L2 Readability

O Coh-Metrix foi adaptado para o português do Brasil nos dois anos de uma Iniciação Científica de Carolina Scarton (SCARTON; ALUÍSIO, 2010), no contexto do projeto PorSimples². O PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) (ALUÍSIO *et al.*, 2008) foi uma iniciativa que visou construir sistemas para promover o acesso a textos escritos em Português Brasileiro por analfabetos funcionais, pessoas com problemas cognitivos (como afasia e dislexia), e crianças e adultos em fase de aprendizado de leitura e escrita.

O projeto produziu duas ferramentas principais: um editor denominado SIMPLIFICA, que visa auxiliar autores a adequarem seus textos ao público alvo pretendido, exibindo possíveis pontos de complexidade léxica e sintática e sugerindo alterações, e uma ferramenta de pós-processamento de textos denominada FACILITA, que sumariza e simplifica automaticamente textos da Web conforme o usuário navega, facilitando seu acesso à informação.

O Coh-Metrix-Port, apesar de ser também uma ferramenta *stand-alone*, se inseriu no projeto PorSimples no editor SIMPLIFICA. Nessa ferramenta, o Coh-Metrix-Port é responsável por analisar a inteligibilidade do texto e classificá-lo conforme o nível de alfabetização necessário para compreendê-lo: rudimentar, básico ou pleno, segundo os critérios do INAF³.

Além de se utilizado no SIMPLIFICA, o Coh-Metrix-Port foi utilizado em diversos cenários de análise e classificação textual. Scarton e Aluísio (2010) avaliaram a primeira versão do Coh-Metrix-Port, na época com 38 métricas, por meio da comparação entre textos escritos para adultos, supostamente mais complexos, e textos escritos para crianças, supostamente mais simples. O cópulo utilizado consistia de textos jornalísticos (textos regulares do jornal ZeroHora⁴

² <www.nilc.icmc.usp.br/porsimples>

³ Indicador de Alfabetismo Funcional. Relatório disponível em <http://www.ibope.com.br/ipm/relatorios/relatorio_inaf_2009.pdf>.

⁴ <<http://zerohora.clicrbs.com.br/>>

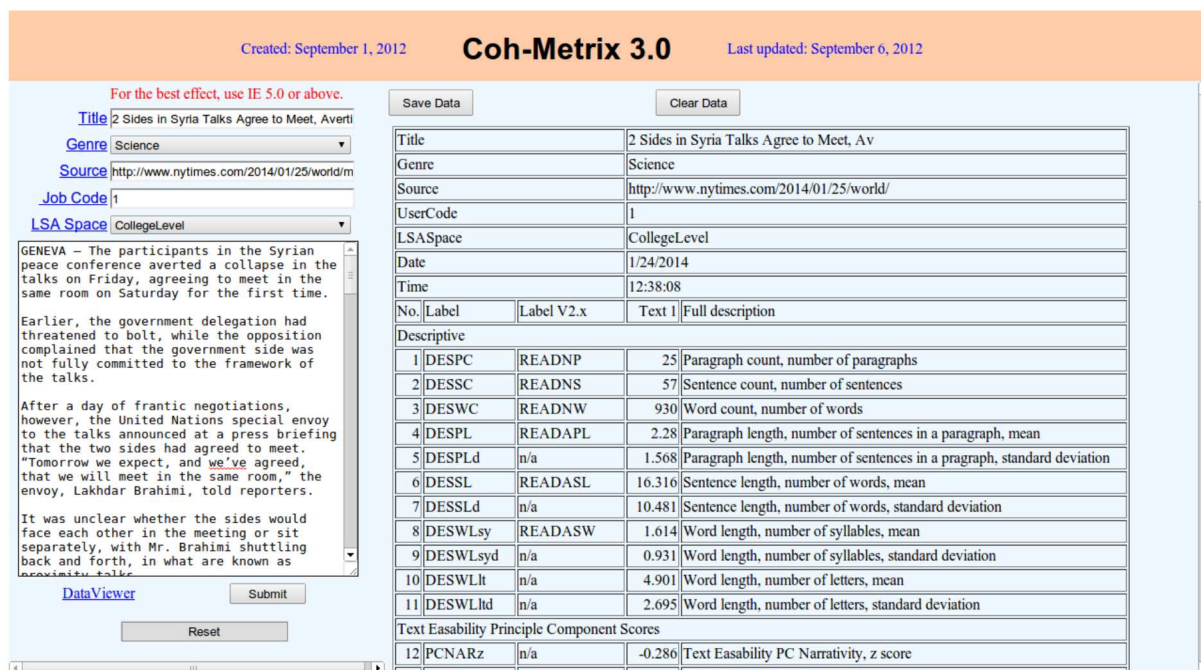


Figura 1 – Tela do Coh-Metrix exibindo métricas de um texto de exemplo.

– complexos – e da seção *Para seu filho ler* do mesmo jornal – simples) e textos de divulgação científica (textos das revistas *Ciência Hoje*⁵ – complexos – e *Ciência Hoje das Crianças*⁶ – simples). As autoras analisaram cada uma das métricas em termos de significância estatística entre as classes e de contribuição na tarefa de classificação, identificando que a maioria delas era distintiva no cópulo considerado.

Com o cópulo compilado, classificadores SVM (Support Vector Machine) foram treinados para as classes “simples” e “complexo” e apresentaram resultados de medida F de 93% considerando-se todos os textos (jornalísticos e científicos), 97% com os textos jornalísticos apenas e 94% com os textos científicos apenas. Por fim, as autoras avaliaram o desempenho dos classificadores considerando textos de outros gêneros; a melhor taxa de acerto foi de 94% e a pior, de 61,3%, de onde concluíram que os classificadores poderiam ter bom desempenho também em outros contextos, mesmo sendo treinados com textos pertencentes a um único domínio.

O trabalho de [Scarton e Aluísio \(2010\)](#) foi expandido por [Scarton, Gasperin e Aluisio \(2010\)](#), que acrescentou 10 novas métricas ao Coh-Metrix-Port (totalizando as 48 métricas atualmente disponíveis) e realizou uma análise mais profunda da tarefa de aprendizado de máquina. Os cópulos utilizados foram os mesmos do trabalho anterior. Dessa vez, foram comparados dois algoritmos de seleção de características e melhor aferido o desempenho dos classificadores quando treinados em um cenário e utilizados em outro.

[Aluisio et al. \(2010\)](#) utilizaram as métricas do Coh-Metrix-Port, com 4 métricas adicionais e métricas de um modelo de língua estatístico, usado para estimar a probabilidade de

⁵ <<http://cienciahoje.uol.com.br/revista-ch>>

⁶ <<http://www.chc.org.br/>>

unigramas, bigramas, e trigramas, para assim classificar textos simplificados. Foi utilizado o cópulus de referência do projeto PorSimples([ALUÍSIO; GASPERIN, 2010](#)), que contém textos originais do jornal ZeroHora e dois conjuntos de textos simplificados manualmente: um considerado simplificação natural e outro, simplificação forte. A hipótese era que os textos originais seriam ideais para pessoas com nível de alfabetização pleno, os textos de simplificação natural para pessoas com nível básico, e os textos de simplificação forte para pessoas com nível rudimentar. Também foram exploradas técnicas de aprendizado de máquina supervisionado: classificação padrão, ordinal (ranking), e regressão. A classificação ordinal apresentou os melhores resultados, com medida F de 90,4%, 48,4% e 73,1% para as classes original, natural, e forte, respectivamente.

[Finatto et al. \(2011\)](#) utilizaram o Coh-Matrix-Port para avaliar textos do jornalismo popular, utilizando, além das métricas da ferramenta, outras relacionadas a elipses, identificadas manualmente. Foram utilizados textos do jornal ZeroHora e do jornal popular Diário Gaúcho⁷. Consideraram-se quatro cenários: um com todas as métricas, outro somente com as do Coh-Matrix-Port (sem métricas de elipses) e outros dois com seleção de atributos. O melhor resultado para medida F foi apresentado pelo SVM em um dos cenários com seleção de atributos e foi de 87,7%.

O trabalho de [Pasqualini, Scarton e Finatto \(2011\)](#) fez uso do Coh-Matrix-Port e do Coh-Matrix original para avaliar traduções de textos literários nas direções Inglês – Português e Português – Inglês. Para tanto, foram considerados contos curtos, alguns originalmente escritos em português, outros em inglês, todos com suas respectivas traduções. A hipótese de trabalho era que as traduções para o português são menos inteligíveis (mais complexas) que os textos originais. Apesar de não haver uma conclusão clara sobre o que acontece no fenômeno da tradução, métricas como o índice Flesch diminuíram do inglês para o português, o que se concluiu ser um indício de que a tradução insere elementos que tornam os textos mais complexos.

A Tabela 2 mostra um sumário do desempenho apresentado pelos classificadores treinados com atributos extraídos pelo Coh-Matrix-Port em três dos trabalhos citados acima. A Figura 2 mostra a saída do Coh-Matrix-Port para um texto de exemplo. Em sua versão mais recente, a ferramenta dispõe de 48 métricas de nível léxico, sintático em nível de sintagmas nominais, semântico, e discursivo ([SCARTON; GASPERIN; ALUISIO, 2010](#)). A Tabela 3 mostra as métricas e uma breve descrição de cada uma.

Tabela 3 – Métricas do Coh-Matrix-Port.

Texto	
Título	Título
Autor	Autor

Continua...

⁷ <<http://diariogaucha.clicrbs.com.br/rs/>>

Tabela 3 – Continuação

Fonte		Fonte
Data de Publicação		Data de Publicação
Gênero		Gênero
Contagens Básicas		
1	Índice Flesch	Índice Flesch
2	Número de Palavras	Número de palavras do texto.
3	Número de Sentenças	Número de sentenças de um texto.
4	Número de Parágrafos	Número de parágrafos de um texto. Parágrafos são apenas onde há quebra de linha (não identações).
5	Palavras por Sentenças	Número de palavras dividido pelo número de sentenças.
6	Sentenças por Paragrafos	Número de sentenças dividido pelo número de parágrafos.
7	Sílabas por Palavras de Conteúdo	Número médio de sílabas por palavras de conteúdo (substantivos, verbos, adjetivos e advérbios).
8	Incidência de Verbos	Incidência de verbos em um texto.
9	Incidência de Substantivos	Incidência de substantivos em um texto.
10	Incidência de Adjetivos	Incidência de adjetivos em um texto.
11	Incidência de Advérbios	Incidência de advérbios em um texto.
12	Incidência de Pronomes	Incidência de pronomes em um texto.
13	Incidência de Palavras de Conteúdo	Incidência de Palavras de Conteúdo (substantivos, adjetivos, advérbios e verbos).
14	Incidência de Palavras Funcionais	Incidência de Palavras Funcionais (artigos, preposições, pronomes, conjunções e interjeições).
Operadores Lógicos		
15	Incidência de Operadores Lógicos	Incidência de operadores lógicos em um texto. Consideramos como operadores lógicos: e, ou, se, negações e um número de condições.
16	Incidência de E	Incidência do operador lógico e em um texto.
17	Incidência de OU	Incidência do operador lógico ou em um texto.
18	Incidência de SE	Incidência do operador lógico se em um texto.

Continua...

Tabela 3 – Continuação

19	Incidência de Negações	Incidência de Negações. Consideramos como negações: não, nem, nenhum, nenhuma, nada, nunca e jamais.
Frequências		
20	Frequências	Média de todas as frequências das palavras de conteúdo encontradas no texto. O valor da frequência das palavras é retirado da lista de frequências do <i>cópus Banco do Português</i> .
21	Mínimo Frequências	Identifica-se a menor frequência dentre todas as palavras de conteúdo em cada sentença. Depois, calcula-se uma média de todas as frequências mínimas. A palavra com a menor frequência é a mais rara da sentença.
Hiperônimos		
22	Hiperônimos de verbos	Hiperônimos de verbos.
Pronomes, Tipos e Token		
23	Incidência de Pronomes Pessoais	Incidência de pronomes pessoais em um texto. Consideramos como pronomes pessoais: eu, tu, ele/ela, nós, nós, eles/elas, você e vocês.
24	Pronomes por Sintagmas	Média do número de pronomes que aparecem em um texto pelo número de sintagmas.
25	Type/Token	Número de palavras únicas dividido pelo número de tokens dessas palavras. Cada palavra única é um tipo. Cada instância desta palavra é um token.
Constituintes		
26	Incidência de Sintagmas	Incidência de sintagmas nominais por 1000 palavras.
27	Modificadores por Sintagmas	Média do número de modificadores por sintagmas nominais, adjetivos, advérbios e artigos, que participam de um sintagma.
28	Palavras antes de verbos principais	Média de palavras antes de verbos principais na cláusula principal da sentença.
Conectivos		
29	Incidência de Conectivos	Incidência de todos os conectivos que aparecem em um texto.

Continua...

Tabela 3 – Continuação

30	Conectivos Aditivos Positivos	Incidência de conectivos classificados como aditivos positivos.
31	Conectivos Aditivos Negativos	Incidência de conectivos classificados como aditivos negativos.
32	Conectivos Temporais Positivos	Incidência de conectivos classificados como temporais positivos.
33	Conectivos Temporais Negativos	Incidência de conectivos classificados como temporais negativos.
34	Conectivos Causais Positivos	Incidência de conectivos classificados como causais positivos.
35	Conectivos Causais Negativos	Incidência de conectivos classificados como causais negativos.
36	Conectivos Lógicos Positivos	Incidência de conectivos classificados como lógicos positivos.
37	Conectivos Lógicos Negativos	Incidência de conectivos classificados como lógicos negativos.
Ambiguidades		
38	Verbos	Ambiguidade de Verbos.
39	Substantivos	Ambiguidade de Substantivos.
40	Adjetivos	Ambiguidade de Adjetivos.
41	Advérbios	Ambiguidade de Advérbios.
Correferência		
42	Sobreposição de argumentos adjacentes	Sobreposição de argumentos em sentenças adjacentes.
43	Sobreposição de argumentos	Sobreposição de argumentos em todos os pares de sentenças.
44	Sobreposição de radicais de palavras adjacentes	Sobreposição de argumentos em sentenças adjacentes.
45	Sobreposição de radicais de palavras	Sobreposição de radicais de palavras em todos os pares de sentenças.
46	Sobreposição de palavras de conteúdo	Sobreposição de palavras de conteúdo em sentenças adjacentes.
Anáforas		
47	Referência anafórica adjacente	Referência anafórica em sentenças adjacentes.

Continua...

The screenshot shows the PorSimples Coh-Metrix-Port interface. At the top, there is a blue header with the PorSimples logo and the text 'Simplificando o Português' on the left, and 'Coh-Metrix-Port' on the right. Below the header, there is a navigation bar with 'Página inicial > Enviar novo texto > Resultados' and a user profile 'Administrador | Sair'. The main content area is titled 'Resultados' and contains a table with two sections: 'Texto' and 'Contagens Básicas'.

Texto		
Título	Dilma diz que países emergentes seguirão sendo "fundamentais"	Título
Autor	Folha de São Paulo	Autor
Fonte	http://www1.folha.uol.com.br/mercado/2014/01/1402246-em-davos-dilma-diz-que-estabilidade-da-moeda-e-valor-central-do-brasil.shtml	Fonte
Data de Publicação	24/01/2014	Data de Publicação
Gênero	Jornalístico	Gênero
Contagens Básicas		
Índice Flesch	39.0816054421769	Índice Flesch
Número de Palavras	392.0	Número de palavras do texto.
Número de Sentenças	15.0	Número de sentenças de um texto.
Número de Parágrafos	14.0	Número de parágrafos de um texto. Parágrafos são apenas onde há quebra de linha (não identações).
Palavras por Sentenças	26.1333333333333	Número de palavras dividido pelo número de sentenças.
Sentenças por Parágrafos	1.07142857142857	Número de sentenças dividido pelo número de parágrafos.
Sílabas por Palavras de Conteúdo	2.90666666666667	Número médio de sílabas por palavras de conteúdo (substantivos, verbos, adjetivos e advérbios).

Figura 2 – Saída do Coh-Metrix-Port para um texto de exemplo.

Tabela 3 – Continuação

48	Referência anafórica	Referência anafórica em até cinco sentenças anteriores.
----	----------------------	---

Como se pode perceber, o Coh-Metrix-Port foi empregado em diversos cenários de classificação textual, obtendo desempenho bastante satisfatório em muitos deles. Como o presente trabalho também diz respeito à classificação textual em um cenário específico, acreditamos que adaptar o Coh-Metrix-Port para lidar com textos de pacientes com DA e CCL, dando origem ao Coh-Metrix-Dementia, é uma abordagem promissora.

3.1.2 A ferramenta Análise de Inteligibilidade de Córpus (AIC)

O AICórpus (Análise de Inteligibilidade de Córpus⁸, ou AIC) (MAZIERO; PARDO; ALUÍSIO, 2008) é uma ferramenta com estrutura e propósito muito similares aos do Coh-Metrix-Port. A ferramenta analisa textos e retorna métricas textuais que podem servir de base à análise da inteligibilidade do texto (daí seu nome). Ele também foi projetado durante o projeto PorSimples para calcular características de córpus de textos simples disponíveis na Web para fundamentar os sistemas de simplificação léxica e sintática criados pelo projeto (ALUÍSIO *et al.*, 2008).

⁸ <<http://www.nilc.icmc.usp.br/porsimples/AIC/>>

Tabela 2 – Desempenho dos classificadores com características extraídas pelo Coh-Metrix-Port.

Trabalho	Cenário	Medida F
Scarton e Aluísio (2010)	Todos os textos	93,0%
	Jornalísticos	97,0%
	Científicos	94,0%
	Outros - melhor	94,0%
	Outros - pior	61,3%
Aluisio <i>et al.</i> (2010)	Original	90,4%
	Natural	48,4%
	Forte	73,1%
Finatto <i>et al.</i> (2011)	Melhor cenário	87,7%

O diferencial do AICórpus é que ele utiliza um analisador sintático total, o PALAVRAS (BICK, 2000). Devido a isso, o AICórpus é capaz de computar características sintáticas do texto, algo que o Coh-Metrix-Port ainda não é capaz de fazer. As métricas que o AICórpus retorna estão organizadas em seis classes (MAZIERO; PARDO; ALUÍSIO, 2008): **estatísticas do texto**, **voz passiva**, **características das orações**, **densidade**, **personalização** e **marcadores discursivos**.

As métricas relativas às **estatísticas do texto** são: número de caracteres, número de palavras, número médio de caracteres por palavra, número médio de palavras por sentença, número de sentenças e número de palavras simples, que foram compiladas durante o projeto PorSimples. A métrica **voz passiva** traz o número de sentenças na voz passiva. As métricas das **características das orações** são : número de orações, número de orações que iniciam com conjunções subordinadas, número de orações que iniciam com conjunções coordenadas, sentenças que contenham um dado número de orações, moda do cálculo anterior, número médio de orações por sentença, número de conjunções subordinadas, número de conjunções coordenadas, número de verbos no gerúndio, número de verbos no particípio, número de verbos no infinitivo, soma dos verbos no gerúndio, particípio e infinitivo.

A classe **densidade** traz características das orações que podem torná-la difícil de ler como: número de objetos preposicionais, número médio de objetos preposicionais por sentença, número médio de objetos preposicionais por oração, número de orações relativas, número de apostos especificadores, número de adjuntos adverbiais, número de advérbios e número de adjetivos. A classe **personalização** traz as seguintes métricas: número de pronomes pessoais, número de pronomes pessoais na segunda pessoa do singular, número de pronomes pessoais na primeira pessoa do plural, número de pronomes pessoais na segunda pessoa do plural, número de pronomes pessoais na terceira pessoa do plural, número de pronomes possessivos, número de pronomes possessivos na segunda pessoa (Teu, Tua, Teus, Tuas), número de pronomes possessivos na segunda pessoa (Vosso, Vossa, Vossos, Vossas), número de pronomes possessivos na primeira pessoa, número de pronomes possessivos na terceira pessoa. Por fim, a classe **marcadores discursivos** traz o número de marcadores discursivos e o número de marcadores

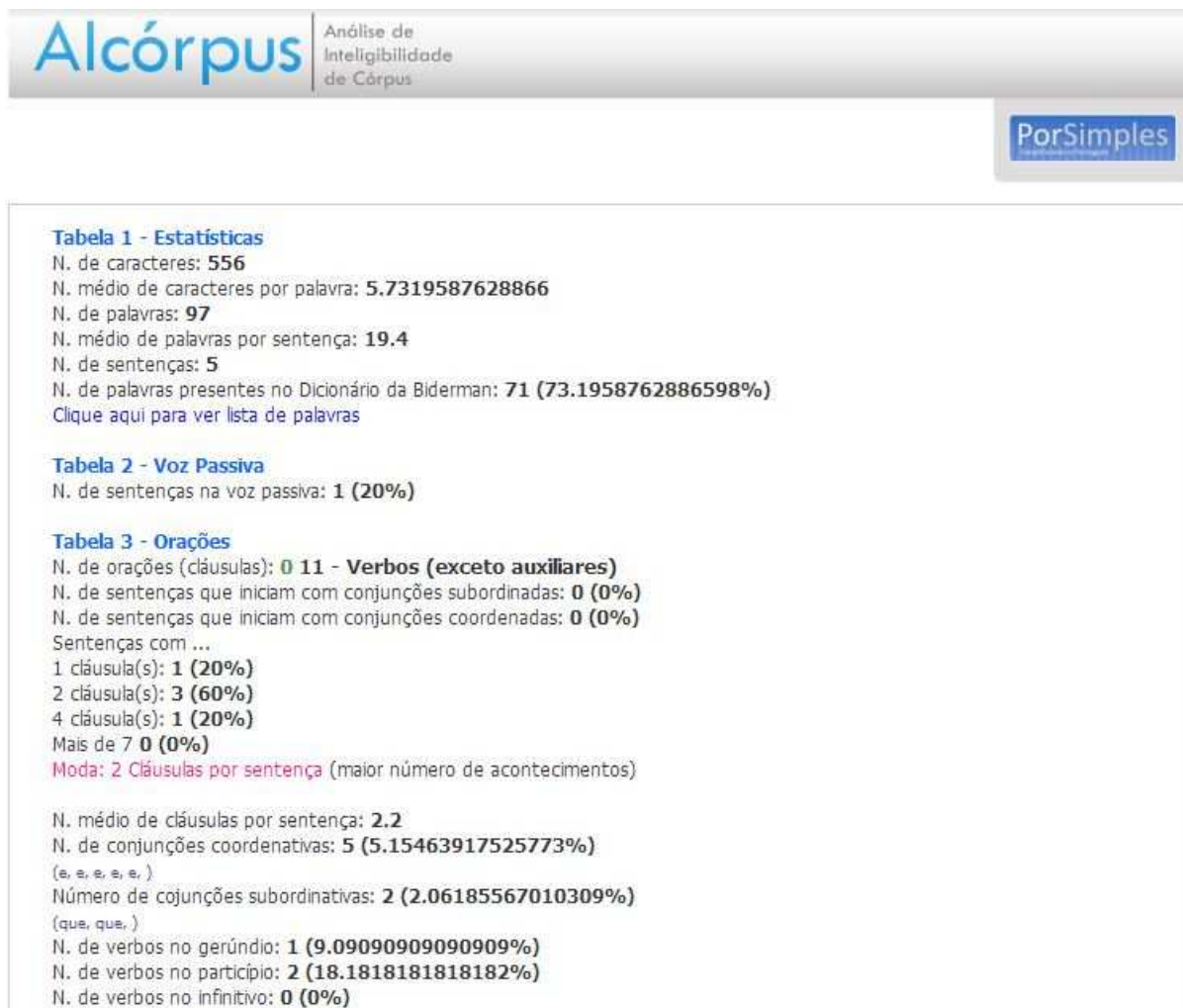


Figura 3 – Exemplo de parte da tela de saída do AIC.

discursivos ambíguos.

A Figura 3 mostra um trecho da saída do AICorpus para um resumo de artigo em português, tomado como exemplo.

3.1.3 O Linguistic Inquiry and Word Count (LIWC)

O LIWC (PENNEBAKER; FRANCIS; J., 2001; PENNEBAKER *et al.*, 2007) é uma ferramenta de análise textual que calcula o grau em que as pessoas utilizam diferentes categorias de palavras no texto. As categorias incluem: **estatísticas comuns do texto, dimensão linguística, processos psicológicos, relatividade, assuntos pessoais e miscelânea**. Além destas, faz a contagem de pontuações, num total de quase 100 métricas.

As **estatísticas comuns** disponibilizadas são: número de palavras, palavras por sentença, palavras pertencentes ao dicionário LIWC, palavras únicas, palavras com mais de 6 caracteres. A **dimensão linguística** analisada pelo LIWC fornece contagens de: pronomes, pronomes pessoais, negações, artigos, preposições, e números. Os **processos psicológicos** analisados pelo

conj	negate	quant	number	swear	social	family	friend	humans	affect	posemo	negemo	anx	anger	sad
0.00	0.00	0.00	0.00	0.00	20.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5.41	2.70	5.41	0.00	0.00	18.92	0.00	0.00	0.00	10.81	2.70	5.41	0.00	2.70	0.00
0.00	3.57	3.57	0.00	0.00	39.29	0.00	0.00	0.00	7.14	7.14	0.00	0.00	0.00	0.00
4.00	4.00	2.00	2.00	0.00	16.00	0.00	0.00	0.00	4.00	4.00	0.00	0.00	0.00	0.00
4.44	2.22	4.44	0.00	0.00	13.33	0.00	0.00	0.00	6.67	4.44	2.22	0.00	0.00	0.00
2.94	2.94	2.94	0.00	0.00	29.41	0.00	0.00	0.00	17.65	11.76	5.88	0.00	0.00	2.94
4.08	6.12	2.04	2.04	0.00	16.33	0.00	0.00	0.00	4.08	4.08	0.00	0.00	0.00	0.00
8.33	5.56	2.78	0.00	0.00	16.67	0.00	0.00	0.00	2.78	0.00	2.78	0.00	0.00	0.00
3.92	3.92	1.96	1.96	0.00	15.69	0.00	0.00	0.00	3.92	3.92	0.00	0.00	0.00	0.00
3.92	3.92	1.96	1.96	0.00	15.69	0.00	0.00	0.00	3.92	3.92	0.00	0.00	0.00	0.00
3.45	3.45	0.00	0.00	0.00	24.14	0.00	0.00	0.00	10.34	10.34	0.00	0.00	0.00	0.00

Figura 4 – Tela de saída do LIWC, com o dicionário para o inglês, versão de 2007, mostrando a análise de uma música.

LIWC são relacionados com reações a emoções, compondo as emoções positivas, as negativas e dentro destas últimas a ansiedade, raiva, tristeza, mecanismos cognitivos, relacionados a sensações/percepções, visão, audição, toque, sociais, comunicativos, relacionados a amigos, família, humanos, dentre outros.

A categoria **relatividade** traz referências a tempo, espaço, movimento, verbos no passado, presente e futuro. A categoria relacionada a **assuntos pessoais** traz referências a ocupação de trabalho, lazer, música, dinheiro, sexo, morte, religião, dentre outros. Quanto à **miscelânea** o LIWC captura palavras de ofensa e xingamento e particularidades da **fala** que são: quantidade de disfluências e quantidade de preenchedores.

O LIWC foi originalmente concebido para a língua inglesa e codifica seu conhecimento na forma de um dicionário, tendo sido traduzido/adaptado para mais de 10 línguas como espanhol, francês, alemão, e russo. O dicionário que dá apoio ao LIWC foi traduzido para o Português do Brasil sob a coordenação da Profa. Sandra Aluísio do NILC/ICMC/USP, que disponibilizou o dicionário no site PortLex⁹. Mais detalhes de seu uso podem ser lidos em Balage Filho, Pardo e Aluísio (2013). A Figura 4 mostra uma tela da saída do LIWC 2007, com o resultado da análise de uma música em inglês, que foi dividida em 11 segmentos, contando seu título, via configuração do LIWC, para análise individual de suas estrofes.

Como se pode observar no Capítulo 4, o LIWC foi empregado em diversos trabalhos relacionados, e como uma tradução para o português está prontamente disponível, acreditamos que empregar seu dicionário ou métricas inspiradas nas extraídas pela ferramenta é importante para alcançar o desempenho das tarefas de classificação para as quais o Coh-Metrix-Dementia venha a ser empregado.

⁹ <<http://www.nilc.icmc.usp.br/portlex/index.php/en/liwc>>

3.2 Medidas Automatizadas

Esta Seção descreve várias medidas automatizadas empregadas na literatura de avaliação de condições clínicas de pacientes a partir de textos transcritos ou escritos, organizadas conforme o nível linguístico de análise envolvido.

3.2.1 Medidas de diversidade lexical

Uma métrica de riqueza lexical já mencionada aqui é a **Relação Tipo por Token** (*Type to Token Ratio, TTR*). A *TTR* é dada pela razão entre o número de tokens distintos presentes no texto (o tamanho do vocabulário, denotado por V) e o número total de tokens (o comprimento do texto, N). Ou seja:

$$TTR = \frac{V}{N} \quad (3.1)$$

Pode também ser utilizado o número de palavras léxicas (substantivos, adjetivos, verbos e advérbios) presentes no texto como V e o número total de tokens (o comprimento do texto, N). A *TTR* é sensível ao comprimento do texto (THOMAS *et al.*, 2005). Uma forma de padronizar esta medida é calculá-la para cada n palavras do texto (por exemplo, mil palavras), e depois calcular a média. Se o texto tiver menos que n palavras, convencionou-se que a *TTR* é zero.

Outra medida, insensível ao comprimento da enunciação, é o **Índice de Brunet W** (BRUNET, 1978), calculado através da seguinte equação (THOMAS *et al.*, 2005):

$$W = N^{V^{-0.165}} \quad (3.2)$$

Na Equação 3.2, N é o número de palavras lexicais, e V é o número total de tokens usados. Os valores de W típicos variam entre 10 e 20, sendo que uma fala mais rica produz valores *menores* (THOMAS *et al.*, 2005). Outra estatística insensível ao comprimento do texto é a **Estatística de Honoré R** (HONORÉ, 1979), calculada como (THOMAS *et al.*, 2005):

$$R = \frac{100 \log N}{1 - \frac{V_1}{V}} \quad (3.3)$$

Na Equação 3.3, N é o número total de tokens, V_1 é número de palavras do vocabulário que aparecem uma única vez, e V é o número de palavras lexicais.

Uma outra medida bastante simples e muito similar à densidade de ideias (seção 3.2.3.1) é a **densidade de conteúdo**, calculada como a razão entre o número de palavras de classe aberta e o número de palavras de classe fechada. Classes abertas são classes gramaticais às quais se pode adicionar palavras indefinidamente, como substantivos e verbos, e normalmente possuem grande quantidade de palavras; classes fechadas são fixas, e normalmente possuem poucas, como pronomes

e preposições (ROARK *et al.*, 2011). Essa medida possui resultados próximos à densidade de ideias, exceto que, na densidade de conteúdo, nomes também são levados em conta (ROARK *et al.*, 2011).

3.2.2 Medidas de complexidade sintática

Medir a complexidade sintática de um fragmento discursivo, oral ou escrito, é de grande interesse em diversas áreas do conhecimento. Tal medida pode ser utilizada, por exemplo, para aferir a inteligibilidade e a facilidade de leitura de um texto, permitindo verificar sua adequação ao público alvo pretendido. Além disso, diversos quadros demenciais apresentam, entre seus sintomas, a redução da complexidade sintática das enunciações, o que torna sua medição importante na discriminação de sujeitos sadios e comprometidos (ROARK; MITCHELL; HOLLINGSHEAD, 2007).

Não existe uma medida única e universalmente aceita de complexidade sintática. Várias medidas foram propostas na literatura, cada uma delas baseada em uma noção primária diferente de complexidade (ROARK; MITCHELL; HOLLINGSHEAD, 2007). Nas próximas sub-seções, são apresentadas algumas dessas medidas, geralmente citadas na literatura de avaliação de quadros demenciais.

3.2.2.1 Complexidade de Yngve

A complexidade de Yngve (YNGVE, 1960) baseia-se na premissa de que as árvores sintáticas das sentenças da língua inglesa tendem a se ramificar para a direita, e que desvios em relação a esse padrão correspondem a uma maior complexidade na linguagem (ROARK; MITCHELL; HOLLINGSHEAD, 2007). Dessa forma, a complexidade de Yngve procura medir o quanto uma árvore sintática se desvia desse padrão de ramificação.

Para calcular a complexidade de uma sentença, é preciso que se tenha sua árvore sintática. A figura 5 traz um exemplo para a sentença “*She found a cat with a red tail.*” (PAKHOMOV *et al.*, 2011). Inicialmente, atribuímos um peso para cada nó não-terminal da árvore, da seguinte maneira: para cada nó, atribuir peso 0 ao seu filho mais à direita, 1 ao segundo filho mais a direita, e assim por diante até chegar ao filho mais a esquerda. Dessa forma, os filhos serão numerados com peso 0, 1, 2, ..., da direita para a esquerda. No exemplo da Figura 5, o peso de cada nó está indicado entre colchetes.

Em seguida, calcula-se a pontuação de cada palavra, dada pela soma dos pesos dos não-terminais existentes no caminho entre a raiz e a palavra. Por exemplo, o peso de *a* em *a cat* é dado pela soma do peso de VP (1), mais o do primeiro NP (0), mais o do segundo NP (1) mais o de DT (1), totalizando peso 3. O peso de cada palavra está indicado, na figura 5, entre colchetes, ao lado da palavra.

Uma vez calculada a pontuação das palavras, a complexidade da sentença pode ser

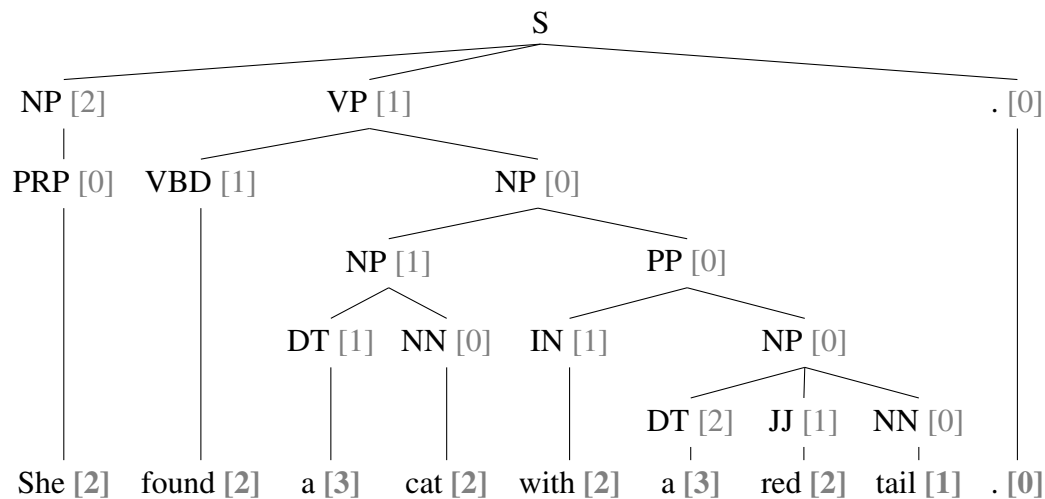


Figura 5 – Exemplo de árvore sintática para cálculo da complexidade de Yngve (extraído de Pakhomov *et al.* (2011)).

calculada de diversas formas: como a média, soma ou máximo das pontuações das palavras. No exemplo, a média dos valores é 1.89, a soma, 17 e o máximo, 3. Roark, Mitchell e Hollingshead (2007) apontam que, em seu estudo, a **média** apresentou resultados melhores.

Existe ainda uma outra forma de se encarar a pontuação de cada nó: a partir de uma pilha utilizada em uma derivação de cima para baixo, da esquerda para a direita (ROARK; MITCHELL; HOLLINGSHEAD, 2007). Nesse caso, o escore de uma palavra é dado pelo número de elementos que ainda permanecem na pilha quando a palavra é finalmente derivada. Utilizando-se o mesmo exemplo, para que se derive a palavra *She*, aplica-se a regra “S → NP VP .”, seguida por “NP → PRP”, e finalmente “PRP → *She*”. Portanto, quando *She* é derivado, ainda existem 2 elementos na pilha (VP e .), resultando em sua pontuação de 2.

Alguns trabalhos da literatura, como Resnik (1992), relacionaram o tamanho da pilha necessário para processar uma sentença à sua demanda de memória operacional, apesar de ele medir diretamente apenas o desvio de uma ramificação à direita (ROARK; MITCHELL; HOLLINGSHEAD, 2007).

3.2.2.2 Complexidade de Frazier

Frazier (1985) propôs uma abordagem *bottom-up* para o cálculo da complexidade sintática de uma sentença, que parte da palavra e sobe na árvore sintática até encontrar um nó que não seja o filho mais à esquerda de seu pai. Cada nó na árvore recebe uma pontuação 1, e nós filhos de nós do tipo sentença, 1.5. A pontuação de cada palavra é dada pela soma das pontuações dos nós pertencentes a seu ramo.

Para ilustrar, considere a figura 6, que mostra a mesma sentença utilizada para o cálculo da complexidade de Yngve com os pesos atribuídos pelo método de Frazier. Um peso marcado com $[x]$ indica fim de ramo. Nesse exemplo, para a derivação da palavra *She*, o nó PRP imediatamente

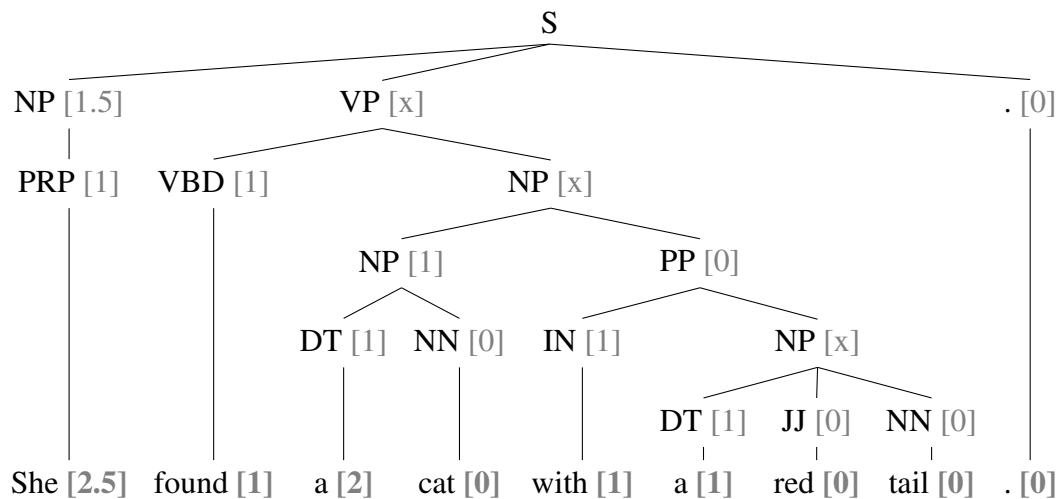


Figura 6 – Exemplo de árvore sintática para cálculo da complexidade de Frazier (extraído de Pakhomov *et al.* (2011)).

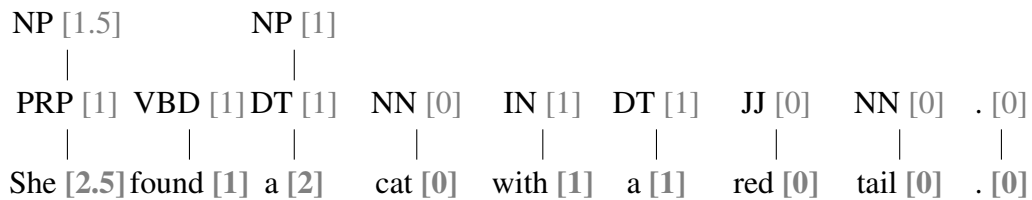


Figura 7 – Exemplo da Figura 6 com os ramos individuais das palavras.

acima é incluído, pois é o filho mais à esquerda de NP, recebendo pontuação 1; PRP é filho de NP, que é o filho mais à esquerda de seu pai, e como seu pai é um nó do tipo sentença, recebe pontuação 1.5; portanto, a pontuação de *She* é $1 + 1.5 = 2.5$.

Para o cálculo da pontuação de *found*, seu nó pai VBD é incluído, pois é o filho mais à esquerda de VP, recebendo pontuação 1; porém, o pai de VBD, VP, não é incluído, pois não é o filho mais à esquerda de S; portanto, a derivação para, e a pontuação de *found* é 1. No caso de *cat*, nem mesmo seu nó pai NN é incluído, pois não é o filho mais à esquerda de NP, recebendo pontuação 0. A Figura 7 mostra os ramos utilizados para o cálculo da pontuação de cada palavra.

Assim como na complexidade de Yngve, aqui também pode-se calcular a complexidade da sentença como a soma, média ou máximo das pontuações das palavras. Porém, Frazier propôs dividir a sentença em trigramas (sequências de três palavras), calcular a soma das pontuações das palavras em cada trigrama e adotar o máximo dessas somas como sendo a complexidade da sentença. Para a sentença de exemplo, a soma seria 7.5, a média, 0.83, o máximo, 2.5, e o máximo entre trigramas, 4.5.

3.2.2.3 Distância de dependência

Note que as medidas de Yngve e de Frazier são computadas com base na **árvore de derivação** da sentença. Porém, é possível calcular a complexidade também sobre uma **estrutura**

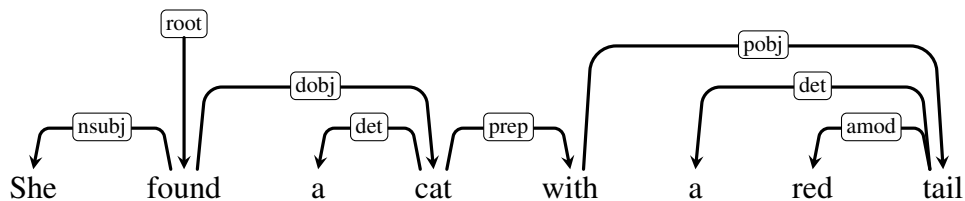


Figura 8 – Exemplo de estrutura de dependências.

de dependências. A Figura 8 mostra a árvore de dependências da mesma sentença dos exemplos anteriores.

A cada relação de dependência está associada uma distância entre as palavras na superfície textual. Por exemplo, na relação **nsubj** entre *She* e *found*, existe distância 1, enquanto na relação **dobj**, entre *cat* e *found*, existe distância 2. A métrica de distância de dependência é calculada como a soma das distâncias associadas às relações. No exemplo cita, esse valor é 11.

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 (PAKHOMOV *et al.*, 2011; GIBSON, 1998; LIN, 1996). Lin (1996) defendeu o uso de estruturas de dependência como forma de medir dificuldades de processamento, afirmando que grandes distâncias entre palavras relacionadas geram *overhead* de memória (ROARK; MITCHELL; HOLLINGSHEAD, 2007).

3.2.2.4 Nível de desenvolvimento (nível D)

O nível de desenvolvimento (em inglês, *Developmental level* ou *D-level*) (ROSENBERG; ABBEDUTO, 1987) é uma escala, originalmente com 7 níveis, que se baseia no nível de desenvolvimento de sentenças complexas em crianças com desenvolvimento normal (ROARK; MITCHELL; HOLLINGSHEAD, 2007). Cheung e Kemper (1992) acrescentaram um nível 0, para representar sentenças simples, contendo apenas uma cláusula, resultando em uma escala de 8 níveis.

Cada nível corresponde à presença de construções gramaticais, cada vez mais complexas, da seguinte maneira (CHEUNG; KEMPER, 1992; ROARK; MITCHELL; HOLLINGSHEAD, 2007):

- O **nível 0** corresponde a sentenças simples, com uma única cláusula.
- O **nível 1**, a sentenças com complementos infinitivos embutidos. Exemplo: *She needs to pay the rent.*
- O **nível 2**, a sentenças complexas contendo como complemento predicados do tipo *wh*, orações coordenadas e sujeitos compostos. Exemplos: *She worked all day and worried all night. The woman and her four children had not eaten for two days.*

- O **nível 3**, a sentenças complexas com cláusulas relativas modificando o sintagma nominal objeto, ou sintagmas nominais como complemento no predicado. Exemplo: *The police caught the man who robbed the woman.*
- O **nível 4**, a sentenças complexas com complementos gerundivos ou construções comparativas. Exemplo: *They were hungrier than her.*
- O **nível 5**, a sentenças complexas contendo cláusulas relativas modificando o sintagma nominal sujeito, complementos do sintagma nominal sujeito ou nominalizações do sujeito. Exemplo: *The woman who worked in the cafeteria was robbed.*
- O **nível 6**, a sentenças complexas contendo orações subordinadas. Exemplo: *He robbed because he was hungry.*
- O **nível 7**, a sentenças contendo múltiplas construções pertencentes aos níveis anteriores.

O trabalho de Toledo (2011), em seu estudo do impacto das variáveis demográficas na produção discursiva de adultos sadios, utilizou um método semelhante, ao classificar a complexidade do discurso do paciente em 4 níveis, conforme a construção mais complexa encontrada:

- Nível 0, quando o texto consiste unicamente de palavras isoladas, sem conexão entre elas.
- Nível 1, quando o texto contém apenas declarativas simples, frases de pouca complexidade que informam ou declaram algo.
- Nível 2, quando o texto possui orações coordenadas.
- Nível 3, quando o texto possui orações subordinadas.

3.2.2.5 Entropia cruzada

Uma maneira de se calcular a riqueza sintática de um texto é procurar por padrões inco-muns de combinações de palavras ou de classes gramaticais (ROARK; MITCHELL; HOLLING-SHEAD, 2007). Para tanto, podemos utilizar um modelo de língua baseado em n-gramas; com isso, conseguimos calcular a probabilidade de ocorrência da elocução e, com esse valor, sua entropia cruzada.

Para exemplificar, considere um modelo de bigramas. Seja w_1, w_2, \dots, w_n uma sequência de palavras, e assuma que w_0 é um marcador de início, e que w_{n+1} é um marcador de fim. Assim, a probabilidade da sequência de palavras é dada por:

$$P(w_1, w_2, \dots, w_n) = \sum_{i=1}^{n+1} P(w_i | w_{i-1}) \quad (3.4)$$

A entropia cruzada da sequência de palavras é então dada por:

$$H(w_1, w_2, \dots, w_n) = -\frac{1}{n} \log P(w_1, w_2, \dots, w_n) \quad (3.5)$$

Quanto maior a entropia cruzada de uma sentença, mais incomum ela é perante o modelo de língua, o que é um indício de maior complexidade.

Existem outras diversas medidas sugeridas na literatura que estão entre os níveis léxico e sintático, pois só necessitam saber o que é uma sentença. Algumas são baseadas no tamanho da sentença, como o **Comprimento Médio da Elocução** (*Mean Length of Utterance*, MLU (BROWN, 1973)) e o **Número Médio de Cláusulas por Elocução** (*Mean Clauses per Utterance*, MCU (KEMPER *et al.*, 1989)), medem o tamanho da sentença, assumindo que esse tamanho – seja em morfemas, palavras, ou cláusulas – está diretamente associado à complexidade (CHEUNG; KEMPER, 1992).

3.2.2.6 Considerações finais

Cada uma das diversas métricas apresentadas acima baseia-se diferentes conceitos de complexidade. Métricas como MCU e MLU baseiam-se no tamanho da sentença, associando maior tamanho a maior complexidade. Tais medidas são superficiais, pois não analisam a estrutura da sentença, mas apenas sua forma superficial; dessa maneira, uma sentença mais curta contendo termos e construções gramaticais raros será considerada mais simples que uma sentença maior contendo palavras comuns e construções diretas, o que pode não representar com acurácia a real dificuldade de compreensão.

Outras medidas, como o nível D, baseiam-se na ordem com que construções gramaticais são adquiridas por crianças aprendendo sua língua nativa. Nesse caso, quanto mais tardia é a aquisição do domínio de uma construção, mais complexa ela é considerada. Medidas como a distância de dependência focam-se na demanda de memória necessária para processar uma sentença, associando-a à distância entre palavras dependentes; aqui, considera-se que quanto maior a distância entre as palavras que apresentam relação dentro da sentença, mais complexa esta se torna.

Medidas como os índices de Yngve e Frazier baseiam-se no comportamento geral de ramificação à direita de alguns idiomas, considerando desvios desse padrão como indícios de complexidade. Outras, como a entropia cruzada, procuram medir o nível de “surpresa” do ouvinte/leitor diante da sentença, com base em um modelo de língua, considerando que frases que causam maior estranhamento são mais complexas.

A literatura mostra que os índices de Yngve e de Frazier, assim como a distância de dependência, estão relacionados à demanda de memória operacional (RESNIK, 1992). Os estágios iniciais da DA são marcados por perdas semânticas, sendo que a fala semanticamente “vazia”, caracterizada pelo uso excessivo de pronomes, se mostrou ser uma das características

distintivas da doença (ALMOR *et al.*, 1999; KEMPLER, 1995), junto a deficits na habilidade de determinar relações de parentesco semântico entre conceitos (ARONOFF *et al.*, 2006). Entretanto, descobriu-se que os declínios cognitivos na DA também estão associados à perda de desempenho em tarefas que envolvem memória operacional, especialmente em estágio avançado (ALMOR *et al.*, 1999; BICKEL *et al.*, 2000; KEMPLER *et al.*, 1998; MACDONALD *et al.*, 2001). Assim sendo, (PAKHOMOV *et al.*, 2011) afirmam que, dada sua possível associação com a memória operacional e com a deterioração das relações semânticas, espera-se que essas medidas de complexidade sintática sejam sensíveis aos efeitos da DA na produção e na compreensão da linguagem.

Segundo Roark, Mitchell e Hollingshead (2007), dois aspectos devem ser considerados quando se trata de escolher como medir a complexidade sintática do discurso falado no contexto de avaliações psicométricas. Primeiramente, é preciso que as medidas apresentem alto **poder discriminativo** entre grupos, uma vez que elas serão usadas para distinguir indivíduos normais de indivíduos com demência, bem como indivíduos com diferentes quadros demenciais.

Em segundo lugar, é preciso que as métricas possam ser **automatizadas confiavelmente**. Isso porque diferentes medidas de complexidade dependem de diferentes níveis de detalhamento da árvore sintática, que são disponibilizados por *parsers* em diferentes níveis de confiabilidade. Assim, se uma métrica depende de uma informação muito detalhada fornecida pelo *parser* com baixa precisão, os resultados automatizados não serão confiáveis. Os autores apontam que, idealmente, medidas simples, fáceis de automatizar e com alto poder discriminativo são preferíveis (ROARK; MITCHELL; HOLLINGSHEAD, 2007).

3.2.3 Medidas de densidade semântica

3.2.3.1 A Análise de Densidade de Ideias (AID)

A densidade de ideias de um fragmento discursivo (também denominada *densidade proposicional* ou *P-density*) consiste no número de **proposições** expressas dividido pelo número de palavras¹⁰ (BROWN *et al.*, 2008). Ela tem por objetivo medir a quantidade de informação que é transmitida em relação ao número de palavras utilizadas para transmiti-la. Dessa forma, ela quantifica a eficiência com que a informação é sintetizada: maior densidade de ideias implica em compactidade no discurso, enquanto menor densidade de ideias implica em repetição e imprecisão (CHAND *et al.*, 2010). Em termos semânticos, a densidade de ideias é o grau ao qual o sujeito realiza asserções ou perguntas, ao invés de simplesmente referir-se a entidades (BROWN *et al.*, 2008).

Proposições consistem, normalmente, de um predicador acompanhado de seus argumentos, de maneira semelhante à lógica formal, mas não sendo tão restritas quanto esta (CHAND *et*

¹⁰ Alguns trabalhos, como Kemper, Thompson e Marquis (2001) e Chand *et al.* (2010), calculam a densidade de ideias como o número de proposições para cada 10 palavras do texto. Outros autores, como Brown *et al.* (2008), dividem pelo número de palavras apenas. Neste trabalho, seguiremos a primeira convenção.

al., 2010). Mais especificamente, uma proposição pode ser de três tipos (CHAND *et al.*, 2010):

- **Predicação:** consiste de um predicador (por exemplo, um verbo simples como *vender* ou um predicado complexo como *tomar conta*) acompanhado de seus argumentos (agente, tema, instrumento, beneficiário, etc.).
- **Modificação:** consiste de um atributo (um adjetivo) que modifica uma entidade (um substantivo).
- **Conexão:** consiste da ligação entre duas proposições, por meio de uma relação que se estabelece entre elas (pode ser realizada, por exemplo, por meio de uma conjunção coordenativa ou subordinativa).

A título de exemplo, a Tabela 4 mostra algumas sentenças, em língua inglesa, seguidas pelas proposições subjacentes a cada uma e pelo valor correspondente de densidade de ideias. No primeiro exemplo, a primeira proposição é uma predicação envolvendo o verbo ser (*is, big, the house*), enquanto a segunda é uma modificação (*house, John's*). Na segunda sentença, vemos um rephraseamento da primeira, em que há manutenção do número de proposições e elevação do número de palavras, resultando em menor densidade de ideias. No último exemplo, vemos uma proposição de conexão, a quarta proposição (*so*), que carrega ideia de causalidade.

Turner e Greene (1977) desenvolveram um manual, apresentando regras que possibilitam extrair de um fragmento discursivo suas ideias. Neste manual, os conceitos são rotulados como agente, tema, objetivo, etc., seguindo-se a gramática de casos de Fillmore (1968), Fillmore (1969). Kemper e seus colaboradores realizaram modificações no manual de Turner e Greene (1977), adaptando-o para a análise de narrativas orais de pacientes com DA. Chand *et al.* (2010) utilizaram ambos estes manuais para criar seu próprio, voltado à análise da produção oral de idosos, tanto saudáveis quanto portadores de DA e CCL, com vistas à detecção do progresso da deterioração semântica, o que o torna adequado a uso neste trabalho e justifica sua escolha.

Conforme já mencionado, a análise de produção discursiva tem ganhado importância no cenário de avaliações clínicas de linguagem com vistas ao diagnóstico de demências. Um dos estudos que comprova esse fato é o Estudo das Freiras (SNOWDON *et al.*, 1996), que acompanhou um conjunto de 678 freiras, realizando exames comportamentais, neurológicos e, após a morte, análises de neuropatologias no cérebro. Esse estudo empregou a medida de densidade de ideias para analisar narrativas escritas produzidas pelas participantes em sua juventude, e comprovou ser esta bastante eficaz na predição do diagnóstico de doenças como a DA 50 anos mais tarde (CHAND *et al.*, 2010).

Ainda nesse estudo, a densidade de ideias se mostrou capaz não apenas de prever com acurácia as chances de o paciente desenvolver a DA no final de sua vida, mas também de determinar a presença de emaranhados neurofibrilares e de detectar mudanças mnésicas e cognitivas

Tabela 4 – Exemplos de proposições e densidade de ideias (extraídos de Chand *et al.* (2010)).

Sentença	Proposições	Palavras	Densidade
John's house is big.	1. is, big, the house 2. house, John's	4	5,0
The house of John is big.	1. is, big, the house 2. of John	6	3,3
It was 80 years ago.	1. was, it, years ago 2. years ago, 80	5	4,0
I had a sister and 3 brothers so we were always busy.	1. had, I, a sister 2. had, I, brothers 3. brothers, 3 4. so 5. were, we, busy 6. busy, always	12	5,0

sutis que, apesar de induzirem mudanças comportamentais, não resultaram em diagnóstico de demência. Além disso, essa medida é a única que computa diretamente a habilidade do sujeito em usar seu conhecimento de mundo para estruturar proposições na fala espontânea (CHAND *et al.*, 2010). Brown *et al.* (2008) apontam ainda estudos na literatura que mostram que a densidade de ideias também se relaciona a aspectos como inteligibilidade (KINTSCH, 1998; KINTSCH; KEENAN, 1973), memória (THORSON; SNYDER, 1984), qualidade de escrita de estudantes (TAKAO; PROTHERO; KELLY, 2002) e ao envelhecimento (KEMPER; THOMPSON; MARQUIS, 2001; KEMPER; SUMNER, 2001).

Manuais como o de Turner e Greene (1977) e o de Chand *et al.* (2010) descrevem mecanismos manuais de contagem de proposições. Porém, essa tarefa exige treinamento extensivo e demanda muito tempo. Por isso, foi desenvolvido um software denominado CPIDR¹¹ (*Computerized Propositional Idea Density Rater*, pronunciado como “spider”) (BROWN *et al.*, 2008), capaz de automatizar a contagem de proposições e o cálculo da densidade de ideias.

O CPIDR emprega um método bastante barato de cálculo da densidade de ideias, que necessita unicamente de um etiquetador morfossintático como ferramenta base. Sabe-se pela literatura que a densidade proposicional de um fragmento textual pode ser aproximada pelo número de verbos, adjetivos, advérbios, preposições e conjunções dividido pelo número total de palavras (SNOWDON *et al.*, 1996). O CPIDR utiliza *regras de reajuste* para refinar o valor dado por essa aproximação, conseguindo, por meio desse pós-processamento, valores precisos de densidade de ideias (BROWN *et al.*, 2008).

A Figura 9 mostra um exemplo de uso do CPIDR; do lado esquerdo, é colocada a entrada, “John's house is big.” (a primeira sentença de exemplo da Figura 4) e, do lado direito, são

¹¹ <<http://www.ai.uga.edu/caspr/>>

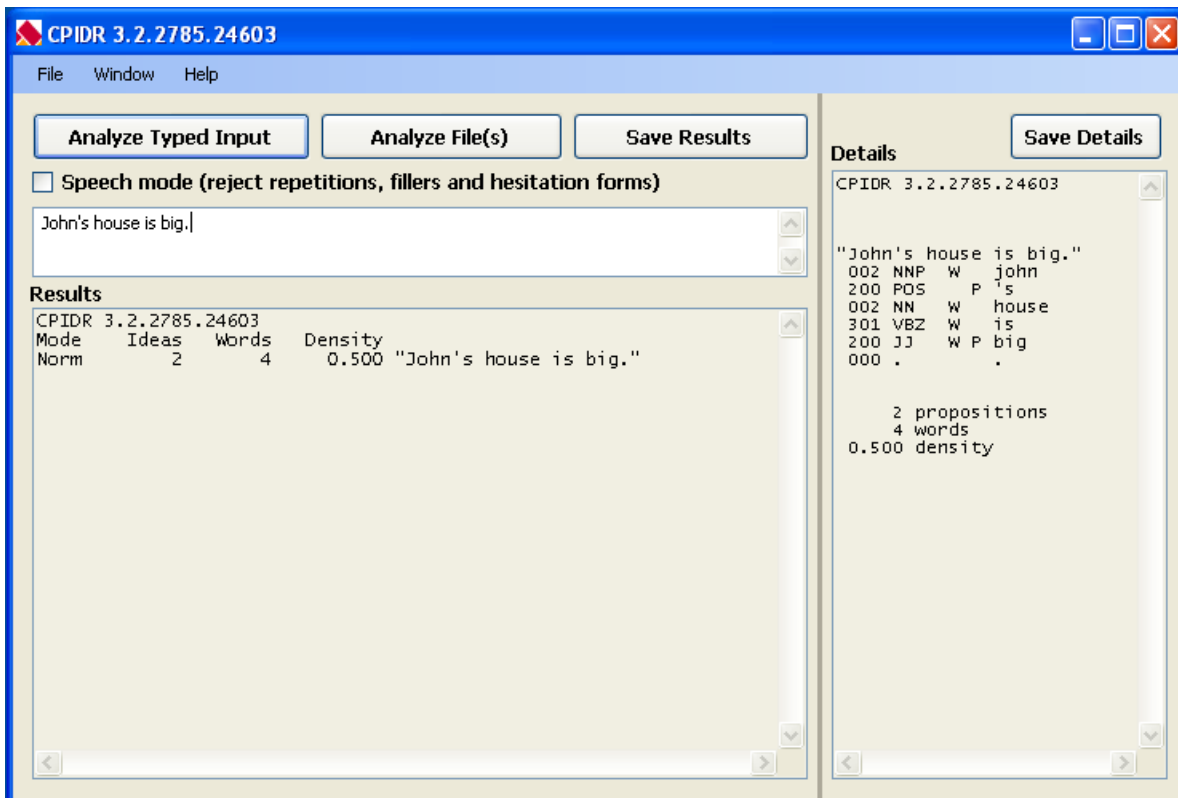


Figura 9 – Exemplo de uso do CPIDR.

exibidos o resultado do etiquetador morfossintático, o número de proposições, o número de palavras e a densidade de ideias. O CPIDR possui também um *modo fala*, que rejeita repetições, preenchedores e hesitações, sendo assim apropriado para transcrição de fala não editada. A Figura 10 mostra um exemplo, contendo a sentença “*John’s house is uh is big.*”; note que o número de proposições não varia, apesar do preenchedor *uh* e da repetição do *is*.

3.2.3.2 A densidade de CSU’s

Uma medida semelhante à densidade de ideias é a **Taxa de CSU’s para cada 100 palavras** (referida aqui apenas como *taxa de CSU’s*). Uma CSU (*Clause-like Semantic Unit*) é definida como uma cadeia de palavras gramaticalmente conectadas, e sua taxa mede a coesão semântica nas sentenças e a habilidade do sujeito em agrupar palavras em sintagmas e dar uma indicação do fluxo da fala (THOMAS *et al.*, 2005; HOLMES; SINGH, 1996). Usa-se o termo *clause-like* porque, na fala agramática (característica de alguns tipos de afasias), diversas cláusulas são deixadas incompletas pelo falante (HOLMES; SINGH, 1996).

Para a separação em CSU’s, são utilizadas 13 regras, apresentadas no apêndice de Bucks *et al.* (2000). Holmes e Singh (1996) apresentam o seguinte exemplo de divisão, onde o sinal “|” é utilizado como separador de CSU’s:

*I went to the market where I met my friend | but I didn’t recognize him at first
| I was hungry | and we had the money | so we went to a restaurant.*

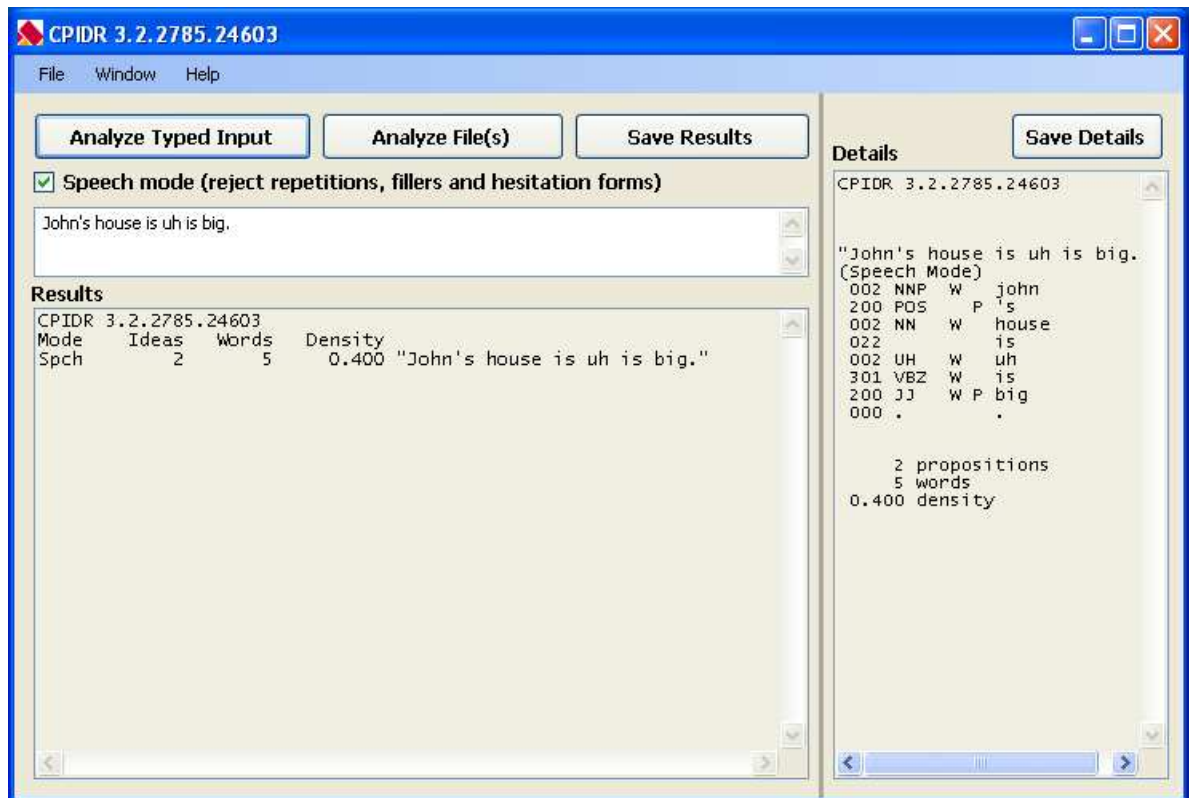


Figura 10 – Exemplo de uso do CPIDR no modo fala.

Quanto maior a taxa de CSU's, menores são essas unidades em número de palavras, o que indica maior dificuldade de articulação por parte do sujeito.

3.2.4 Medidas de semântica latente

A Análise de Semântica Latente (*Latent Semantic Analysis*, LSA) (DUMAIS *et al.*, 1988; DEERWESTER *et al.*, 1990) é uma técnica originalmente concebida para melhorar o desempenho de sistemas de recuperação de informação (DUMAIS, 2004). O principal problema de sistemas como buscadores na Web é a apresentação de resultados irrelevantes para o usuário, não correlacionados à intenção subjacente à cadeia de busca. Grande parte dos sistemas de busca utiliza, como base, um casamento lexical entre a cadeia de busca do usuário e as páginas indexadas (DUMAIS, 2004). Uma possibilidade de geração de resultados irrelevantes, nesse caso, vem da relação muitos-para-muitos existente, nas línguas naturais, entre as **palavras** e os **conceitos** ou **significados** a que elas podem remeter; essa relação pode ser descrita em termos de dois fenômenos linguísticos, a **sinonímia** e a **polissemia**.

A *sinonímia* acontece quando várias palavras se associam ao mesmo conceito, enquanto a *polissemia* ocorre quando uma mesma palavra se refere a diversos conceitos. São exemplos de sinônimos *casa* e *moradia*, que são palavras que se referem basicamente à mesma ideia, e são exemplos de polissemia palavras como *banco* e *manga*, que podem cada uma se referir a diversos conceitos (no primeiro caso, por exemplo, a palavra pode se referir a um assento ou

a uma instituição financeira). Se apenas um casamento lexical simples, literal, é realizado, um usuário buscando por *casa* pode não encontrar um documento relevante que utiliza o termo *moradia*, e outro buscando por *banco* como instituição financeira pode encontrar referências ao objeto. O principal objetivo da LSA é detectar automaticamente relações de sinonímia e polissemia entre palavras, melhorando a qualidade dos resultados obtidos na busca.

Uma característica muito importante da LSA é que a técnica não utiliza nenhuma ferramenta ou recurso de PLN, como etiquetadores, analisadores, dicionários, WordNet's, redes semânticas ou representações de conhecimento. Trata-se de um método puramente estatístico, não-supervisionado, que recebe como entrada apenas uma grande quantidade de textos, gera uma matriz termo-por-documento e procura estabelecer relações de semelhança úteis a tarefas de recuperação de informação ou tarefas semelhantes (DUMAIS, 2004).

A LSA consiste basicamente em 4 passos (DUMAIS, 2004):

1. *Geração da matriz termo-por-documento (TxD)*: inicialmente, uma grande quantidade de textos é transformada em uma matriz, onde cada linha representa uma palavra e cada coluna representa uma unidade maior, como uma sentença, um parágrafo ou mesmo o texto todo, dependendo da aplicação desejada. O valor de cada célula é a frequência de aparição da palavra no fragmento correspondente. Note que a ordem em que as palavras aparecem no texto não é utilizada na análise.
2. *Transformação da matriz termo-por-documento*: ao invés de utilizar a matriz TxD pura, com os valores originais de frequência, os valores podem ser transformados, para obtenção de melhor desempenho. A transformação mais comumente utilizada é a *tf-idf* (frequência do termo vezes o inverso da frequência do documento).
3. *Redução de dimensionalidade*: o próximo passo consiste em aplicar uma Decomposição em Valores Singulares (*Singular Value Decomposition*, SVD) sobre a matriz TxD. Nesse passo, os k maiores valores singulares da matriz TxD são mantidos, enquanto os outros são colocados como 0. Ao final desse processo, a representação resultante é a melhor aproximação k -dimensional da matriz original, segundo o critério dos mínimos quadrados.
4. *Recuperação no espaço reduzido*: em seguida, as palavras e os documentos são representados como vetores no mesmo espaço de dimensão reduzida. Como estão no mesmo espaço, pode-se calcular a distância palavra-palavra, palavra-documento e documento-documento. Além disso, pode-se também representar as cadeias de busca do usuário no mesmo espaço, com base nas palavras que constituem a busca (pode-se calcular o centroide ponderado das palavras constituintes, por exemplo, para obter um vetor para a busca), e então obter os documentos mais próximos segundo alguma métrica de distância. A medida mais utilizada, segundo (DUMAIS, 2004), é a distância do cosseno, por ter apresentado melhores desempenhos na prática.

Além da aplicação originalmente pretendida na recuperação de informações, a LSA encontra utilidade em diversas outras tarefas (DUMAIS, 2004).

Uma delas é a **recuperação multilíngue de informações**. Nesse caso, deseja-se que uma cadeia de busca em um idioma consiga encontrar documentos relevantes escritos em outro idioma. Esse é o caso extremo da divergência de vocabulário entre usuários de sistemas de busca e escritores de documentos, divergência que a LSA originalmente se propunha a resolver (DUMAIS, 2004). Para essa tarefa, são empregados córpus paralelos, onde as sentenças originais e suas respectivas traduções encontram-se alinhadas, bem como uma noção ligeiramente diferente da matriz termo-por-documento (DUMAIS, 2004).

A LSA foi aplicada ainda a tarefas relativas à **modelagem da memória humana**, como pontuação de redações (LANDAUER *et al.*, 1997; FOLTZ; LAHAM; LANDAUER, 1999), testes de vocabulário (LANDAUER; DUMAIS, 1997; DUMAIS, 2004), para a medição de coerência textual, entre outras. A medição de coerência textual é de particular interesse neste trabalho. Já abordamos na seção 3.2.3.1 que Kintsch foi responsável por criar métodos para representar textos por meio de proposições; Foltz, Kintsch e Landauer (1998) utilizou a LSA para medir a coerência textual automaticamente, encontrando alta correlação entre a pontuação da LSA e a de testes humanos (DUMAIS, 2004). Além disso, ela foi utilizada por Dunn *et al.* (2002) para medir a capacidade de relembrar fatos em testes de memória; nesse estudo, a LSA foi comparada a outros métodos de pontuação já estabelecidos, baseados em unidades temáticas que o sujeito conseguiu lembrar, e se mostrou altamente correlacionada a eles (DUMAIS, 2004). Nesse mesmo trabalho, a LSA foi capaz de detectar problemas de memória em pacientes com déficits cognitivos, o que a torna de grande interesse para este trabalho.

A LSA já está disponível no Coh-Metrix original. As métricas relacionadas a ela encontram-se na Tabela 1, e estão reproduzidas na Tabela 5 para facilitar a leitura do texto. Como se pode notar, é medida a correlação entre sentenças e parágrafos adjacentes, e entre todas as sentenças de um parágrafo, entre outras.

Tabela 5 – Métricas de LSA do Coh-Metrix 3.0.

LSA			
40	LSASS1	LSAassa	LSA overlap, adjacent sentences, mean
41	LSASS1d	LSAassd	LSA overlap, adjacent sentences, standard deviation
42	LSASSp	LSApssa	LSA overlap, all sentences in paragraph, mean
43	LSASSpd	LSApssd	LSA overlap, all sentences in paragraph, standard deviation
44	LSAPP1	LSAppa	LSA overlap, adjacent paragraphs, mean
45	LSAPP1d	LSAppd	LSA overlap, adjacent paragraphs, standard deviation
46	LSAGN	LSAGN	LSA given/new, sentences, mean
47	LSAGNd	n/a	LSA given/new, sentences, standard deviation

TRABALHOS RELACIONADOS

Este capítulo apresenta trabalhos que realizaram análises semelhantes às realizadas neste trabalho de mestrado, agrupados conforme o nível de análise linguística empregado.

4.1 Abordagem lexical

O trabalho de [MacWhinney et al. \(2010\)](#) (*Automated analysis of the Cinderella story*) utilizou ferramentas do TalkBank¹, um projeto amplo que provê métodos e ferramentas para o estudo de uma variedade de tipos de linguagem, para a análise automatizada da fala de pacientes afásicos, utilizando textos do projeto AphasiaBank².

O objetivo do TalkBank é construir uma base de dados compartilhada de recursos multimídia sobre a comunicação humana ([MACWHINNEY et al., 2010](#)). O projeto envolve outros sub-projetos, como o BillingBank³, de aquisição de segunda língua, o CABank⁴, de análise conversacional, entre outros. Nestes, inclui-se o AphasiaBank, um projeto que coleta e analisa amostras em áudio e vídeo de discurso afásico e não-afásico ao longo de diversas tarefas, com o objetivo de melhorar o tratamento da afasia ([MACWHINNEY et al., 2010](#)).

O estudo de [MacWhinney et al. \(2010\)](#) utilizou narrações da história da Cinderela, produzidas por sujeitos normais (n = 25) e por sujeitos afásicos (n = 24), transcritas utilizando o formato CHAT e um conjunto de programas denominado CLAN ([MACWHINNEY, 2000](#)). O formato de transcrição CHAT tem sido desenvolvido nos últimos 30 anos para capturar características relevantes do uso da linguagem para análise em diversas disciplinas, como aquisição de primeira e segunda linguagem, análise conversacional, etc.; os programas CLAN, projetados para lidarem com o formato CHAT, permitem analisar diversas estruturas linguísticas

¹ <<http://talkbank.org/>>

² <<http://talkbank.org/AphasiaBank/>>

³ <<http://talkbank.org/BillingBank>>

⁴ <<http://talkbank.org/CABank>>

@G: Cinderella
 *PAR: &guh a little bit I think, yeah.
 *PAR: was [//] what was the name ?
 PAR: Secerundid [: Cinderella] [nk].
 *PAR: she was &guh &b angel for legwood@n. [+ jar]
 *PAR: she was &guh &f for fendle@n for someone else. [+ jar]
 *PAR: the other children [/] &gr &d children for her are three children or whatever . [+ es]
 PAR: with her it was very closed [wu] walking [* wu] in generalis@n . [+ jar]
 *PAR: >h &th &p pezzels@n are going for the party.
 PAR: and she was &gf fen@n people [wu] for prezzled@n (.) for the present [* wu]. [+ jar]
 *PAR: the present > (...) was s(up)posed to be &uh thirty [/] &t &uh thirty or something. [+ es]
 PAR: she &gch &er had a ranned@n from home she &ha huddled [wu]. [+ jar]
 *PAR: the &guh (..) people were +//.
 *PAR: they found her letter.
 PAR: and <the pezzes@n> [//] &gw the other people wed [wu] they found her.
 *PAR: found her for the prezzled@n and the calls this one so. [+ jar]

Figura 11 – Transcrição no formato CHAT da história da Cinderela (extraído de (MACWHINNEY *et al.*, 2010)).

e discursivas (MACWHINNEY *et al.*, 2010).

A Figura 11 mostra um exemplo de transcrição no formato CHAT. O formato inclui diversos códigos de erros (como [* wu], que indica que o erro é uma palavra real e que a palavra pretendida é desconhecida), códigos ao nível da enunciação (como [+ jar], que indica uso de jargão), e códigos para repetição ([/]), revisão ([//]), fragmentos de palavras e preenchedores (&), trocas ([: *palavra pretendida*]) e pausas (.) (MACWHINNEY *et al.*, 2010).

Com base nas transcrições, os autores utilizaram o CLAN para extrair características dos textos e, por meio delas, comparar os sujeitos normais com os afásicos. A primeira característica analisada foi a relação **Tipo por Token**, de onde se concluiu que os sujeitos afásicos possuem um discurso lexicalmente pobre quando comparado ao dos sujeitos normais (os sujeitos afásicos produziram 526 tipos de palavras, num total de 5330 tokens, enquanto os sujeitos normais produziram 839 tipos, para 13309 *tokens*).

O CLAN foi utilizado também para levantar os 10 substantivos e os 10 verbos mais frequentes tanto na fala normal quanto na afásica. Com base nos substantivos, os autores notaram que o discurso afásico, apesar de capturar os tópicos principais da história narrada, é mais vago, mais abstrato, que o dos normais (os afásicos, por exemplo, utilizam palavras como *man*, *shoe*, *girl*, menos intimamente ligadas à história da Cinderela que palavras como *dress*, *fairy* e *stepdaughter*, encontradas no discurso dos normais). Com base nos verbos, notou-se que os afásicos fazem uso mais intenso de *light verbs*, isto é, verbos com pouco conteúdo semântico por si só e assim formam um predicado com uma expressão adicional, por exemplo, *tomar conta*,

indicando uma diversidade mais limitada para verbos (MACWHINNEY *et al.*, 2010).

Os autores também exibiram o comando, que pode ser dado ao CLAN, para computar o número de erros, mas não compararam os textos com base nesses erros.

O site do AphasiaBank possui links para um documento de duas páginas resumindo as orientações gerais para transcrição, um documento explicando a codificação de erros, um manual mais detalhado para treinamento de transcrição, e os manuais completos do CHAT e do CLAN (MACWHINNEY *et al.*, 2010). Esses documentos poderão ser de grande valia para a definição do protocolo de transcrição utilizado neste trabalho. Os autores utilizaram um etiquetador morfosintático para anotar os textos e relatam que, apesar de ter sido treinado em um domínio de textos bem formados, o etiquetador apresentou bom desempenho nos textos orais afásicos. Apesar de utilizarem análise computadorizada dos textos transcritos, os autores não empregam *classificação* automática dos textos, o que é um diferencial do presente trabalho de mestrado.

O trabalho de Thomas *et al.* (2005) (*Automatic Detection and Rating of Dementia of Alzheimer Type through Lexical Analysis of Spontaneous Speech*) apresenta diversas abordagens lexicais para a detecção e quantificação da DA, com o objetivo de explorar se técnicas automáticas baseadas na análise de fala espontânea podem fornecer medidas objetivas do nível de demência de pacientes com DA (THOMAS *et al.*, 2005).

No estudo, o córpus utilizado foi composto com transcrições do *Atlantic Canada Alzheimer's Disease Investigation of Expectations* (ACADIE), um estudo sobre a droga donepezil (THOMAS *et al.*, 2005). O estudo empregou oito métricas, extraídas dos textos do córpus:

1. Taxa de adjetivos (número de adjetivos dividido pelo número total de palavras).
2. Taxa de substantivos.
3. Taxa de pronomes.
4. Taxa de verbos.
5. Relação Tipo por *Token* (*TTR*).
6. Índice de Brunét (*W*).
7. Estatística de Honoré (*R*).
8. Densidade de CSU's.

Os valores dessas métricas foram utilizados em duas tarefas de classificação, uma com duas classes (alta e baixa) e outra com quatro classes (normal, leve, moderada, severa). As classes foram tomadas com base na pontuação de cada paciente no Mini-Exame do Estado

Cenário	Melhor acurácia (%)
Duas classes: alto e baixo	69,6
Quatro classes	50,0
Duas classes: normal e severo	94,5
Duas classes: normal e leve	75,3

Tabela 6 – Sumário dos resultados de acurácia de classificação de [Thomas et al. \(2005\)](#).

Mental ([FOLSTEIN; FOLSTEIN; MCHUGH, 1975](#)). Adicionalmente, foram feitas duas outras tarefas de classificação binária, utilizando-se sub-conjuntos das quatro classes. Foram realizados testes utilizando-se diversos cenários de seleção de atributos, e os melhores resultados são apresentados na tabela 6.

Os autores do trabalho asseveram que, quando se deseja desenvolver novos testes de quantificação de demências que superem as deficiências dos métodos atuais, os pesquisadores devem procurar por métodos automáticos e objetivos que façam uso da análise de fala espontânea ([BUCKS et al., 2000](#)). Os autores concluem que soluções puramente computacionais oferecem uma alternativa viável às abordagens padrões de diagnóstico do nível de demência dos pacientes, embora reconheçam que mais deve ser feito para melhorar a acurácia desses métodos.

4.2 Abordagem baseada em complexidade sintática

O trabalho de [Roark, Mitchell e Hollingshead \(2007\)](#) (*Syntactic complexity measures for detecting Mild Cognitive Impairment*) analisou gravações em áudio de 55 exames neuropsicológicos administrados no *Layton Aging & Alzheimer's Disease Center*. Os sujeitos foram divididos entre saudáveis e com CCL segundo sua *Clinical Dementia Rating* (CDR, [Morris \(1993\)](#)). Foram aplicados diversos testes nos sujeitos, mas apenas o resultado dos testes *Wechsler Logical Memory I e II* foram utilizados. Esses testes consistem em ouvir uma história curta (três sentenças) e recontá-la duas vezes: uma imediatamente após a escuta e outra após 30 minutos de atividades descorrelacionadas ao teste. Cada narrativa foi transcrita e analisada, manual e automaticamente.

No trabalho, foram utilizadas cinco medidas de complexidade sintática:

1. Complexidade de Yngve.
2. Complexidade de Frazier.
3. Distância de dependência.
4. Nível de desenvolvimento.
5. Entropia cruzada.

O primeiro experimento realizado no trabalho consistiu em computar as métricas de complexidade sobre os textos de três corpuses, de maneira manual e automática, e calcular as correlações entre as medidas. Conclui-se que havia, em todas as medidas, alta correlação entre a medida calculada de maneira manual e de maneira automática, o que é um bom indicativo das perspectivas do presente trabalho, que lida com análises automáticas.⁵

No segundo experimento, verificou-se a existência de diferença estatisticamente significativa entre os grupos (normais e CCL), em ambos os testes (I e II), para cada uma das cinco medidas de complexidade. Notou-se que medidas diferentes apresentam padrões de comportamento distintos quando aplicadas às amostras de linguagem; por exemplo, para a anotação manual, a complexidade de Frazier conseguiu distinguir os grupos no teste I, mas não no teste II, enquanto a complexidade de Yngve não conseguiu distinguir no teste I, mas conseguiu no teste II. Os autores concluem, então, que essas medidas são complementares, e que pode-se beneficiar do uso de múltiplas métricas. Esse resultado também é de grande interesse para este trabalho, uma vez que o Coh-Matrix-Dementia fará uso conjunto de diversas métricas para melhorar sua acurácia.

O trabalho de Roark *et al.* (2007) (*Automatically derived spoken language markers for detecting Mild Cognitive Impairment*) realizou uma análise muito semelhante à de Roark, Mitchell e Hollingshead (2007). Foram utilizadas gravações em áudio de 44 exames também administrados no *Layton Aging & Alzheimer's Disease Center*. Os sujeitos foram novamente divididos entre saudáveis e com CCL segundo sua CDR e foram utilizados os recontos narrativos dos testes *Wechsler Logical Memory I e II*.

O estudo utilizou medidas divididas em duas categorias: medidas de **complexidade sintática** e medidas **fonológicas**. As medidas de complexidade sintática utilizadas foram o número de palavras por sentença e a complexidade de Yngve. As medidas fonológicas utilizadas foram a taxa verbal, a taxa de fonação, a duração média das pausas e a taxa de pausas padronizada, calculadas da seguinte maneira: seja W o número de palavras, N o número de pausas, P o tempo total de pausa e T o tempo total da amostra; assim:

$$\begin{aligned} \text{Taxa verbal} &= \frac{W}{T} \\ \text{Taxa de fonação} &= \frac{(T - P)}{T} \\ \text{Duração média das pausas} &= \frac{P}{N} \\ \text{Taxa de pausas padronizada} &= \frac{W}{N} \end{aligned}$$

⁵ É importante notar que as ferramentas de PLN básicas para a língua portuguesa (etiquetadores, analisadores sintáticos, etc.) não possuem, na data de escrita deste trabalho, desempenho comparável aos das ferramentas correspondentes para o inglês. Porém, quando ferramentas melhores se tornarem disponíveis, o Coh-Matrix-Dementia será beneficiado sem esforços, já que é alimentado diretamente pela saída desses sistemas.

O experimento do trabalho consistiu em verificar a existência de diferença estatisticamente significativa entre os grupos (normal e CCL) para cada uma das seis medidas adotadas, tanto na extração manual quanto na automática. O número de palavras por cláusula apresentou diferença estatisticamente significativa entre normais e CCLs em ambos os testes, tanto na extração manual quanto na automática. A complexidade de Yngve apresentou resultados apenas no teste II, tanto na análise manual quanto na automática. Dentre as métricas fonológicas, o único caso em que houve diferença estatisticamente significativa foi com a taxa de pausas padronizada na análise manual no teste I; todas as outras métricas em todos os outros casos não produziram diferenças significativas.

Os autores concluem, mais uma vez, que a extração automática dessas métricas pode ser efetiva, pois a diferença estatisticamente significativa foi mantida em todos os casos, exceto um. Apesar de os resultados utilizando as métricas fonológicas não apresentarem diferença estatística, os autores acreditam que esse fato se deva ao tamanho reduzido do cópulo. De qualquer maneira, o trabalho fornece mais evidências de que as medidas de complexidade sintática tratadas aqui têm potencial de sucesso na separação entre sujeitos normais e sujeitos com CCL.

Os trabalhos de [Roark et al. \(2007\)](#) e [Roark, Mitchell e Hollingshead \(2007\)](#) foram estendidos por [Roark et al. \(2011\)](#) (*Spoken language derived measures for detecting Mild Cognitive Impairment*), incluindo um número maior de participantes (74 sujeitos, 37 normais e 37 com CCL), mais métricas e o treinamento de um classificador com base nessas medidas. As medidas de complexidade sintática utilizadas foram⁶:

1. Complexidade de Yngve.
2. Complexidade de Frazier.
3. Distância de dependência.
4. Entropia cruzada.
5. Densidade de ideias.
6. Densidade de conteúdo.

Os autores empregaram ainda 10 medidas fonológicas (pausas por reconto, tempo total de pausa, taxa de pausas padronizada, tempo total de fonação, tempo total de locução, taxa de fonação, taxa de fonação transformada e taxa verbal). Mais uma vez, os autores calcularam a correlação entre o cômputo manual e o automático de cada métrica, alcançando grande correlação entre eles (mais que 87%).

⁶ Os autores utilizaram, ainda, métricas calculadas sobre versões transformadas das árvores sintáticas, mas relatam que não houve grandes diferenças de comportamento com as alterações realizadas nas árvores.

Novamente, os autores verificaram se havia de diferença estatisticamente significativa entre sujeitos normais e com CCL em cada um dos dois testes para cada atributo. Dentre os atributos sintáticos, notou-se o mesmo padrão de complementaridade observado nos trabalhos anteriores; por exemplo, a complexidade de Yngve e a distância de dependência mostraram diferença entre os grupos no teste II, mas não no teste I, enquanto a entropia cruzada mostrou diferença no teste I, mas não no teste II. Duas medidas mostraram diferença significativa em ambos os casos: palavras por cláusula e densidade de conteúdo. Os outros atributos não exibiram diferenças estatisticamente significativas.

4.3 Abordagem baseada em densidade de ideias

Bryant *et al.* (2013) (*Propositional Idea Density in aphasic discourse*) conduziram um estudo com o objetivo de investigar o quanto a densidade de ideias era diferente entre os discursos afásico e não-afásico, bem como determinar como a densidade de ideias se relaciona com a intensidade da afasia e com medidas já estabelecidas de outros aspectos da informatividade. Foram analisadas entrevistas de 50 participantes com afasia e de 49 controles saudáveis; toda a fala do entrevistador foi excluída, e os textos foram analisados utilizando o CPIDR 3.2 em modo fala e o SALT⁷ (*Systematic Analysis of Language Transcripts*) versão 8. Além da densidade de ideias, outras medidas foram extraídas dos textos: a relação tipo por *token*, o número de palavras diferentes (os tipos), o comprimento médio da elocução (MLU) e o número de elocuições.

Os autores, inicialmente, procuraram estabelecer a porcentagem de concordância entre a análise do CPIDR e a análise manual quando aplicadas sobre textos afásicos, uma vez que a confiabilidade do CPIDR comparada aos humanos não havia sido testada nesse cenário. Dentre as 50 transcrições de sujeitos afásicos, o percentual de concordância variou entre 87,34% e 100%, com porcentagem total de concordância de 99,57%, sendo que 48 das 50 transcrições apresentaram concordância acima de 95%; dentre os 49 controles sadios, a concordância variou entre 98,25% e 100%, com porcentagem total de 99,74%. Esses dados mostram que o método empregado pelo CPIDR é bastante robusto ao lidar com textos mal-formados, como os dos afásicos, o que reforça a justificativa de seu uso no presente trabalho.

Em seguida, realizou-se um teste de significância estatística para determinar se havia diferença significativa entre os afásicos e os controles com relação às métricas analisadas. A hipótese de pesquisa era que a densidade de ideias era reduzida em sujeitos afásicos, quando comparados com sujeitos normais. No teste, a densidade de ideias, o número de palavras diferentes, a relação tipo por *token* e o MLU apresentaram diferenças; o número de elocuições não apresentou diferença estatisticamente significativa.

Os autores realizaram ainda um teste de correlação, para determinar a existência de correlação entre as métricas utilizadas e a intensidade da afasia. A hipótese, nesse caso, era que,

⁷ <<http://www.saltsoftware.com/>>

Tabela 7 – Sumário dos testes de significância de Bryant *et al.* (2013) (X sinaliza significância).

Métrica	Diferença	Correlação
Densidade de ideias	X	X
Número de palavras diferentes	X	X
Relação tipo por <i>token</i>	X	
Comprimento médio da elocução	X	X
Número de elocuições		

quanto mais intensa era a afasia, menor era a densidade de ideias presente no discurso. No teste de correlação, a densidade de ideias, o número de palavras diferentes e o MLU apresentaram correlação estatisticamente significativa, enquanto a relação tipo por *token* e o número de elocuições não apresentaram correlação.

A Tabela 7 mostra um resumo dos resultados de significância apresentados acima; a primeira coluna mostra o resultado do teste de diferença afásicos vs. não-afásicos, e a segunda mostra o resultados do teste de correlação com a intensidade da afasia. Esses resultados mostram o potencial apresentado pela densidade de ideias e pelo comprimento médio da enunciação na análise do discurso de pacientes com quadros demenciais, tanto na identificação da doença quanto em sua quantificação.

4.4 Abordagem baseada em traços semânticos e categorias morfossintáticas

O trabalho de Peintner *et al.* (2008) (*Learning diagnostic models using speech and language measures*) realiza análises lidando com a demência frontotemporal (DFT). Os participantes foram divididos em quatro grupos: DFT comportamental, APPNF, DS e controles. Os participantes tiveram sua fala gravada e transcrita, e tanto o áudio quanto a transcrição foram analisados, dando origem a diversas características **fonológicas**, **morfossintáticas** e de **traços semânticos**, essas últimas extraídas do LIWC (veja seção 3.1.3).

As características **fonológicas** empregadas no trabalho são: média e desvio padrão de fricativas vozeadas, vogais, nasais, soantes, fones, aproximantes, consoantes desvozeadas, fricativas desvozeadas, grupos soante + fricativa vozeada, obstruintes, consoantes, consoantes soantes, fricativas, obstruintes vozeadas, consoantes vozeadas, oclusivas desvozeadas, oclusivas vozeadas, oclusivas; fonemas por segundo; pausas por trecho de fala ininterrupta; hesitações; hesitações por trecho de fala ininterrupta; fonemas por segundo de fala ininterrupta.

As características do LIWC empregadas são as frequências de: interjeições, verbos, advérbios, adjetivos, pronomes, determinantes por nome, verbos por nome, pronomes por nome, palavras funcionais, nomes, todas as outras, palavras de seis letras, palavras funcionais, pronomes pessoais, eu, nós, você/vocês, ele/ela, eles/elas, artigos, tempo passado, tempo presente, tempo

Tabela 8 – Resultados (aproximados) obtidos por Peintner *et al.* (2008).

Tarefa	Melhor acurácia (%)	Algoritmo
<i>Cont. x APPNF x DFT x DS</i>	72	MLP
Cont. x Doentes	97	MLP
<i>Cont. x DFT</i>	88	MLP
Cont. x APPNF	100	J48
Cont. x DS	99	MLP
Cont. x APPNF x DS	96	LR

Tabela 9 – Resultados obtidos por Jarrold *et al.* (2010).

Cenário	Acurácia
Alzheimer x Controles	73,0%
Comp. Cognitivo x Controles	82,6%
Depressão x Controles	97,6%

futuro. Outras ainda foram incluídas: xingamento, social, família, amigo, afeto, emoção positiva, emoção negativa, raiva, tristeza, introspecção, sexual, movimento, morte, dinheiro.

Os autores realizaram diversos experimentos, com combinações de **algoritmos** de classificação, **tarefas** de separação e **conjuntos de atributos**. Os **algoritmos** empregados foram Regressão Logística (LR), Multilayer Perceptron (MLP) e J48; as **tarefas** de separação foram: Controles x APPNF x DSF x DS, Controles x Doentes, Controles x DFT, Controles x DS e Controles vs APPNF vs DS; os **conjuntos de atributos** foram: um subconjunto das características do LIWC, as características morfossintáticas, as características fonêmicas e todas. Os resultados (aproximados) encontram-se sumarizados na tabela 8; nela, é apresentado somente o resultado do cenário que apresentou melhor desempenho em cada tarefa.

As linhas marcadas em itálico na tabela (a saber, a primeira e a terceira) são cenários que não são de interesse ao presente trabalho, uma vez que envolvem a DFT, que não será tratada aqui. Nos outros cenários, podemos verificar um desempenho próximo a 100%, o que revela um cenário promissor de pesquisa.

O trabalho de Jarrold *et al.* (2010) (*Language Analytics for Assessing Brain Health: Cognitive Impairment, Depression and Pre-symptomatic Alzheimer's Disease*) tratou de três distúrbios cerebrais: Doença de Alzheimer, Comprometimento Cognitivo e Depressão Clínica. Foram utilizadas as características do LIWC e de **densidade de ideias**, extraídas de transcrições de entrevistas com os sujeitos. Foram realizados experimentos de classificação, separando pacientes com cada uma das três distúrbios dos controles saudáveis. Os algoritmos utilizados foram os mesmos do trabalho de Peintner *et al.* (2008), e os resultados obtidos encontram-se sumarizados na Tabela 9.

4.5 Abordagem em vários níveis

O trabalho de Fraser *et al.* (2012) (*Automated classification of primary progressive aphasia subtypes from narrative speech transcripts*) compartilha, em grande medida, dos objetivos e métodos deste trabalho de mestrado. Nele, os autores analisaram textos produzidos por sujeitos de três grupos: controles saudáveis e duas variantes de APP: APPNF e DS. Os sujeitos foram orientados a produzir narrações da história de Cinderela, que foram então transcritas manualmente e analisadas computacionalmente.

A transcrição seguiu os procedimentos da *Quantitative Production Analysis* (BERNDT *et al.*, 2000), com a exceção de algumas adaptações, que foram feitas para melhor desempenho das ferramentas de PLN. Tais modificações incluem a remoção de pausas e vírgulas e a inclusão de pontos e letras maiúsculas para delimitação de sentenças, entre outras. Os textos modificados dessa maneira foram analisados por ferramentas de PLN, produzindo um total de 58 métricas, capazes de extrair características léxicas, sintáticas e semânticas.

Dentre as 58 características escolhidas para análise nesse trabalho incluem-se:

- o número de palavras da fala transcrita;
- 22 características, baseadas no trabalho de Lu (2010), para avaliar a complexidade sintática de textos;
- 4 outras medidas de complexidade sintática para detectar o declínio cognitivo advindo da idade;
- 13 medidas extraídas de um etiquetador morfossintático⁸;
- frequência de verbos *light* e de verbos *heavy*, distinguidos entre si por sua complexidade semântica;
- 11 características relacionadas a cálculos de frequência, imageabilidade, idade de aquisição e familiaridade;
- 6 características que avaliam fluência e riqueza vocabular, incluindo a conhecida medida *type/token*, utilizada para detectar repetições de palavras;
- 3 características correspondentes a pausas preenchidas, comuns da fala;
- e a velocidade da fala, calculada como o número de palavras dividido pelo tempo do discurso.

⁸ A diferença na produção de substantivos e verbos já havia sido utilizada em estudos sobre os dois tipos de afasia considerados e, além disso, é sabido que pacientes com APPNF omitem palavras de conteúdo.

Tabela 10 – Acurácia (%) dos três classificadores de Fraser *et al.* (2012).

Método	DS x Cont.	APPNF x Cont.	DS x APPNF
Naïve Bayes	92,3	90,0	79,2
Regressão Logística	96,2	93,3	70,8
SVM	100	96,7	75,0
<i>Baseline</i>	61,5	53,3	58,3

Com base nessas métricas, foram treinados três classificadores de paradigmas diferentes: Naïve Bayes, Regressão Logística e SVM. Os resultados obtidos estão reproduzidos na tabela 10. Com base nela, nota-se que foi possível separar controles de afásicos com precisão próxima a 100%, e separar os afásicos entre si com precisão em torno de 80%, com *baseline* dado pela classe majoritária.

COH-METRIX-DEMENTIA: IMPLEMENTAÇÃO

Neste capítulo, são apresentadas as etapas realizadas para atingir os objetivos do trabalho. Na Seção 5.1, é apresentada a arquitetura interna do Coh-Metrix-Dementia; na Seção 5.2, é apresentado, de forma breve, como as métricas já existentes no Coh-Metrix-Port foram recodificadas na nova versão; finalmente, na Seção 5.3, é descrita, em maior detalhe, a forma como as novas métricas propostas foram codificadas.

5.1 Arquitetura do Coh-Metrix-Dementia

A arquitetura geral do Coh-Metrix-Dementia é apresentada na Figura 12. Como entrada para o sistema temos amostras de fala que podem ser transcritas, e tanto a transcrição quanto o áudio são fornecidos como entrada ao Coh-Metrix-Dementia. Na versão atual do sistema, apenas a transcrição é analisada, mas a ferramenta foi projetada de maneira a suportar a extensão por meio de ferramentas de análise de fala. O Coh-Metrix-Dementia, então, fornece como saída um conjunto de métricas, que podem ser utilizadas tanto para análises manuais por parte de profissionais da saúde, quanto para treinamento de classificadores automáticos.

Internamente à ferramenta, os textos de entrada fornecidos são armazenados por objetos da classe `Text`. Cada uma das métricas textuais consiste em uma subclasse da classe base `Metric`. Métricas são agrupadas em categorias, cada uma correspondendo a uma subclasse da classe `Category`. Categorias podem, ainda, ser agrupadas em conjuntos, objetos da classe `MetricsSet`. Objetos da classe `Metric` possuem um método `value_for_text`, que recebe um objeto da classe `Text` como entrada e retorna um valor real, correspondente ao valor da métrica para aquele texto. Objetos das classes `Category` e `MetricsSet` possuem um método `values_for_text`, que retorna um objeto da classe `ResultSet`, responsável por agrupar os valores das métricas e facilitar o acesso a estes.

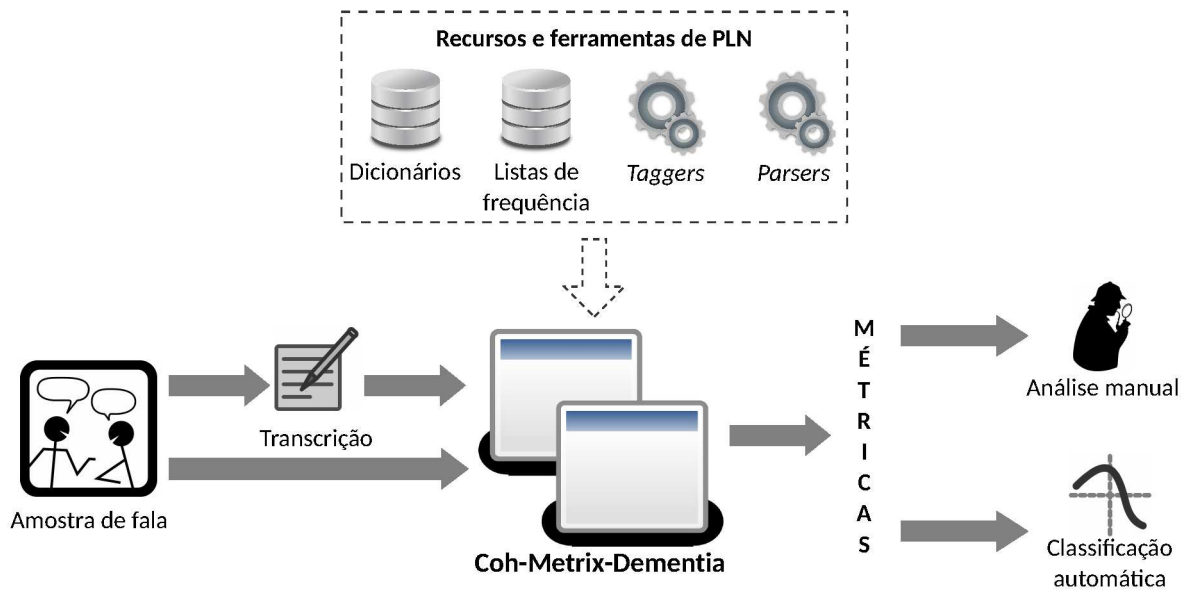


Figura 12 – Arquitetura geral do Coh-Metrix-Dementia.

A fim de que facilitar o cálculo das métricas, o sistema mantém um cache de ferramentas e recursos, codificado como um objeto da classe `ResourcePool`. Esse objeto, acessível durante o cálculo das métricas, calcula e fornece informações acerca do texto, como as etiquetas morfossintáticas, diferentes árvores sintáticas (de constituintes e de dependência), entre outras, bem como recursos de interesse para o cálculo das métricas (como espaços LSA, listas de itens lexicais, entre outros). Esses recursos são calculados de maneira *on demand* pela ferramenta, e armazenados para uso posterior, caso outras métricas necessitem da mesma informação.

Além da biblioteca base, foi construída uma interface Web, que permite que profissionais da área médica e o público em geral possam acessar, de maneira mais amigável, as funcionalidades de análise textual do Coh-Metrix-Dementia. A Figura 13 mostra a tela inicial da interface, onde o usuário pode realizar login, ou se cadastrar, caso ainda não o tenha feito. Quando é feito o login, o usuário é redirecionado para a tela de resumo, exibida na Figura 14 que mostra a lista dos textos já analisados por ele. Quando o usuário deseja submeter um texto novo a ser analisado, a tela da Figura 15 é exibida, onde o usuário pode entrar com os dados do texto. Quando o usuário clica sobre uma linha na lista de textos, é exibida uma visão geral dos valores das métricas, bem como é dada a opção de exportar os dados em diversos formatos; essa tela é exibida parcialmente na Figura 16.

5.2 As Métricas Herdadas do Coh-Metrix-Port

A Tabela 3 mostra as 48 métricas disponíveis no Coh-Metrix-Port. Todas elas foram recodificadas e inseridas no Coh-Metrix-Dementia.

Para métricas que necessitam de etiquetas morfossintáticas, foi utilizada a biblioteca

Welcome!

This is Coh-Metrix-Dementia

Coh-Metrix-Dementia is a system that can analyze texts in terms of several metrics, which are related to cohesion, coherence, and readability, among other textual aspects.

Login

Email

Password ?

Remember me

[Login](#)

Or register

Your full name

Your email

Your institution

Password

Confirm password

[Register](#)

Figura 13 – Interface web do Coh-Metrix-Dementia: tela inicial.

Coh-Metrix-Dementia [Analyze](#) [Explore](#) [Classify](#) [Help](#) [About](#) [John Doe](#) ▾

Your texts

[Search](#) [Submit text](#)

Title	Author	Source	Date	Genre
Hansel and Gretel	Brothers Grimm	https://en.wikipedia.org/wiki/Hansel_and_Gretel	2015-08-14	Fairy tale
Brasil	Diversos	https://pt.wikipedia.org/wiki/Brasil	2015-08-10	Jornalístico

Figura 14 – Interface web do Coh-Metrix-Dementia: lista dos textos do usuário.



Figura 15 – Interface web do Coh-Metrix-Dementia: tela para submissão de um novo texto.

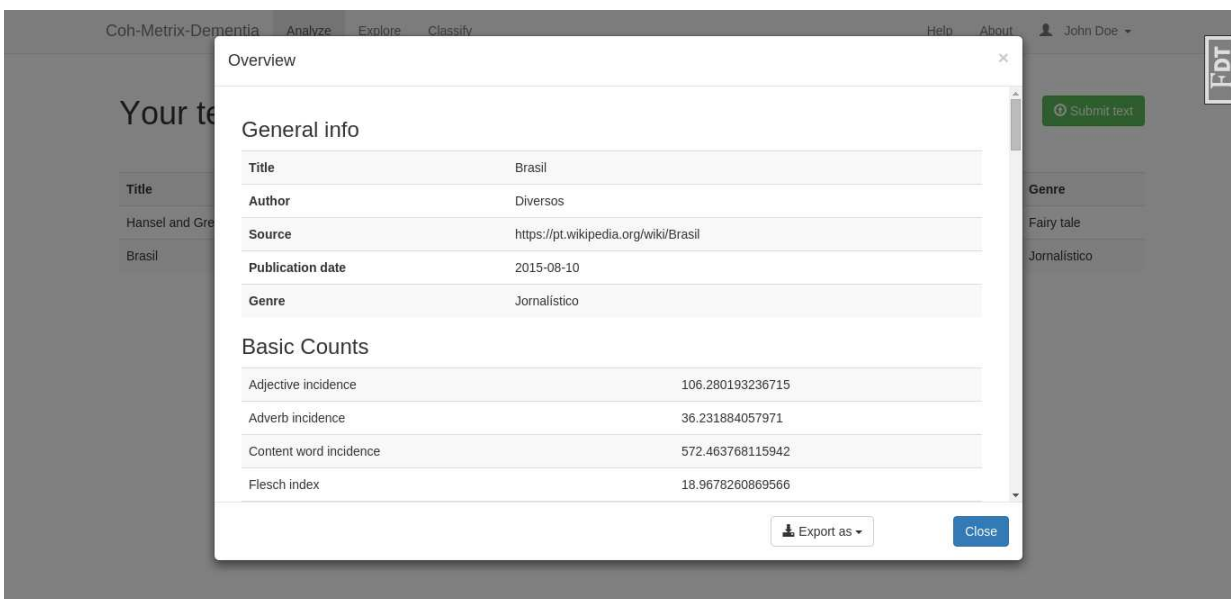


Figura 16 – Interface web do Coh-Metrix-Dementia: visão geral de um texto analisado.

nlpnet¹, de Erick Fonseca (FONSECA; ROSA, 2013a; FONSECA; ROSA, 2013b). Essa biblioteca apresentou resultados em estado da arte na tarefa de etiquetagem morfosintática, pois usa word embeddings (FONSECA; ROSA; ALUÍSIO, 2015) e uma versão revisada do corpus MAC-Morpho (junção de etiquetas e correção da anotação) para garantir um desempenho melhor para tarefa. O etiquetador morfosintático utilizado no Coh-Matrix-Port é o MXPOST, treinado com o corpus MAC-Morpho disponibilizado pelo projeto Lácio-Web², em sua versão original.

A métrica número 24, Pronomes por Sintagmas, originalmente empregava uma ferramenta de extração de sintagmas nominais (um analisador sintático superficial). Como, para a inclusão das métricas de complexidade sintática, foi necessário incluir um analisador sintático pleno (vide Seção 5.3.3), esse analisador foi usado para a extração dos sintagmas.

Para as tarefas comuns de segmentação de sentenças e de itens lexicais, foram utilizadas as ferramentas padrão para a língua portuguesa da biblioteca NLTK³. As listas de palavras e de frequências utilizadas no Coh-Matrix-Port foram mantidas.

5.3 As Novas Métricas

Esta seção trata as métricas que foram adicionadas às já existentes no Coh-Matrix-Port. As métricas dividem-se em categorias, descritas nas próximas subseções.

5.3.1 Métricas de disfluência

As transcrições das narrativas dos pacientes utilizadas neste trabalho contém marcas de oralidade que podem ser significativas na análise destes textos, o que torna importante extraí-las.

A Tabela 11 mostra os eventos de interesse, presentes nas transcrições, que foram incorporados como métricas no Coh-Matrix-Dementia. A primeira coluna apresenta o nome da métrica, e a segunda exemplifica como o evento correspondente encontra-se anotado nas transcrições originais.

A duração das pausas é calculada somando-se a duração de cada uma e dividindo-se pelo total de palavras do texto. A métrica de emissões vazias é calculada como o número de palavras dentro destas emissões dividido pelo total de palavras do texto, o mesmo valendo para a métrica de disfluências. As outras métricas são calculadas como o número de ocorrências do evento dividido pelo total de palavras.

¹ Disponível em <<http://nilc.icmc.sc.usp.br/nlpnet/>>.

² <http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html>

³ <http://www.nltk.org/>

Tabela 11 – Métricas de disfluência acrescentadas ao Coh-Metrix-Dementia.

Métrica	Exemplo
Duração das pausas	((pausa X segundos))
Número de pausas curtas	(...)
Prolongamentos de vogais	a:::gora
Repetições	Ela <i>foi foi</i> para casa.
Preenchedores	uh, um
Emissões vazias	<empty> ... </empty>
Disfluências	<disf> ... </disf>

Tabela 12 – Córpus utilizados para geração do espaço LSA.

Córpus	Gênero	Docs	Tokens	Docs limpos	Tokens limpos
Wikipédia	Enciclopédico	923.375	96.183.280	923.375	96.183.280
PLN-BR	Jornalístico	103.870	39.967.672	103.564	18.003.126
LácioWeb	Misto	5934	8.020.769	5900	3.723.109
FAPESP	Divulgação Cient.	3806	6.056.843	3806	2.904.105
<i>Total</i>	—	<i>1.036.985</i>	<i>150.228.564</i>	<i>1.036.645</i>	<i>120.813.620</i>

5.3.2 Métricas de Análise de Semântica Latente

Conforme mencionado anteriormente, a LSA é uma técnica que tem por objetivo mapear um conjunto de documentos em um espaço vetorial, representando termos e documentos como vetores nesse espaço. A representação vetorial permite calcular a similaridade entre documentos, entre termos, e entre termos e documentos, podendo ser utilizada como base para sistemas de recolha de informação, seu uso originalmente pretendido. Porém, como a LSA permite calcular a similaridade entre documentos (que podem ser sentenças, parágrafos, ou mesmo textos inteiros), ela pode ser – e já foi – usada para medir a coesão e a coerência textuais.

Matematicamente, a LSA consiste em criar uma matrix termo por documento M , decompô-la em um produto de matrizes, e então zerar alguns dos auto-valores da matrix M (facilmente obtidos após a decomposição), deixando apenas os auto-valores de maior magnitude. Esse processo tem o efeito de criar um espaço vetorial de dimensões reduzidas, efetivamente aglutinando termos com padrões de uso parecido. Isso torna a LSA mais robusta, enquanto medida de similaridade textual, que técnicas de aferição de coerência que medem, por exemplo, a sobreposição de palavras entre sentenças adjacentes, uma vez que LSA é capaz de lidar com sinônimos.

O primeiro passo na criação de um espaço LSA consiste em obter uma grande quantidade de documentos, para a construção da matrix M . Para tanto, utilizamos quatro córpus de textos em Português: (i) a Wikipédia completa para o português; (ii) o córpus PLN-BR; (iii) o córpus do projeto LácioWeb; e (iv) o córpus Revista FAPESP. A Tabela 12 mostra dados sobre esses córpus.

Em seguida, é preciso gerar um dicionário de termos, com base nos textos presentes no *cópus*. Para tanto, foram removidos sinais de pontuação, e todas as letras foram postas minúsculas; além disso, foram removidas as *stopwords*, com base na lista destas presente na biblioteca NLTK. A Tabela 12 mostra a quantidade de documentos e de itens lexicais antes e depois desse processamento⁴. Por fim, foi levantada a lista de palavras únicas presentes no *cópus*, e então realizado um corte, após o qual as 350.000 palavras mais frequentes que tenham aparecido pelo menos cinco vezes no *cópus* foram tomadas. Esse corte é necessário para não tornar a matriz termo por documento excessivamente grande, tornando a geração do modelo intratável.

Esse dicionário foi, então, utilizado para gerar a matriz termo por documento M (tem-se, portanto uma matriz $350.000 \times 1.036.645$). Aplicou-se, então, uma transformação $tf \cdot idf$ à matriz. Em sua versão mais simples, essa transformação consiste em normalizar a frequência $f(t, d)$ em que um termo t aparece em um documento d , dividindo-se tal número pelo logaritmo do inverso da fração de documentos no qual esse termo aparece. Ou seja:

$$f(t, d) \leftarrow f(t, d) \cdot \log \frac{N}{N_t} \quad (5.1)$$

Na Equação 5.1, N_t é o número de documentos nos quais o termo t aparece. O objetivo da transformação $tf \cdot idf$ é diminuir o peso de palavras que ocorram com muita frequência, e aumentar o peso de palavras mais raras, complementando, assim, a remoção de *stopwords*. O valor de idf foi apresentado primeiramente em Jones (1972) como *especificidade do termo*; o raciocínio que embasa tal fórmula é que, quanto maior o número de documentos no qual um termo aparece, menor sua especificidade, e menor sua relevância na recolha de informação (ou mesmo na medição de similaridade).

O próximo passo na geração do modelo LSA é a decomposição da matriz termo por documento utilizando-se a *decomposição em valores singulares*, com o posterior descarte dos auto-valores menos significativos da matriz. O número de auto-valores mantidos é denominado *número de tópicos* do modelo. Para o modelo gerado neste trabalho, foram utilizados 400 tópicos.

Ao final desse processo, tem-se um modelo de LSA pronto para ser utilizado. Para o uso realizado por este trabalho – ou seja, a medição de similaridade entre sentenças de um texto, duas sentenças s_1 e s_2 cuja similaridade se deseja saber são mapeadas para o espaço 400-dimensional do modelo, produzindo-se os vetores v_1 e v_2 . A similaridade entre eles é medida, então, utilizando-se o cosseno do ângulo entre os vetores, dado pela razão entre o produto interno dos dois vetores e o produto de suas normas:

$$\cos \alpha = \frac{v_1 \cdot v_2}{|v_1| |v_2|} \quad (5.2)$$

⁴ Alguns documentos foram removidos por não conterem informações relevantes à tarefa ou por não estarem bem estruturados.

O valor de similaridade é, portanto, um número real entre -1 e 1. Com base nesse conceito, pode-se entender como são extraídas oito das dez métricas de LSA presentes no Coh-Metrix-Dementia. As oito métricas estão explicadas abaixo.

LSA sentence adjacent mean

Essa medida é a média da similaridade entre os pares de sentenças adjacentes presentes no texto. Ela tem por objetivo quantificar o quão conceitualmente similar cada sentença do texto é em relação à sentença anterior.

LSA sentence adjacent std

Desvio padrão da similaridade entre os pares de sentenças adjacentes presentes no texto. Essa métrica quantifica o quão consistente é a intersecção semântica entre sentenças adjacentes no texto.

LSA sentence all mean

Média da similaridade entre todos os pares de sentenças do texto, não apenas as adjacentes. O objetivo aqui é quantificar a coesão do texto como um todo.

LSA sentence all (within paragraph) std

Essa medida é o desvio padrão da similaridade entre todos os pares de sentenças dentro do mesmo parágrafo no texto.

LSA paragraph adjacent mean

Essa métrica consiste na média de similaridade entre parágrafos adjacentes no texto.

LSA paragraph adjacent std

Desvio padrão da similaridade entre parágrafos adjacentes no texto.

LSA sentence givenness mean

Média da similaridade entre cada sentença e todo o texto que a precede.

LSA sentence givenness std

Desvio padrão da similaridade entre cada sentença e todo o texto que a precede.

As duas últimas métricas apresentadas acima procuram medir, em uma dada sentença, a parcela da informação nela presente que já foi apresentada anteriormente no texto. Para tanto, todo o texto anterior à sentença é mapeado num único vetor, e então calcula-se a similaridade do cosseno entre o vetor da sentença atual e esse vetor.

Os autores sugerem um método alternativo, denominado *span*, que permite capturar melhor a presença de informações dadas na sentença atual. O método consiste em, ao invés de

transformar as k sentenças anteriores à sentença atual em um único vetor no espaço, utiliza-se essas sentenças como a base de um sub-espaço vetorial. Em seguida, decompõe-se a sentença atual em duas componentes: uma componente pertencente ao sub-espaço das sentenças anteriores, e uma componente perpendicular a esse espaço. A componente que pertence ao espaço é interpretada como informação dada (*givenness*), e a componente perpendicular, como informação nova (*newness*).

Matematicamente, sejam v_1, v_2, \dots, v_k os vetores correspondentes às sentenças anteriores à sentença atual (cujo vetor correspondente é, portanto, v_{k+1}). Seja $A = [v_1 v_2 \dots v_k]$ a matriz que possui esses vetores como colunas. Define-se, então, a *matriz de projeção* P_A sobre o sub-espaço gerado por v_1, v_2, \dots, v_k como sendo:

$$P_A = A \cdot (A^T \cdot A)^{-1} \cdot A^T \quad (5.3)$$

Na Equação 5.3, A^T indica a matriz transposta de A , e A^{-1} , sua inversa. Assim, a projeção de v_{k+1} no sub-espaço gerado por v_1, v_2, \dots, v_k é dada por:

$$v'_{k+1} = P_A \cdot v_{k+1} \quad (5.4)$$

Define-se, portanto, o *span* da sentença s_{k+1} como sendo a similaridade do cosseno entre v_{k+1} e v'_{k+1} . Com isso, duas novas métricas foram inseridas no sistema:

LSA sentence span mean

Média do *span* de cada sentença do texto.

LSA sentence span std

Desvio padrão do *span* de cada sentença do texto.

5.3.3 ***Diversidade lexical e complexidade sintática***

Como medidas de diversidade lexical, foram adicionados o Índice de Brunét e a Estatística de Honoré (vide Seção 3.2.1). Tais medidas são fórmulas simples, envolvendo apenas contagens, e sua adição ao sistema é direta.

Como medidas de complexidade sintática, foram adicionados o número médio de cláusulas por sentença (*Mean Clauses per Utterance*, MCU), as complexidades de Yngve e Frazier, a distância de dependência, e a entropia cruzada. As complexidades de Yngve e Frazier partem de uma estrutura de constituintes da sentença, enquanto a distância de dependência utiliza uma estrutura de dependências. Portanto, foi necessário realizar um levantamento de ferramentas disponíveis para a realização destas tarefas em textos em português.

Na categoria de *analisadores de dependências*, destacam-se o **MaltParser**⁵ e o **MST-Parser**⁶. O MaltParser (NIVRE; HALL; NILSSON, 2006) constrói estruturas de dependência utilizando uma abordagem linear, que percorre a sentença da direita para a esquerda, utilizando uma pilha para armazenar *tokens* cuja estrutura de dependentes ainda não está completa. A cada passo, o analisador decide entre colocar o próximo *token* no topo da pilha – estabelecendo ou não uma relação de dependência entre o *token* no topo da pilha e o novo *token* sendo adicionado, ou remover um *token* da pilha – estabelecendo ou não uma relação de dependência entre o próximo *token* da entrada e o *token* recém-removido da pilha. Para decidir qual ação deve ser realizada, o analisador utiliza um classificador baseado em memória, treinado a partir de treebanks, utilizando atributos de *tokens* dentro de uma janela de tamanho 6 na sentença [Nivre and Nilssohn, 2005].

O MSTParser (MCDONALD; LERMAN; PEREIRA, 2006) segue uma abordagem baseada em dois estágios: no primeiro, é gerada uma estrutura de dependências não-rotulada da sentença, e, no segundo estágio, são atribuídos os rótulos de cada arco. Na primeira etapa, a ferramenta segue uma abordagem baseada em árvores geradoras máximas (maximum spanning trees, MSTs): produz-se um grafo contendo todas as possíveis relações de dependência na sentença, e atribui-se um peso a cada arco com base em um conjunto de atributos das palavras envolvidas e em um vetor de pesos, aprendido automaticamente. Em seguida, a estrutura de dependências da sentença é dada pela MST desse grafo. Na segunda etapa, o sistema trata a atribuição de rótulos como um problema de etiquetamento sequencial, empregando o algoritmo de Viterbi para seleção da sequência com maior pontuação.

Tanto o MaltParser quanto o MSTParser participaram da tarefa compartilhada em análise de dependências multilíngue da CoNLL-X 2006 [Buchholz and Marsi, 2006]. Nessa tarefa, os sistemas foram treinados e testados utilizando treebanks de 13 línguas, inclusive o português (mas exclusive o inglês). O desempenho foi medido em termos de *Labeled Attachment Score* (LAS), que consiste na porcentagem de *tokens* pontuáveis aos quais o sistema encontrou o núcleo e o rótulo corretos para a relação de dependência correspondente. Nesta tarefa, tanto o MaltParser quanto o MSTParser foram os sistemas vencedores, obtendo o melhor desempenho tanto no total (80,2% e 80,3%, respectivamente) quanto na língua portuguesa (87,6% e 86,8%, respectivamente).

Encontra-se em desenvolvimento no NILC um analisador de dependências baseado em redes neurais profundas, inspirado no sistema SENNA⁷ [Collobert et al., 2011; Collobert, 2011]. Nesse analisador, os *tokens* são mapeados para vetores numéricos previamente gerados por métodos não-supervisionados, que são dados como entrada para duas redes neurais. Ambas realizam uma operação de convolução para converter sentenças de tamanho arbitrário em uma sequência de valores de tamanho fixo. A primeira rede funciona como um analisador baseado em MST, que emite pontuações para cada combinação entre dois *tokens*, nesse contexto uma

⁵ Disponível em <<http://www.maltparser.org/>>.

⁶ Disponível em <<http://www.seas.upenn.edu/~strctlrn/MSTParser/MSTParser.html>>.

⁷ <<http://ml.nec-labs.com/senna/>>.

candidata a ser considerada uma aresta de dependência. Extraindo-se a árvore geradora com maior pontuação nas somas das arestas, obtém-se a estrutura completa não-rotulada de dependência. Em seguida, a segunda rede classifica cada aresta com um rótulo.

No momento da realização deste trabalho de mestrado, o analisador supra-citado não se encontrava pronto para uso, e, portanto, não pôde ser empregado. Como o MaltParser apresentou desempenho ligeiramente superior ao MSTParser para a língua portuguesa, e por ser mais fácil de treinar e usar, optou-se por utilizá-lo neste trabalho.

Para a *análise de constituintes*, o **analisador de Berkeley** [Petrov et al., 2006; Petrov and Klein, 2007] é capaz de aprender uma gramática a partir de dados de um treebank de treinamento. O analisador é inicializado com uma gramática *baseline* simples, que é então refinada iterativamente com base no conjunto de treinamento, por meio de operações de separação, junção, e suavização. Em termos de desempenho, para a língua inglesa, os autores relatam uma medida-F de 90,2% na seção 23 do Penn Treebank.

O **analisador de Stanford** [Socher et al., 2013] utiliza uma técnica denominada *Compositional Vector Grammar* (CVG), que melhora as gramáticas livres de contexto probabilísticas (PCFG), utilizadas pelo analisador, por meio de redes neurais recursivas. O analisador é capaz de combinar informações sintáticas e semânticas ao acessar representações vetoriais ricas de palavras e computar vetores semânticos posicionais para sintagmas longos [Socher et al., 2013]. Os autores relatam uma medida-F de 90,44% na seção 23 do Penn Treebank.

Silva et al. (2010) comparam o desempenho de três analisadores distintos: o analisador de Bikel [2002], o de Stanford, e o de Berkeley. Os sistemas foram treinados em um cópulo de língua portuguesa, sem nenhum ajuste de parâmetros, e seu desempenho foi aferido em diversos cenários, por meio de diversas medidas de desempenho. De maneira geral, os analisadores de Stanford e de Berkeley apresentaram desempenho superior ao de Bikel, com o analisador de Berkeley apresentando desempenho superior ao de Stanford em boa parte dos cenários e medidas.

Com base nesses resultados, os autores desenvolveram um analisador denominado LX-Parser⁸, uma versão (originalmente) do analisador de Berkeley treinada com o cópulo CINTIL⁹. Ao verificarmos o website atual do projeto LX-Parser, verificamos que a versão que lá se encontra é uma versão do analisador de Stanford treinada sobre o CINTIL.

Ainda para a *análise de constituintes*, o analisador mais amplamente utilizado é o PALAVRAS (BICK, 2000); porém, trata-se de uma ferramenta proprietária, paga, e de código fechado, o que impossibilita seu uso neste trabalho. Como a única alternativa gratuita encontrada foi o LX-Parser, optou-se por utilizá-lo.

Desta maneira, as complexidades de Yngve e Frazier são tomadas a partir da árvore sintática gerada pelo LX-Parser, enquanto a distância de dependência utiliza a árvore de de-

⁸ Disponível em <<http://lxcenter.di.fc.ul.pt/tools/en/conteudo/LXParser.html>>.

⁹ <<http://cintil.ul.pt/pt/index.jsp>>

pendências gerada pelo MaltParser. O número médio de cláusulas por sentença foi estimado a partir da árvore sintática gerada pelo LX-Parser. Para o cálculo da entropia cruzada, foi utilizado um modelo de trigramas com suavização de Kneser-Ney modificada (CHEN *et al.*, 1998), gerado pela ferramenta KenLM¹⁰ sobre o mesmo cópulo utilizado para a LSA. A ordem 3 do modelo foi escolhida por ser adequada ao tamanho do cópulo de treinamento, e a suavização de Kneser-Ney é a mais utilizada para modelos *n-gram*, por apresentar melhores resultados em análises comparativas (CHEN *et al.*, 1998).

5.3.4 Densidade semântica

Nesta categoria, a principal métrica adicionada à ferramenta foi a **Densidade de Ideias**. Conforme já mencionado, não havia, na data de início deste trabalho, uma ferramenta automatizada capaz de extrair automaticamente as proposições (ou ideias) a partir de um fragmento textual. Portanto, foi necessário desenvolver tal ferramenta. Descreve-se abaixo a metodologia utilizada, bem como resultados de avaliações realizadas sobre a ferramenta.

A metodologia que optamos por adotar consiste em três passos:

1. Confecção de uma ferramenta de extração de proposições para a língua inglesa.
2. Generalização dessa ferramenta, produzindo uma *versão universal*.
3. Refinamento da versão universal para a língua portuguesa.

Esse fluxo de trabalho está esquematizado na Figura 17. Optamos por essa metodologia porque a língua inglesa dispõe de um manual de extração de densidade de ideias (CHAND *et al.*, 2010), especificamente voltado à análise proposicional de textos produzidos por pacientes com DA. Tal manual pôde, então, ser empregado como referência durante a codificação da ferramenta para o inglês. Além disso, o inglês possui ferramentas base de PLN (etiquetadores morfossintáticos, analisadores sintáticos, entre outras) que geralmente possuem qualidade melhor que as correspondentes disponíveis para o português.

Para a criação da ferramenta para o inglês, que nomeamos IDD3 (*Propositional Idea Density from Dependency Trees*), optamos por criar um sistema **baseado em regras** que age sobre **estruturas de dependência**. A razão da escolha de um sistema baseado em regras e não em aprendizado de máquina se deu pela facilidade de extração das proposições a partir de árvores de dependências e da inexistência de um cópulo anotado com proposições para se gerar classificadores via aprendizado de máquina.

Para ilustrar o funcionamento do sistema, a Figura 18 mostra uma sentença de exemplo, bem como sua estrutura de dependências, suas quatro proposições correspondentes, e a saída do

¹⁰ Disponível em <<https://github.com/kpu/kenlm>>.

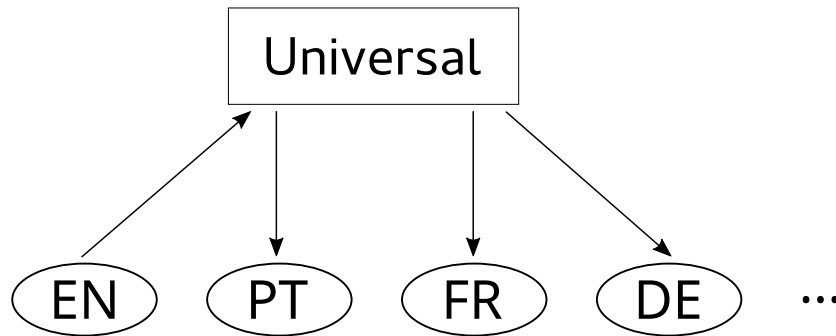
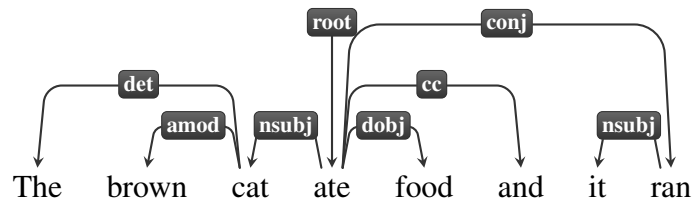


Figura 17 – Fluxo de trabalho para criação da ferramenta IDD3.

Sentença

The brown cat ate food and it ran.

Árvore de dependência



Proposições

1. ate, the cat, food
2. the cat, brown
3. ran, it
4. and, 1, 3

Saída

- 1 The cat, brown [M]
- 2 ate, The cat, food [P]
- 3 ran, it [P]
- 4 and, 2, 3 [C]

Figura 18 – Proposições geradas pelo IDD3 para uma sentença de exemplo.

IDD3 para essa sentença. Na saída, o sistema etiqueta as proposições de acordo com seu tipo ([P] significa predicção, [M], modificação, e [C], conexão).

As regras empregadas pelo sistema dividem-se em *conjuntos de regras (rule sets)*, um para cada rótulo de relação (**nsubj**, **dobj**, **amod**, e assim por diante). A Figura 19 mostra um excerto simplificado das regras que se aplicam à sentença de exemplo dada.

A árvore de dependências é percorrida recursivamente, iniciando-se pela relação **root**. O sistema determina automaticamente qual função process deve ser invocada, com base no rótulo da relação sendo processada. No pseudo-código apresentado, cada conjunto de regras possui acesso a algumas variáveis e funções. As variáveis são:

- **word**: a palavra correspondente ao nó atual.
- **pos**: a etiqueta morfofossintática da palavra correspondente ao nó atual.

Conjunto de regras para root e conj:

```
function process()
  subj = process(children.nsubj)
  dobj = process(children.dobj)
  my_id = emit (word, subj, dobj)
  if 'conj' in children
    cc = process(children.cc)
    conj_id = process(children.conj)
    emit (cc, my_id, conj_id)
  end if
  return my_id
end function
```

conjunto de regras para nsubj:

```
function process()
  det = process(children.det)
  amod = process(children.amod)
  this = det + ' ' + word
  emit (this, amod)
  return this
end function
```

Conjunto de regras para det:

```
function process()
  if word in ['the', 'a', 'an']
    return word
  end if
end function
```

Conjunto de regras para amod e cc:

```
function process()
  return word
end function
```

Figura 19 – Conjuntos de regras para processamento da sentença de exemplo da Figura 18.

- **head**: o nó pai do nó atual na árvore.
- **children**: lista contendo os nós filhos do nó atual.

As funções e diretivas disponíveis são:

- **process**: analisa recursivamente uma subárvore. O valor de retorno dessa função é flexível, dependendo do rótulo da relação sendo processada.
- **return**: retorna um valor para quem invocou o método (nesse caso, o conjunto de regras que chamou `process`).
- **emit**: emite uma proposição, adicionando-a a uma lista de proposições que pode ser recuperada após o processo de análise estar concluído, e retorna o identificador numérico da proposição.

Uma classe base `Ruleset`, a que `process` pertence, é definida, e uma subclasse dela é definida para cada rótulo de relação. Assim, por exemplo, uma classe `NsubjRuleset` é criada, como subclasse de `Ruleset`, onde é definida a função `process` adequada para processar a relação **nsubj**. Conjuntos de regras são agrupados conforme o tipo de subestrutura que eles esperam; por exemplo, a relação **root** espera uma estrutura do tipo sintagma verbal, enquanto as relações **nsubj** e **dobj** esperam estruturas do tipo sintagma nominal. Um objeto da classe `Engine` é responsável por invocar o método `process` do conjunto de regras correto cada vez que uma relação é processada, bem como por coletar e organizar as proposições emitidas.

A fim de criar a versão universal, tomamos por base as diretivas e dados disponibilizados pelo projeto *Universal Dependencies* (MCDONALD *et al.*, 2013). O projeto visa a criar um conjunto de *treebanks* em diversos idiomas seguindo um mecanismo homogêneo de anotação, baseado nas Dependências (Universais) de Stanford (MARNEFFE; MANNING, 2008; MARNEFFE *et al.*, 2014), nas etiquetas morfossintáticas universais do Google (PETROV; DAS; MCDONALD, 2012), e na *InterSet*, uma espécie de interlíngua para etiquetas e atributos morfosintáticos (ZEMAN *et al.*, 2014). A filosofia geral do projeto é produzir um conjunto universal de categorias e diretrizes que facilite a anotação multilíngue consistente de fenômenos comuns, ao mesmo tempo em que permite extensões específicas à língua quando necessário¹¹.

A adaptação das regras do inglês para produzir a versão universal ocorreu em três etapas. Inicialmente, as regras foram adaptadas para aceitar o novo conjunto de rótulos de relações de dependência. Em seguida, as regras foram adaptadas para lidar com o conjunto universal de etiquetas morfossintáticas. Finalmente, algumas dependências em nível lexical foram transformadas em parâmetros mínimos; essas dependências incluem, por exemplo, listas de artigos que não emitem proposições, entre outras listas. Assim, ao se informar essa lista

¹¹ Extraído de <<http://universaldependencies.github.io/docs/introduction.html>>.

mínima de parâmetros para o conjunto universal de regras, pode-se obter uma versão inicial da extração de proposições para uma língua alvo desejada. Pode-se, em seguida, refinar as regras que constam no conjunto universal, a fim de capturar fenômenos específicos da nova língua.

COH-METRIX-DEMENTIA: AVALIAÇÃO

Este capítulo apresenta as avaliações realizadas do sistema Coh-Metrix-Dementia. Mostram-se tanto avaliações intrínsecas, que aferem a correção das ferramentas utilizadas para a extração das métricas, quanto avaliações extrínsecas, que medem a qualidade das métricas na tarefa de classificação multiclasse de narrativas, sendo as classes de interesse: normais, com DA, e com CCL. Na Seção 6.1, é descrito um experimento piloto, que trata textos de sujeitos sadios em uma tarefa de descrição de figuras, e mostra a utilidade de se empregar ferramentas de PLN na automatização da análise de textos com objetivos clínicos. Na Seção 6.2, encontram-se os resultados de avaliações feitas sobre ferramentas desenvolvidas como componentes do Coh-Metrix-Dementia. Na Seção 6.3, encontra-se uma análise do desempenho de métodos de classificação automática atuando sobre métricas textuais extraídas pelo Coh-Metrix-Dementia.

6.1 Experimento Piloto

Grande parte das pesquisas com lesados cerebrais compara o desempenho destes com o de indivíduos sadios (SMITH; IVNIK, 2003). Apesar disso, pouca ênfase tem sido dada à produção discursiva de indivíduos normais, embora se reconheça a necessidade de obter normas de referência. Caracterizar o desempenho desses indivíduos é importante para diagnosticar, avaliar e reabilitar sujeitos com alterações de linguagem, bem como para apreciar aspectos preservados nos quais podem se apoiar orientações a cuidadores e familiares.

Uma justificativa importante para o estudo de indivíduos normais é a ampla diversidade prevista para sua produção discursiva. Entre os fatores responsáveis por tal diversidade estão a idade e a escolaridade. Muitos dos estudos se referem aos efeitos da idade na extensão do material produzido, conteúdo informativo, precisão semântica, foco, coerência e fluência da enunciação (MARINI *et al.*, 2005; WILLS; CAPILOUTO; WRIGHT, 2012; WRIGHT; CAPILOUTO; KOUTSOFTAS, 2013). O discurso dos indivíduos com maior escolaridade tem sido apontado como mais extenso e com maior densidade de conteúdo quando comparado ao dos sujeitos

com escolaridade restrita (Le Dorze; BÉDARD, 1998; MACKENZIE, 2000). Além disso, a escolaridade influencia positivamente a capacidade de decisão lexical, o conhecimento fonológico e as habilidades visual-espaciais (ARDILA *et al.*, 2010).

Dados sobre as características discursivas da população adulta são escassos. Na ausência de referências da população normal, os clínicos geralmente julgam o desempenho de seus pacientes por critérios subjetivos e variados.

Toledo (2011) desenvolveu um estudo buscando verificar o impacto das variáveis sociodemográficas (idade, escolaridade e gênero) na produção de discurso descritivo de sujeitos sadios. Foram avaliados 200 indivíduos sadios de nacionalidade brasileira, de ambos os gêneros. Os sujeitos foram divididos em 8 grupos de igual tamanho, de acordo com suas idades (30 – 60 anos e 61 anos ou mais) e escolaridades (3 e 4 anos; 5 a 8 anos; 9 a 15 anos e mais de 15 anos de estudo).

Os sujeitos foram instruídos a fazer uma descrição escrita de duas figuras, cada uma retratando uma cena diferente: uma figura simples e uma complexa. Foram definidas algumas **variáveis de impacto** para a análise das descrições, todas contabilizadas manualmente e submetidas a um critério de confiabilidade interjuízes.

Para analisar a **extensão do discurso**, foram verificados o *número total de palavras escritas* e o *número de palavras e expressões com função fática*. Para analisar as **dificuldades no resgate verbal**, foram verificados o *número de aproximações semânticas* (substituição de uma palavra por outra semanticamente relacionada), o *número de aproximações visuais* (substituição de uma palavra por outra visualmente semelhante) e o *número de aproximações grafêmicas* (seleção ou combinação incorreta de grafemas na escrita, na forma de trocas, omissões ou acréscimos).

Para analisar o **conteúdo informativo**, foram verificados o *número de frases irrelevantes* (declarações alheias à figura, como comentários pessoais), o *número de termos indefinidos* (termos vagos ou indeterminados), *dificuldades por problemas visuais* e a *quantidade de informação* (verificada como a quantidade de subtemas de cada figura que o sujeito foi capaz de identificar). Para analisar as **habilidades sintáticas**, foram verificadas a *extensão da frase* (número de palavras) e a *complexidade frasal* (maior complexidade encontrada no discurso, dentre quatro níveis: palavras isoladas, declarativas simples, orações coordenadas e orações subordinadas). Verificou-se também o **tempo de descrição**, medido em segundos.

O estudo apontou a **escolaridade** como a variável sociodemográfica de maior influência no desempenho dos sujeitos sadios. Observou-se impacto no número de palavras, na quantidade de informação, na extensão da frase, na complexidade frasal e no tempo de escrita, sendo os maiores valores encontrados nos sujeitos com escolaridade de mais de 15 anos. O número de erros ortográficos e a falta de pontuações foram marcantes nas descrições de indivíduos de 3 a 8 anos de estudo.

Indivíduos com mais escolaridade (15 a mais anos, principalmente) algumas vezes trazem, em suas descrições, analogias ("*Final de tarde*", "*São Paulo as 18:00 horas.*"), julgamentos ("*Falta de atenção*", "*A falta de humanidade de ajudar o próximo.*", "*A intolerância uns com os outros.*", "*Família meio viciada*"), listas de observações simples sobre as figuras ("*Centro, avenida, congestionamento*", "*Estresse, caos*") e títulos que resumem a figura ("*Caos urbano*", "*Confusão no trânsito*", "*O estress do dia a dia*"). Por serem textos curtos, muitas vezes sem verbos, tais descrições se assemelham às de indivíduos com menor escolaridade (Le Dorze; BÉDARD, 1998; MACKENZIE, 2000), dificultando sua análise automática por ferramentas computacionais e, por consequência, sua classificação.

Essas discrepâncias nos levaram a excluir descrições que não atendiam à tarefa em sua forma mais prototípica. Devido a isso, as 400 descrições originais do estudo de Toledo (2011) foram reduzidas a 242 neste experimento, sendo maior a exclusão de descrições de participantes com 3 a 4 anos de estudo. Após a exclusão, as 242 descrições restantes ficaram assim divididas: 43 descrições de participantes com 3 a 4 anos de estudo, 64 com 5 a 8, 61 com 9 a 14 e 74 com 15 a mais anos de estudo.

Além de exclusões, foram realizadas pequenas modificações nos textos para este experimento: foram incluídas vírgulas em listas de tópicos e pontos finais antes de capitalizações ou no final das descrições. Essas modificações foram realizadas para que o analisador sintático pudesse ter um melhor desempenho e calcular as características corretamente, operando da mesma maneira que a análise humana de Toledo (2011), que pressupôs essas modificações na contagem de palavras por sentenças e do número de sentenças.

Essas modificações foram realizadas em textos *escritos*; em textos *transcritos*, caso com o qual lidaremos durante nosso projeto de mestrado, serão levadas em conta essas e outras modificações, descritas em Fraser *et al.* (2012), como (i) contagem e remoção das marcas de pausas preenchidas transcritas; (ii) exclusão de *tokens* que não são considerados palavras da língua; (iii) contagem e remoção de falsos começos e repetições; (iv) inclusão de pontuações e capitalizações, com ajuda de características semânticas, sintáticas e prosódicas; (v) exclusão de neologismos e fala incompreensível.

As características que extraímos dos textos dividem-se em quatro grupos. No primeiro, temos 46 das 48 características extraídas pelo Coh-Matrix-Port; duas características envolvendo parágrafos foram eliminadas, pois todas as descrições utilizam um único parágrafo.

No segundo grupo, temos 21 características extraídas com uso do AIC. Foram eliminadas características que já eram capturadas pelo Coh-Matrix-Port. No terceiro grupo, estão três características adicionais, que inserimos para este experimento:

- **Polaridade positiva:** proporção de palavras com conotação positiva, segundo o dicionário do LIWC.

- **Polaridade negativa:** proporção de palavras com conotação negativa, segundo o dicionário do LIWC.
- **Erros ortográficos:** quantidade de erros ortográficos (apenas léxicos, não sintáticos) presentes no texto.

Finalmente, no quarto grupo, temos 6 características que foram computadas manualmente por Toledo (2011):

- **Tempo de escrita:** tempo médio de escrita, em segundos, de cada palavra do texto.
- **Idade:** faixa etária do sujeito (30-60 ou mais de 60 anos).
- **Ordem de apresentação:** ordem em que as figuras foram apresentadas, podendo ser complexa-simples ou simples-complexa.
- **Figura:** figura a que corresponde a descrição atual, podendo ser simples ou complexa.
- **Quantidade de informação:** quantos subtemas de cada figura o sujeito conseguiu identificar em sua descrição (a figura simples possui 7 subtemas, e a complexa, 11).
- **Subtema essencial:** característica binária, que representa se o sujeito conseguiu ou não identificar o subtema principal da figura.

Temos, portanto, um total de 76 características que foram empregadas nos experimentos. Em todos os experimentos, foi utilizado o pacote WEKA (HALL *et al.*, 2009). As questões de pesquisa deste estudo podem ser enunciadas da seguinte forma:

- a. Há um método de aprendizado de máquina multiclasse mais adequado para a tarefa de diferenciar grupos de indivíduos sadios em relação à escolaridade na tarefa de descrição escrita de figuras? Se sim, qual o número de classes que permite melhor desempenho?
- b. Métodos automáticos de seleção de características conseguem recuperar aquelas (ou os equivalentes automáticos delas) que se mostraram estatisticamente significativas em trabalhos de análise tradicional da literatura? Qual dos métodos traz características que geram o classificador de melhor desempenho?

Neste estudo, empregamos seis algoritmos de classificação, cada um deles pertencente a um paradigma diferente: Naive Bayes, baseado em probabilidades; J48, baseado em árvores; Multilayer Perceptron (MLP), baseada em redes neurais; Simple Logistic (SL), baseado em máxima entropia; JRip (implementação do WEKA do *Repeated Incremental Pruning to Produce Error Reduction* - RIPPER), baseado em regras; e SVM com *kernel* RBF, baseado em hiperplanos de máxima margem.

Empregamos, ainda, três métodos de seleção de atributos: no primeiro, utilizamos dois métodos de *ranking* de atributos, o ganho de informação (denominado `InfoGainAttributeEval` no WEKA) e o SVM (denominado `SVMAttributeEval` no WEKA), selecionamos os 38 atributos mais significativos segundo cada um dos dois e tomamos a intersecção entre esses conjuntos (utilizamos 38 atributos por ser metade da quantidade de atributos original); no segundo, empregamos uma seleção manual de atributos, selecionando aqueles que mais se assemelhavam aos utilizados na análise manual de Toledo (2011), resultando em 21 atributos; no terceiro, utilizamos o método *Correlation-based Feature Selection* (CFS) (HALL, 1998), denominado `CfsSubsetEval` no WEKA), que procura eliminar características redundantes e manter aquelas com maior poder preditivo

O objetivo da seleção manual foi verificar se os métodos automáticos de extração e seleção de atributos são substitutos confiáveis para a análise e a seleção manuais; ou seja, desejamos saber se a seleção intuitiva, feita pelos clínicos, dos atributos relevantes e sua posterior contagem manual podem ser substituídas pela análise computadorizada de diversos atributos seguida pela seleção, também automatizada, dos atributos mais relevantes. Avaliar esse aspecto é de grande importância aos propósitos do Coh-Matrix-Dementia.

Inicialmente, treinamos um classificador quaternário, com as quatro classes originais de Toledo (2011), obtendo desempenho bastante ruim, sendo o melhor classificador o MLP, com 42,3% de medida F e *baseline* de 30,6%, dado pela classe majoritária. Esse resultado nos levou a trabalhar apenas com classificadores binários, a fim de contarmos com maior número de instâncias nas classes.

A esse experimento inicial, seguiram quatro conjuntos de experimentos. No primeiro conjunto, verificamos a possibilidade de agrupar os dados em duas classes, 3-8 anos de escolaridade e 9+ anos, em quatro cenários de conjuntos de atributos: todos os atributos e cada um dos três métodos de seleção apresentados. Os resultados encontram-se na tabela 13. Com base nela, pode-se perceber que não foi possível obter bom desempenho, e que os algoritmos de seleção implicaram ganho pequeno.

Tabela 13 – Medida F (%) dos métodos para o primeiro conjunto de experimentos (3-8 versus 9+).

Algoritmo	Todos	IG \cap SVM	Manual	CFS
<i>Naive Bayes</i>	66,3	71,8	68,0	68,2
<i>SVM</i>	41,8	41,8	41,6	64,1
<i>MLP</i>	64,4	66,0	69,7	62,8
<i>SimpleLogistic</i>	69,8	69,4	71,3	71,2
<i>JRip</i>	65,7	69,7	67,2	65,4
<i>J48</i>	63,6	65,1	68,8	66,0
Baseline	55,8	55,8	55,8	55,8

No segundo conjunto de experimentos, treinamos classificadores utilizando apenas as

classes extremas (3-4 e 15+), e depois apenas as intermediárias (5-8 e 9-15). Nossa hipótese era que as classes extremas formavam grupos mais bem separados, enquanto as classes intermediárias apresentavam grande intersecção entre os grupos, não havendo separação tão pronunciada. Os resultados encontram-se na tabela 27, e mostram que, de fato, o desempenho com classes extremas foi bastante superior ao desempenho com classes intermediárias; os algoritmos de seleção de atributos foram aplicados no cenário com classes intermediárias, para verificar se era possível melhorar significativamente o desempenho, o que não ocorreu (obtivemos um máximo em torno de 71%, valor próximo aos resultados obtidos no primeiro conjunto de experimentos).

Tabela 14 – Medida F (%) dos métodos com classes extremas e intermediárias.

Algoritmo	Extr.	Intr.	Manual	IG \cap SVM	CFS
<i>Naive Bayes</i>	79,5	55,0	61,3	58,2	60,1
<i>SVM</i>	52,8	34,7	37,3	34,7	34,7
<i>MLP</i>	81,9	58,4	51,9	60,8	62,2
<i>SimpleLogistic</i>	84,6	56,8	67,2	52,7	56,4
<i>JRip</i>	81,3	52,8	51,9	63,2	71,2
<i>J48</i>	81,9	58,3	53,6	58,4	71,0
<i>Baseline</i>	63,2	51,2	51,2	51,2	51,2

Com base nesses experimentos, concluímos que alguma das classes intermediárias (5-8 ou 9-15) não era um grupo bem formado, sendo responsável pelo desempenho ruim obtido até então. Para averiguar, no terceiro conjunto de experimentos, eliminamos cada uma delas de cada vez; os valores encontram-se na tabela 15. Primeiramente, removemos a classe 9-15, ficando com 3-8 e 15+, e não obtendo bom desempenho; em seguida, removemos a classe 5-8, ficando com 3-4 e 9+, e obtendo o melhor desempenho até o momento. Apesar de o desempenho para o cenário 3-4 e 9+ ter sido o melhor, percebemos que as classes estavam bastante desbalanceadas, o que pode ser percebido pelo *baseline* elevado; por isso, replicamos os dados da classe minoritária (3-4), obtendo o cenário que denominamos 3-4r x 9+, onde obtivemos um desempenho bastante elevado (97,7%).

Tabela 15 – Medida F (%) dos métodos de classificação com remoção de classes intermediárias.

Algoritmo	3-8 x 15+	3-4 x 9+	3-4r x 9+
<i>Naive Bayes</i>	70,9	74,2	75,7
<i>SVM</i>	43,9	68,0	97,7
<i>MLP</i>	69,0	81,7	93,5
<i>SimpleLogistic</i>	73,4	86,7	86,3
<i>JRip</i>	73,6	87,9	88,6
<i>J48</i>	66,2	79,3	93,2
<i>Baseline</i>	59,1	75,8	51,1

Para concluirmos, realizamos uma última rodada de experimentos, para verificar o

impacto dos algoritmos de seleção de atributos no último cenário (3-4r e 9+). Os resultados encontram-se na tabela 16, e mostram que, pelo menos para o melhor algoritmo (SVM), os métodos de seleção não alteraram o desempenho do classificador.

Tabela 16 – Medida F (%) dos métodos de classificação e seleção de atributos com remoção de classes intermediárias.

Algoritmo	Todos	IG \cap SVM	Manual	CFS
<i>Naive Bayes</i>	75,7	81,3	74,2	80,4
<i>SVM</i>	97,7	97,7	97,7	97,7
<i>MLP</i>	93,5	93,2	91,6	93,5
<i>SimpleLogistic</i>	86,3	85,2	81,8	84,8
<i>JRip</i>	88,6	86,3	87,8	88,3
<i>J48</i>	93,2	91,2	90,5	92,8
Baseline	51,1	51,1	51,1	51,1

Com isso, concluímos que a melhor divisão de classes seria de 3-4 anos e mais de 9 anos de escolaridade e que, portanto, os indivíduos da classe 5-8 anos não possuem um comportamento característico. Isso significa que alguns deles escrevem como os de 3-4 anos e outros escrevem como os de mais de 9 anos. Portanto, um indivíduo nessa faixa deve ser realocado para alguma das outras; ou seja, deve-se classificá-lo e então tratá-lo como da classe resultante.

Esses resultados corroboram os dados do INAF, relatório de 2012, que destaca que 59% dos brasileiros que completaram ao menos um ano/série do segundo ciclo do ensino fundamental atinge o nível básico de alfabetismo, o que de fato dificulta a definição de uma classe coesa para os indivíduos com 5-8 anos de estudo. Essa conclusão é um resultado importante desse experimento e que nos auxiliará, durante a execução do projeto, a realizar comparações mais precisas entre afásicos e controles.

Concluimos, ainda, que podemos utilizar métodos automáticos de seleção de atributos em nossas análises, o que se acredita que evite *overfitting* e confira maior poder de generalização ao modelo gerado, e ainda assim garantir desempenho aceitável. Indicamos o método CFS para ser empregado como substituto à seleção manual de atributos, e elegemos o SVM, com *kernel* RBF, como o melhor algoritmo para classificação no cenário proposto.

Com isso, respondemos às questões postas neste experimento. Apesar de ser apenas um experimento inicial, essa investigação é pioneira em vários aspectos. Em primeiro lugar, por aplicar um método computadorizado com propósitos clínicos no Português Brasileiro. Em segundo lugar, por destacar a escolaridade, variável de inequívoca importância no cenário nacional. Em terceiro lugar, por adaptar construtos elaborados em outras línguas para o estudo de sujeitos brasileiros.

6.2 Das Ferramentas Desenvolvidas

A subseção 6.2.1 apresenta os resultados da avaliação realizada sobre a ferramenta de extração de proposições. As outras métricas criadas não foram avaliadas, por serem simples fórmulas léxicas ou não apresentarem opções de escolha, como as medidas dependentes de analisadores sintáticos.

6.2.1 Extração de Proposições

A avaliação da ferramenta IDD3 apresentada abaixo foi publicada em inglês no *28th IEEE International Symposium on Computer-Based Medical Systems* (CUNHA *et al.*, 2015).

A fim de avaliar o desempenho do sistema IDD3, nós executamos dois conjuntos de experimentos: (i) uma comparação, em termos de número de proposições, do IDD3 e do CPIDR *versus* a extração manual apresentada por Chand *et al.* (2010); e (ii) uma avaliação sobre textos anotados manualmente por uma linguísta treinada na tarefa *versus* as proposições geradas automaticamente pelo nosso sistema (*cf.* Figura 18). Em todas essas avaliações, foram utilizadas árvores de dependência manualmente revisadas; ou seja, efetivamente criou-se um *treebank* para os experimentos. Originalmente, as árvores foram produzidas pelo Analisador de Stanford Chen e Manning (2014) e depois revisadas.

A primeira avaliação tem como objetivo comparar o número de proposições emitido pelo CPIDR e pelo IDD3 a um *gold standard*, utilizando-se elocuições extraídas do manual de Chand *et al.* (2010). 91 exemplos de elocuições curtas, gramaticais, e completas foram extraídos do manual para uso nos experimentos. O manual conta com cerca de 160 exemplos, incluindo emissões mais longas e complexas, assim como emissões incompletas.

Na Figura 20, pode-se observar uma comparação, em termos de contagem de proposições, entre o IDD3, o CPIDR, e o manual de Chand *et al.* (2010) para as 91 sentenças de exemplo. Uma vez que o IDD3 foi construído para refletir a contagem apresentada pelo manual, ocorre um casamento perfeito entre as contagens. O CPIDR, concebido para refletir a contagem de proposições apresentada por Turner e Greene (1977), apresenta resultados menos consistentes para as sentenças utilizadas.

Como exemplos de erros cometidos pelo CPIDR nas sentenças, o sistema computa conjunções, como *and*, em sintagmas nominais compostos, enquanto o manual de Chand *et al.* (2010) não o faz. Assim, a sentença “*So I helped raise my brothers and sisters*” possui quatro proposições correspondentes, enquanto o CPIDR emite cinco, com a proposição extra correspondendo ao *and* entre *brothers* e *sisters*. Outro erro está relacionado a adjetivos possessivos, que não contam como proposições para Chand *et al.* (2010), mas são computados pelo CPIDR. Por exemplo, a sentença “*We went to his house*” corresponde a duas proposições, enquanto o CPIDR gera três, a outra correspondendo a *his*.

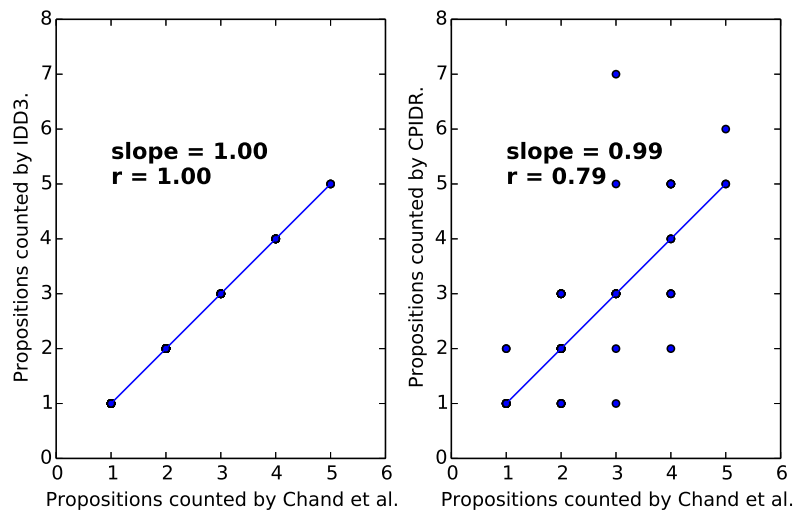


Figura 20 – Correlação, em termos de contagem de proposições, entre Chand et al., IDD3, e CPIDR para 91 sentenças de exemplo (CUNHA *et al.*, 2015).

Para a segunda avaliação, utilizamos quatro textos de dois gêneros diferentes: textos enciclopédicos escritos, e transcrições de fala. Os textos foram agrupados em quatro classes: (i) um texto enciclopédico bem formado e simples; (ii) um trecho de fala transcrita de um paciente com DA descrevendo duas figuras, extraído do projeto TalkBank¹; (iii) um trecho de fala transcrita de um paciente com Afasia Progressiva Primária Semântica (PPA-s); e (iv) um trecho de fala transcrita de um paciente com Afasia Progressiva Primária Não-fluente/agramatical (PPA-g).

O texto simples utilizado consiste nos primeiros seis parágrafos da entrada sobre Doença de Alzheimer na Simple English Wikipedia². As duas descrições de figuras do projeto TalkBank são de um paciente em estágio três, que possui quinze anos de escolaridade. As duas últimas descrições foram extraídas de Fraser *et al.* (2012) (Apêndice A; PPA-s, denominada SD no artigo, paciente 29; PPA-g, denominada PNFA, paciente 41). As três transcrições de fala foram incluídas a fim de avaliar o comportamento do IDD3 nesses textos problemáticos, bem como para levantar as modificações que devem ser realizadas a fim de se analisar tais textos com a ferramenta. Com isso, avaliamos a aplicabilidade do nosso método em contextos clínicos.

A Tabela 17 mostra algumas estatísticas do corpus de avaliação, divididas pelo tipo de proposição³. Nessa tabela, o número de sentenças se refere aos textos adaptados (veja abaixo),

¹ <<http://talkbank.org/browser/index.php?url=DementiaBank/Holland/tele01a.cha>> e <<http://talkbank.org/browser/index.php?url=DementiaBank/Holland/tele01b.cha>>

² <http://simple.wikipedia.org/wiki/Alzheimer%27s_disease>

³ WHAT se refere a um tipo relativamente raro de proposição que aparece com verbos bitransitivos atributivos. Por

enquanto o número de palavras se refere ao texto original, sem adaptações.

Tabela 17 – Estatísticas do corpus de avaliação (CUNHA *et al.*, 2015).

Texto	Sentenças	Palavras	Pred	Mod	Con	What	Total
Simples	18	360	51	107	21	2	181
DA	17	230	37	52	6	0	95
PPA-g	16	173	22	32	3	0	57
PPA-s	23	289	51	43	7	1	102
Total	74	1052	161	234	37	3	435

A fim de que o IDD3 pudesse analisar os textos transcritos, algumas adaptações foram necessárias, uma vez que a ferramenta espera sentenças bem-formadas. Foram removidos preenchedores lexicais (tais como *uh*, *um*), marcas de prolongamento de vogais, repetições que não têm a intenção de indicar ênfase, falsos inícios de palavras, marcas de pausas, e fala ininteligível. Também foram removidas conjunções, tais como *and*, que eram usadas repetidamente e não geravam proposições; às vezes, foi necessário também dividir elocuições em múltiplas sentenças quando da ocorrência dessas conjunções. Também foram inseridas vírgulas antes de correções de pudessem gerar proposições, e substituídas construções orais por sua forma em inglês padrão (por exemplo, *gonna* foi substituído por *going to*). Finalmente, perguntas do tipo *tag questions* foram colocadas em sentenças próprias. É importante notar que essas modificações são utilizadas apenas para extração de proposições, não afetando a contagem de palavras.

Para essa avaliação, a anotação do sistema foi comparada à anotação manual utilizando-se a **medida F**. A Tabela 18 relata a medida F para cada tipo de proposição, bem como o coeficiente total. A Tabela 19 mostra a média da medida F ao longo das sentenças de cada texto.

Nos resultados para o texto jornalístico, foi possível observar que o sistema alcançou seu melhor desempenho para predicções, com apenas um erro em 51 proposições. Modificações vêm em seguida; para essa categoria, todos os erros do sistema se deveram ao fato de “*Alzheimer’s Disease*” não ter sido reconhecido como uma entidade nomeada – e portanto um elemento atômico, o que fez com que o sistema gerasse proposições extras. Para conexões, o sistema apresentou um desempenho relativamente baixo, causado pela não-emissão de proposições em

exemplo, a sentença *They called me John* gera duas proposições: **1. called, they, me [P]**, e **2. John [WHAT]**.

Tabela 18 – Medida F total para as anotações (CUNHA *et al.*, 2015).

Texto	Pred	Mod	Con	What	Total
Simples	0,98	0,96	0,80	0,67	0,95
DA	0,78	0,93	0,73	1,00	0,86
PPA-g	0,98	0,94	0,50	1,00	0,94
PPA-s	0,88	0,86	0,33	0,00	0,83
Total	0,90	0,93	0,68	0,50	0,90

Tabela 19 – Medida F média para as anotações (CUNHA *et al.*, 2015).

Texto	Pred	Mod	Con	What	Total
Simples	0,97	0,95	0,83	0,94	0,94
DA	0,90	0,97	0,91	1,00	0,94
PPA-g	0,98	0,93	0,88	1,00	0,94
PPA-s	0,81	0,83	0,80	0,96	0,79
Total	0,92	0,92	0,85	0,98	0,90

Tabela 20 – Valores de correlação para a contagem de proposições entre o IDD3 e a anotação manual, no corpus de avaliação (CUNHA *et al.*, 2015).

Tipo	Correlação (r)
Predicações	0,99
Modificações	0,98
Conexões	0,89
What	0,57
Total	0,98

algumas circunstâncias. Proposições WHAT apareceram apenas duas vezes na anotação manual, e o sistema errou uma delas, o que explica a medida F baixa de 0,67. No total, o texto possui 181 proposições, e o IDD3 emitiu 180, das quais 171 estão corretas, gerando uma medida F total de 0,95.

Uma vez que o Coh-Matrix-Dementia está interessado apenas na densidade de ideias, e não nas proposições em si, é de interesse avaliar se o IDD3 é capaz de se desempenhar bem na contagem de proposições, mesmo se o sistema não emitir as proposições corretas. A Tabela 20 mostra a correlação entre as anotações manual e automática para as 74 sentenças existentes nos quatro textos analisados. Esses resultados mostram que, em geral, existe uma alta correlação em termos de número de proposições entre o IDD3 e a anotação manual, o que indica que a ferramenta pode ser usada de forma confiável para a contagem de proposições seguindo o manual de Chand *et al.* (2010).

Os resultados apresentados aqui indicam que o IDD3 pode ser usado para levantar as proposições tanto de textos escritos quanto transcritos, contanto que devidas adaptações sejam feitas aos textos transcritos para adequá-los ao processamento feito pelo sistema. A confiabilidade do sistema varia dependendo do tipo de proposição, sendo mais preciso para predicações e modificações, e menos preciso para conexões e WHATs. O sistema também apresentou alta correlação com a anotação manual em termos de contagem de proposições, o que é um resultado promissor para aplicações cujo interesse se restringe à densidade de ideias.

Como uma desvantagem da abordagem escolhida, uma vez que ela usa informações mais refinadas (árvores de dependência), ela necessita de uma ferramenta de análise mais sofisticada (um analisador de dependências), que é mais passível de erros que ferramentas mais simples,

como o etiquetador morfossintático empregado pelo CPIDR. Isso torna o sistema mais sensível a construções sintáticas não convencionais que induzam o analisador a erro, demandando esforços humanos de adaptação para textos transcritos.

Finalmente, vale notar que o IDD3 é um software livre e de código aberto, que pode ser encontrado em [<https://github.com/andrecunha/idd3/>](https://github.com/andrecunha/idd3/).

6.3 Da Classificação Textual

Nesta Seção, são apresentados os resultados de algoritmos de classificação automática, quando treinados sobre o *córpus* de estudo desta pesquisa. A Subseção 6.3.1 descreve o *córpus* utilizado nesta análise, que é composto por 60 sujeitos, sendo 20 em cada grupo de pesquisa: DA, CCL, e controles sadios; a Subseção 6.3.2 analisa os resultados obtidos no treinamento de classificadores sobre o *córpus* de pesquisa, utilizando-se como atributos as métricas do Coh-Metrix-Dementia.

6.3.1 *Córpus de pesquisa*

Os textos utilizados nesta análise foram produzidos por pacientes provenientes do ambulatório de linguagem do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, ou do ambulatório do Centro de Referência em Distúrbios Cognitivos da Faculdade de Medicina da Universidade de São Paulo. A amostra consiste em 60 indivíduos: 20 sujeitos saudáveis, 20 sujeitos com CCL amnésico, e 20 indivíduos com DA leve. Os indivíduos possuem o Português como primeira língua, idade igual ou superior a 60 anos, e escolaridade igual ou superior a 3 anos. A Tabela 21 mostra os critérios de inclusão e exclusão adotados (tabela retirada de Toledo (2015), trabalho que deve ser consultado case se deseje obter informações detalhadas e referências quanto aos critérios mencionados na tabela).

Os textos analisados são compostos por narrativas prototípicas da história da Cinderela, elicitadas de acordo com o seguinte procedimento: inicialmente, o examinador questiona o participante se este se lembra da história da Cinderela. Em seguida, o examinador apresenta ao participante um livro, com cenas em sequência, sem palavras, contando a história. As cenas apresentadas para o paciente encontram-se na Figura 21.

Quando o participante termina de ver cada cena, o livro é retirado, e o examinador solicita que o participante conte a história com suas próprias palavras. Não é estipulado tempo limite para a narração, durante a qual o examinador não faz nenhum comentário, a menos que o participante tenha dificuldades em iniciar ou continuar a história, fazendo uma pausa longa; nesse caso, o examinador intervém com questões genéricas, como “O que aconteceu depois?”, para encorajar o participante a continuar a narração. Caso ele não consiga, o examinador pergunta se ele se lembra das cenas da história; caso não, é permitido que o participante veja o livro novamente. Todo o

Tabela 21 – Critérios de inclusão e exclusão dos sujeitos (TOLEDO, 2015).

	Critérios de inclusão		
	Controles	CCL	DA
Idade	Acima de 60 anos	Acima de 60 anos	Acima de 60 anos
Escolaridade	Acima de 3 anos	Acima de 3 anos	Acima de 3 anos
Critério MOANS	Cumprir todos os requisitos	Não é critério para inclusão	Não é critério para inclusão
Diagnóstico sindrômico da demência – critérios do DSM-IV	Estadiamento 0	Estadiamento 0,5	Estadiamento 1 ou 2
Critérios do NINCDS-ARDRA	Não é critério para inclusão	Não é critério para inclusão	Critérios de probabilidade – DA provável
Critérios do consenso Winblad, 2004	Não é critério para inclusão	Subtipo amnésico	Não é critério para inclusão
Mínixame do estado mental	Desempenho aceitável para a escolaridade	Não é critério para inclusão	Não é critério para inclusão
Teste de fluência verbal – critério semântico – animais	Desempenho aceitável para a escolaridade	Não é critério para inclusão	Não é critério para inclusão
Avaliação neuropsicológica abrangente	Não aplicada – não é critério para inclusão	Definiu a inclusão e o subtipo de CCL	Não aplicada – não é critério para inclusão
	Critérios de exclusão		
	Controles	CCL	DA
Escala de Cornell para Depressão em Demência	Não aplicada a este grupo	Não aplicada a este grupo	Escore igual ou superior a 7 (indicativo de presença de sintomas depressivos)
Escala de Depressão Geriátrica – GDS	Escore igual ou superior a 10 (indicativo de presença de sintomas depressivos)	Escore igual ou superior a 10 (indicativo de presença de sintomas depressivos)	Não aplicada a este grupo
Questionário sobre Declínio Cognitivo no Idoso IQCode	Nota inferior a 3,41	Não é critério para exclusão	Não é critério para exclusão
Doenças psiquiátricas prévias (DSM – IV, 1994)	Histórico de diagnóstico prévio	Histórico de diagnóstico prévio	Histórico de diagnóstico prévio
Nível de escolaridade	Analfabetos foram excluídos	Analfabetos foram excluídos	Analfabetos foram excluídos
Acuidade visual	Compatível com funcionalidade e desempenho da tarefa alvo	Compatível com funcionalidade e desempenho da tarefa alvo	Compatível com funcionalidade e desempenho da tarefa alvo
Acuidade auditiva	Compatível com funcionalidade e desempenho da tarefa alvo	Compatível com funcionalidade e desempenho da tarefa alvo	Compatível com funcionalidade e desempenho da tarefa alvo

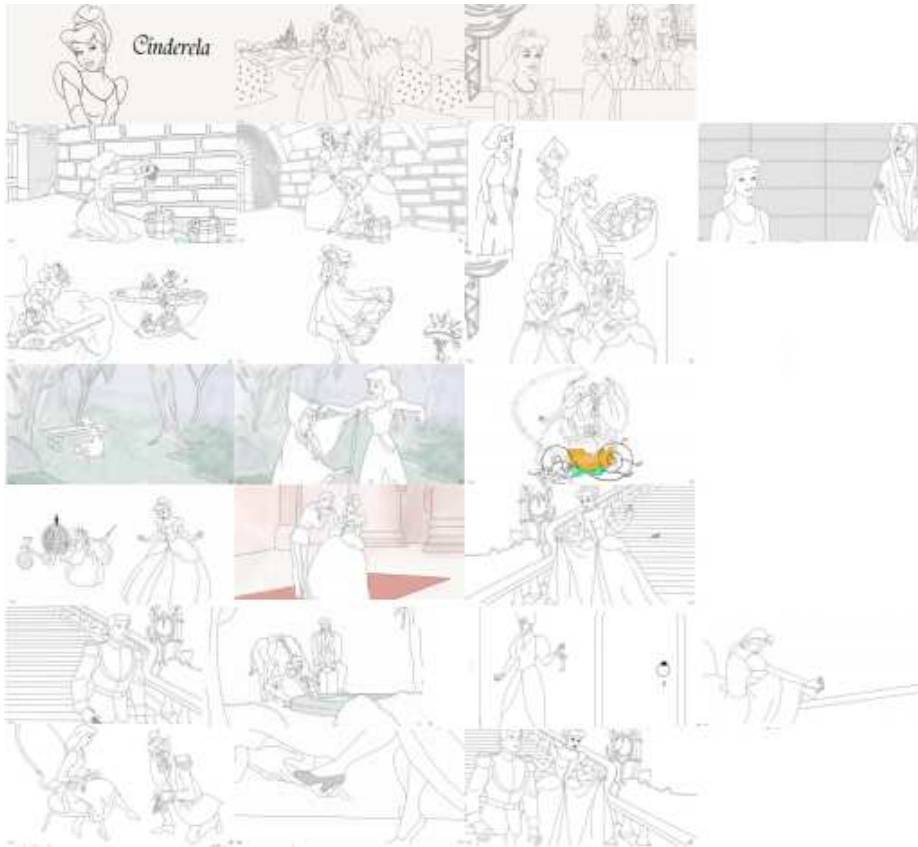


Figura 21 – Cenas da história da Cinderela apresentadas aos sujeitos.

teste é registrado com câmera de vídeo e com gravador de áudio. O áudio é, então, transcrito manualmente, gerando os textos utilizados neste trabalho.

Algumas anotações são feitas sobre esta transcrição bruta. A saber, são anotados segmentos em que houve hesitação ou repetição de mais de uma palavra ou de segmentos de uma palavra. Para tanto, utiliza-se a notação **<disf>...</disf>**. Repetições de palavras inteiras são capturadas automaticamente pela ferramenta, não necessitando de anotação manual. São anotadas também emissões vazias, que são aquelas que não estão relacionadas ao tópico da narração (neste caso, a história da Cinderela). Emissões vazias são delimitadas por **<empty>** e **</empty>**. São anotados ainda prolongamentos de vogais (indicados com **::**), pausas breves (indicadas com **...**), e pausas longas (indicadas com **((pausa XX segundos))**). A Figura 22a mostra um exemplo dessa transcrição bruta.

A ferramenta IDD3, descrita no capítulo anterior e responsável por extrair as proposições de um texto e calcular a densidade de ideias, espera como entrada textos bem formados, o que impossibilita que a transcrição bruta seja informada diretamente à ferramenta. Portanto, faz-se necessário editar as transcrições brutas, a fim de torná-las mais próximas de um texto escrito.

Para esse fim, as transcrições foram segmentadas em sentenças, majoritariamente seguindo-se um critério semântico-estrutural. Trechos como pausas (breves e longas) e orações vazias foram utilizados como indicadores na segmentação das sentenças. Na adaptação, marcas de

uma mocinha agradando o cavalo... <empty>mais?</empty> ... ((pausa 14 segundos)) mocinha que vai na festa vai ... ou não ... ((pausa 8 segundos)) a mocinha ta fazendo limpeza da parede... ((pausa 12 segundos)) <empty>como que fala esse</empty> ... filha do:: ... <empty>como que é</empty> ... ((pausa 5 segundos)) sogra e fala <empty>será?</empty> Ficou brava ... ((pausa 5 segundos)) não... ((pausa 9 segundos)) <empty>é::</empty> mo mocinha <empty>é::</empty> ganhou convite ... ((pausa 3 segundos)) <empty>só</empty> ... ((pausa 3 segundos)) <empty>é::</empty> <empty>como que fala</empty> ... madrasta que queria ir na festa não deixou ... ((pausa 11 segundos)) <empty>esse eu não sei</empty> ... ((pausa 16 segundos)) e a <disf>mo</disf> mocinha quer ir no baile e fez o:: vestido bonitinho ... ((pausa 3 segundos)) mas ela ... <empty>como que fala</empty> ... ((pausa 5 segundos)) <empty>sempre esqueço</empty> ... ((pausa 18 segundos)) <empty>a::</empty> ... ir irmão não quer deixar ir na festa e rasgou ... ((pausa 4 segundos))

(a) Fragmento da transcrição bruta.

Uma mocinha agradando o cavalo.
Mocinha que vai na festa.
Vai ou não.
A mocinha está fazendo limpeza da parede.
Ficou brava.
Ganhou convite.
Madrasta que queria ir na festa não deixou.
E a mocinha quer ir no baile.
E fez o vestido bonitinho.
Mas ela...
Irmão não quer deixar ir na festa.
E rasgou.

(b) Fragmento correspondente na transcrição adaptada.

Figura 22 – Exemplos de transcrição bruta e adaptada, para um sujeito com DA leve.

oralidade presentes na transcrição foram substituídas pela sua forma ortográfica padrão (por exemplo, *pro* foi substituído por *para o*).

Durante a codificação do IDD3, notou-se que o analisador de dependências tende a cometer mais erros quando as sentenças são muito grandes, por exemplo contendo diversas orações coordenadas. Além disso, Chand *et al.* (2010) indicam que conjunções como *e* que ocorrem com muita frequência não devem originar proposições. Portanto, quando sequências grandes de orações são concatenadas por conjunções “vazias”, como o *e*, elas são editadas como orações separadas, para permitir melhor desempenho do IDD3. Além disso, marcações de pausas e de prolongamentos de vogais foram removidas. Emissões vazias e trechos de hesitação também foram removidos.

A Figura 22b mostra um fragmento de transcrição editada, correspondendo ao fragmento bruto apresentado na Figura 22a. A Tabela 22 mostra estatísticas do corpus. Nela, o número de palavras corresponde ao total nas transcrições brutas, o número de sentenças corresponde ao total nas transcrições revisadas, o número de emissões vazias corresponde ao número de segmentos

Tabela 22 – Estatísticas do córpus utilizado na avaliação.

Categoria	Palavras	Sentenças	Em. vazias	Pal. vazias
<i>Controles</i>	8012	929	141	211
<i>CCL</i>	7782	641	220	503
<i>DA</i>	7814	593	355	2184
<i>Total</i>	<i>23608</i>	<i>2163</i>	<i>716</i>	<i>2898</i>

contíguos marcados como tal na transcrição bruta, e o número de palavras vazias corresponde ao total de palavras presentes nesses segmentos.

6.3.2 Treinamento dos Classificadores

Utilizando-se o córpus descrito na Subseção 6.3.1, deseja-se responder às três questões de pesquisa apresentadas na Introdução e retomadas aqui:

1. Existem características, que possam ser obtidas a partir de ferramentas de PLN disponíveis para o português, capazes de (i) distinguir indivíduos saudáveis de indivíduos com DA e CCL, e também (ii) distinguir indivíduos saudáveis dos com CCL, a fim de detectar precocemente o início da deterioração, com precisão aceitável para dar suporte à decisão médica? Se sim, quais são elas?
2. Qual a contribuição de métricas de complexidade sintática, densidade de ideias, e coerência textual, via semântica latente, aqui propostas para serem adaptadas ao português, para as duas classificações propostas?
3. Quais, dentre os métodos de classificação considerados neste trabalho, desempenham melhor a tarefa de predizer a classe de um indivíduo (sadio, DA, ou CCL) com base nas características linguísticas de seu texto?

Foram realizados dois conjuntos de experimentos utilizando-se diferentes algoritmos de classificação do pacote Weka (HALL *et al.*, 2009), em suas configurações padrão, e diferentes conjuntos de métricas. Foram utilizados algoritmos de diferentes paradigmas de aprendizado de máquina: Naive Bayes (NB), *Support Vector Machines* (SVM) com *kernel* linear⁴, *Multilayer Perceptron* (MLP), regressão logística, JRip (algoritmo baseado em regras), e J48 (algoritmo baseado em árvore de decisão), para determinar qual o melhor algoritmo de classificação para o cenário da pesquisa, e assim responder à questão de pesquisa 3. O *baseline* utilizado foi a proporção de pacientes na classe majoritária.

⁴ Alguns pesquisadores da área de aprendizado de máquina sugerem que o *kernel* linear é o mais indicado para o cenário em que há poucos dados e muitos atributos, como é o caso em nossos experimentos. Para mais informações, vide <<https://www.youtube.com/watch?v=i25MEJeX0Eg>>.

Tabela 23 – Medida F (%) dos métodos de classificação para diferentes conjuntos de métricas, com três classes (cenário 1).

Algoritmo	CMP	Todas	Novas	CFS
<i>NB</i>	65,1	73,3	76,7	81,7
<i>SVM</i>	66,9	71,5	73,1	75,3
<i>MLP</i>	56,6	53,6	63,3	60,1
<i>RL</i>	61,6	70,1	71,8	75,0
<i>JRip</i>	50,0	69,9	75,0	73,2
<i>J48</i>	49,8	66,6	63,3	74,8
Baseline	33,3	33,3	33,3	33,3

Os algoritmos foram testados em diferentes conjuntos de métricas: apenas as 48 métricas derivadas do Coh-Matrix-Port (conjunto denominado CMP), com todas as 73 métricas (48 do Coh-Matrix-Port + 25 adicionadas, formando o conjunto Todas), com apenas as 25 desenvolvidas nesta pesquisa (conjunto Novas), e com um subconjunto das métricas selecionado pelo algoritmo *Correlation-based Feature Selection* (CFS) (HALL, 1998), no esquema de validação cruzada de 10 *folds*. O método CFS parte da premissa de que um bom conjunto de atributos é formado por atributos altamente correlacionados à classe dos objetos, porém pouco correlacionados entre si, e foi escolhido aqui por ter sido indicado como melhor método de seleção de atributos no trabalho de Toledo *et al.* (2014), embora o trabalho trate de descrição de figuras, e aqui de narrativas. Com esse esquema de testes, pretende-se avaliar quais métricas possuem maior impacto no desempenho dos algoritmos de classificação, a fim de responder à questão de pesquisa 2. Também é possível, desta maneira, responder à questão 1, que diz respeito à existência de métricas automatizadas capazes de fornecer um desempenho aceitável na classificação.

Utilizam-se aqui dois cenários de classificação. No cenário 1, os algoritmos são treinados com três classes: controles (20 sujeitos), DA (20 sujeitos), e CCL (20 sujeitos). No cenário 2, utilizam-se duas classes: controles (20 sujeitos), e indivíduos com CCL (20 sujeitos). A métrica de avaliação utilizada é a **medida F**. Antes da aplicação dos classificadores, os atributos foram normalizados para possuírem média nula e variância unitária. Os resultados de classificação foram calculados no esquema *leave-one-out*, uma vez que o conjunto de dados é pequeno.

Para o cenário 1, as 17 métricas selecionadas pelo CFS encontram-se na Tabela 24. A Tabela mostra o número de *folds*, no esquema *10-fold cross-validation*, em que a métrica foi escolhida como relevante, e o respectivo valor em porcentagem. As métricas em itálico, num total de 11, foram desenvolvidas neste trabalho, não estando presentes no Coh-Matrix-Port. Dentre as onze métricas selecionadas pelo CFS, três se enquadram na categoria de complexidade sintática, duas na categoria de densidade semântica, duas de LSA na categoria de coerência textual via semântica latente, e quatro na categoria de disfluências. Nota-se que boa parte das métricas selecionadas pelo CFS foi desenvolvida neste trabalho (11 de 17), indicando que são métricas que de fato impactam positivamente na tarefa de classificação, colaborando para a resposta à

Tabela 24 – Atributos selecionados pelo CFS para três classes (cenário 1).

# de folds	% de folds	Atributo
5	50%	Palavras por sentença
2	20%	Número de palavras
6	60%	MCU
2	20%	Incidência de conectivos aditivos negativos
10	100%	Distância de dependência
10	100%	Complexidade de Yngve
1	10%	Densidade de conteúdo
1	10%	Palavras antes do verbo principal
1	10%	Sobreposição de radical em sentenças adjacentes
5	50%	Sobreposição de radical
10	100%	LSA: média com todas as sentenças
1	10%	LSA: desvio padrão de givenness
10	100%	Emissões vazias
2	20%	Duração média das pausas
10	100%	Número médio de pausas curtas
1	10%	Média de prolongamentos de vogais
9	90%	Densidade de ideias

Tabela 25 – Média e desvio padrão (normalizados) para as métricas selecionadas pelo CFS para três classes (cenário 1).

Atributo	μ_{CTRL}	σ_{CTRL}	μ_{CCL}	σ_{CCL}	μ_{DA}	σ_{DA}
Palavras por sentença	-0.28	0.51	0.77	0.75	-0.48	1.16
Número de palavras	0.24	0.92	0.12	0.95	-0.36	1.07
MCU	-0.40	0.58	0.84	0.79	-0.45	1.01
Inc. de conec. aditivos negativos	0.19	0.69	0.23	1.15	-0.42	1.02
Distância de dependência	-0.38	0.41	0.85	0.77	-0.48	1.09
Complexidade de Yngve	-0.34	0.45	0.86	0.77	-0.52	1.05
Densidade de conteúdo	-0.29	0.90	-0.36	0.42	0.65	1.21
Palavras antes do verbo principal	-0.22	0.53	0.70	1.02	-0.48	0.99
Sobreposição de radical em sent. adj.	-0.32	0.64	0.48	0.92	-0.15	1.22
Sobreposição de radical	-0.55	0.52	0.35	0.87	0.20	1.26
LSA: média com todas as sentenças	-0.17	0.59	-0.35	0.57	0.52	1.41
LSA: desvio padrão de givenness	-0.35	0.60	-0.14	0.91	0.50	1.22
Emissões vazias	-0.55	0.19	-0.27	0.30	0.82	1.37
Duração média das pausas	-0.36	0.08	-0.28	0.18	0.64	1.55
Número médio de pausas curtas	-0.47	0.43	-0.51	0.38	0.98	1.12
Média de prolongamentos de vogais	-0.16	0.68	-0.05	0.91	0.21	1.31
Densidade de ideias	0.44	0.49	0.37	0.57	-0.81	1.22

questão de pesquisa 2.

A Tabela 25 mostra a média e o desvio padrão normalizados dos valores das métricas, ao longo dos 60 textos, divididos por categorias. Como exemplo de análise dos atributos, na métrica de *palavras por sentença*, os controles apresentaram escores maiores que os pacientes com DA. Isso se deve ao fato de que os indivíduos saudáveis tenderam a empregar com mais intensidade orações subordinadas, que, durante a segmentação manual, não puderam ser separadas de suas orações principais. Enquanto isso, pacientes com DA utilizavam períodos com orações coordenadas com mais frequência; esses períodos foram divididos em orações separadas na fase de adaptação dos textos, conforme discutido na Seção 6.3.1 (veja Figura 22b); assim, obtém-se uma menor quantidade de palavras por sentença.

Para a *LSA givenness*, tem-se uma variância maior para pacientes com DA, em comparação a pacientes com CCL e controles, com CCLs tendo uma variância ligeiramente maior que os controles. Tal dado é um indício de que, conforme existe progressão do comprometimento cognitivo, as mudanças de assunto começam a se tornar mais frequentes, o que é consistente com os dados encontrados na literatura (veja o Capítulo 2).

Para as *emissões vazias*, nota-se que os pacientes com DA apresentam uma ocorrência significativamente maior dessas emissões, seguidos por pacientes com CCL e controles, sendo que pacientes com CCL apresentam uma densidade um pouco maior que os controles. Isso indica que, para esse tipo de discurso narrativo, a progressão da demência é acompanhada por um discurso menos focado, mais prolixo, e com mais informações irrelevantes, o que também é coerente com os resultados presentes na literatura sobre DA.

Para o *número médio de pausas curtas*, o mesmo padrão de emissões vazias é percebido, o que indica que o discurso de pacientes com DA é globalmente mais disfluyente e esforçado. Para a duração média das pausas, nota-se que pacientes com DA tendem a apresentar uma duração média maior que os pacientes com CCL e os controles, que não apresentam diferença entre si.

Para a *densidade de ideias* (DI), é possível perceber que os pacientes com DA apresentam DI nitidamente inferior aos pacientes com CCL e os controles, e que pacientes com CCL apresentam uma DI ligeiramente inferior aos controles. Tal resultado mostra uma progressiva redução da eficiência comunicativa, medida pela DI, conforme o quadro demencial se torna mais pronunciado. Uma análise manual dos textos mostra que, nos textos de pacientes com DA, nota-se grande quantidade de nomeações de entidades (frequentemente durante tentativas, por parte do paciente, de lembrar o nome exato daquilo que se deseja referir), porém sem que se faça nenhuma afirmação sobre elas. Esse resultado também é coerente com o conhecimento já estabelecido na literatura médica.

Como um exemplo de métrica não selecionada pelo CFS, mas que ainda assim permite extrair informações relevantes, cite-se a *entropia cruzada*. Para essa métrica, pacientes com DA em geral apresentam entropia maior, seguidos por controles, e então por pacientes com CCL, que

apresentam entropia menor. Analisando manualmente os textos dos pacientes, percebeu-se que textos de controles tendem a ser mais complexos, porém ainda com sentido; textos de pacientes com CCL tendem a ser mais simples, mas ainda apresentam fluência textual; e textos de pacientes com DA tendem a ser mais fragmentados e de difícil compreensão. Os textos fragmentados e com sentido difícil de compreender dos pacientes com DA geram combinações de palavras muito raras, aumentando a entropia; os textos dos controles são mais complexos que os de CCL, mas ainda fazem sentido, explicando uma entropia intermediária; os textos de pacientes com CCL apresentam construções simples, de uso comum, o que explica a baixa entropia.

O número de *referências anafóricas em sentenças adjacentes* foi maior entre os pacientes com CCL, seguidos pelos controles, e então por pacientes com DA. Isso indica que o discurso de pacientes com CCL é coeso porém repetitivo, com grande número de referências entre as sentenças; os controles apresentam um discurso também coeso, porém não tão repetitivo; por fim, os pacientes com DA apresentam menor conexão no discurso.

Quanto aos algoritmos de classificação, pode-se notar, com base na Tabela 23, que o classificador Naive Bayes obteve, em geral, o melhor desempenho (ou um desempenho próximo ao melhor). Quando se compara a primeira coluna à segunda, nota-se um ganho significativo de desempenho em todos menos um algoritmo (MLP). Quando se compara os algoritmos com todas as métricas e com apenas as métricas atuais, nota-se que todos os algoritmos exceto o J48 apresentam um desempenho similar ou melhor utilizando-se apenas as métricas adicionadas neste trabalho. O melhor desempenho no total é obtido com as métricas selecionadas pelo CFS, utilizando-se o algoritmo Naive Bayes, que apresentou 81,7 % de medida F.

A Figura 23 exibe um gráfico de dispersão do conjunto de dados normalizado, utilizando-se os atributos da Tabela 24 que foram selecionados em pelo menos 50% dos *folds*. A Tabela 26 mostra a matrix de confusão para o algoritmo Naive Bayes utilizando-se os atributos selecionados pelo CFS no cenário 1 (três classes). Para os controles, houve quatro exemplos classificados incorretamente: três foram classificados como DA, e um como CCL; para os pacientes com CCL, um cenário semelhante: três foram classificados como controles, um como DA; para os pacientes com DA, o grupo com maior acerto, todos os três erros foram classificados como CCL. Com base nessas análises, o erro mais grave cometido pela ferramenta foram os três casos em que sujeitos saudáveis foram classificados como DA; tais erros podem estar relacionados à forma como os textos foram editados para melhor desempenho da ferramenta IDD3 e das outras ferramentas de PLN. Esta edição pode ter removido das transcrições elementos relevantes e ter tido um efeito de homogeneização que confundiu os métodos de classificação. A Figura 24 mostra a árvore gerada pelo algoritmo J48 para este cenário, utilizando-se os atributos selecionados pelo CFS. Nesse diagrama, note-se a grande importância das emissões vazias (nó raiz da árvore) na identificação da DA.

No segundo cenário de avaliação, os pacientes com DA foram eliminados da tarefa de classificação. O objetivo desse experimento é determinar se é possível utilizar as métricas do

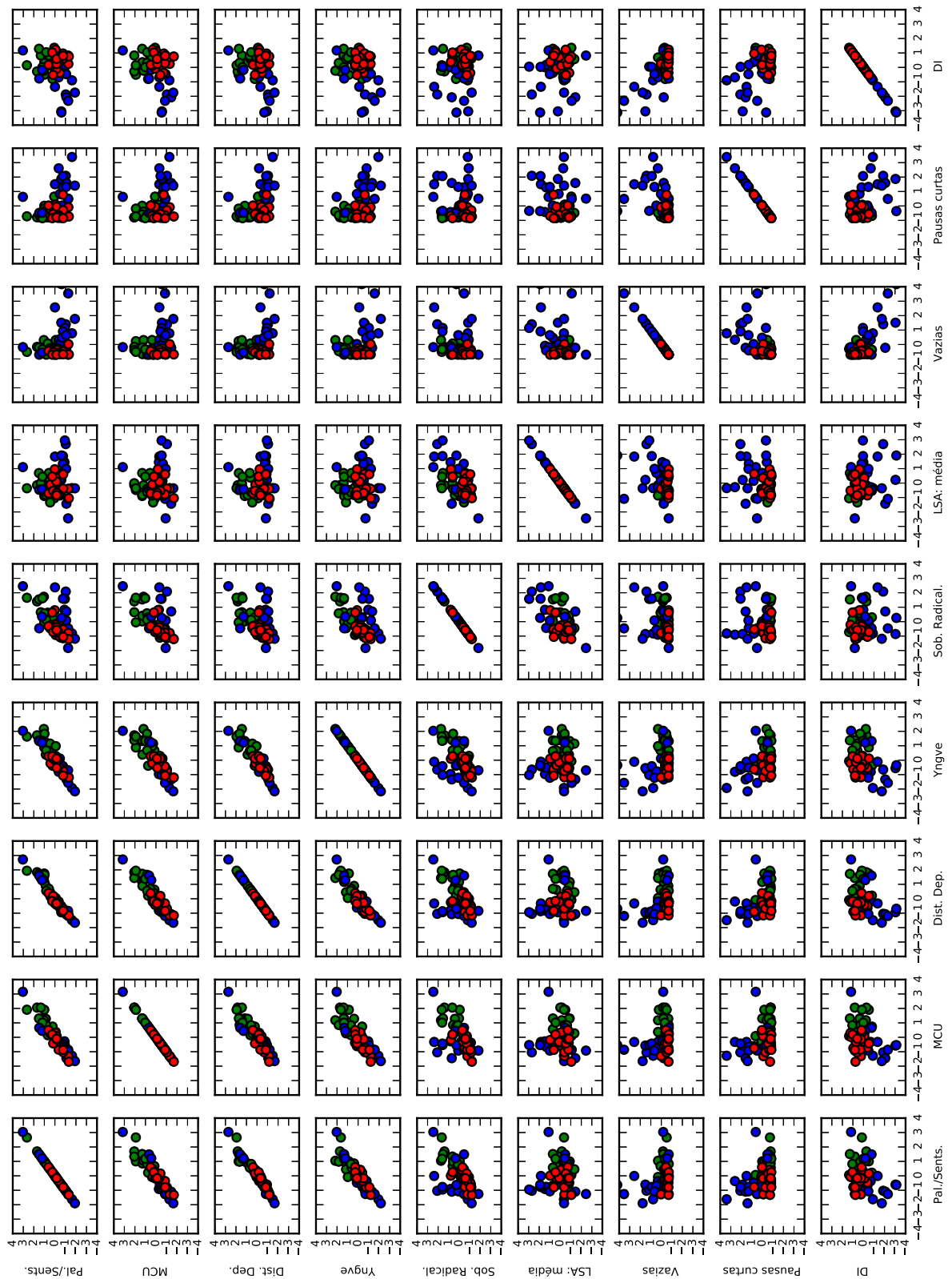


Figura 23 – Gráfico de dispersão do conjunto de dados normalizado. Dados de controles em vermelho, CCL em verde, DA em azul.

Tabela 26 – Matriz de confusão para o algoritmo Naive Bayes com três classes (cenário 1).

	CTRL	CCL	DA
CTRL	16	1	3
CCL	3	16	1
DA	0	3	17

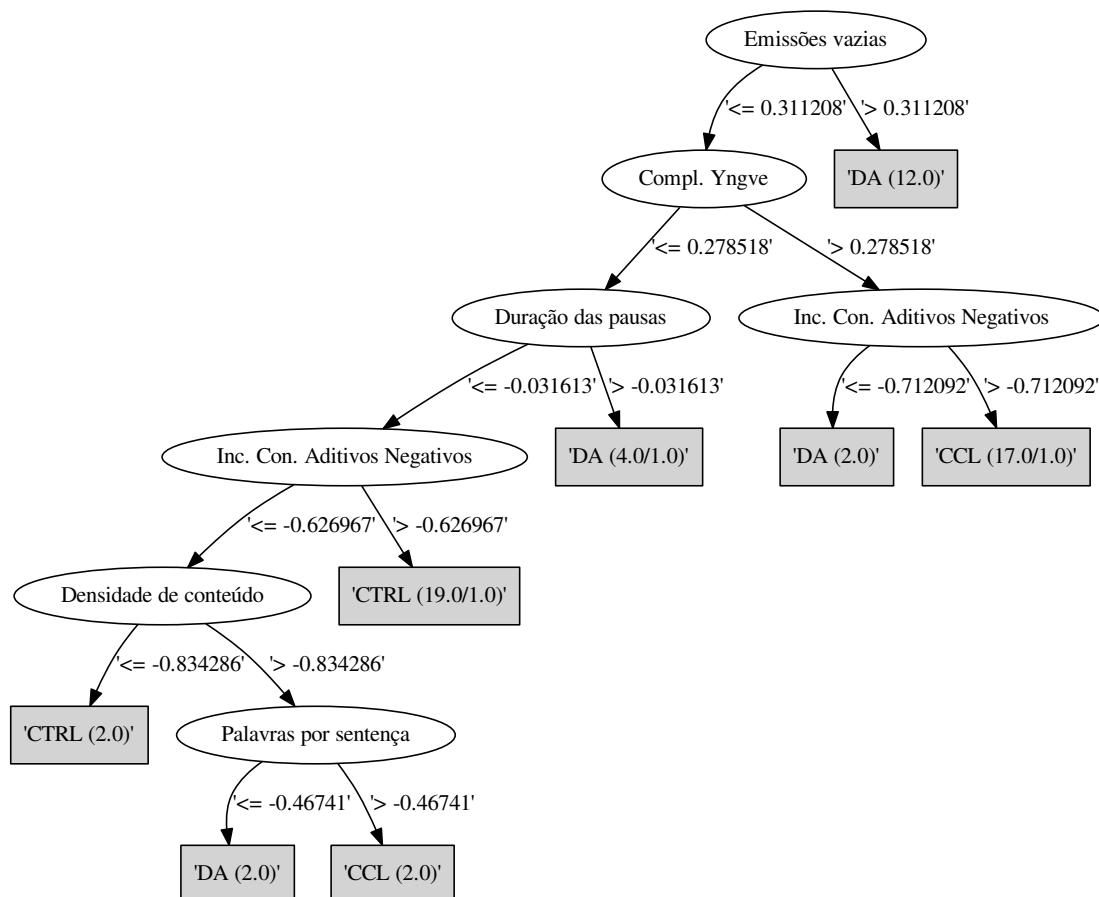


Figura 24 – Árvore de decisão gerada pelo algoritmo J48 para três classes (cenário 1) utilizando-se os atributos selecionados pelo CFS.

Tabela 27 – Medida F (%) dos métodos de classificação para diferentes conjuntos de métricas, com duas classes (cenário 2).

Algoritmo	CMP	Todas	Novas	CFS
<i>NB</i>	72,5	82,5	85,0	82,5
<i>SVM</i>	77,5	74,7	79,8	84,8
<i>MLP</i>	72,5	69,9	77,5	82,5
<i>RL</i>	74,9	74,7	69,7	74,9
<i>JRip</i>	69,7	87,5	80,0	77,5
<i>J48</i>	59,6	90,0	90,0	90,0
Baseline	50,0	50,0	50,0	50,0

Tabela 28 – Atributos selecionados pelo CFS para duas classes (cenário 2).

# de folds	% de folds	Atributo
4	40%	Índice Flesch
8	80%	<i>MCU</i>
1	10%	Inc. de conec. causais positivos
10	100%	<i>Distância de dependência</i>
1	10%	<i>Complexidade de Frazier</i>
10	100%	<i>Complexidade de Yngve</i>
6	60%	Palavras antes do verbo principal
1	10%	Referências anafóricas
1	10%	Sobreposição de argumentps
2	20%	Sobreposição de radical em sent. adj.
4	40%	Sobreposição de radical
1	10%	Sobreposição de pal. de conteúdo em sent. adj.
4	40%	<i>LSA: desvio padrão de span</i>
6	60%	<i>Emissões vazias</i>
1	10%	<i>Duração média das pausas</i>
1	10%	<i>Prolongamentos de vogais</i>

Coh-Matrix-Dementia para detectar os primeiros sinais da deterioração, e assim colaborar com o diagnóstico precoce das demências. A Tabela 27 mostra o desempenho da classificação para cada algoritmo e conjunto de métricas, e a Tabela 28 mostra os atributos selecionados pelo CFS. Neste cenário, o melhor desempenho é obtido nos cenários que incluem métricas novas (Todas, Novas, e CFS), com o algoritmo J48, que produziu 90,0% de medida F.

Para este cenário, o CFS selecionou 16 métricas, das quais 8 foram adicionadas neste trabalho. Dentre as métricas novas selecionadas, quatro enquadram-se na categoria de complexidade sintática, uma na de LSA, e três na de disfluências. Note-se, dentre as métricas selecionadas, a importância de métricas diretamente relacionadas à coesão na superfície textual, dado que foram selecionadas quatro métricas de correferência, e uma de referências anafóricas. Note-se ainda que, interessantemente, a densidade de ideias não foi selecionada dentre as medidas mais relevantes.

Tabela 29 – Matriz de confusão para o algoritmo J48 para duas classes (cenário 2).

	CTRL	CCL
CTRL	17	3
CCL	1	19

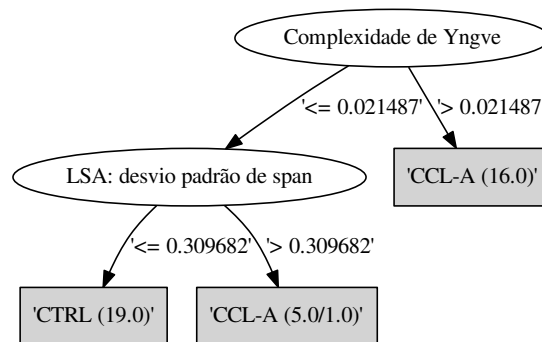


Figura 25 – Árvore de decisão gerada pelo algoritmo J48 para duas classes (cenário 2) utilizando-se os atributos selecionados pelo CFS.

A Tabela 29 mostra a matriz de confusão para o algoritmo J48, utilizando-se os atributos selecionados pelo CFS. Pela tabela, o maior número de erros (três) consistiu em indivíduos saudáveis sendo classificados como possuindo CCL; dentre os pacientes com CCL, apenas um foi incorretamente classificado como saudável. A Figura 25 mostra a árvore de decisão gerada pelo algoritmo J48, utilizando-se os atributos selecionados pelo CFS. Na árvore, os atributos presentes são a complexidade de Yngve e o desvio padrão de *span* da LSA.

A Tabela 30 mostra o *p*-valor do teste de Wilcoxon *signed rank* pareado unilateral, utilizado para comparar o desempenho utilizando-se as 73 métricas do Coh-Metrix-Dementia com o desempenho utilizando-se as 48 métricas do Coh-Metrix-Port. Com esse teste, pretende-se agregar evidências à hipótese de que as métricas adicionadas neste trabalho efetivamente impactaram o desempenho na classificação. Para o cenário com três classes, pode-se notar que uma diferença estatisticamente significativa (para $p < 0,05$) é obtida com os algoritmos Naive Bayes, JRip, e J48. Para os algoritmos SVM, MLP, e regressão logística, foi obtido um *p*-valor muito maior que 0,05; assim, não há evidências para rejeitar a hipótese nula, que declara que não existem condições de tendência ou padrões no experimento. Para o cenário com duas classes, apenas os algoritmos JRip e J48 apresentaram diferença estatisticamente significativa; para os outros algoritmos, em geral, o *p*-valor foi bastante superior a 0,05, levando a conclusões semelhantes.

Concluimos, com base nos resultados apresentados e em resposta às questões de pesquisa,

Tabela 30 – p -valor para teste de Wilcoxon pareado unilateral, no cenário 1, Coh-Metrix-Dementia versus Coh-Metrix-Port.

Algoritmo	p-valor
<i>NB</i>	0.036
<i>SVM</i>	0.214
<i>MLP</i>	0.690
<i>RL</i>	0.118
<i>JRip</i>	0.003
<i>J48</i>	0.021

Tabela 31 – p -valor para teste de Wilcoxon pareado unilateral, no cenário 2, Coh-Metrix-Dementia versus Coh-Metrix-Port.

Algoritmo	p-valor
<i>NB</i>	0.091
<i>SVM</i>	0.725
<i>MLP</i>	0.725
<i>RL</i>	0.579
<i>JRip</i>	0.012
<i>J48</i>	0.000

que é possível separar os sujeitos em três grupos (controles, DA, e CCL) com medida F de 81,7%, utilizando-se o algoritmo Naive Bayes, e também separar os sujeitos em duas classes (controles e CCL), com medida F de 90,0%, utilizando-se o algoritmo J48. Com relação aos atributos utilizados, notou-se que os atributos mais relevantes para as tarefas de classificação, selecionados pelo algoritmo CFS (vide Tabelas 24 e 28), são compostos em grande medida por métricas adicionadas neste trabalho, o que mostra o seu valor e impacto na classificação.

Acreditamos que os valores de medida F aqui obtidos indicam que os atributos atualmente suportados pelo Coh-Metrix-Dementia têm o potencial de detectar declínios nas habilidades linguísticas dos sujeitos ainda em estágio inicial; entretanto, acreditamos que nosso cópulo de teste tem um tamanho reduzido, e que testes maiores, incluindo mais textos de mais pacientes, devem ser realizados, antes que se conclua em definitivo a favor dessa hipótese.

CONCLUSÕES

Neste Capítulo, são apresentadas as conclusões obtidas neste trabalho, suas limitações, e sugestões de trabalhos futuros.

7.1 Conclusões e Contribuições

Os objetivos desta pesquisa de mestrado, apresentados na introdução e reproduzidos abaixo, são:

1. Contribuir para a detecção precoce automática de alterações linguístico-cognitivas, bem como para a identificação de características linguísticas distintivas no diagnóstico da DA e do CCL amnésico.
2. Fornecer à equipe de saúde a classe a que um indivíduo pertence (sadio, DA, ou CCL), com base em algoritmos de classificação da área de aprendizado de máquina, treinados com uso de atributos da área de PLN, para auxílio ao diagnóstico final.

Esta pesquisa de mestrado produziu a ferramenta Coh-Matrix-Dementia, uma biblioteca na linguagem Python, em código aberto, liberada sob licença GPL versão 3¹, que permite a extração de atributos textuais de textos escritos e transcritos, contanto que os textos transcritos sejam adequadamente adaptados. O código fonte da biblioteca, juntamente com a documentação para instalação e *download* dos modelos necessários para execução, encontra-se disponível em <https://github.com/andrecunha/coh-matrix-dementia>.

Produziu também uma interface Web, igualmente de código aberto, capaz de facilitar o acesso às funcionalidades básicas do Coh-Matrix-Dementia para profissionais da saúde. O código fonte da interface Web encontra-se disponível em <https://github.com/andrecunha/cmd-webint/>.

¹ Disponível em <http://www.gnu.org/licenses/gpl-3.0.en.html>.

A interface ainda se encontra em desenvolvimento ativo, e uma versão de demonstração pode ser acessada em <<http://143.107.183.175:22380/>>. Foi disponibilizada, ainda, a biblioteca IDD3, capaz de extrair proposições de textos em Português e em Inglês (com a possibilidade de extensão para outros idiomas). O código fonte da biblioteca IDD3 pode ser obtido em <<https://github.com/andrecunha/idd3/>>.

O Coh-Matrix-Dementia é capaz de produzir os valores das métricas desejadas para os textos analisados, e exportar esses dados em diversos formatos; tais dados podem ser informados como entrada a ferramentas de treinamento de classificadores (como a WEKA, utilizada na avaliação do capítulo anterior, ou outras bibliotecas na própria linguagem Python), produzindo classificadores capazes de prever a qual das classes de interesse o texto pertence. Além disso, os resultados da análise automática podem ser submetidos ao crivo humano, como ilustrado na avaliação deste trabalho no Capítulo 6, o que permite verificar a consistência da análise automática com a análise humana já disponível na literatura.

Com base nos resultados obtidos nas tarefas de classificação, notamos que houve uma diferença marginal entre o desempenho do Coh-Matrix-Dementia e do Coh-Matrix-Port para os textos dos pacientes, com o Coh-Matrix-Dementia apresentando resultados ligeiramente superiores. Concluímos que os experimentos dão indícios positivos de que se pode efetivamente obter desempenho aceitável na tarefa de classificação, especialmente com o emprego das métricas atuais, e que as métricas do Coh-Matrix-Dementia têm o potencial de detectar declínios nas habilidades de linguagem ainda em estágio inicial. Porém, ressaltamos que o corpus de treinamento utilizado aqui é bastante pequeno, o que torna estes resultados preliminares, não sendo considerados conclusivos para propósitos clínicos.

Desta forma, acreditamos que a biblioteca Coh-Matrix-Dementia, principalmente em seu uso combinado com ferramentas de classificação, permite alcançar os objetivos deste trabalho, contribuindo com uma análise rápida e sistematizada dos textos dos pacientes, bem como com a informação de a que classe o sujeito provavelmente pertence, com medida F de 81,7 % para a tarefa com três classes, resultado similar ao do trabalho de Jarrold *et al.* (2010). Acreditamos que esses recursos podem ser de grande utilidade para a pesquisa sobre a deterioração da linguagem nas demências, para o Português Brasileiro.

7.2 Limitações e Trabalhos Futuros

Conforme já mencionado, não foi possível obter resultados com alto valor de medida F para a classificação ternária. Porém, julgamos que os resultados aqui obtidos indicam que a abordagem proposta e adotada é promissora, e sugere-se que, no futuro, o sistema seja testado com uma quantidade consideravelmente maior de textos de pacientes, para que se possa obter classificadores e medidas de avaliação mais confiáveis.

A ferramenta IDD3, responsável pela extração de proposições de um texto, comporta-

se melhor para a língua inglesa do que para a língua portuguesa. A saber, alguns fenômenos característicos do português não são tratados pela ferramenta, e ela não é capaz de lidar, mesmo na língua inglesa, com sentenças mal formadas (sentenças sem verbo, ou com elipses que impactem significativamente a estrutura da frase, como elipses de verbos).

Como trabalho futuro no IDD3, sugerem-se refinamentos na emissão de conexões como um fator importante para melhorar a confiabilidade geral do sistema. Sugere-se ainda refinar a saída do sistema por meio de ferramentas como um Reconhecedor de Entidades Nomeadas (e.g., o sistema Rembrandt²), o que evitaria a emissão de proposições desnecessárias. É importante também avaliar o desempenho do sistema para árvores de dependências não revisadas, comparando-se diferentes softwares de análise, o que não foi realizado neste trabalho. Sugere-se também um levantamento dos fenômenos da língua portuguesa que diferem dos padrões sintático-semânticos presentes na língua inglesa, e a posterior adaptação das regras do IDD3 para tais casos.

As métricas relacionadas à complexidade sintática (complexidade de Yngve, de Frazier, e distância de dependência) partiram de árvores geradas por analisadores automáticos (LX-Parser para árvores de constituintes, MaltParser para árvores de dependência). Tais analisadores não foram avaliados neste trabalho, nem se mediu o impacto de seus erros nos valores das métricas obtidos ao final. Acreditamos que esta avaliação é importante para que o sistema final seja considerado confiável e robusto.

Sugerimos ainda a adição ao Coh-Metrix-Dementia de métricas criadas com base no LIWC, descrito no capítulo de trabalhos relacionados. Além disso, ferramentas melhores para o cálculo de radicais (*stemmers*) podem ser incorporadas no Coh-Metrix-Dementia; nessa categoria, cite-se o *stemmer* desenvolvido por Orengo e Huyck (2001).

Nos nossos experimentos, os algoritmos de classificação foram executados utilizando-se suas configurações padrão na ferramenta WEKA. Para estas configurações, o algoritmo SVM não apresentou os melhores resultados, mas sabe-se que, com a escolha correta dos parâmetros, é possível que se obtenha resultados melhores com essa técnica. Assim, sugerimos que se explore melhor os parâmetros do classificador SVM, por exemplo testando-se o uso de um *kernel* polinomial de segunda ordem, como em Roark *et al.* (2011).

Tal como se encontra hoje, o sistema depende de intervenção humana em diversos de seus aspectos operacionais. Em particular, a transcrição do áudio da entrevista com o paciente, a anotação dos eventos de interesse na transcrição (a duração das pausas, os prolongamentos de vogais, as emissões vazias), e a adaptação da transcrição, para análise por parte das ferramentas automatizadas aqui empregadas, depende exclusivamente de trabalho manual humano. Acreditamos que a automatização das tarefas supramencionadas é bastante complexa, e que soluções eficientes e operativas a curto prazo dificilmente surgirão, mas apontamos que a automatização

² <<http://xldb.di.fc.ul.pt/Rembrandt/>>

de tais processos é uma tarefa, apesar de desafiadora, de potencial interesse acadêmico, clínico, e comercial.

REFERÊNCIAS

- ALBERT, M. L.; SPIRO, A.; SAYERS, K. J.; COHEN, J. A.; BRADY, C. B.; GORAL, M.; OBLER, L. K. Effects of Health Status on Word Finding in Aging. **Journal of the American Geriatrics Society**, Blackwell Publishing Inc, v. 57, n. 12, p. 2300–2305, 2009. ISSN 1532-5415. Disponível em: <<http://dx.doi.org/10.1111/j.1532-5415.2009.02559.x>>. Citado na página 30.
- ALMOR, A.; KEMPLER, D.; MACDONALD, M. C.; ANDERSEN, E. S.; TYLER, L. K. Why Do Alzheimer Patients Have Difficulty with Pronouns? Working Memory, Semantics, and Reference in Comprehension and Production in Alzheimer's Disease. **Brain and Language**, v. 67, n. 3, p. 202–227, 1999. ISSN 0093-934X. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0093934X99920552>>. Citado na página 71.
- 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.
- ALUISIO, S.; SPECIA, L.; GASPERIN, C.; SCARTON, C. Readability assessment for text simplification. In: **Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (IUNLPBEA '10), p. 1–9. Disponível em: <<http://dl.acm.org/citation.cfm?id=1866795.1866796>>. Citado 2 vezes nas páginas 55 e 61.
- ALUÍSIO, S. M.; GASPERIN, C. Fostering digital inclusion and accessibility: the PorSimples project for simplification of Portuguese texts. In: **Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (YIWCALA '10), p. 46–53. Citado na página 56.
- ALUÍSIO, S. M.; SPECIA, L.; PARDO, T. A. S.; MAZIERO, E. G.; CASELI, H. M.; FORTES, R. P. M. A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps Towards Text Simplification Systems. In: **Proceedings of the 26th Annual ACM International Conference on Design of Communication**. New York, NY, USA: ACM, 2008. (SIGDOC '08), p. 15–22. ISBN 978-1-60558-083-8. Disponível em: <<http://doi.acm.org/10.1145/1456536.1456540>>. Citado 2 vezes nas páginas 54 e 60.
- AMARAL-CARVALHO, V.; CARAMELLI, P. Normative Data for Healthy Middle-Aged and Elderly Performance on the Addenbrooke Cognitive Examination-Revised. **Cognitive and Behavioral Neurology**, v. 25, n. 2, p. 72–76, 2012. ISSN 1543-3633. Citado na página 31.
- ANDERSON, S.; WHITE-SCHWOCH, T.; PARBERRY-CLARK, A.; KRAUS, N. A dynamic auditory-cognitive system supports speech-in-noise perception in older adults. **Hearing Research**, v. 300, n. 0, p. 18–32, 2013. ISSN 0378-5955. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0378595513000749>>. Citado na página 29.
- ANDRADE, C. R. F.; MARTINS, V. O. Variação da fluência da fala em idosos. **Pro-Fono Revista de Atualização Científica**, v. 22, n. 1, p. 13–18, 2010. Citado 2 vezes nas páginas 28 e 30.

ANDREETTA, S.; CANTAGALLO, A.; MARINI, A. Narrative discourse in anomic aphasia. **Neuropsychologia**, v. 50, n. 8, p. 1787–1793, 2012. ISSN 0028-3932. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0028393212001546>>. Citado na página 24.

ARDILA, A.; BERTOLUCCI, P. H.; BRAGA, L. W.; CASTRO-CALDAS, A.; JUDD, T.; KOSMIDIS, M. H.; MATUTE, E.; NITRINI, R.; OSTROSKY-SOLIS, F.; ROSSELLI, M. Illiteracy: The Neuropsychology of Cognition Without Reading. **Archives of Clinical Neuropsychology**, v. 25, n. 8, p. 689–712, 2010. Disponível em: <<http://acn.oxfordjournals.org/content/25/8/689.abstract>>. Citado na página 108.

ARONOFF, J. M.; GONNERMAN, L. M.; ALMOR, A.; ARUNACHALAM, S.; KEMPLER, D.; ANDERSEN, E. S. Information content versus relational knowledge: Semantic deficits in patients with Alzheimer's disease. **Neuropsychologia**, v. 44, n. 1, p. 21–35, 2006. ISSN 0028-3932. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0028393205001752>>. Citado na página 71.

Balage Filho, P.; PARDO, T.; ALUÍSIO, S. An Evaluation of the Brazilian Portuguese LIWC Dictionary. **To be published in the Proceedings of STIL 2013**, p. 5, 2013. Citado na página 63.

BAYLES, K. A.; TOMOEDA, C. K.; TROSSET, M. W. Relation of linguistic communication abilities of Alzheimer's patients to stage of disease. **Brain and Language**, v. 42, n. 4, p. 454–472, 1992. ISSN 0093-934X. Disponível em: <<http://www.sciencedirect.com/science/article/pii/0093934X9290079T>>. Citado na página 35.

BERNDT, R. S.; WAYLAND, S.; ROCHON, E.; SAFFRAN, E.; SCHWARTZ, M. **Quantitative Production Analysis: A Training Manual for the Analysis of Aphasic Sentence Production**. Hove, UK: Psychology Press, 2000. Citado na página 88.

BICK, E. **The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. [S.l.]: Aarhus University Press, 2000. Citado 2 vezes nas páginas 61 e 101.

BICKEL, C.; PANTEL, J.; EYSENBACH, K.; SCHRÖDER, J. Syntactic Comprehension Deficits in Alzheimer's Disease. **Brain and Language**, v. 71, n. 3, p. 432–448, 2000. ISSN 0093-934X. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0093934X99922770>>. Citado na página 71.

BOEUF, C. **Raconte...: 55 historiettes en images**. Ecole, 1971. Disponível em: <<http://books.google.com.br/books?id=bpe7nQEACAAJ>>. Citado na página 39.

BROWN, C.; SNODGRASS, T.; KEMPER, S.; HERMAN, R.; COVINGTON, M. Automatic measurement of propositional idea density from part-of-speech tagging. **Behavior Research Methods**, Springer-Verlag, v. 40, n. 2, p. 540–545, 2008. ISSN 1554-351X. Disponível em: <<http://dx.doi.org/10.3758/BRM.40.2.540>>. Citado 2 vezes nas páginas 71 e 73.

BROWN, R. **A first language**. Cambridge, MA: Harvard University Press, 1973. Citado na página 70.

BRUCKI, S.; ROCHA, M. Category fluency test: effects of age, gender and education on total scores, clustering and switching in Brazilian Portuguese-speaking subjects. **Brazilian Journal of Medical and Biological Research**, scielo, v. 37, n. 12, p. 1771–1777, 2004. ISSN 0100-879X. Citado na página 31.

BRUCKI, S. M. D.; MALHEIROS, S. M. F.; OKAMOTO, I. H.; BERTOLUCCI, P. H. F. Dados normativos para o teste de fluência verbal categoria animais em nosso meio. **Arq Neuropsiquiatr**, v. 55, n. 1, p. 56–61, 1997. Citado na página 31.

BRUNET, É. **Le Vocabulaire de Jean Giraudoux: structure et évolution : statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la langue française**. Slatkine, 1978. (Travaux de linguistique quantitative). Disponível em: <<http://books.google.com.br/books?id=J6o2AAAAIAAJ>>. Citado na página 64.

BRYANT, L.; SPENCER, E.; FERGUSON, A.; CRAIG, H.; COLYVAS, K.; WORRALL, L. Propositional Idea Density in aphasic discourse. **Aphasiology**, v. 27, n. 8, p. 992–1009, 2013. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/02687038.2013.803514>>. Citado 4 vezes nas páginas 17, 25, 85 e 86.

BSCHOR, T.; KÜHL, K.-P.; REISCHIES, F. M. Spontaneous Speech of Patients With Dementia of the Alzheimer Type and Mild Cognitive Impairment. **International Psychogeriatrics**, v. 13, p. 289–298, 9 2001. ISSN 1741-203X. Disponível em: <http://journals.cambridge.org/article_S1041610201007682>. Citado na página 41.

BUCKS, R. S.; SINGH, S.; CUERDEN, J. M.; WILCOCK, G. K. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. **Aphasiology**, v. 14, n. 1, p. 71–91, 2000. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/026870300401603>>. Citado 2 vezes nas páginas 74 e 82.

CANNIZZARO, M. S.; COELHO, C. A. Analysis of Narrative Discourse Structure as an Ecologically Relevant Measure of Executive Function in Adults. **Journal of Psycholinguistic Research**, Springer US, p. 1–23, 2012. ISSN 0090-6905. Disponível em: <<http://dx.doi.org/10.1007/s10936-012-9231-5>>. Citado na página 24.

CARAMELLI, P.; MANSUR, L. L.; NITRINI, R. Language and Communication Disorders in Dementia of the Alzheimer Type. In: STEMMER, B.; WHITAKER, H. (Ed.). **Handbook of Neurolinguistics**. [S.l.]: Academic Press, 1998. Citado na página 30.

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 38.

CARVALHO, I. A. de; MANSUR, L. L. Validation of ASHA FACS-functional assessment of communication skills for Alzheimer disease population. **Alzheimer disease and associated disorders**, v. 22, n. 4, p. 375–381, 2008. Validation Studies, Research Support, Non-U.S. Gov't,. Disponível em: <<http://europepmc.org/abstract/MED/19068501>>. Citado na página 21.

CERA, M. L.; ORTIZ, K. Z.; BERTOLUCCI, P. H.; MINETT, T. S. Speech and orofacial apraxias in Alzheimer's disease. **Int Psychogeriatr**, v. 25, n. 10, p. 1679–1685, 2013. Citado na página 36.

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.

- CHAND, V.; BAYNES, K.; BONNICI, L.; FARIAS, S. T. **Analysis of Idea Density (AID): A Manual**. [S.l.], 2010. Citado 10 vezes nas páginas 17, 24, 25, 71, 72, 73, 102, 114, 117 e 121.
- CHEN, D.; MANNING, C. A fast and accurate dependency parser using neural networks. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Association for Computational Linguistics, 2014. p. 740–750. Disponível em: <<http://aclweb.org/anthology/D14-1082>>. Citado na página 114.
- CHEN, S. F.; CHEN, S. F.; GOODMAN, J.; GOODMAN, J. **An Empirical Study of Smoothing Techniques for Language Modeling**. [S.l.], 1998. Citado na página 102.
- CHEUNG, H.; KEMPER, S. Competing complexity metrics and adults' production of complex sentences. **Applied Psycholinguistics**, v. 13, p. 53–76, 1 1992. ISSN 1469-1817. Disponível em: <http://journals.cambridge.org/article_S0142716400005427>. Citado 2 vezes nas páginas 68 e 70.
- COOPER, P. V. Discourse Production and Normal Aging: Performance on Oral Picture Description Tasks. **Journal of Gerontology**, v. 45, n. 5, p. 210–214, 1990. Citado na página 24.
- CREAMER, S.; SCHMITTER-EDGEcombe, M. Narrative comprehension in Alzheimer's disease: assessing inferences and memory operations with a think-aloud procedure. **Neuropsychology**, v. 24, n. 3, p. 279–290, 2010. Citado na página 35.
- CROOK, T.; BARTUS, R. T.; FERRIS, S. H.; WHITEHOUSE, P.; COHEN, G. D.; GERSHON, S. Age-associated memory impairment: Proposed diagnostic criteria and measures of clinical change — report of a national institute of mental health work group. **Developmental Neuropsychology**, v. 2, n. 4, p. 261–276, 1986. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/87565648609540348>>. Citado na página 27.
- 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.; SOUSA, L. Bender de; MANSUR, L.; ALUISIO, S. Automatic proposition extraction from dependency trees: Helping early prediction of alzheimer's disease from narratives. In: **Computer-Based Medical Systems (CBMS), 2015 IEEE 28th International Symposium on**. [S.l.: s.n.], 2015. p. 127–130. Citado 6 vezes nas páginas 15, 17, 114, 115, 116 e 117.
- DEERWESTER, S.; DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K.; HARSHMAN, R. Indexing by latent semantic analysis. **JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE**, v. 41, n. 6, p. 391–407, 1990. Citado na página 75.
- 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 37.
- DUMAIS, S. T. Latent semantic analysis. **Annual Review of Information Science and Technology**, Wiley Subscription Services, Inc., A Wiley Company, v. 38, n. 1, p. 188–230, 2004. ISSN 1550-8382. Disponível em: <<http://dx.doi.org/10.1002/aris.1440380105>>. Citado 3 vezes nas páginas 75, 76 e 77.

- DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K.; DEERWESTER, S.; HARSHMAN, R. Using Latent Semantic Analysis to Improve Access to Textual Information. In: **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. New York, NY, USA: ACM, 1988. (CHI '88), p. 281–285. ISBN 0-201-14237-6. Disponível em: <<http://doi.acm.org/10.1145/57167.57214>>. Citado na página 75.
- DUNN, J. C.; ALMEIDA, O. P.; BARCLAY, L.; WATERREUS, A.; FLICKER, L. Latent Semantic Analysis: A New Method to Measure Prose Recall. **Journal of Clinical and Experimental Neuropsychology**, v. 24, n. 1, p. 26–35, 2002. PMID: 11935421. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1076/jcen.24.1.26.965>>. Citado na página 77.
- FACAL-MAYO, D.; JUNCOS-RABADÁN, O.; ÁLVAREZ, M.; PEREIRO-ROZAS, A. X.; Díaz Fernández, F. Efectos del envejecimiento en el acceso al léxico: el fenómeno de la punta de la lengua ante los nombres propios. **Rev Neurol**, v. 43, p. 719–723, 2006. Citado na página 31.
- FERGADIOTIS, G.; WRIGHT, H. H.; CAPILOUTO, G. J. Productive vocabulary across discourse types. **Aphasiology**, v. 25, n. 10, p. 1261–1278, 2011. ISSN 0268-7038. Disponível em: <<http://www.biomedsearch.com/nih/Productive-vocabulary-across-discourse-types/22904592.html>>. Citado na página 32.
- FERGUSON, A.; SPENCER, E.; CRAIG, H.; COLYVAS, K. Propositional Idea Density in women's written language over the lifespan: Computerized analysis. 2013. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0010945213001408>>. Citado na página 33.
- FILLMORE, C. J. The Case for Case. In: BACH, E.; HARMS, R. T. (Ed.). **Universals in Linguistic Theory**. New York: Holt, Rinehart and Winston, 1968. p. 0–88. Citado na página 72.
- _____. Toward a modern theory of case. In: REIBEL, D. A.; SCHANE, S. A. (Ed.). **Modern Studies in English**. [S.l.]: Prentice Hall, 1969. p. 361–375. Citado na página 72.
- FINATTO, M. J. B.; SCARTON, C. E.; ROCHA, A.; ALUÍSIO, S. M. Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. In: **The 8th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2011. v. 1. Citado 2 vezes nas páginas 56 e 61.
- FISCHMAN, H. C.; FERNANDES, C. S.; LOURENÇO, R. A.; PARADELA, E. M. P.; CARTHERY-GOULART, M. T.; CARAMELLI, P. Age and educational level effects on the performance of normal elderly on category fluency tasks. **Dement Neuropsychol**, v. 3, n. 1, p. 49–54, 2009. Citado na página 31.
- 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**, v. 12, n. 3, p. 189–198, nov. 1975. ISSN 0022-3956. Disponível em: <<http://view.ncbi.nlm.nih.gov/pubmed/1202204>>. Citado na página 82.
- FOLTZ, P. W.; KINTSCH, W.; LANDAUER, T. K. The measurement of textual coherence with latent semantic analysis. **Discourse Processes**, v. 25, n. 2-3, p. 285–307, 1998. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/01638539809545029>>. Citado na página 77.
- FOLTZ, P. W.; LAHAM, D.; LANDAUER, T. K. The Intelligent Essay Assessor: Applications to Educational Technology. **Interactive Multimedia Education Journal of Computer enhanced learning On-line journal**, v. 1(2), 1999. Citado na página 77.

FONSECA, E.; ROSA, J. Mac-morpho revisited: Towards robust part-of-speech tagging. In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2013. p. 98–107. Citado na página 95.

_____. A two-step convolutional neural network approach for semantic role labeling. In: **Proceedings of the 2013 International Joint Conference on Neural Networks**. [S.l.: s.n.], 2013. p. 2955–2961. ISSN 2161-4393. Citado na página 95.

FONSECA, E.; ROSA, J. G.; ALUÍSIO, S. Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. **Journal of the Brazilian Computer Society**, Springer London, v. 21, n. 1, 2015. ISSN 0104-6500. Disponível em: <<http://dx.doi.org/10.1186/s13173-014-0020-x>>. Citado na página 95.

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 2 vezes nas páginas 38 e 41.

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**, 2012. ISSN 0010-9452. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0010945212003413>>. Citado 9 vezes nas páginas 17, 22, 23, 24, 25, 88, 89, 109 e 115.

FRAZIER, L. Syntactic Complexity. In: DOWTY, D. R.; KARTTUNEN, L.; ZWICKY, A. M. (Ed.). **Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives**. Cambridge: Cambridge University Press, 1985. p. 129–189. Citado na página 66.

FRIED, L. P. Epidemiology of Aging. **Epidemiologic Reviews**, v. 22, n. 1, p. 95–106, 2000. Disponível em: <<http://epirev.oxfordjournals.org/content/22/1/95.short>>. Citado na página 22.

FROTA, N. A. F.; NITRINI, R.; DAMASCENO, B. P.; FORLENZA, O.; DIAS-TOSTA, E.; SILVA, A. B. da; Herrera Junior, E.; MAGALDI, R. M. Critérios para o diagnóstico de doença de Alzheimer. **Dement Neuropsychol**, v. 5, n. 1, p. 5–10, 2011. Citado na página 33.

GARCIA, F. H. A.; MANSUR, L. L. Habilidades funcionais de comunicação: idoso saudável. **Acta fisiatr**, v. 13, n. 2, p. 87–89, 2006. Citado na página 28.

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.

GIBSON, E. Linguistic complexity: Locality of syntactic dependencies. **Cognition**, v. 68, n. 1, p. 1–76, August 1998. Disponível em: <[http://dx.doi.org/10.1016/S0010-0277\(98\)00034-1](http://dx.doi.org/10.1016/S0010-0277(98)00034-1)>. Citado na página 68.

GOODGLASS, H.; KAPLAN, E.; BARRESI, B. **The Assessment of Aphasia and Related Disorders**. 3. ed. [S.l.]: Lippincott Williams & Wilkins, 2001. 102 p. ISBN 978-0683036046. Citado na página 38.

GORNO-TEMPINI, M. L.; HILLIS, A. E.; WEINTRAUB, S.; KERTESZ, A.; MENDEZ, M.; CAPPA, S. F.; OGAR, J. M.; ROHRER, J. D.; BLACK, S.; BOEVE, B. F.; MANES, F.; DRONKERS, N. F.; VANDENBERGHE, R.; RASCOVSKY, K.; PATTERSON, K.; MILLER, B. L.; KNOPMAN, D. S.; HODGES, J. R.; MESULAM, M. M.; GROSSMAN, M. Classification of primary progressive aphasia and its variants. **Neurology**, Lippincott Williams & Wilkins, v. 76, n. 11, p. 1006–1014, mar. 2011. ISSN 1526-632X. Disponível em: <<http://dx.doi.org/10.1212/wnl.0b013e31821103e6>>. Citado na página 22.

GRAESSER, A.; MCNAMARA, D.; LOUWERSE, M.; CAI, Z. Coh-Metrix: Analysis of text on cohesion and language. **Behavior Research Methods, Instruments, & Computers**, Springer-Verlag, v. 36, n. 2, p. 193–202, 2004. ISSN 0743-3808. Disponível em: <<http://dx.doi.org/10.3758/BF03195564>>. Citado na página 46.

GRAESSER, A. C.; CAI, Z.; LOUWERSE, M. M.; DANIEL, F. Question Understanding Aid (QUAID) A Web Facility that Tests Question Comprehensibility. **Public Opinion Quarterly**, AAPOR, v. 70, n. 1, p. 3–22, 2006. Citado na página 48.

GRAESSER, A. C.; MCNAMARA, D. S.; KULIKOWICH, J. M. Coh-Metrix: Providing Multi-level Analyses of Text Characteristics. **Educational Researcher**, v. 40, n. 5, p. 223–234, 2011. Disponível em: <<http://edr.sagepub.com/content/40/5/223.abstract>>. Citado 4 vezes nas páginas 46, 47, 48 e 49.

GROSSMAN, M.; KOENIG, P.; DEVITA, C.; GLOSSER, G.; MOORE, P.; GEE, J.; DETRE, J.; ALSOP, D. Neural basis for verb processing in Alzheimer's disease: an fMRI study. **Neuropsychology**, v. 17, n. 4, p. 658–74, 2003. ISSN 0894-4105. Disponível em: <<http://www.biomedsearch.com/nih/Neural-basis-verb-processing-in/14599278.html>>. Citado na página 41.

GROSSMAN, M.; RHEE, J. Cognitive resources during sentence processing in Alzheimer's disease. **Neuropsychologia**, v. 39, n. 13, p. 1419–1431, 2001. ISSN 0028-3932. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0028393201000598>>. Citado na página 35.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA data mining software: an update. **SIGKDD Explor. Newsl.**, ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, nov. 2009. ISSN 1931-0145. Disponível em: <<http://doi.acm.org/10.1145/1656274.1656278>>. Citado 2 vezes nas páginas 110 e 122.

HALL, M. A. **Correlation-based Feature Subset Selection for Machine Learning**. Tese (Doutorado) — University of Waikato, Hamilton, New Zealand, 1998. Citado 2 vezes nas páginas 111 e 123.

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.

HENRY, M. L.; GORNO-TEMPINI, M. L. The logopenic variant of primary progressive aphasia. **Current Opinion in Neurology**, v. 23, n. 6, p. 633–637, 2010. ISSN 1350-7540. Disponível em: <http://journals.lww.com/co-neurology/Fulltext/2010/12000/The_logopenic_variant_of_primary_progressive.14.aspx>. Citado na página 22.

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.

HOLMES, D. I.; SINGH, S. A Stylometric Analysis of Conversational Speech of Aphasic Patients. **Literary and Linguistic Computing**, v. 11, n. 3, p. 133–140, 1996. Citado na página 74.

HONORÉ, A. Some Simple Measures of Richness of Vocabulary. **Association for Literary and Linguistic Computing Bulletin**, v. 7, p. 172–177, 1979. Citado na página 64.

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.

HORTON, W. S.; SPIELER, D. H. Age-related effects in communication and audience design. **Psychol Aging**, v. 22, p. 281–290, 2007. Citado na página 30.

HOWARD, D.; PATTERSON, K.; COMPANY, T. V. T. **The Pyramids and Palm Trees Test: A Test of Semantic Access from Words and Pictures : [manual]**. Thames Valley Test Company, 1992. ISBN 9781874261155. Disponível em: <<http://books.google.com.br/books?id=dykONQAACA AJ>>. Citado na página 41.

Jacob Filho, W. Envelhecimento e atendimento domiciliário. In: DUARTE, Y. A. O.; DIOGO, M. J. D. (Ed.). **Atendimento domiciliar: um enfoque gerontológico**. São Paulo: Atheneu, 2000. p. 19–25. Citado na página 21.

JARROLD, W.; PEINTNER, B.; YEH, E.; KRASNOW, R.; JAVITZ, H.; SWAN, G. 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.). **Brain Informatics**. [S.l.]: Springer Berlin Heidelberg, 2010, (Lecture Notes in Computer Science, v. 6334). p. 299–307. ISBN 978-3-642-15313-6. Citado 6 vezes nas páginas 17, 22, 23, 25, 87 e 134.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of Documentation**, v. 28, p. 11–21, 1972. Citado na página 97.

KALPOUZOS, G.; EUSTACHE, F.; SAYETTE, V.; VIADER, F.; CHÉTELAT, G.; DESGRANGES, B. Working memory and FDG-PET dissociate early and late onset Alzheimer disease patients. **Journal of Neurology**, Steinkopff-Verlag, v. 252, n. 5, p. 548–558, 2005. ISSN 0340-5354. Disponível em: <<http://dx.doi.org/10.1007/s00415-005-0685-3>>. Citado na página 36.

KAPLAN, E.; GOODGLASS, H.; WEINTRAUB, S. **Boston Naming Test**. 2. ed. Philadelphia: Lippincott Williams & Wilkins, 2001. Citado na página 40.

KEMPER, S.; GREINER, L. H.; MARQUIS, J. G.; PRENOVOST, K.; MITZNER, T. L. Language decline across the life span: findings from the Nun Study. **Psychol Aging**, v. 16, n. 2, p. 227–239, jun 2001. Citado na página 31.

KEMPER, S.; HERMAN, R.; LIAN, C. Age Differences in Sentence Production. **The Journals of Gerontology Series B: Psychological Sciences and Social Sciences**, v. 58, n. 5, p. 260–268, 2003. Disponível em: <<http://psychsocgerontology.oxfordjournals.org/content/58/5/P260.abstract>>. Citado na página 31.

KEMPER, S.; KYNETTE, D.; RASH, S.; O'BRIEN, K.; SPROTT, R. Life-span changes to adults' language: Effects of memory and genre. **Applied Psycholinguistics**, v. 10, p. 49–66, 1989. Citado na página 70.

KEMPER, S.; SUMNER, A. The structure of verbal abilities in young and older adults. **Psychology and Aging**, v. 16, n. 2, p. 312–322, jun. 2001. Citado na página 73.

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 4 vezes nas páginas 31, 36, 71 e 73.

KEMPLER, D. Language changes in dementia of the Alzheimer's type. In: LUBINSKY (Ed.). **Dementia and communication: Research and clinical implications**. San Diego: Singular, 1995. p. 98–114. Citado na página 71.

KEMPLER, D.; ALMOR, A.; TYLER, L. K.; ANDERSEN, E. S.; MACDONALD, M. C. Sentence Comprehension Deficits in Alzheimer's Disease: A Comparison of Off-Line vs. On-Line Sentence Processing. **Brain and Language**, v. 64, n. 3, p. 297–316, 1998. ISSN 0093-934X. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0093934X98919800>>. Citado na página 71.

KENDALL, K. Presbyphonia: a review. **Current Opinion in Otolaryngology & Head and Neck Surgery**, v. 15, n. 3, 2007. ISSN 1068-9508. Disponível em: <http://journals.lww.com/co-otolaryngology/Fulltext/2007/06000/Presbyphonia__a_review.2.aspx>. Citado na página 29.

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.

_____. **Comprehension: A Paradigm for Cognition**. Cambridge: Cambridge University Press, 1998. Citado na página 73.

KINTSCH, W.; KEENAN, J. Reading Rate and Retention as a Function of the Number of Propositions in the Base Structure of Sentences. **Cognitive Psychology**, v. 5, p. 257–274, 1973. Citado na página 73.

LAM, B.; MASELLIS, M.; FREEDMAN, M.; STUSS, D. T.; BLACK, S. E. Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome. **Alzheimers Res Ther**, v. 5, n. 1, p. 1, 2013. ISSN 1758-9193. Citado na página 23.

LANDAUER, T. K.; DUMAIS, S. T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. **Psychological review**, v. 104, n. 2, p. 211–240, 1997. Disponível em: <<http://www.bibsonomy.org/bibtex/264a15051245f4da439e8c315d8edd681/infospace>>. Citado na página 77.

LANDAUER, T. K.; LAHAM, D.; REHDER, B.; SCHREINER, M. E. How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. **Cognitive Science**, 1997. Citado na página 77.

- Le Dorze, G.; BÉDARD, C. Effects of age and education on the lexico-semantic content of connected speech in adults. **Journal of Communication Disorders**, v. 31, n. 1, p. 53–71, 1998. ISSN 0021-9924. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0021992497000518>>. Citado 2 vezes nas páginas 108 e 109.
- LIN, D. On the Structural Complexity of Natural Language Sentences. In: **COLING**. [S.l.: s.n.], 1996. p. 729–733. Citado na página 68.
- 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.
- LU, X. Automatic analysis of syntactic complexity in second language writing. **International Journal of Corpus Linguistics**, v. 15, n. 4, p. 474–496, 2010. Disponível em: <<http://www.ingentaconnect.com/content/jbp/ijcl/2010/00000015/00000004/art00002>>. Citado na página 88.
- MACDONALD, M. C.; ALMOR, A.; HENDERSON, V. W.; KEMPLER, D.; ANDERSEN, E. S. Assessing Working Memory and Language Comprehension in Alzheimer's Disease. **Brain and Language**, v. 78, n. 1, p. 17–42, 2001. ISSN 0093-934X. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0093934X00924362>>. Citado na página 71.
- MACKAY, A.; CONNOR, L.; ALBERT, M.; OBLER, L. K. Noun and verb retrieval in healthy aging. **J Int Neuropsychol Soc**, v. 8, n. 6, p. 764–770, 2002. Citado na página 30.
- MACKENZIE, C. Adult spoken discourse: the influences of age and education. **Int. J. Lang. Comm. Dis.**, v. 35, n. 2, p. 269–285, apr-jun 2000. Citado 2 vezes nas páginas 108 e 109.
- MACWHINNEY, B. **The CHILDES Project: Tools for Analyzing Talk**. 3rd. ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000. Citado na página 79.
- MACWHINNEY, B.; FROMM, D.; HOLLAND, A.; FORBES, M.; WRIGHT, H. Automated analysis of the Cinderella story. **Aphasiology**, v. 24, n. 6-8, p. 856–868, 2010. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/02687030903452632>>. Citado 4 vezes nas páginas 15, 79, 80 e 81.
- MANNING, W. H.; SHIRKEY, E. A. Fluency and the Aging Process. In: BEASLEY, D. S.; DAVIS, G. A. (Ed.). **Aging Communication Processes and Disorders**. Nova Iorque: Grune & Straton, 1981. Citado na página 30.
- MANSUR, L. L.; CARTHERY, M. T.; CARAMELLI, P.; NITRINI, R. Linguagem e Cognição na doença de Alzheimer. **Psicologia, Reflexão e Crítica**, v. 18, n. 3, p. 300–307, 2005. Citado na página 28.
- MARINI, A.; BOEWE, A.; CALTAGIRONE, C.; CARLOMAGNO, S. Age-related Differences in the Production of Textual Descriptions. **Journal of Psycholinguistic Research**, Kluwer Academic Publishers-Plenum Publishers, v. 34, n. 5, p. 439–463, 2005. ISSN 0090-6905. Disponível em: <<http://dx.doi.org/10.1007/s10936-005-6203-z>>. Citado 2 vezes nas páginas 32 e 107.
- MARNEFFE, M.-C. de; DOZAT, T.; SILVEIRA, N.; HAVERINEN, K.; GINTER, F.; NIVRE, J.; MANNING, C. D. Universal stanford dependencies: A cross-linguistic typology. In: **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)**. European Language Resources Association (ELRA), 2014. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf>. Citado na página 105.

MARNEFFE, M.-C. de; MANNING, C. D. **Stanford typed dependencies manual**. 2008. Disponível em: <http://nlp.stanford.edu/software/dependencies_manual.pdf>. Citado na página 105.

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.

MAZIERO, E.; PARDO, T.; ALUÍSIO, S. **Ferramenta de análise automática de inteligibilidade de corpus (AIC)**. NILC - ICMC - USP, 2008. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional). Citado 2 vezes nas páginas 60 e 61.

MCDONALD, R.; LERMAN, K.; PEREIRA, F. Multilingual Dependency Analysis with a Two-stage Discriminative Parser. In: **Proceedings of the Tenth Conference on Computational Natural Language Learning**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. (CoNLL-X '06), p. 216–220. Disponível em: <<http://dl.acm.org/citation.cfm?id=1596276.1596317>>. Citado na página 100.

MCDONALD, R. T.; NIVRE, J.; QUIRMBACH-BRUNDAGE, Y.; GOLDBERG, Y.; DAS, D.; GANCHEV, K.; HALL, K. B.; PETROV, S.; ZHANG, H.; TÄCKSTRÖM, O.; BEDINI, C.; CASTELLÓ, N. B.; LEE, J. Universal dependency annotation for multilingual parsing. In: **ACL (2)**. [S.l.]: The Association for Computer Linguistics, 2013. p. 92–97. ISBN 978-1-937284-51-0. Citado na página 105.

MCKENNA, P.; WARRINGTON, E. **Graded naming test**. NFER-Nelson, 1983. Disponível em: <<http://books.google.com.br/books?id=y2LdRwAACAAJ>>. Citado na página 40.

MCKHANN, G. M.; KNOPMAN, D. S.; CHERTKOW, H.; HYMAN, B. T.; JACK, C. R.; KAWAS, C. H.; KLUNK, W. E.; KOROSHETZ, W. J.; MANLY, J. J.; MAYEUX, R.; MOHS, R. C.; MORRIS, J. C.; ROSSOR, M. N.; SCHELTENS, P.; CARRILLO, M. C.; THIES, B.; WEINTRAUB, S.; PHELPS, C. H. 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 : the journal of the Alzheimer's Association**, Elsevier, Inc., v. 7, n. 3, p. 263–269, may 2011. ISSN 1552-5260. Citado na página 22.

MCNAMARA, D.; KINTSCH, E.; SONGER, N. B.; KINTSCH, W. Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning from Text. **Cognition and Instruction**, v. 14, n. 1, p. 1–43, 1996. Citado na página 46.

MCNAMARA, D.; LOUWERSE, M.; GRAESSER, A. **Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension**. 2002. Grant Proposal. <http://cohmetrix.memphis.edu/cohmetrixpr/publications.html>. Citado na página 46.

MENDEZ, M. F.; LEE, A. S.; JOSHI, A.; SHAPIRA, J. S. Nonamnesic Presentations of Early-Onset Alzheimer's Disease. **American Journal of Alzheimer's Disease and Other Dementias**, v. 27, n. 6, p. 413–420, 2012. Disponível em: <<http://aja.sagepub.com/content/27/6/413.abstract>>. Citado na página 34.

MORRIS, J. C. The Clinical Dementia Rating (CDR): current version and scoring rules. **Neurology**, v. 43, n. 11, p. 2412–2414, nov. 1993. ISSN 0028-3878. Disponível em: <<http://view.ncbi.nlm.nih.gov/pubmed/8232972>>. Citado na página 82.

NASREDDINE, Z. S.; PHILLIPS, N. A.; BÉDIRIAN, V.; CHARBONNEAU, S.; WHITEHEAD, V.; COLLIN, I.; CUMMINGS, J. L.; CHERTKOW, H. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. **Journal of the American Geriatrics Society**, Blackwell Science Inc, v. 53, n. 4, p. 695–699, 2005. ISSN 1532-5415. Disponível em: <<http://dx.doi.org/10.1111/j.1532-5415.2005.53221.x>>. Citado na página 41.

NIPPOLD, M. A.; CRAMOND, P. M.; HAYWARD-MAYHEW, C. Spoken language production in adults: Examining age-related differences in syntactic complexity. PMID: 24093162. 2013. Disponível em: <<http://informahealthcare.com/doi/abs/10.3109/02699206.2013.841292>>. Citado na página 32.

NIVRE, J.; HALL, J.; NILSSON, J. MaltParser: A data-driven parser-generator for dependency parsing. In: **In Proc. of LREC-2006**. [S.l.: s.n.], 2006. p. 2216–2219. Citado na página 100.

ORENGO, V.; HUYCK, C. A stemming algorithm for the portuguese language. In: **String Processing and Information Retrieval, 2001. SPIRE 2001. Proceedings.Eighth International Symposium on**. [S.l.: s.n.], 2001. p. 186–193. Citado na página 135.

PAKHOMOV, S.; CHACON, D.; WICKLUND, M.; GUNDEL, J. Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing. **Behavior Research Methods**, Springer-Verlag, v. 43, n. 1, p. 136–144, 2011. Citado 6 vezes nas páginas 15, 65, 66, 67, 68 e 71.

PARK, D. C.; BISCHOF, G. N. The aging mind: neuroplasticity in response to cognitive training. **Dialogues Clin Neurosci**, v. 15, n. 1, p. 109–119, mar 2013. Citado na página 28.

PASCHOAL, S. M. P. Epidemiologia do Envelhecimento. In: Papaléo Netto, M. (Ed.). **Gerontologia**. São Paulo: Atheneu, 1996. p. 26–43. Citado 2 vezes nas páginas 21 e 22.

PASQUALINI, B.; SCARTON, C. E.; FINATTO, M. J. B. Comparando Avaliações de Inteligibilidade Textual entre Originais e Traduções de Textos Literários. In: **The 8th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2011. v. 1. Citado na página 56.

PEINTNER, B.; JARROLD, W.; VERGYRI, D.; RICHEY, C.; TEMPINI, M. L. G.; OGAR, J. Learning diagnostic models using speech and language measures. In: **Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE**. [S.l.: s.n.], 2008. p. 4648–4651. ISSN 1557-170X. Citado 4 vezes nas páginas 17, 25, 86 e 87.

PENNEBAKER, J. W.; CHUNG, C. K.; IRELAND, M.; GONZALES, A.; BOOTH, R. J. **The development and psychometric properties of LIWC2007**. Austin, TX, 2007. Citado na página 62.

PENNEBAKER, J. W.; FRANCIS, M. E.; J., B. R. **Linguistic Inquiry and Word Count (LIWC): LIWC2001**. Mahwah: Lawrence Erlbaum Associates, 2001. Citado na página 62.

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 2 vezes nas páginas 36 e 38.

PETERSEN, R. C.; SMITH, G. E.; WARING, S. C.; IVNIK, R. J.; TANGALOS, E. G.; KOKMEN, E. Mild cognitive impairment: Clinical characterization and outcome. **Archives of Neurology**, v. 56, n. 3, p. 303–308, 1999. Disponível em: <<http://dx.doi.org/10.1001/archneur.56.3.303>>. Citado na página 40.

PETROV, S.; DAS, D.; MCDONALD, R. A universal part-of-speech tagset. In: CHAIR), N. C. C.; CHOUKRI, K.; DECLERCK, T.; DOĞAN, M. U.; MAEGAARD, B.; MARIANI, J.; MORENO, A.; ODIJK, J.; PIPERIDIS, S. (Ed.). **Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)**. Istanbul, Turkey: European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7. Citado na página 105.

PICHORA-FULLER, M. K.; LEVITT, H. Speech Comprehension Training and Auditory and Cognitive Processing in Older Adults. **American Journal of Audiology**, v. 21, n. 2, p. 351–357, 2012. Disponível em: <[http://dx.doi.org/10.1044/1059-0889\(2012/12-0025\)](http://dx.doi.org/10.1044/1059-0889(2012/12-0025))>. Citado na página 28.

PINHEIRO, M. M. C.; DESGUALDO, P. L. Processamento auditivo em idosos: estudo da interação por meio de testes com estímulos verbais e não-verbais. **Revista Brasileira de Otorrinolaringologia**, scielo, v. 70, n. 2, p. 209–214, 04 2004. ISSN 0034-7299. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-72992004000200011&nrm=iso>. Citado 2 vezes nas páginas 28 e 29.

RESNIK, P. Left-Corner Parsing And Psychological Plausibility. In: **COLING**. [S.l.: s.n.], 1992. p. 191–197. Citado 2 vezes nas páginas 66 e 70.

RIVEROL, M.; LÓPEZ, O. Biomarkers in Alzheimer's disease. **Frontiers in Neurology**, v. 2, n. 46, 2011. ISSN 1664-2295. Disponível em: <<http://www.frontiersin.org/dementia/10.3389/fneur.2011.00046/abstract>>. Citado na página 22.

ROARK, B.; HOSOM, J.-P.; MITCHELL, M.; KAYE, J. A. Automatically derived spoken language markers for detecting mild cognitive impairment. In: **2nd International Conference on Technology and Aging (ICTA)**. [s.n.], 2007. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.63.1182>>. Citado 3 vezes nas páginas 25, 83 e 84.

ROARK, B.; MITCHELL, M.; HOLLINGSHEAD, K. Syntactic complexity measures for detecting Mild Cognitive Impairment. In: **Biological, translational, and clinical language processing**. Prague, Czech Republic: Association for Computational Linguistics, 2007. p. 1–8. Disponível em: <<http://www.aclweb.org/anthology/W/W07/W07-1001>>. Citado 9 vezes nas páginas 25, 65, 66, 68, 69, 71, 82, 83 e 84.

ROARK, B.; MITCHELL, M.; HOSOM, J.; HOLLINGSHEAD, K.; KAYE, J. Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. **Audio, Speech, and Language Processing, IEEE Transactions on**, v. 19, n. 7, p. 2081–2090, Sept 2011. ISSN 1558-7916. Citado 4 vezes nas páginas 25, 65, 84 e 135.

ROHRER, J. D.; ROSSOR, M. N.; WARREN, J. D. Alzheimer's pathology in primary progressive aphasia. **Neurobiology of aging**, Elsevier, v. 33, n. 4, p. 744–752, apr 2012. ISSN 0197-4580. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0197458010002344?showall=true>>. Citado na página 22.

- ROSENBERG, S.; ABBEDUTO, L. Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. **Applied Psycholinguistics**, v. 8, p. 19–32, 1987. ISSN 1469-1817. Citado na página 68.
- SALING, L. L.; LAROO, N.; SALING, M. M. When more is less: Failure to compress discourse with re-telling in normal ageing. **Acta Psychologica**, v. 139, n. 1, p. 220–224, 2012. ISSN 0001-6918. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0001691811001995>>. Citado na página 32.
- 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.
- SALTHOUSE, T. **Theoretical Perspectives on Cognitive Aging**. Erlbaum, 1991. ISBN 9780805811704. Disponível em: <<http://books.google.com.br/books?id=kT1hG7IVBD4C>>. Citado na página 27.
- SAUR, D.; SCHELTER, B.; SCHNELL, S.; KRATOCHVIL, D.; KÜPPER, H.; KELLMEYER, P.; KÜMMERER, D.; KLÖPPEL, S.; GLAUCHE, V.; LANGE, R.; MADER, W.; FEES, D.; TIMMER, J.; WEILLER, C. Combining functional and anatomical connectivity reveals brain networks for auditory language comprehension. **NeuroImage**, v. 49, n. 4, p. 3187–3197, 2010. ISSN 1053-8119. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1053811909011938>>. Citado na página 29.
- SCARTON, C.; ALUÍSIO, S. 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. 45–62, abr. 2010. ISSN 1647–0818. Citado 4 vezes nas páginas 46, 54, 55 e 61.
- SCARTON, C.; GASPERIN, C.; ALUISIO, S. Revisiting the Readability Assessment of Texts in Portuguese. In: KURI-MORALES, A.; SIMARI, G. (Ed.). **Advances in Artificial Intelligence – IBERAMIA 2010**. Springer Berlin Heidelberg, 2010, (Lecture Notes in Computer Science, v. 6433). p. 306–315. ISBN 978-3-642-16951-9. Disponível em: <http://dx.doi.org/10.1007/978-3-642-16952-6_31>. Citado 2 vezes nas páginas 55 e 56.
- SILVA, J.; BRANCO, A.; CASTRO, S.; REIS, R. Out-of-the-box robust parsing of portuguese. In: PARDO, T. A. S.; BRANCO, A.; KLAUTAU, A.; VIEIRA, R.; LIMA, V. L. S. de (Ed.). **Computational Processing of the Portuguese Language**. Springer Berlin Heidelberg, 2010, (Lecture Notes in Computer Science, v. 6001). p. 75–85. ISBN 978-3-642-12319-1. Disponível em: <http://dx.doi.org/10.1007/978-3-642-12320-7_10>. Citado na página 101.
- SILVA, T. B. L.; YASSUDA, M. S.; GUIMARÃES, V. V.; FLORINDO, A. A. Fluência verbal e variáveis sociodemográficas no processo de envelhecimento: um estudo epidemiológico. **Psicol Reflex Crit**, v. 24, n. 4, p. 739–746, 2011. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-79722011000400014&lng=pt&nrm=iso&tlng=en>. Citado na página 31.
- SKA, B.; DUONG, A. Communication, discours et démence. **Psychol NeuroPsychiatr Vieil**, v. 3, n. 2, p. 125–133, 2005. Citado na página 39.
- SKA, B.; JOANETTE, Y. Vieillissement normal et cognition. **Médecine sciences**, v. 22, n. 3, p. 284–287, mar 2006. Citado 3 vezes nas páginas 21, 27 e 28.

SMITH, E.; IVNIK, R. J. Normative neuropsychology. In: PETERSEN, R. C. (Ed.). **Mild cognitive impairment**. New York: Oxford University Press, 2003. p. 63–88. Citado na página 107.

SNOWDON, D.; KEMPER, S.; MORTIMER, J.; GREINER, L.; WEKSTEIN, D.; MARKESBERY, W. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. **JAMA**, v. 275, n. 7, p. 528–532, 1996. Disponível em: <<http://dx.doi.org/10.1001/jama.1996.03530310034029>>. Citado 3 vezes nas páginas 43, 72 e 73.

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 36.

SOARES, L. M.; CACHIONI, M.; FALCÃO, D. V. d. S.; BATISTONI, S. S. T.; LOPES, A.; NERI, A. L.; YASSUDA, M. S. Determinants of cognitive performance among community dwelling older adults in an impoverished sub-district of São Paulo in Brazil. **Archives of Gerontology and Geriatrics**, v. 54, n. 2, p. 187–192, 2012. ISSN 0167-4943. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S016749431100330X>>. Citado na página 31.

SPERLING, R. A.; KARLAWISH, J.; JOHNSON, K. A. Preclinical Alzheimer disease - the challenges ahead. **Nat Rev Neurol**, Nature Publishing Group, v. 9, n. 1, p. 54–58, jan 2013. ISSN 1759-4758. Disponível em: <<http://dx.doi.org/10.1038/nrneurol.2012.241>>. Citado na página 22.

STEINER, V. A. G. **Fluência semântica, fonêmica e de verbos em indivíduos com comprometimento cognitivo leve**. Tese (Doutorado) — Faculdade de Medicina da Universidade de São Paulo, 2012. Citado na página 40.

STEINER, V. A. G.; MANSUR, L. L.; BRUCKI, S. M.; NITRINI, R. Phonemic verbal fluency and age: a preliminary study. **Dement Neuropsychol**, v. 2, n. 4, p. 328–332, 2008. Citado na página 31.

TAKAO, A. Y.; PROTHERO, W. A.; KELLY, G. J. Applying Argumentation Analysis To Assess the Quality of University Oceanography Students' Scientific Writing. **Journal of Geoscience Education**, v. 50, n. 1, p. 40–48, 2002. ISSN 1089-9995. Disponível em: <<http://www.editlib.org/p/92998>>. Citado na página 73.

TALER, V.; PHILLIPS, N. A. Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. **Journal of Clinical and Experimental Neuropsychology**, v. 30, n. 5, p. 501–556, 2008. PMID: 18569251. Citado 3 vezes nas páginas 23, 35 e 41.

THOMAS, C.; KESELJ, V.; CERCONI, N.; ROCKWOOD, K.; ASP, E. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In: **Mechatronics and Automation, 2005 IEEE International Conference**. [S.l.: s.n.], 2005. v. 3, p. 1569–1574 Vol. 3. Citado 6 vezes nas páginas 17, 25, 64, 74, 81 e 82.

THORSON, E.; SNYDER, R. Viewer recall of television commercials: Prediction from the propositional structure of commercial scripts. **Journal of Marketing Research**, v. 21, p. 127–136, 1984. Citado na página 73.

TOGHER, L. Discourse sampling in the 21st century. **Journal of Communication Disorders**, v. 34, n. 1–2, p. 131–150, 2001. ISSN 0021-9924. Citado na página 24.

TOLEDO, C. M. **Variáveis sociodemográficas na produção do discurso em adultos sadios**. Dissertação (Mestrado) — Faculdade de Medicina da Universidade de São Paulo, São Paulo, 2011. Citado 6 vezes nas páginas 33, 69, 108, 109, 110 e 111.

TOLEDO, C. M. **Análise de aspectos micro e macrolinguísticos da narrativa de indivíduos com doença de Alzheimer, comprometimento cognitivo leve e sem comprometimentos cognitivos**. Tese (Doutorado) — Faculdade de Medicina, 2015. Citado 3 vezes nas páginas 17, 118 e 119.

TOLEDO, C. M.; CUNHA, A.; SCARTON, C. E.; ALUÍSIO, S. M. Automatic classification of written descriptions by healthy adults: an overview of the application of natural language processing and machine learning techniques to clinical discourse analysis. In: **Dementia & Neuropsychologia**. [S.l.: s.n.], 2014. v. 8, p. 227–235. Citado na página 123.

TRAYKOV, L.; RIGAUD, A.-S.; CESARO, P.; BOLLER, F. Le déficit neuropsychologique dans la maladie d'Alzheimer débutante. **L'Encéphale**, v. 33, n. 3, p. 310–316, 2007. ISSN 0013-7006. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0013700607920448>>. Citado na página 34.

TURNER, A.; GREENE, E. **The Construction and Use of a Propositional Text Base**. [S.l.], 1977. Citado 3 vezes nas páginas 72, 73 e 114.

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 39.

VERHAEGEN, C.; PONCELET, M. Changes in Naming and Semantic Abilities With Aging From 50 to 90 years. **Journal of the International Neuropsychological Society**, v. 19, p. 119–126, 2 2013. ISSN 1469-7661. Disponível em: <http://journals.cambridge.org/article_S1355617712001178>. Citado na página 30.

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.

WILLIAMS, L. J.; ABDI, H.; FRENCH, R.; ORANGE, J. B. A Tutorial on Multiblock Discriminant Correspondence Analysis (MUDICA): A New Method for Analyzing Discourse Data From Clinical Populations. **Journal of Speech, Language, and Hearing Research**, v. 53, n. 5, p. 1372–1393, 2010. Disponível em: <[+http://dx.doi.org/10.1044/1092-4388\(2010/08-0141\)](http://dx.doi.org/10.1044/1092-4388(2010/08-0141))>. Citado na página 28.

WILLS, C.; CAPILOUTO, G.; WRIGHT, H. Attention and Off-Topic Speech in the Recounts of Middle-Age and Elderly Adults: A Pilot Investigation. **Contemporary issues in communication science and disorders**, niversity of Kentucky, Lexington, v. 39, p. 105–112, 2012. Disponível em: <<http://europepmc.org/abstract/MED/23420319>>. Citado 3 vezes nas páginas 24, 32 e 107.

WINBLAD, B.; PALMER, K.; KIVIPELTO, M.; JELIC, V.; FRATIGLIONI, L.; WAHLUND, L.-O.; NORDBERG, A.; BÄCKMAN, L.; ALBERT, M.; ALMKVIST, O.; ARAI, H.; BASUN, H.; BLENNOW, K.; De Leon, M.; DECARLI, C.; ERKINJUNTTI, T.; GIACOBINI, E.; GRAFF, C.; HARDY, J.; JACK, C.; JORM, A.; RITCHIE, K.; Van Duijn, C.; VISSER, P.; PETERSEN, R. Mild cognitive impairment – beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. **Journal of Internal Medicine**, Blackwell Science Ltd, v. 256, n. 3, p. 240–246, 2004. ISSN 1365-2796. Disponível em: <<http://dx.doi.org/10.1111/j.1365-2796.2004.01380.x>>. Citado na página 40.

WINGFIELD, A.; GROSSMAN, M. Language and the Aging Brain: Patterns of Neural Compensation Revealed by Functional Brain Imaging. **Journal of Neurophysiology**, v. 96, n. 6, p. 2830–2839, 2006. Disponível em: <<http://jn.physiology.org/content/96/6/2830.abstract>>. Citado na página 28.

WRIGHT, H. H.; CAPILOUTO, G. J.; KOUTSOFTAS, A. Evaluating measures of global coherence ability in stories in adults. **International Journal of Language & Communication Disorders**, v. 48, n. 3, p. 249–256, 2013. ISSN 1460-6984. Disponível em: <<http://dx.doi.org/10.1111/1460-6984.12000>>. Citado na página 107.

WRIGHT, H. H.; KOUTSOFTAS, A. D.; CAPILOUTO, G. J.; FERGADIOTIS, G. Global coherence in younger and older adults: Influence of cognitive processes and discourse type. **Aging, Neuropsychology, and Cognition**, v. 21, n. 2, p. 174–196, 2014. PMID: 23656430. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/13825585.2013.794894>>. Citado na página 32.

YASSUDA, M. S.; DINIZ, B. S.; FLAKS, M. K.; VIOLA, L. F.; PEREIRA, F. S.; NUNES, P. V.; FORLENZA, O. V. Neuropsychological Profile of Brazilian Older Adults with Heterogeneous Educational Backgrounds. **Archives of Clinical Neuropsychology**, v. 24, n. 1, p. 71–79, 2009. Disponível em: <<http://acn.oxfordjournals.org/content/24/1/71.abstract>>. Citado na página 31.

YNGVE, V. H. A Model and an Hypothesis for Language Structure. **Proceedings of the American Philosophical Society**, v. 104, p. 444–466, 1960. Citado na página 65.

ZEMAN, D.; DUŠEK, O.; MAREČEK, D.; POPEL, M.; RAMASAMY, L.; ŠTĚPÁNEK, J.; ŽABOKRTSKÝ, Z.; HAJIČ, J. Hamletd: Harmonized multi-language dependency treebank. **Language Resources and Evaluation**, Springer Netherlands, v. 48, n. 4, p. 601–637, 2014. ISSN 1574-020X. Disponível em: <<http://dx.doi.org/10.1007/s10579-014-9275-2>>. Citado na página 105.

ZWAAN, R. A.; RADVANSKY, G. A. Situation Models in Language Comprehension and Memory. **Psychological Bulletin**, v. 123, n. 2, p. 162–185, 1998. Disponível em: <http://psych.wisc.edu/ugstudies/Psych733/Zwann_Radvansky_1998.pdf>. Citado na página 49.