Coh-Metrix-Dementia: Análise Automática de Distúrbios de Linguagem nas Demências utilizando Processamento de Línguas Naturais

Andre Luiz Verucci da Cunha

Coh-Metrix-Dementia: Análise Automática de Distúrbios de Linguagem nas Demências utilizando Processamento de Línguas Naturais

Andre Luiz Verucci da Cunha

Orientadora: ***Prof.ª Dr.ª Sandra Maria Aluísio***
Coorientadora: ***Prof.ª Dr.ª Letícia Lessa Mansur***

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. *EXEMPLAR DE QUALIFICAÇÃO*

USP – São Carlos
Fevereiro de 2014

Resumo

(Contexto) Segundo a Organização Mundial da Saúde, a demência é 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. Mas neste trabalho também temos interesse em uma 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 amnéstico, evoluem para a DA a uma taxa de 15% por ano, versus 1 a 2% 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, principalmente 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, 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, chamado Coh-Metrix-Port, a ferramenta LIWC, cujo dicionário também foi adaptado para o português e a ferramenta AIC, todas elas desenvolvidas no NILC e incluíndo a adaptação para o português de dezoito novas ferramentas para calcular a complexidade sintática, a densidade de ideias e a coerência textual, via semântica latente, é possível classificar narrativas de pessoas normais, com DA em vários graus e com CCL, em uma abordagem de aprendizado de máquina, com precisão comparável a dos testes clássicos para distinção de normais versus DA e CCL e entre DA e CCL. Todas as métricas já desenvolvidas e as dezoito novas a serem criadas durante este mestrado serão implementadas em um ambiente único, de uso adaptado a profissionais e pesquisadores da área médica. *(Conclusão)* Esperamos que o ambiente Coh-Metrix-dementia supra uma lacuna significativa nas ferramentas de PLN para o português do Brasil com aplicação na área de avaliações de desordens neurodegenerativas e forneça a médicos e pesquisadores análises discursivas rápidas e sistematizadas, fomentando assim a pesquisa em detecção e gerenciamento de demências.

Sumário

1 Introdução **1**

2 Fundamentação Teórica **7**

2.1 Envelhecimento e linguagem . . . 7

2.1.1 O envelhecimento normal . . . 7

2.1.2 A linguagem nos indivíduos acometidos por doença de Alzheimer (DA) . . . 13

2.1.3 A linguagem nos indivíduos acometidos por Comprometimento Cognitivo Leve (CCL) 19

2.1.4 Extratos de entrevistas de pacientes acometidos com DA: rupturas e reformulações . 20

2.1.5 Considerações finais . . . 30

2.2 Ferramentas automáticas para análise de características textuais disponíveis para o português 32

2.2.1 O Coh-Metrix e o Coh-Metrix-Port . . . 32

2.2.2 A ferramenta Análise de Inteligibilidade de Córpus (AIC) . . . 45

2.2.3 O *Linguistic Inquiry and Word Count* (LIWC) . . . 46

3 Análise Automatizada de alterações de linguagem em doenças neurológicas degenerativas **49**

3.1 Medidas automatizadas usadas na identificação de condições clínicas a partir de amostras de linguagem . . . 49

3.1.1 Medidas de diversidade lexical . . . 49

3.1.2 Medidas de complexidade sintática . . . 50

3.1.3 Medidas de densidade semântica . . . 56

3.1.4 Medidas de semântica latente . . . 60

3.2 Trabalhos relacionados . . . 62

3.2.1 Abordagem lexical . . . 62

3.2.2 Abordagem baseada em complexidade sintática . . . 65

3.2.3 Abordagem baseada em densidade de ideias . . . 67

3.2.4 Abordagem baseada em traços semânticos e categorias morfossintáticas . . . 68

3.2.5 Abordagem em vários níveis . . . 69

4 Proposta de pesquisa: Coh-Metrix-Dementia **72**

4.1 Córpus de pesquisa . . . 72

4.2 Arquitetura do Coh-Metrix-Dementia . . . 74

4.3 Avaliação . . . 76

4.4 Experimentos piloto . . . 76

4.4.1 Experimento 1 . . . 76

4.4.2 Experimento 2: desempenho das ferramentas de PLN no corpus DA-PLN-EVAL . . . 82

5 Plano de Trabalho e Cronograma **92**

6 Considerações finais **94**

REFERÊNCIAS **94**

A Extratos de entrevistas compilados do Anexo IV da tese de Mansur [1996] **111**

Lista de Tabelas

2.1 Sujeitos: condição cognitiva e grau de escolaridade. 21
2.2 Normas para a transcrição utilizadas. 22
2.3 Métricas do Coh-Metrix 3.0. 35
2.4 Desempenho dos classificadores com características extraídas pelo Coh-Metrix-Port. 42
2.5 Métricas do Coh-Metrix-Port. 42

3.1 Exemplos de proposições e densidade de ideias (extraídos de Chand et al. [2010]). 58
3.2 Métricas de LSA do Coh-Metrix 3.0. 61
3.3 Sumário dos resultados de acurácia de classificação de Thomas et al. [2005]. 65
3.4 Sumário dos testes de significância de Bryant et al. [2013] (X sinaliza significância). 68
3.5 Resultados (aproximados) obtidos por Peintner et al. [2008]. 69
3.6 Resultados obtidos por Jarrold et al. [2010]. 70
3.7 Acurácia (%) dos três classificadores de Fraser et al. [2012]. 71

4.1 Medida F (%) dos métodos para o primeiro conjunto de experimentos (3-8 versus 9+). 80
4.2 Medida F (%) dos métodos com classes extremas e intermediárias. 81
4.3 Medida F (%) dos métodos de classificação com remoção de classes intermediárias. 81
4.4 Medida F (%) dos métodos de classificação e seleção de atributos com remoção de classes intermediárias. 82
4.5 Estatísticas básicas das entrevistas. 82
4.6 Trecho da transcrição original do sujeito com DA leve - EFA, simplificado por E. 84
4.7 Segunda (esquerdo) e terceira (direito) versões do trecho apresentado na tabela 4.6. 85
4.8 Conjunto de Etiquetas do *tagger* MXPOST treinado com o Mac-Morpho. 85

5.1 Cronograma previsto para o projeto. 93

Lista de Figuras

2.1 Tela do Coh-Metrix exibindo métricas de um texto de exemplo. 35
2.2 Saída do Coh-Metrix-Port para um texto de exemplo. 46
2.3 Exemplo de parte da tela de saída do AIC. 47
2.4 Tela de saída do LIWC, com o dicionário para o inglês, versão de 2007, mostrando a análise de uma música. 48

3.1 Exemplo de árvore sintática para cálculo da complexidade de Yngve (extraído de Pakhomov et al. [2011]). 51
3.2 Exemplo de árvore sintática para cálculo da complexidade de Frazier (extraído de Pakhomov et al. [2011]). 52
3.3 Exemplo da Figura 3.2 com os ramos individuais das palavras. 52
3.4 Exemplo de estrutura de dependências. 53
3.5 Exemplo de uso do CPIDR. 59
3.6 Exemplo de uso do CPIDR no modo fala. 59
3.7 Transcrição no formato CHAT da história da Cinderela (extraído de [MacWhinney et al., 2010]). 63

4.1 Arquitetura geral do Coh-Metrix-Dementia. 74

1. 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 2000[1] e 2010[2]. Segundo o censo de 2000, havia 19.221 homens de 95 a 99 anos e 36.977 mulheres nesta faixa; em 2010 eram 31.529 homens e 66.806 mulheres. Já para a faixa de 90 a 94 anos, 65.117 homens e 115.309 mulheres em 2000. Em 2010 havia, na faixa de 90 a 94 anos, 114.964 homens e 211.595 mulheres. Na faixa de 85 a 89 anos, o aumento nestes 10 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 aumenta no seu topo. A estimativa para o país é de que, em 2025, existam 31,8 milhões de idosos com sessenta anos 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 4ª ou 5ª década de vida, graduais e inexoráveis [Ska and 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 [de Carvalho and Mansur, 2008]. Neste sentido, faz-se necessário ampliar o conhecimento sobre os idosos brasileiros, sob o foco da linguagem.

O aumento da expectativa de vida está associado aos avanços na área de saúde e de saneamento básico e refletem seus efeitos sobre a taxa de mortalidade [Paschoal, 1996]. No entanto, trará sérios problemas em termos de recursos financeiros e sociais, sobretudo na área da saúde, pela repercussão nos diversos níveis assistenciais e demanda por novos recursos e estruturas. Citem-se as doenças crônico-degenerativas, como a doença de Alzheimer (DA), que acompanham a maior sobrevida, com sequelas e complicações [Fried, 2000; Paschoal, 1996]. Desta forma, cresce em importância as pesquisas em neurociência sobre detecção e gestão de doenças que normalmente afetam pessoas com idade avançada, como a **demência**, em geral.

A demência é, em geral, o resultado de uma desordem neurodegenerativa progressiva e irreversível, que se manifesta de diversas formas.

O principal desafio no gerenciamento da demência 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 et al., 2013]. O melhor tratamento atualmente disponível para a demência consiste em retardar a progressão da doença quando da detecção de seus primeiros sinais. Apesar de não haver tratamentos que 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

[1] http://www.ibge.gov.br/home/estatistica/populacao/perfilidoso/perfidosos2000.pdf
[2] http://censo2010.ibge.gov.br/sinopse/webservice/frm_piramide.php

[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 da demência [Riverol and 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 correspondência entre os problemas cognitivos apresentados e a patologia molecular subjacente é, entretanto, imperfeita, uma vez que os sintomas cognitivos são geralmente determinados pelo local do dano cerebral, não pelas causas subjacentes. Sendo assim, nossa capacidade de identificar os primeiros sinais da demência se beneficiará de uma melhor compreensão da relação entre o local da disfunção e a progressão do comprometimento cognitivo.

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. Como diferentes aspectos da linguagem são afetados por disfunções neurais em regiões distintas do cérebro, uma caracterização mais sofisticada dos prejuízos à linguagem presentes nos diferentes tipos de demência auxiliará seu diagnóstico precoce.

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 (APPNF) caracteriza-se pela fala esforçada, hesitante e disfluente, com dificuldades de resgate verbal. Além desses, agramatismo e apraxia de fala[3] também são considerados sintomas centrais. Esse quadro de perda gradual da competência gramatical e da fluência é semelhante ao da afasia de Broca e associa-se à degeneração do lobo frontal esquerdo. A Demência Semântica (DS, também denominada "APP fluente") 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. Pacientes com DS apresentam anomia[4] severa, apesar de a produção oral permanecer fluente, bem articulada e gramaticalmente correta, com prosódia normal [Fraser et al., 2012]. A DS está associada à degeneração bilateral de áreas anteriores do lobo temporal. Tanto a APPNF quanto a DS são consideradas subtipos de *Demência Frontotemporal* (DFT), uma desordem neurodegenerativa distinta da *Doença de Alzheimer* (DA). Uma terceira forma de APP, chamada Afasia Progressiva Primária Logopênica (APPL), é, em vez disso, relacionada à DA na sua patologia subjacente [Rohrer et al., 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 and Gorno-Tempini, 2010]. Esses sintomas, relacionados à neurodegeneração no lobo temporoparietal, 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

[3] O paciente sabe as palavras que deve usar, mas sente dificuldades em coordenar os movimentos necessários para pronunciá-las; ainda assim, reconhece o erro e pode ser capaz de corrigi-lo em outras tentativas.

[4] 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.

DS, anomia e maior nível de déficits no discurso que impedem a conversação, como mudanças frequentes de assunto [Taler and Phillips, 2008].

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, mas é frequentemente acompanhada por outros danos corticais, cuja localização é menos consistente nos pacientes [Stopford et al., 2008].

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 severidade 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]. 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 et al., 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 et al., 2012; Cannizzaro and 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 narrativo[5] induzido por figuras é útil para pesquisas, pois elicia a fala de forma padronizada e permite a comparação entre indivíduos e grupos [Cooper, 1990]. Nos estudos sobre lesados cerebrais, tem-se analisado os indivíduos com lesões focais (afasias) [Andreetta et al., 2012] e difusas (traumatismos crânio-encefálicos e processos degenerativos como doença de Alzheimer e demências do espectro lombar frontotemporal, entre outras) [Ash et al., 2006].

Segundo Chand et al. [2010], testes padronizados de linguagem, como o de fluência verbal, frequentemente são capazes de detectar declínios na DA ainda em momento 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 lingua-

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

gem tal como é usada para comunicação nem sobre a complexa interação entre conhecimentos linguísticos e semânticos necessária para tanto. Por fim, os autores apontam 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 cada variante da APP [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: portadores de DS, portadores de APPNF e controles sadios. 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, 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 APPNF *vs* DS 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 sadios e portadores de três subtipos de demência frontotemporal (DFT comportamental, APPNF e DS). 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 desordens cerebrais: DA, comprometimento cognitivo leve (CCL) e depressão clínica. Utilizando métricas textuais semelhantes às de Peintner et al. [2008](características morfossintáticas, características extraídas do LIWC e Densidade de Ideias [Chand et al., 2010]), os classificadores conseguiram separam pacientes pré-DA de controles sadios 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. [2007a], Roark et al. [2007b] 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 3.1.3) no discurso de sujeitos normais e de sujeitos afásicos. 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 intensidade da demência.

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 deficiências, por exemplo, não relatarem quais características foram decisivas na classificação dos sujeitos, como bem 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 identificação de **características distintivas** no diagnóstico da DA e CCL e confrontar esses dados com pesquisas já existentes na medicina quanto ao tópico.

2. Fornecer à equipe de saúde **a classe a que um indivíduo pertence** (sadio, DA e CCL), com base em algoritmos de classificação, 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 com DA dos com CCL, 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. Existem métodos de classificação que desempenham melhor na tarefa de predizer as classes de um indivíduo sadio, com DA e com CCL, conforme as características linguísticas de seu texto? Se sim, quais são eles?

Os capítulos que se seguem desenvolvem melhor as questões e objetivos apresentados acima e organizam-se da seguinte maneira: no capítulo 2, é apresentada a fundamentação teória deste trabalho; na seção 2.1, são apresentadas informações sobre o envelhecimento, tanto o natural quanto o patológico na DA e no CCL; na seção 2.2, são apresentadas ferramentas de processamento de línguas naturais (PLN) disponíveis para o português brasileiro que serão usadas como base para este trabalho. No capítulo 3, são descritas medidas automatizadas usadas na identificação de condições clínicas com base em amostras de linguagem (seção 3.1), bem como trabalhos na literatura que automatizaram a avaliação de doenças neurológicas (seção 3.2).

No capítulo 4, é apresentada a proposta de pesquisa, com dois experimentos piloto, um realizado com discursos de sujeitos sadios (seção 4.4.1) e outro realizado para demonstar as dificuldades de se lidar com textos de pacientes com demência (seção 4.4.2). No capítulo 5, é apresentado o cronograma previsto para a realização das atividades propostas. No capítulo 6, são feitas as últimas considerações.

2. Fundamentação Teórica

2.1 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 and Joanette, 2006]. Frequentemente são relatadas alterações de atenção, memória e habilidades visual-espaciais ao lado de relativa preservação da linguagem.

Nesta seção, apresentamos os fundamentos sobre cognição e linguagem de indivíduos normais e em condições patológicas. A revisão, abaixo exposta, foi organizada segundo relevância temática à qual se subordina a apresentação cronológica. A base de dados PubMed[1] foi a principal fonte de consulta para obtenção dos textos.

2.1.1 O envelhecimento normal

Idoso 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. Por outro lado, a heterogeneidade também se aplica ao ritmo 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 and Joanette, 2006].

Habilidades linguístico-comunicativas de idosos sadios

As descrições da linguagem de idosos ampliaram-se consideravelmente nas últimas décadas, porém ainda há aspectos a investigar, 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

[1] http://www.ncbi.nlm.nih.gov/pubmed

episódica e habilidades visual-espaciais, a linguagem mostra-se relativamente preservada e resistente ao processo de envelhecimento [Park and Bischof, 2013].

Mudanças na linguagem e comunicação

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 and Mansur, 2006].

Estilos prolixo ou simplificado podem ocorrer como efeito de tarefas. Nas tarefas que requerem o reconto 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 nacionais [Mansur et al., 2005]. Numa investigação de caráter transversal longitudinal [Andrade and Martins, 2010], foram examinados 128 idosos brasileiros, incluindo nonagenários. Os autores constataram "tendência a decréscimo na taxa de fala e maior 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 and Desgualdo, 2004; Wingfield and Grossman, 2006; Pichora-Fuller and 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 neurossensorial, porém em muitos casos os déficits no processamento da fala são desproporcionais à perda auditiva para tons puros [Pinheiro and Desgualdo, 2004].

A queixa dos idosos para "entender" sobrepuja à de dificuldades para "ouvir". Além de fatores auditivos centrais, respondem por essas dificuldades, fatores de outras naturezas: linguísticos, tais como a complexidade sintático-semântica do material a ser tratado; mecanismos relacionados ao tratamento do significado – como realização de inferências e fatores extralinguísticos, relacionados ao caráter temporal do processamento e 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 tem mostrado que os modelos de processamento da linguagem falada são interessantes porém insuficientes para que se entenda a audição de idosos e também 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 perceptuais e de processamento no idoso, pouco se sabe sobre o impacto de mudanças da idade na produção da linguagem.

a. Nível fonético-fonológico

Nos estudos a respeito da produção de linguagem não constatam, 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, alterações no timbre vocal e ritmo de emissão, que conferem à produção oral do idoso características próximas às da fala de pacientes com disartria[2]. 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 que a do idoso 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.

São citadas ainda hesitações, repetições de sons, fenômenos de disfluência [Andrade and Martins, 2010], uso frequente de preenchedores (“uh”, “aham”) quando comparados com jovens [Horton and Spieler, 2007]. A respeito das disfluências do idoso, Manning and Shirkey [1981] conceberam um modelo que distingue disfluências “motoras” e “formulativas”. As primeiras compreenderiam aquelas “rupturas” (incluindo aí as repetições) intravocábulos e sílabas acompanhadas de aumento de tônus global da emissão, enquanto as segundas abrangeriam as pausas intervocábulos e repetições de segmentos. Para o autor, a repetição de sílabas é considerada como 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 “formulativas”. No modelo, a previsão é de que ocorra diminuição de disfluências “motoras” e o aumento das disfluências “formulativas” (pausas preenchidas ou não, repetições de palavras e segmentos) à medida que o indivíduo envelhece. Para eles, o aumento de disfluências do tipo “formulativas” relacionadas à idade, evidenciaria hesitações devidas a falhas psicolinguísticas (de acesso lexical, de formulação sintática, entre outras).

Há necessidade de estudos sobre as disfluências presentes na produção oral dos idosos uma vez que

[2] Disartria é a 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.

esses fenômenos não estão totalmente esclarecidos.

b. Nível lexical

Dificuldades no nível lexical são queixas frequentes entre os idosos e objeto de constante investigação.

A tendência é de 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 et al. [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 jovens, quando devem decidir se um segmento é ou não uma "palavra" porém melhoram seu desempenho quando 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 and Poncelet [2013] mostra que já aos 50 anos, há declínio nessa habilidade. Por outro lado, diferenças entre verbos e substantivos tem 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 de "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 itemalvo, como o número de sílabas, significado, porém falham no acesso a forma fonológica. Decorrido algum tempo, frequentemente podem recordar espontaneamente o item [Facal-Mayo et al., 2006].

Se por um lado, sabe-se que o acesso lexical é um problema para os idosos, sua natureza, etiologia e concomitância de fatores cognitivos permanecem obscuros. Os estudos tem investigado correlações com vocabulário, memória operacional, memória episódica, velocidade de processamento da informação e monitoramento, sem alcançar conclusões definitivas [Facal et al., 2012; Salthouse and Mandell, 2013].

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 and Rocha

[2004], Steiner et al. [2008], Fischman et al. [2009], Yassuda et al. [2009], Silva et al. [2011], Amaral-Carvalho and Caramelli [2012] e Soares et al. [2012]. Os autores constataram que tanto a idade quanto a escolaridade influenciaram os resultados. Os estudos mais recentes tem encontrado efeito de 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 et al. [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 tipos 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. Esses últimos também fizeram maior número de erros e foram mais lentos nas respostas.

O estudo das freiras [Kemper et al., 2001a] 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 idea density para caracterizar os efeitos da idade.

A mesma autora em outra investigação [Kemper et al., 2001b] dirigiu-se à produção de sentenças em situação controlada com e sem uma tarefa motora competitiva. Os autores observaram as fases de planejamento e a 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. Planejar ou produzir sentenças com longos sintagmas nominais foi especialmente difícil para os idosos.

A ideia de que habilidades sintáticas ficam estáveis no envelhecimento não é consensual. Nippold et al. [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 sinaliza a necessidade de estudos adicionais sobre o tema. Além disso, desconhecemos 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 em geral esses eventos são associados a dificuldades cognitivas do tipo atencionais. Wills et al. [2012] que estudaram idosos sadios com idades entre 40-80 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.

Mar [2004] dedicou-se à investigação de compreensão e produção de narrativas e buscou suas bases neurofuncionais, processos cognitivos e a dinâmica interação entre produção e compreensão.

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 semânticas, paragramatismos[3] e complexidade sintática e grau de coerência global e local , como erros do tipo referenciação ambígua e no nível de informatividade veiculado nas histórias. Para os autores, o abrupto declínio no grupo mais idoso sinaliza a perda de habilidades relacionadas ao envelhecimento. O tipo de estímulo (prancha única versus prancha em sequência) influencia as medidas de "informatividade" e aspectos macrolinguísticos.

Os efeitos de idade no reconto de histórias foi notado no estudo de Saling et al. [2012]. Esses autores constataram que ao contrário dos jovens que tendiam a sintetizar histórias quando a enunciação ocorria em intervalos repetidos, os idosos mantinham o estilo menos conciso e prolixo nos vários recontos.

Wright et al. [2014] estudaram ainda o efeito de 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 reconto 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 reconto de histórias e discursos de procedimento.

Fergadiotis et al. [2011] estudaram aspectos microlinguísticos. Eles examinaram o efeito do tipo de discurso na diversidade lexical testando quatro tipos de discurso (discursos de procedimento, eventos, contação de histórias, recontos) produzidos por 86 adultos e idosos cognitivamente sadios. Os idosos estavam na faixa de 70-89 anos. As amostras de discurso foram analisadas pelo software voc-D[4] para estimar a diversidade lexical. Os resultados indicaram que a diversidade lexical é um dos índices influenciados pelo tipo de discurso e idade.

A pesquisa realizada por Toledo [2011] examinou aspectos microlinguísticos e macrolinguísticos. Foram avaliados 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). Elegeu-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, fonológicos, repetições imediatas de palavras, sentenças); conteúdo da descrição (emissões irrelevantes, vagas, dificuldades de interpretação visual e quantidade de informação) e habilidades sintáticas (extensão e complexidade das sentenças). O tempo de descrição foi registrado. A autora verificou significativa influência da escolaridade e idade nos discursos. O efeito da idade foi verificado no número de palavras, repetição imediata de palavras, emprego de termos indefinidos, complexidade frasal e no aumento do tempo para a produção do discurso.

Ferguson et al. [2013] também avaliaram a influência da idade e escolaridade, porém na linguagem escrita. A informatividade do discurso de larga coorte[5] de 19.512 participantes foi analisada pelo Computerized Propositional Idea Density Rater 3 – (CPIDR-3 - http://www.ai.uga.edu/caspr/). Os autores encontraram

[3] Parafasias semânticas são trocas lexicais, que ocorrem num mesmo campo semântico (Por exemplo, a troca de mesa por cadeira). Paragramatismos são 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).

[4] http://ltj.sagepub.com/content/24/4/459.abstract

[5] Em estatística, **coorte** é um conjunto de pessoas que têm em comum um evento que se deu no mesmo período. Por exemplo: coorte de pessoas que nasceram em 1960, coorte de mulheres casadas em 1990, etc. (Fonte: http://pt.wikipedia.org/wiki/Coorte)

pequena mas significativa diminuição da densidade de proposições na faixa etária acima de 78 anos e concluíram que essa medida mantem-se relativamente estável no envelhecimento. Por outro lado, os autores não verificaram efeitos de escolaridade.

2.1.2 A linguagem nos indivíduos acometidos por doença de Alzheimer (DA)

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, 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, funções executivas, linguagem, personalidade ou comportamento.[6]

A doença de Alzheimer 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 amnéstica (quando predomina o comprometimento de memória, associado ao comprometimento em outro domínio cognitivo) ou não-amnéstica (deve haver outro domínio afetado) com alterações em linguagem (lembranças de palavras), aspectos visual-espaciais, agnosia (dificuldade de reconhecer objetos ou faces, e dificuldade de leitura relacionada a aspectos visual-espaciais), funções executivas (alteração do raciocínio, julgamento e solução de problemas).

Os exames de neuro-imagem são utilizados para excluir outros diagnósticos.

Assim, embora a perda de memória seja o traço cognitivo mais frequente na doença de Alzheimer, 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

[6] **Memória** - da capacidade para adquirir ou evocar informações recentes. Os sintomas de quadro demencial incluem: repetição das mesmas perguntas ou assuntos, esquecimento de eventos, 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, redução da capacidade para cuidar das finanças, de tomar decisões e de planejar atividades complexas ou sequenciais.

Habilidades visuo-espaciais - capacidade de reconhecer faces ou objetos comuns, encontrar objetos no campo visual, manusear utensílios, 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, 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 incaracterísticas), agitação, apatia, desinteresse, isolamento social, perda de empatia, desinibição, comportamentos obsessivos, compulsivos ou socialmente inaceitáveis.

de idade. O exame de neuro imagem nesses casos mostra que a variante amnéstica tem comprometimento predominante no hipocampo, enquanto na variante linguagem o predomínio de comprometimento é nas regiões parietais esquerdas e na variante visual-espacial em hemisfério direito, regiões parietal e occipital [Mendez et al., 2012].

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, tem surgido estudos de linguagem em situações funcionais.

Apresentaremos a seguir uma breve revisão da linguagem na DA enfatizando a produção da linguagem, interesse de nossa investigação.

Recepção e compreensão

Para os sujeitos com DA, a possibilidade de compreensão da linguagem oral está relacionada não só à complexidade da tarefa. Pesquisadores detectaram vários níveis de comprometimento, inclusive dificuldades em compreensão de prosódia afetiva e gramatical, já ao início da doença [Taler and 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 and 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 and 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 só 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 et al., 1992]. Recentemente essa ideia tem sido revista. Em primeiro lugar, tende-se a diferenciar do ponto de vista cognitivo os indivíduos com Alzheimer cuja apresentação é pré-senil daqueles em que a apresentação ocorre após os 65 anos de idade. Os indivíduos com apresentação pré-senil da DA apresentam déficits proeminentes de linguagem, principalmente aqueles associados à memória operacional, como os fonético-fonológicos [Kalpouzos et al., 2005]. Além disso, a investigação de aspectos relacionados à produção da linguagem e praxias bucofaciais e de fala[7] 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 imageabilidade[8]. 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 e 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 reduzir o conteúdo das proposições e a linguagem fica reduzida a sentenças curtas, familiares, repetitivas ou fragmentos, chegando ao mutismo [Kemper et al., 2001b].

A redução das habilidades sintáticas está relacionada à perda das bases semânticas da linguagem. É o que se observou no seminal estudo de Snowdon et al. [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

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

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

pos-mortem confirmou o diagnóstico de DA, já na juventude apresentavam traços 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 constituam no cerne das perdas da linguagem, causadas pelo processo degenerativo da doença de Alzheimer. No que diz respeito à produção da linguagem, esses déficits têm sido estudados principalmente em tarefas de nomeação (por confrontação visual e por definição) e fluência verbal.

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

A fluência verbal semântica entre outras habilidades depende da integridade da bagagem semântica, razão pela qual se supõe que o déficit de memória semântica na DA reflita uma degradação desse repertório [Henry et al., 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 clusteres (agrupamentos de itens de determinada categoria) menores quando comparados com idosos saudáveis.

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

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

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

Uma questão interessante proposta nos estudos sobre nomeação em pacientes com DA é se existe vantagem na nomeação de verbos de ação quando comparada a nomeação de 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 and McClung, 2008].

Finalmente, cabe pontuar que alterações em memória semântica acham-se comprometidas já em fase pré-clínica da DA [Cuetos et al., 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 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, semânticos e pragmáticos.

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, organização).

Os portadores de DA tornam-se repetitivos, esquecem o que ouviram ou leram, perdem o tópico. Ao longo do tempo, o discurso torna-se empobrecido, fragmentado, caracterizado por falta de coerência. Nota-se ainda tangencialidade e perseverações [Hooper and Bayles, 2007].

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

Forbes-McKay and 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 et al., 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 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.

de 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, repetição de palavras, 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, uso de termos indefinidos e revisões.

Ska and 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 pranchas em sequência. O objetivo do estudo era determinar níveis de representação discursivos comprometidos nos pacientes com DA, quando comparados a sujeitos normais. O modelo de construção-integração do discurso inclui quatro níveis de representação desde a superfície na qual se analisam componentes linguísticos do discurso (índice lexical, índice sintático e índice referencial) até a organização dos esquemas narrativos abstratos. Os autores verificaram que a prancha única provocou maior número de dificuldades para gerar discursos, entre os pacientes. Além disso, constataram que embora todos os níveis estivessem comprometidos na DA, eles diferiram dos controles em três níveis: nível de superfície, o modelo de situação e a organização da estrutura narrativa.

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

Outro estudo utilizando fala encadeada baseou-se na conversação entre pacientes e seus cônjuges [Williams et al., 2010]. A base de dados para análise dos pontos de ruptura do discurso de 17 pacientes, bem como as revisões realizadas foi construída manualmente. O diferencial do trabalho foi o modelo de análise estatística dos dados, baseado em análise fatorial.

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

2.1.3 A linguagem nos indivíduos acometidos por Comprometimento Cognitivo Leve (CCL)

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: a) queixa subjetiva de perda de memória, b) perdas objetivas de habilidades, c) preservação global de funções cognitivas, d) atividades de vida diária estão preservadas; e) o indivíduo não preenche critérios para o diagnóstico de demência. O Consenso de Stockholm propôs revisão do critério de Petersen [Winblad et al., 2004]: 1. O indivíduo não é nem normal nem demente; 2. 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. 3. 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, tem mais risco para 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 tem 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 et al., 2001] também são utilizados para detectar comprometimentos e diagnosticar CCL. Instrumentos mais sensíveis do que os testes tradicionais tem sido cogitados, como o Graded Naming Test (GNT) [McKenna and 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 and Phillips, 2008].

Estudos isolados com outros testes como o Pyramids and Palm Trees [Howard et al., 1992] e baterias abrangentes como o Montreal Cognitive Assessment [Nasreddine et al., 2005] para detectar CCL tem sido desenvolvidos.

A descrição da prancha do Roubo dos Biscoitos também foi aplicada a indivíduos com CCL. Os autores Bschor et al. [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, nosso foco de interesse, Hodges et al. [1996] examinaram o desempenho de controles saudáveis e indivíduos com diversos graus de comprometimento de DA e 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 estudos com fala espontânea em CCL.

Os estudos sobre descrição (oral e escrita) de figuras simples e complexas realizado por Forbes-McKay and 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.

Em resumo, para avaliar a linguagem de indivíduos com CCL é importante dispor de instrumentos sensíveis para detectar déficits sutis. Além disso, o monitoramento dessas dificuldades também carece de instrumentos acurados.

A análise do discurso mostra-se interessante pois abrange os diferentes componentes da linguagem, numa perspectiva linguístico-cognitiva, cujo declínio é típico das condições mencionadas – envelhecimento saudável, comprometimento cognitivo leve e doença de Alzheimer. Porém seu emprego somente será viável se dispusermos de instrumentos que permitam a organização de uma base de dados com número representativo de informantes.

2.1.4 Extratos de entrevistas de pacientes acometidos com DA: rupturas e reformulações

A tese de doutorado de Letícia Mansur [Mansur, 1996] teve como objetivo o estudo das rupturas da formulação (frases incompletas, digressões, frases ininteligíveis e confusas) e os processos de reformulação da produção oral (repetições de sílabas e palavras, reformulações por paráfrases e por correções) de indivíduos com doença de Alzheimer (DA), em estágios leve e moderado, na situação de discurso, do tipo entrevista, especificamente do tipo conversação. A pesquisa realizou estudos de seis sujeitos com DA (3 em estágio leve

Tabela 2.1: Sujeitos: condição cognitiva e grau de escolaridade.

	Sujeito	Cond. Cognitiva	Sexo	Idade	Ocupação	Escolaridade
1	OB	leve	M	73	vendedor	4 anos
2	EFA	leve	F	70	bibliotecária	15 anos
3	JAM	leve	M	76	dentista	15 anos
4	LMVG	moderado	M	61	advogado	16 anos
5	RCLO	moderado	F	59	pedagoga	15 anos
6	CAP	moderado	F	71	professora	12 anos

e 3 em estágio moderado) e nove idosos sadios.

Os seis sujeitos com DA foram diagnosticados pelo Ambulatório de Demências do Grupo de Neurologia Cognitiva e do Comportamento, coordenado pelo Dr. Ricardo Nitrini, no Hospital das Clínicas da FMUSP. No contexto da avaliação médica foram submetidos a baterias de exames, incluindo exames clínicos e de neuroimagem, sendo excluídos outros diagnósticos relativos a demência do tipo vasculares-cerebrais. A idade dos sujeitos variou entre 59 e 76 anos para os indivíduos idosos com DA e todos estão aposentados. A tabela 2.1 apresenta os seis sujeitos, indicando sexo, idade, profissão ou ocupação prévia e grau de escolaridade.

Para a realização das entrevistas, um roteiro de questões garantiu relativa uniformidade ao desenvolvimento do diálogo. Tal roteiro permitiu analisar os processos de reformulação em situações discursivas e induziu a produção de discursos narrativos, argumentativos e de procedimento, bem como uma variante de procedimento relacionada à solução de problemas.

Foram avaliadas as condições cognitivas dos sujeitos, através do Mini-exame do Estado Mental (*Mini-mental State Examination*, MMSE) [Folstein et al., 1975] . Os indivíduos com DA em grau leve obtiveram a pontuação entre 17 e 24 no MMSE, enquanto os que apresentavam a doença em grau moderado obtiveram entre 12 e 17 pontos no MMSE. O status cognitivo dos sujeitos foi verificado através de testes neuropsicológica, envolvendo a bateria *Mattis Dementia Rating Scale* (Mattis DRS) [Mattis, 1988], que compreende o exame de habilidades cognitivas de atenção, iniciativa, construção, conceituação e memória.

As avaliações de linguagem foram gravadas e transcritas integralmente, segundo convenção de análise de conversação disponibilizada pelo Projeto NURC[9]. A tabela 2.2 apresenta na coluna da esquerda, os fenômenos anotados e, à direita, as marcas de anotação, e logo abaixo as observações com instruções sobre situações específicas de transcrição.

Com o conjunto das entrevistas dos seis pacientes com DA, que estão disponibilizados no Anexo IV da tese de Mansur (1996), se constituiu um corpus, chamado de DA-PLN-EVAL, para as avaliações de ferramentas básicas do PLN como etiquetadores morfossintáticos (*taggers*) e parsers, que dão base às métricas a serem implementadas no *Coh-Metrix-Dementia*.

Além disso, foi possível analisar, frente às características da linguagem dos sujeitos com DA, quais seriam os pós-processamentos necessários após a transcrição, para que as ferramentas de PLN citadas acima, que normalmente são treinadas com córpus jornalísticos, pudessem ter um desempenho bom nos textos dos pacientes acometidos por desordens neurodegenerativas. O trabalho de Fraser et al. [2012] ajudou a definir

[9] http://www.letras.ufrj.br/nurc-rj/

Tabela 2.2: Normas para a transcrição utilizadas.

Fenômeno	**Marca**
Incompreensão de palavras ou segmentos	( )
Truncamento[10]	/
Entoação enfática	maiúsculas
Prolongamento de vogal e consoante	:::::
Silabação	-
Interrogação	?
Qualquer pausa	...
Comentários descritivos do transcritor	(( ))
Superposição, simultaneidade de vozes	{

Observações:

1. Iniciais maiúsculas: só para nomes próprios ou par siglas (USP, etc).
2. Fáticos: ah, éh, ahn, ehn, uhn, tá.
3. Nomes de obras ou nomes comuns estrangeiros são grifados.
4. Números são transcritos por extenso.
5. Não se indica o ponto de exclamação (frase exclamativa).
6. Não se anota o cadenciamento da frase.
7. Podem-se combinar sinais. Por exemplo: oh ::: ... (alongamento e pausa).
8. Não se utilizam sinais de pausa, típicos da língua escrita, como ponto-e-vírgula, ponto final, dois pontos, vírgula. As reticências marcam qualquer tipo de pausa.

os tarefas de pós-processamento da transcrição, relatada na Seção 4.4.2 (Experimento Piloto 2), resultando no corpus DA-PLN-EVAL-pós_processado.

Assim, o objetivo desta seção é ilustrar a linguagem utilizada por sujeitos em estágios diferentes da DA, para antecipar os desafios de processamento computacional desse tipo de texto. São apresentados vinte quatro extratos ilustrando as treze características:

1. Repetições de informação.
2. Falhas de acesso lexical de substantivos e nomes próprios.
3. Hesitações (pausas, prolongamentos e repetições).
4. Uso de termos vagos (na literatura aparece como “linguagem vazia”).
5. Uso de onomatopéias em substituição a itens lexicais.
6. Falhas de formulação (frases abandonadas nas quais a paciente se apoia no contexto ou no saber compartilhado com o entrevistador).
7. Repetição de segmentos com apoio na fala do interlocutor.
8. Respostas evasivas e vagas.
9. Repetição da questão formulada (dificuldade de compreensão).
10. Dificuldade na apreensão do sentido metafórico.
11. Dificuldades de compreensão e produção do texto oral (dificuldade para reter a pergunta dificuldade em compreender e organizar a enunciação segundo eixo semântico).
12. Dificuldades na organização de scripts (estabelecimento da sequencia de passos, estabelecimento do eixo semântico básico para organização - macroestrutura).
13. Dificuldade de acesso à forma fonológica.

As características que são ilustradas nos trechos aparecem negritadas e são agrupadas em seis seções, uma para cada sujeito do córpus analisado, indicado pelas iniciais do nome, para evitar exposição de sujeitos de análise no córpus. O anexo A contém todos os extratos coletados do corpus DA-PLN-EVAL.

2.1.4.1 SUJEITO 1 (DA leve) OB

1. REPETIÇÕES DE INFORMAÇÃO

EXEMPLO 1

E- o senhor correu muito pra chegar até aqui seu Osvaldo?

O- não... viemos tranquilos

E- não correu não?

O- {**nós moramos em Mairiporã né?**

E- ah em Mairiporã?

O- é... tomamos o: ônibus pra podê saltar no metrô aqui... tranquilão/

{hum}

E- e o senhor anda bastante aqui em são Paulo seu seu Osvaldo?

O- aqui não porque eu não moro aqui... só quando eu estou aqui... **nós moramos em Mairiporã né?** ... lá é minha vida né?

EXEMPLO 2

O- eu gostava muito de vendas sempre fui homem de vendas ... já ouviu falá em Kelsons?

Kelsons é a maior fábrica de bolsas de senhoras do Brasil ... no Rio de Janeiro...

então eles tinham uma: ... várias lojas deles aí também ... trabalhei vinte e cinco anos como gerente de vendas... tinha sessenta pessoas... sessenta vendedores...

1ª REPETIÇÃO DAS INFORMAÇÕES FORNECIDAS NO TRECHO 2 (ACIMA EM ITÁLICO)

O- antigamente eu tomava muito café.. porque eu trabalhava eu trabalhava em vendas... em vendas pra iniciar a venda e ser bem sucedido você precisa convidar o o (companheiro) prum café... né?... (porque bom)... deixa... a:... as infelicidades pra trás... e vamos comprar... e do café sai aquela (conversa) gostosa e tal... cê tá vendendo tal...**eu fui:... gerente da Kelsons... gerente de vendas**... **aqui em São Paulo porque a fábrica é no Rio... eu era gerente de São Paulo... cheguei a ter até sessenta pessoas...**

2. FALHAS DE ACESSO LEXICAL

SUBSTANTIVOS COMUNS

EXEMPLO 1

Descrição: dificuldade em acessar o termo “prédios”.

Pistas para análise: pausa de hesitação diante de substantivo [F0E0?] emprego de termo correlato (apartamentos) [F0E0?] substituição brusca pelo termo desejado (prédios)

O- cê nunca ouviu fala?... clube de Mairiporã?

E- eu num num conheço

O- é uma beleza

E- é?

O- tem/ é um é privilegiado ... tem tudo tem três piscinas... mais o terreno é que é próprio é adequado ... uma sede maravilhosa tal **e: existe um:...um:...ô meu Deus deixa pra lá...** apartamentos dentro do próprio

terreno... o terreno tem uma divisão...o clube é uma depois lá eles fizeram... três... **apartamentos mui/ três prédios** muito bons... nós já moramos lá: quando eu deixei de trabalhar

NOMES PRÓPRIOS

EXEMPLO 2

Pistas para a análise:

Pausa de hesitação diante de predicativo ("apelidado") do verbo de ligação, repetição (foi – foi), pausa de hesitação e repetição diante de nome próprio (Diamante Negro)

E- e quem eram os grandes jogadores?

O- bom como o grande Leônidas né?... o: inclusive na na Europa no campeonato:... mundial ele foi... **foi apelidado de o:... o:... de Diamante Negro...**

3. OUTRAS HESITAÇÕES - PAUSAS, PROLONGAMENTOS E REPETIÇÕES

EXEMPLO 1

E- como é que está o futebol agora que eu não tô muito:

O {é está: está bem... está no:... **no auge do do do do do do** campeonato... sei lá... agora termino aqui em São Paulo... o Corinthians foi campeão eu sou:... corinthiano... então

2.1.4.2 SUJEITO 2 (DA leve) EFA

1. USO DE TERMOS VAGOS (NA LITERATURA APARECE COMO "LINGUAGEM VAZIA")
TERMOS VAZIOS EMPREGADOS: "COISA" , "COISINHA", "NEGÓCIO", "TUDO", "TAL", "VÁRIOS"

EXEMPLO 1

EFA- ... bom ... eu já morei aqui em São Paulo quando era minina ... depois meus pais mudaram pra Santos ... i: adoro Santos ... num tem ... pode se dizê qui me criei ... mesmo lá ... eu estudei ... fiz tudo lá em Santos ... eu jamais quiria voltá morá aqui em São paulo ... por causa dessa correria ... **essa coisa assim** ... ADORO São Paulo ... de vir pra São Paulo pra passeá assim ... tá tudo tá ... tá comigo ... mas : morá não ... de jeito nenhum ...

EXEMPLO 2

E- mas como é que é o café? ... como que é a receita?... se a senhora tiver que ensinar prum estrangeiro que chega

EFA- ah ... como é que eu faço café bom ... eu fervo água ... quando a água já tá fervendo (depois) aí eu ponho o pó ... mexo fê/ mexo bem ... deixo ... quando começa a querê subi eu desligo ... passo pro coador ... né? ... côo o café ... fica u cafezinho bom ... aí pego depois u coadinho ... ponho em banho maria ... o bule pra ficá quentinho ... né? ... pego leite ... pãozinho biscoito ... **essas coisas mais**

2. USO DE ONOMATOPÉIAS EM SUBSTITUIÇÃO A ITENS LEXICAIS

EXEMPLO 1

E- e dus outros jogadores que penduraram a chuteira ... a senhora lembra quem eram

EFA- {não porque eu num sô muito di marca sabe? ... nome de de futebol ... eu vejo futebol assim ... um tá assistindo futebol ... eu sento assisto futebol ... torço ... vai ganhá **vá vá va** ... mas eu num fixo assim nus ... nus jogadores

EXEMPLO 2

E- i na segunda feira? ... qual é sua rotina di segunda feira?

EFA- bom segunda feira eu tô com a empregada em casa ... então eu já tô di beleza

E- ((inint.)) aí é que a senhora se diverte né?

EFA- {aí ... é que eu me divirto... eu só saio pra fazê compra ... já vejo o que que ela vai fazê de almoço ... **(pé pé pé pé pé pé)** dô as ord as ordens lá ... ela se vira

EXEMPLO 3

EFA-... aquele belo molho junto com a carne ... aquele molhinho a gente cozinha o macarrão separado ((faz o barulho do cozimento)) **ch ch...** queijo bastante

3. FALHAS DE FORMULAÇÃO – FRASES ABANDONADAS NAS QUAIS A PACIENTE SE APOIA NO CONTEXTO OU NO SABER COMPARTILHADO COM O ENTREVISTADOR

EXEMPLO 1

EFA- Santos é mais calmo né ...

E- é mais calmo

EFA- gente sai : ... é mais tranqui:lo ... condução é mais fácil ... num tem **esse** problema di ... como aqui em são Paulo **essa** loucura né?

2.1.4.3 SUJEITO 3 (DA leve) JAM

1. REPETIÇÃO DE SEGMENTOS COM APOIO NA FALA DO INTERLOCUTOR

Pistas para análise:

O paciente repete o que o entrevistador pergunta como forma de apoio e também repete as frases emitidas, como forma de preenchimento.

Além disso, mantém as hesitações e prolongamentos.

EXEMPLO 1

E- Santos agora já é cidade grande né?

J- ah é ... isso aqui ... isso aqui ... já é:: muito **grande** ... i já:: é: tem uma vida::

E- {já é uma cidade grande ...

J- ... vamos dizê assim é:: independente né? ... muito bom ...

EXEMPLO 2

J- u meu papel é:: si pricisá alguma coisa assim di di di

E- {retaguarda ...

J- é **retaguarda** ... entã:o me consulta mi:

2. RESPOSTAS EVASIVAS e VAGAS

Pista para análise:

"NÃO SEI" , "NÃO GOSTO..." "ACHO QUE FOI BOM" "NÃO SENTI NADA" , "EU NUM PRECISO FAZÊ:: NADA MAIS DO QUE AQUILO QUI EU QUE EU TÔ ACOSTUMADO FAZÊ

EXEMPLO 1

E- morreram alguns mendigos qui tavam (morando)

J- {é qui tavam lá ... é:: ... como se diz? tavam deitados lá ... pra pra::: é **num sei si...o que ... tavam fazendo lá** né?

2.1.4.4 SUJEITO 4 (DA moderada) CAP

1. REPETIÇÃO DA QUESTÃO FORMULADA – DIFICULDADE DE COMPREENSÃO

EXEMPLO 1

E- acha difícil? como é que é a sua cidade?

*C- **a minha cidade ?*** ãhm: agora eu moro no interior né?

EXEMPLO 2

E- ih:... a senhora sai muito?

C- ***se eu saio?*** eu saio sim... um pouco...

EXEMPLO 3

E- que que a senhora faz no domingo?

C- ***nu domingo?*** agora a gente... come muito pouco... a a tudo que queé cum carne com isso... com aquilo...

2. DIFICULDADE NA APREENSÃO DO SENTIDO METAFÓRICO

EXEMPLO 1

E- Dona C.? que que significa apertá o cinto pra senhora?

C- apertá o cinto?

E- é...

C- é apertá o cinto né ?

E- como assim?

C- **si tá fazendo muita arte com o cinto... deixa ele um pouco de lado i i discansá um pouco...**

2.1.4. 5 SUJEITO 5 (DA moderada) RLCO

1. DIFICULDADES DE COMPREENSÃO E PRODUÇÃO DO TEXTO ORAL.

DIFICULDADE PARA RETER A PERGUNTA DIFICULDADE EM COMPREENDER E ORGANIZAR A ENUNCIAÇÃO SEGUNDO EIXO SEMÂNTICO

EXEMPLO 1

E- qual o ... qual o ... que tipo de artista a sra prefere ... os que fazem papel de ...

E- mocinho ... ou os que fazem papel de bandido? ...

R- depende de cada um né?... **fala otra vez a pergunta ...**

E- é:: quê ... que tipo de artista a sra gosta? ... artista de novela? ... por exemplo ... os que

{não tenho paciência ...eu não tenho...}

E- fazem...

E- não? ...

EXEMPLO 2

E- já aconteceu alguma coisa desagradável pra sra aqui na cidade grande? ...

R- a quando na na :: ... ((ininteligível)) onde nós fomos ... no iscã nu iscãd qual que foi?... o que é que foi mesmo? u que queu ia falá ? ... pera aí fal... fala otra vez...

2. DIFICULDADES NA ORGANIZAÇÃO DE SCRIPTS (ESTABELECIMENTO DA SEQUENCIA DE PASSOS, ESTABELECIMENTO DO EIXO SEMÂNTICO BÁSICO PARA ORGANIZAÇÃO - MACROESTRUTURA)

EXEMPLO 1

E- i cafezinho? ... como que a sra. ... gosta du café? ...

R- {ele ele ele que ele gosta eu também gosto ...

E- é?

R- mas eu faço pouco né porque num é bom pra mim... (logo)

E- {mas como que a sra gosta du café? ...

R- mais forte ...

E- forte? ... hum hum ... eu também prefiro forte ...

R- é ...

E- igual o SEU jeito de fazer café? ...

R- eu pego a água ... pego o café (vai) **vai fica fica a** água né ? ... depois eu pego ... um pouco do pó.... **daí fica meio assim assim** ... **depois põe na na coisa**...

3. DIFICULDADE DE ACESSO À FORMA FONOLÓGICA

EXEMPLO 1

E- que que a sra assistiu ... leu de importante nos últimos tempos? ...

R- {ah: foi foi o ***elipse*** ... foi o ***elipse***:

E- ah:: eclipse? ...

R- é ...eclipse ...

SUJEITO 6 (DA moderada) LMVG

4. DIFICULDADES PARA FORMULAR

DIFICULDADES NA ORGANIZAÇÃO DE FRAMES E SCRIPTS (ESTABELECIMENTO DA SEQUENCIA DE PASSOS, ESTABELECIMENTO DO EIXO SEMÂNTICO BÁSICO PARA ORGANIZAÇÃO - MACROESTRUTURA)

LMVG responde em estilo sintético e necessita da intervenção constante do interlocutor para expandir a resposta.

EXEMPLO 1

E- u senhor sai muito?

L- **eu saio** ...

E- é... ih ... u senhor costuma sair sozinho?

L- **sozinho** ...

E- é ... que que o senhor faz?

L- ando ...

E- anda?

L- **ando** ando ando ... ando ando ... ando ...

E- todos os dias?

L- **todos os dias**...

E- é.?

L- é ...

E- qui bom ... faz bem né...

5. ALTERAÇÕES LEXICAIS – FORMA FONOLÓGICA

EXEMPLO 1

E- o senhor vota?

L- claro ... *o* ***meu pau*** ... o meu pai ... *o* ***meu pau*** ... não/ ... calma ... calma ...

E- {ahm}

L-u meu pai ... é velho né... eu num queria ... i ele fala/ não cê tem que i lá ((imita a voz do pai agravada))

2.1.5 Considerações finais

O exame do cenário acima exposto nos permite destacar as seguintes lacunas no conhecimento de idosos sadios, com alterações cognitivas e 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 sadios, falantes do Inglês, porém essa visão não é consensual.

Além disso, desconhecemos o desempenho de idosos brasileiros nesse aspecto e a influência de fatores relacionados ao Português Brasileiro caracterizado por amplas marcas morfológicas e redundância.

Estudos do discurso de idosos sadios 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 sadios é importante, para constituir padrões de referência ao diagnóstico de detecção e evolução de alterações cognitivas, como os comprometimentos cognitivos leves e doença de Alzheimer, 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: doença de Alzheimer

Desconhecemos a exata interação de fatores relacionados ao envelhecimento normal e patológico. Em outras palavras, as manifestações da doença de Alzheimer seriam uma expressão exacerbada do envelhecimento normal? 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, aspectos léxico-semânticos, sintáticos e discursivos) nos subgrupos 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?) mais sensíveis ao processo demencial.

3. Comprometimento cognitivo leve

O comprometimento cognitivo leve 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 comprometimento cognitivo leve.

Um aspecto importante é a taxa de conversão de indivíduos com alteração cognitiva para quadros demenciais. Os fatores de risco (linguagem ??) 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.

2.2 Ferramentas automáticas para análise de características textuais disponíveis para o português

Nesta seção, são apresentadas ferramentas já implementadas, que dão a base ao Coh-Metrix-Dementia. Na seção 2.2.1, é apresentado o Coh-Metrix e sua adaptação ao português brasileiro, o Coh-Metrix-Port, no qual este trabalho fortemente se baseia; na seção 2.2.2, é apresentada a ferramenta AIC (Análise de Inteligibilidade de Córpus), desenvolvida durante o projeto PorSimples que também servirá de base a este trabalho, por conter métricas sintáticas; na seção 2.2.3, é descrita a ferramenta LIWC e seu dicionário que foi traduzido para o português.

2.2.1 O Coh-Metrix e o Coh-Metrix-Port

O Coh-Metrix[11] [Graesser et al., 2004; McNamara et al., 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 and 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 et al.,

[11] http://cohmetrix.memphis.edu/cohmetrixpr/index.html

2011].

1. Palavras

O conhecimento de vocabulário possui um impacto substancial no tempo de leitura e compreensão [Perfetti, 2007; Rayner et al., 2001; Stanovich, 1986; Graesser et al., 2011]. 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 [Perfetti, 2007; Graesser et al., 2011]. 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 et al., 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 na 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 et al., 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 et al., 1994, 2004, 2011; Zwaan and Radvansky, 1998].

2. Sintaxe

Teorias sintáticas atribuem palavras a categorias gramaticais ou morfossintáticas (e.g., nomes, verbos, adjetivos), agrupam palavras em sintagmas (nominais, verbais, preposicionais), e atribuem uma estrutura em árvore sintática para as sentenças [Jurafsky and Martin, 2009]. Algumas sentenças são curtas, seguem (por exemplo) a estrutura sujeito-verbo-objeto, possuem poucas orações subordinadas e usam a voz ativa ao invés da passiva. Sentenças assim são mais próximas da expressão oral [Tannen, 1982], e tendem ser mais fáceis de processar. Outras, em contrapartida, contém sintagmas com grande número de modificadores, colocam muitas palavras antes do verbo principal (“enterram” o verbo) da oração principal (o que impõe esforço 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 et al., 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-Metrix é capaz ainda de medir a frequência de aparição de voz passiva (que é mais difícil de ser processada que a voz ativa [Just and Carpenter, 1987]) e a similaridade sintática entre sentenças do texto, uma vez que o paralelismo sintático facilita a compreensão [Graesser et al., 2011].

3. Base textual

A base textual se refere às ideias explicitadas no texto: o significado, ao invés da superfície textual [Kintsch, 1998; van Dijk and Kintsch, 1983]. 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 [Halliday and Hasan, 1976; McNamara and Kintsch, 1996]. 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 [McNamara and Kintsch, 1996; O'Brien et al., 1998]. O Coh-Metrix analisa diversos tipos de correferência, como **sobreposição de palavras de conteúdo**, **sobreposição de nomes**, **sobreposição de argumentos** e **sobreposição de radical** [Graesser et al., 2011].

Um outro fator importante, também analisado pelo Coh-Metrix 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 et al., 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

O modelo situacional é o conteúdo do assunto sendo tratato 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 [Graesser et al., 2011]. Zwaan and 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 [Graesser et al., 2011]. 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 [O'Brien et al., 1998; Rapp et al., 2007; Zwaan and Radvansky, 1998]. 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

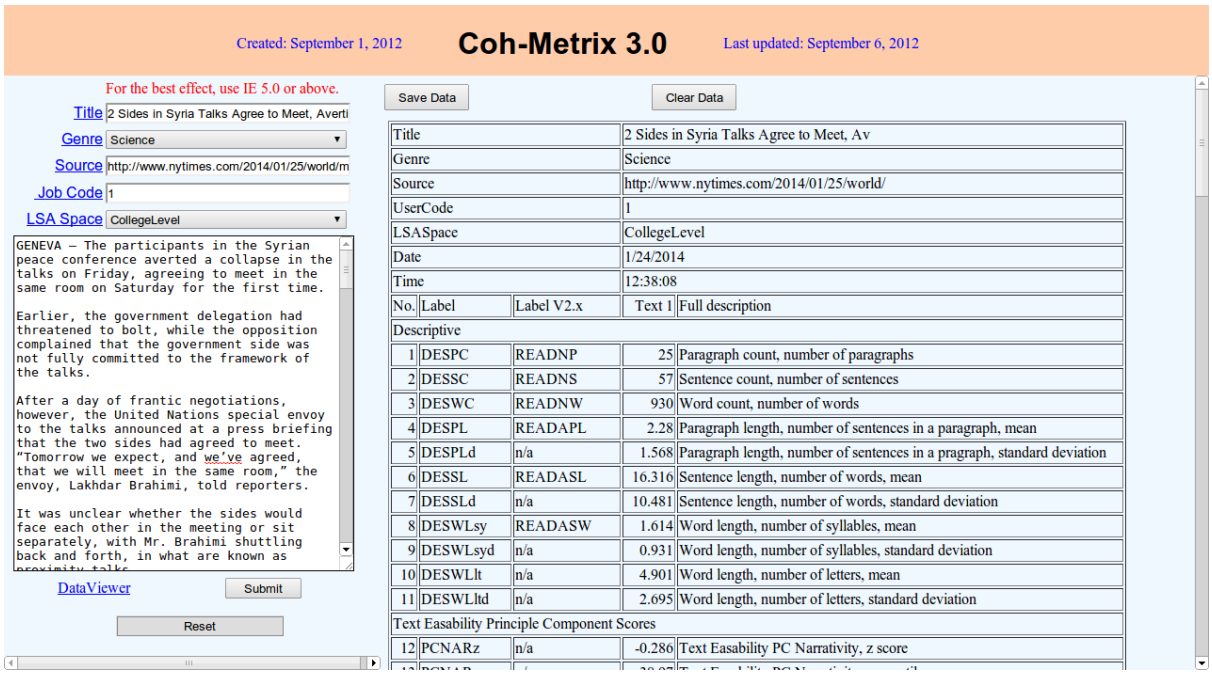


Figura 2.1: Tela do Coh-Metrix exibindo métricas de um texto de exemplo.

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 [McNamara and Kintsch, 1996; O'reilly and McNamara, 2007; Ozuru et al., 2009].

5. Gênero e estrutura retórica

O gênero textual se refere à categoria do texto, que pode ser narração, exposição, persuasão ou descrição [Biber, 1988; Pentimonti et al., 2010; Graesser et al., 2011]. Essas categorias podem ser divididas ainda em sub-categorias, formando uma taxonomia, sendo que o texto narrativo substancialmente mais fácil de ler, compreender e relembrar que o texto informacional [Graesser and McNamara, 2011; Haberlandt and Graesser, 1985]. Treinar os leitores para distinguir o gênero e as estruturas globais do texto os ajuda a melhorar a compreensão [Meyer et al., 2010]. 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 et al., 2011].

A Figura 2.1 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 2.3. São 108 métricas, agrupadas em 11 categorias, além de meta-dados acerca do texto.

Tabela 2.3: Métricas do Coh-Metrix 3.0.

Title	Title

Continua…

Tabela 2.3 – *Continuação*

Genre			Genre
Source			Source
UserCode			UserCode
LSASpace			LSASpace
Date			Date
Time			Time
	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	DESWLsy	READASW	Word length, number of syllables, mean
9	DESWLsyd	n/a	Word length, number of syllables, standard deviation
10	DESWLlt	n/a	Word length, number of letters, mean
11	DESWLltd	n/a	Word length, number of letters, standard deviation
Text Easability Principal Component Scores			
12	PCNARz	n/a	Text Easability PC Narrativity, z score
13	PCNARp	n/a	Text Easability PC Narrativity, percentile
14	PCSYNz	n/a	Text Easability PC Syntactic simplicity, z score
15	PCSYNp	n/a	Text Easability PC Syntactic simplicity, percentile
16	PCCNCz	n/a	Text Easability PC Word concreteness, z score
17	PCCNCp	n/a	Text Easability PC Word concreteness, percentile
18	PCREFz	n/a	Text Easability PC Referential cohesion, z score
19	PCREFp	n/a	Text Easability PC Referential cohesion, percentile
20	PCDCz	n/a	Text Easability PC Deep cohesion, z score
21	PCDCp	n/a	Text Easability PC Deep cohesion, percentile
22	PCVERBz	n/a	Text Easability PC Verb cohesion, z score

Continua…

Tabela 2.3 – *Continuação*

23	PCVERBp	n/a	Text Easability PC Verb cohesion, percentile
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	CRFBSaum	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
46	LSAGN	LSAGN	LSA given/new, sentences, mean
47	LSAGNd	n/a	LSA given/new, sentences, standard deviation
Lexical Diversity			

Continua…

Tabela 2.3 – *Continuação*

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, MTLD, all words
51	LDVOCDa	LEXDIVVD	Lexical diversity, VOCD, all words
Connectives			
52	CNCAll	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
72	SYNMEDwrd	MEDawm	Minimal Edit Distance, all words
73	SYNMEDlem	MEDalm	Minimal Edit Distance, lemmas
74	SYNSTRUTa	STRUTa	Sentence syntax similarity, adjacent sentences, mean.

Continua…

Tabela 2.3 – *Continuação*

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
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

Continua…

Tabela 2.3 – *Continuação*

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 and Aluísio [2010], no contexto do projeto PorSimples[12]. O PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) (Aluisio et al., 2008a, 2008b; Caseli et al., 2009, Candido Jr. et al., 2009) 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 INAF[13].

Além de se utilizado no SIMPLIFICA, o Coh-Metrix-Port foi utilizado em diversos cenários de análise e classificação textual. Scarton and 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 córpus utilizado consistia de textos jornalísticos (textos regulares do jornal ZeroHora[14] - complexos - e da seção *Para seu filho ler* do mesmo jornal - simples) e textos de divulgação científica (textos das revistas Ciência Hoje[15] - complexos - e Ciência Hoje das Crianças[16] - 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 córpus considerado.

Com o córpus 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

[12] www.nilc.icmc.usp.br/porsimples

[13] Indicador de Alfabetismo Funcional. Relatório disponível em http://www.ibope.com.br/ipm/relatorios/relatorio_inaf_2009.pdf.

[14] http://zerohora.clicrbs.com.br/

[15] http://cienciahoje.uol.com.br/revista-ch

[16] http://www.chc.org.br/

(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 and Aluísio [2010] foi expandido por Scarton et al. [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 córpus 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 unigramas, bigramas e trigramas, para assim classificar textos simplificados. Foi utilizado o córpus de referência do projeto PorSimples[Aluísio and 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-Metrix-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 Gaúcho[17]. Consideraram-se quatro cenários: um com todas as métricas, outro somente com as do Coh-Metrix-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 et al. [2011] fez uso do Coh-Metrix-Port e do Coh-Metrix 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 2.4 mostra um sumário do desempenho apresentado pelos classificadores treinados com atributos extraídos pelo Coh-Metrix-Port em três dos trabalhos citados acima. A Figura 2.2 mostra a saída do Coh-Metrix-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 et al., 2010]. A tabela 2.5 mostra as métricas e uma breve descrição de cada uma.

[17] http://diariogaucho.clicrbs.com.br/rs/

Tabela 2.4: Desempenho dos classificadores com características extraídas pelo Coh-Metrix-Port.

Trabalho	Cenário	Medida F
Scarton and Aluísio [2010]	Todos os textos	**93,0%**
	Jornalísticos	**97,0%**
	Científicos	**94,0%**
	Outros - melhor	**94,0%**
	Outros - pior	**61,3%**
Aluisio et al. [2010]	Original	**90,4%**
	Natural	**48,4%**
	Forte	**73,1%**
Finatto et al. [2011]	Melhor cenário	**87,7%**

Tabela 2.5: Métricas do Coh-Metrix-Port.

Texto		
Título		Título
Autor		Autor
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 Sentencas	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.

Continua...

Tabela 2.5 – *Continuação*

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.
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 córpus Banco do Português.
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		

Continua…

Tabela 2.5 – *Continuação*

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.
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.

Continua…

Tabela 2.5 – *Continuação*

44	Sobreposição de radicais de palavras adjacentes	Sobreposição de argumentos em sentenças adjcentes.
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.
48	Referência anafórica	Referência anafórica em até cinco sentenças anteriores.

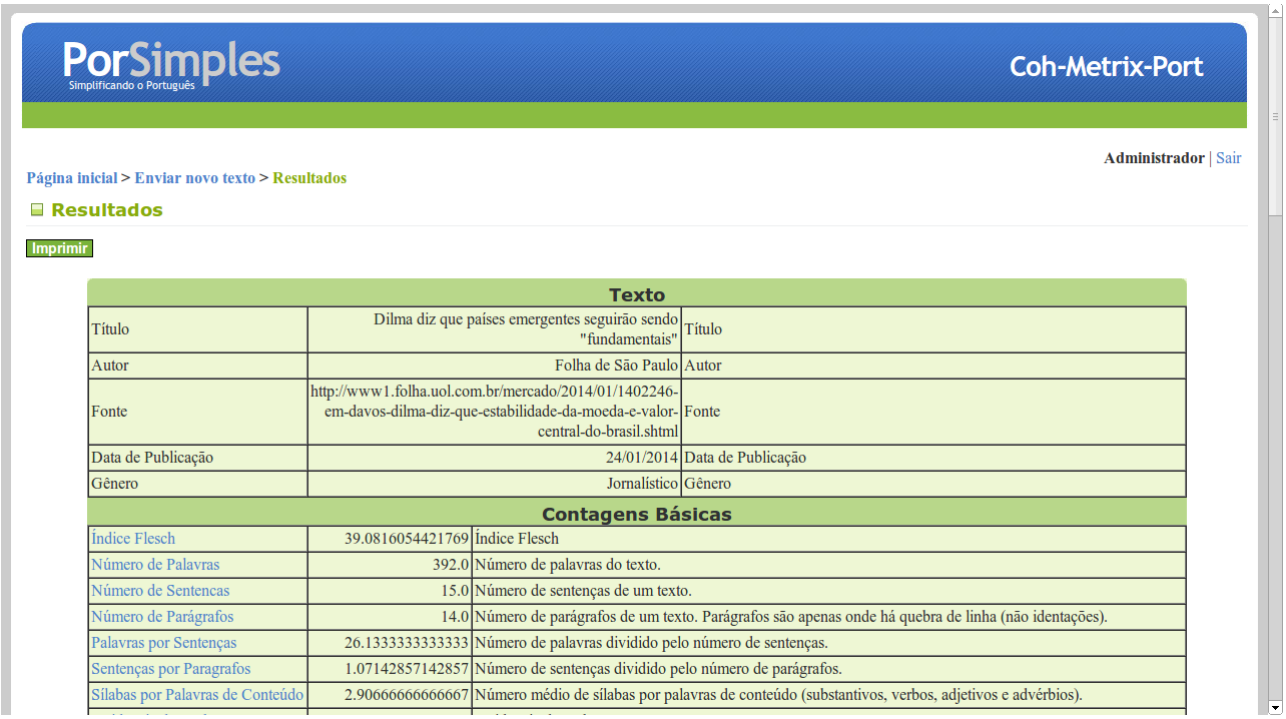
Como se pode perceber, o Coh-Metrix-Port foi empregado em diversos cenários de classificação textual, obtendo desempenho bastante satisfatório em muitos deles. Como o presente trabalho também diz respeito à classificação textual em um cenário específico, acreditamos que adaptar do Coh-Metrix-Port para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

2.2.2 A ferramenta Análise de Inteligibilidade de Córpus (AIC)

O AIcórpus (Análise de Inteligibilidade de córpus[18] ou AIC [Maziero et al., 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]. Além disso, foi usado para realizar o Experimento Piloto 1 desta pesquisa, apresentado no capítulo 4.

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 são [Maziero et al., 2008] estão organizadas em seis classes: **estatísticas do texto**, **voz passiva**, **características das orações**, **densidade**, **personalização** e **marcadores discursivos**. As métricas que trazem as **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

[18] http://www.nilc.icmc.usp.br/porsimples/AIC/



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 Sentencas	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 Paragrafos	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 2.2: Saída do Coh-Metrix-Port para um texto de exemplo.

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 discursivos ambíguos.

A Figura 2.3 mostra um exemplo de um trecho da saída do AIC (saída parcial) para um dado resumo de artigo em português, tomado como exemplo, somente.

2.2.3 O *Linguistic Inquiry and Word Count* (LIWC)

O LIWC [Pennebaker et al., 2001, 2007] é uma ferramenta de análise textual que calcula o grau em que as pessoas utilizam diferentes categorias de palavras no texto.

O LIWC [Pennebaker et al., 2001, 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**

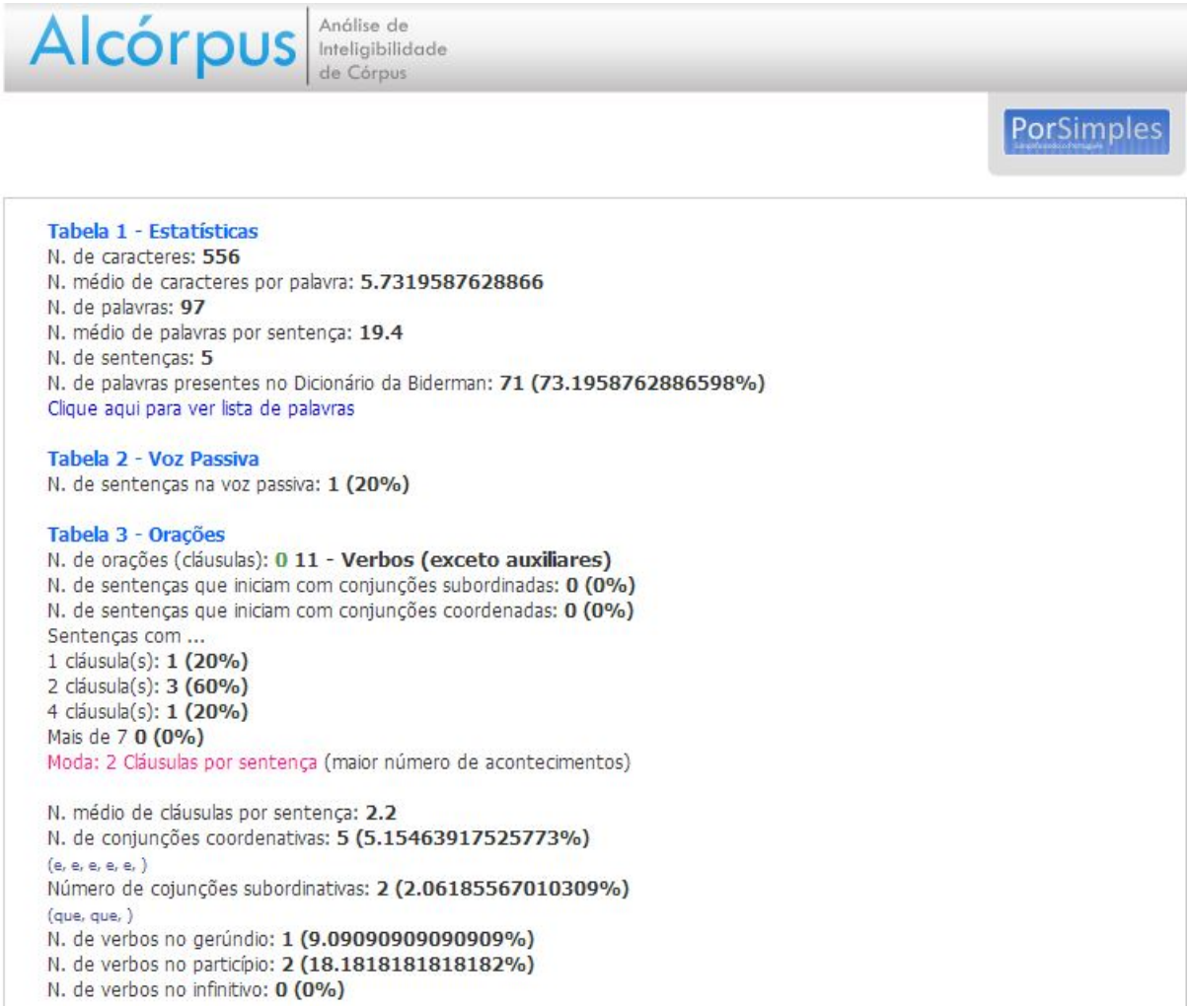


Figura 2.3: Exemplo de parte da tela de saída do AIC.

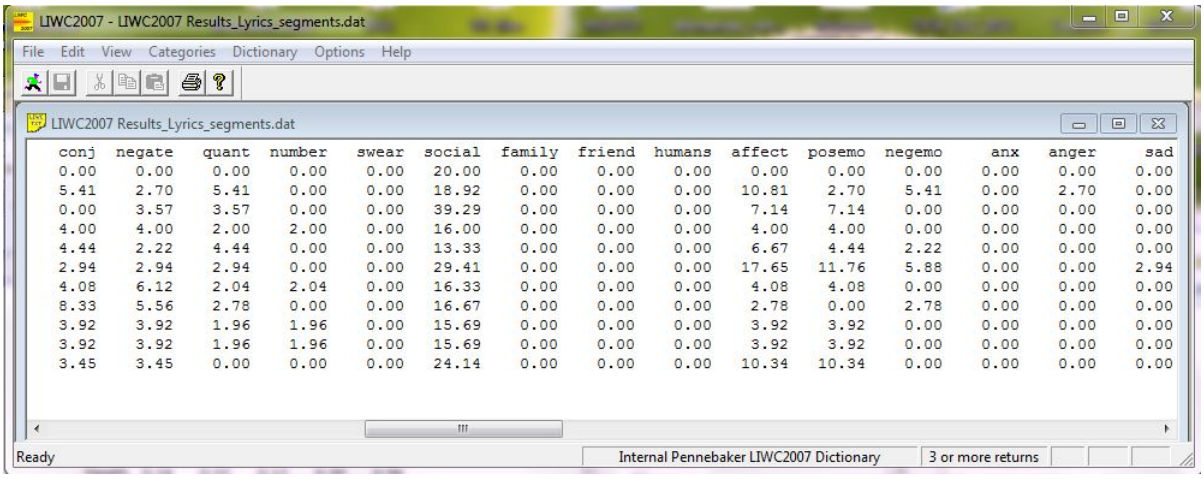


Figura 2.4: Tela de saída do LIWC, com o dicionário para o inglês, versão de 2007, mostrando a análise de uma música.

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 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 PortLex[19]. Mais detalhes de seu uso podem ser lidos em Balage Filho et al. [2013]. A Figura 2.4 mostra uma tela da saída do LIWC 2007, com o resultado da análise de uma música em inglês a qual foi dividida em 11 segmentos, contando seu título, via setup do LIWC, para análise individual de suas estrófes.

Como se pode observar na sção 3.2, o LIWC foi empregado em diversos trabalhos relacionados, e como uma tradução para o português está prontamente disponível, acreditamos que empregá-lo será importante para alcançar o desempenho das tarefas de classificação, nesta pesquisa de mestrado.

[19] http://www.nilc.icmc.usp.br/portlex/index.php/en/liwc

3. Análise Automatizada de alterações de linguagem em doenças neurológicas degenerativas

Este capítulo traz, na seção 3.1, várias medidas automatizadas, principalmente para a língua inglesa, usadas para 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 coerência textual, via semântica latente. Na seção 3.2, uma revisão da literatura de trabalhos que automatizaram a avaliação de doenças neurológicas como a Doença de Alzheimer (DA), o Comprometimento Cognitivo Leve (CCL) e as Afasias Primárias Progressivas (APP). Os nove trabalhos aqui analisados estão organizados com relação ao nível da língua com que as avaliações são realizadas; assim, trazemos trabalhos que tratam: (1) da avaliação de riqueza lexical, usando também a análise de Part-of-Speech (PoS) (ou categoria gramatical ou morfossintática) e da taxa de unidades semânticas similares a cláusulas - *Clause-like Semantic Unit (CSU) rate* ; (2) da complexidade sintática; (3) da densidade de ideias; (4) do uso de métricas derivadas de avaliação de processos psicológicos, com ajuda do software LIWC; (5) de uma análise varrendo todos os níveis linguísticos.

3.1 Medidas automatizadas usadas na identificação de condições clínicas a partir de amostras de linguagem

3.1.1 Medidas de diversidade lexical

Uma métrica de riqueza lexical já mencionada aqui é a **Relação Tipo por *Token*** (*Type to Token Ratio*, TTR). Há várias formas de calculá-la. 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:

$$TTR = \frac{V}{N} \tag{3.1}$$

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 do texto), começando as primeiras n palavras e depois calcular a média. Se o texto tiver menos que mil palavras, convenciona-se que o TTR é zero.

Outra medida, insensível ao comprimento da enunciação, é o **Índice de Brunét** W [Brunet, 1978], calculado através da seguinte equação [Thomas et al., 2005]:

$$W = N^{V^{-0.165}} \tag{3.2}$$

onde 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]:

$$R = \frac{100 \log N}{1 - \frac{V_1}{V}} \tag{3.3}$$

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 3.1.3) 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 pequenas, 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].

3.1.2 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 et al., 2007b].

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 et al., 2007b]. Nas próximas sub-seções, são apresentadas algumas dessas medidas, geralmente citadas na literatura de avaliação de quadros demenciais.

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 et al., 2007b]. 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 3.1 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 segunte 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 3.1, o peso de cada nó está indicado entre colchetes.

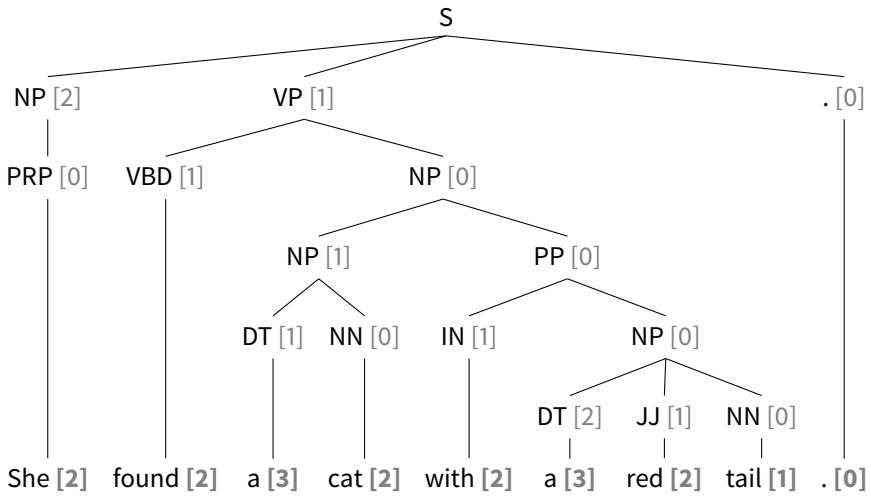


Figura 3.1: Exemplo de árvore sintática para cálculo da complexidade de Yngve (extraído de Pakhomov et al. [2011]).

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 3.1, entre colchetes, ao lado da palavra.

Uma vez calculada a pontuação das palavras, a complexidade da sentença pode ser 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 et al. [2007b] 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 et al., 2007b]. 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 $\rightarrow$ NP VP .", seguida por "NP $\rightarrow$ PRP", e finalmente "PRP $\rightarrow$ *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 et al., 2007b].

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 3.2, 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

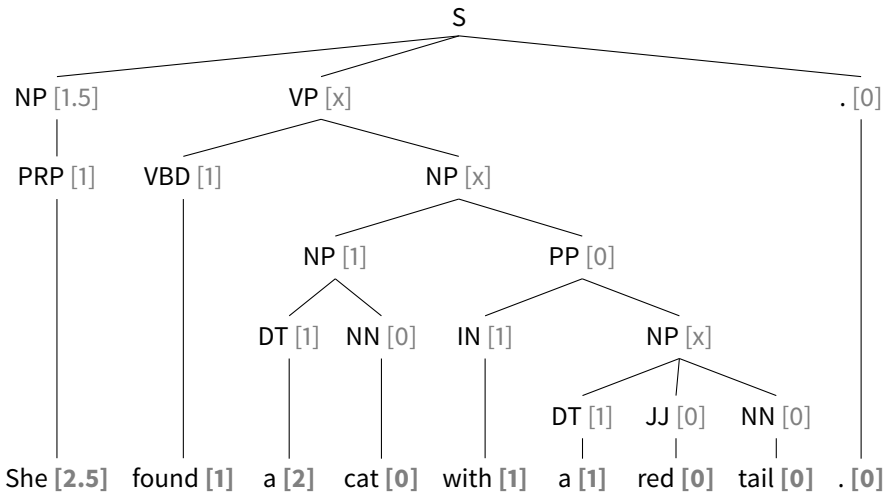


Figura 3.2: Exemplo de árvore sintática para cálculo da complexidade de Frazier (extraído de Pakhomov et al. [2011]).

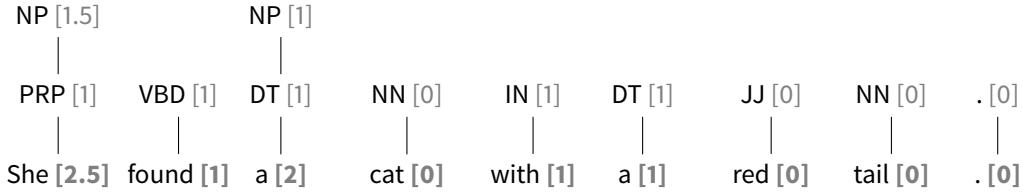


Figura 3.3: Exemplo da Figura 3.2 com os ramos individuais das palavras.

ramo. Nesse exemplo, para a derivação da palavra *She*, o nó PRP imediatamente 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 3.3 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.

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 de dependências**. A Figura 3.4 mostra a árvore de dependências da mesma sentença dos exemplos anteriores.

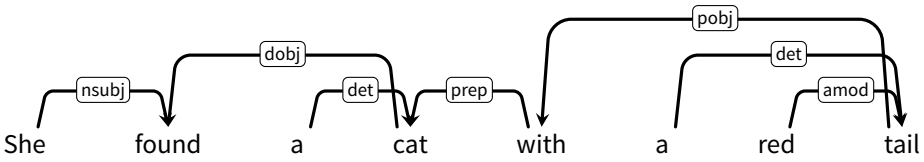


Figura 3.4: Exemplo de estrutura de dependências.

A cada relação de dependência está associada uma distância entre as palavras na superfícies 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 et al., 2007b].

Nível de desenvolvimento (nível D)

O nível de desenvolvimento (em inglês, *Developmental level* ou *D-level*) [Rosenberg and 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 et al., 2007b]. Cheung and 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 and Kemper, 1992; Roark et al., 2007b]:

- 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.

Entropia cruzada

Uma maneira de se calcular a riqueza sintática de um texto é procurar por padrões incomuns de combinações de classes gramaticais [Roark et al., 2007b]. 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 sobre classes gramaticais. Seja τ_i a classe gramatical (ou etiqueta morfossintática) da palavra w_i em uma sequência de palavras $w_1, w_2, \cdots, w_n$, e assuma que τ_0 é um marcador de início e que τ_{n+1} é um marcador de fim. Assim, a probabilidade da sequência de etiquetas é dada por:

$$P(\tau_1, \tau_2, \cdots, \tau_n) = \sum_{i=1}^{n+1} P(\tau_i|\tau_{i-1}) \tag{3.4}$$

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

$$H(\tau_1, \tau_2, \cdots, \tau_n) = -\frac{1}{n} \log P(\tau_1, \tau_2, \cdots, \tau_n) \tag{3.5}$$

Nesse caso, quanto maior a entropia cruzada, mais incomum é a elocução perante o modelo de língua, o que é um indício de maior complexidade. Uma das vantagens dessa medida é que ela necessita apenas de um etiquetador morfossintático e de um modelo de língua treinado em um grande córpus, recursos de PLN prontamente disponíveis para o português e que foram utilizados no Coh-Metrix-Port e seus trabalhos relacionados.

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 and Kemper, 1992].

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 analizam 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 estagios 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 et al. [2007b], 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 et al., 2007b].

3.1.3 Medidas de densidade semântica

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 palavras[1] [Brown et al., 2008]. Ela tem por objetivo medir a quantidade de informação que é transmitida em relação ao número de palavras utilizadas para transmití-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 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 3.1 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 refraseamento da primeira, em que há manutenção do número de proposições e elevação do número de palavras, resultando em menor densidade de

[1] Alguns trabalhos, como Kemper et al. [2001b] 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 segunda convenção.

ideias. No último exemplo, vemos uma proposição de conexão, a quarta proposição (*so*), que carrega ideia de causalidade.

Os conceitos são rotulados como agente, tema, objetivo, etc., seguindo-se a gramática de casos de Fillmore [1968, 1969]. Turner and Greene [1977] desenvolveram um manual, apresentando regras que possibilitam extrair de um fragmento discursivo suas ideias. Kemper e seus colaboradores realizaram modificações no manual de Turner and 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 sadios quanto portadores de DA e CCL, 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 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 and Keenan, 1973], memória [Thorson and Snyder, 1984], qualidade de escrita de estudantes [Takao et al., 2002] e ao envelhecimento [Kemper et al., 2001b; Kemper and Sumner, 2001].

Manuais como o de Turner and 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 CPIDR[2] (*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 3.5 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 3.1) e, do lado direito, são exibidos o resultado do etique-

[2] `http://www.ai.uga.edu/caspr/`

Tabela 3.1: 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	0,50
The house of John is big.	1. is, big, the house 2. of John	6	0,33
It was 80 years ago.	1. was, it, years ago 2. years ago, 80	5	0,40
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	0,50

tador 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 3.6 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*.

O CPIDR 3 um software livre, liberado sob licença GPL. Esses fatores o tornam adequado a utilização como uma das ferramenta de referência para a codificação do Coh-Metrix-Dementia, após adaptação das regras para o português do Brasil.

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 and 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 and Singh, 1996].

Para a separação em CSU's, são utilizadas 13 regras, apresentadas no apêndice de Bucks et al. [2000]. Holmes and 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.*

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.

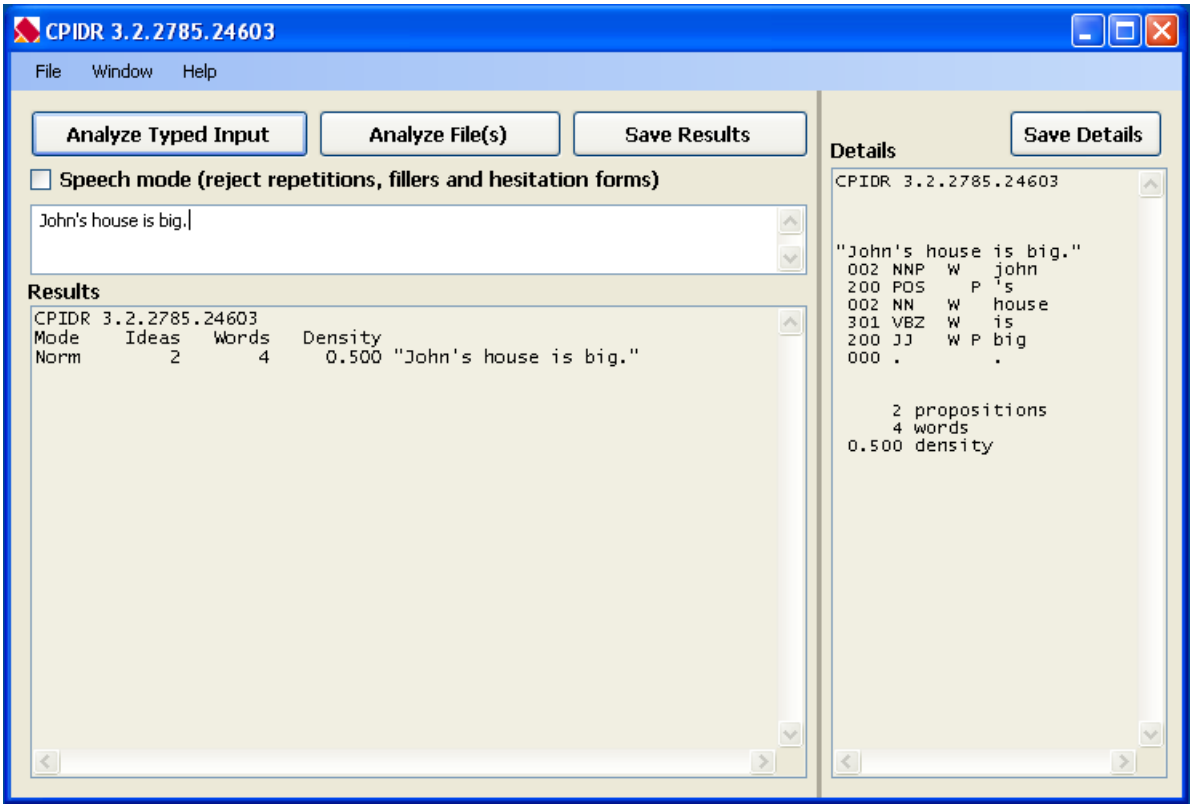


Figura 3.5: Exemplo de uso do CPIDR.

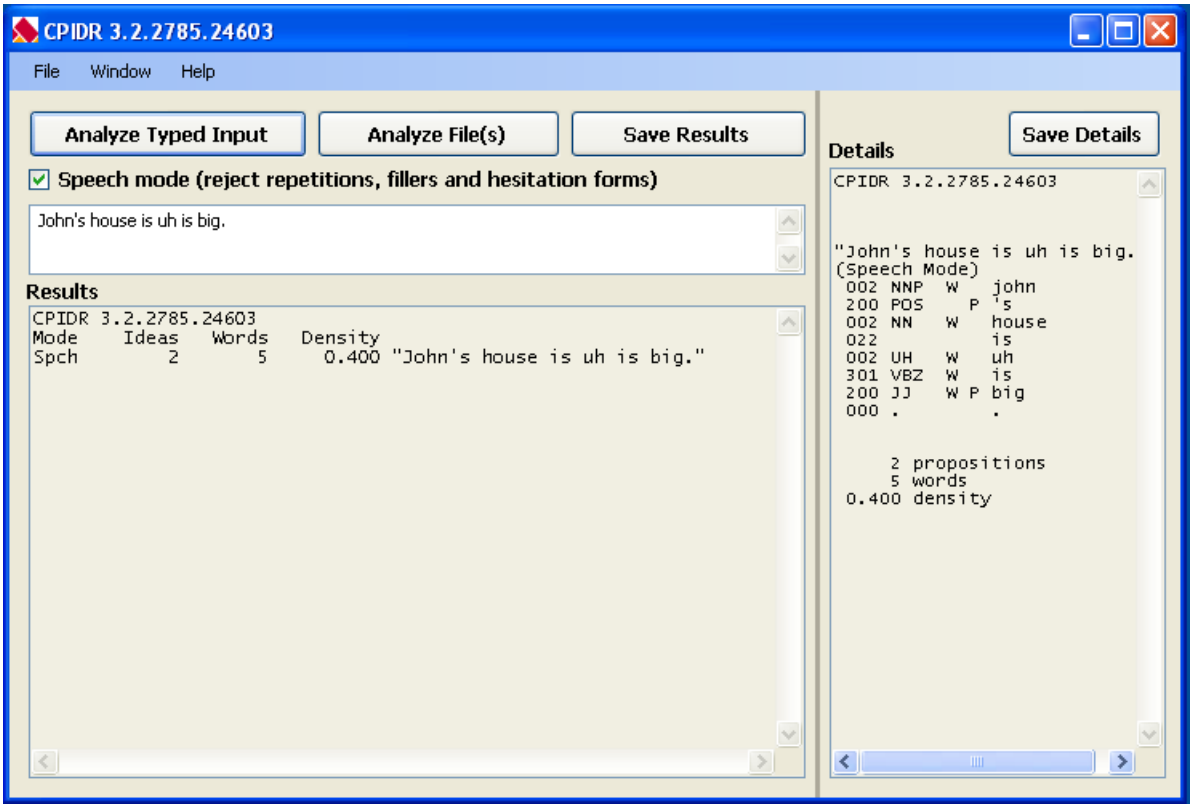


Figura 3.6: Exemplo de uso do CPIDR no modo fala.

3.1.4 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 de 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. Segundo [Dumais, 2004], o melhor desempenho é obtido com transformações sublineares, como a função logaritmo.

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 palava-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 et al., 1999], testes de vocabulário [Landauer and 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 3.1.3 que Kintsch foi responsável por criar métodos para representar textos por meio de proposições; Foltz et al. [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 déficts 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 2.3 e estão reproduzidas na tabela 3.2 para facilitar a leitura do texto. Como se pode notar, é medida a correlação entre sentenças e parágrafos adjacentes, entre todas as sentenças de um parágrafo, entre outras.

Tabela 3.2: 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

Desta forma, a proposta de implementação neste trabalho de mestrado é a mesma sugerida no Coh-Metrix, para assim contribuirmos com o projeto original do Coh-Metrix-Port com mais 8 métricas e também avaliarmos a contribuição delas no cenário do projeto de mestrado.

3.2 Trabalhos relacionados

3.2.1 Abordagem lexical

O trabalho de MacWhinney et al. [2010] (*Automated analysis of the Cinderella story*) utilizou ferramentas do TalkBank[3], 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 AphasiaBank[4].

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 BillingBank[5], de aquisição de segunda língua, o CABank[6], 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,

[3] http://talkbank.org/
[4] http://talkbank.org/AphasiaBank/
[5] http://talkbank.org/BillingBank
[6] 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: &gth &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 &gt (....) 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 3.7: Transcrição no formato CHAT da história da Cinderela (extraído de [MacWhinney et al., 2010]).

projetados para lidarem com o formato CHAT, permitem analisar diversas estruturas linguísticas e discursivas [MacWhinney et al., 2010].

A Figura 3.7 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 sumarizando 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 morfossintático 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 Mental [Folstein et al., 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 3.3.

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.

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 3.3: Sumário dos resultados de acurácia de classificação de Thomas et al. [2005].

3.2.2 Abordagem baseada em complexidade sintática

O trabalho de Roark et al. [2007b] (*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 córpus, 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. [7]

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,

[7] 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-Metrix-Dementia será beneficiado sem esforços, já que é alimentado diretamente pela saída desses sistemas.

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-Metrix-Dementia fará uso conjunto de diversas métricas para melhorar sua acurácia.

O trabalho de Roark et al. [2007a] (*Automatically derived spoken language markers for detecting Mild Cognitive Impairment*) realizou uma análise muito semelhante à de Roark et al. [2007b]. 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:

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

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 córpus. 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. [2007a] e Roark et al. [2007b] 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 foram[8]:

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%).

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.

3.2.3 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 SALT[9] versão 8 (*Systematic Analysis of Language Transcripts*). 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 elocuções.

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

[8] 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.

[9] http://www.saltsoftware.com/

Tabela 3.4: 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 *token*	X	
Comprimento médio da elocução	X	X
Número de elocuções		

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 empregrado 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 elocuções 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, 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 elocuções não apresentaram correlação.

A tabela 3.4 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 na análise do discurso de pacientes com quadros demenciais, tanto na identificação da doença quanto em sua quantificação. Reforçam também o potencial do comprimento médio da enunciação, que também tem sua inclusão prevista no Coh-Metrix-Dementia (veja seção 4.2).

3.2.4 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 2.2.3).

As características **fonológicas** empregradas 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,

Tabela 3.5: Resultados (aproximados) obtidos por Peintner et al. [2008].

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

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 empregradas 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 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 3.5; 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 desordens 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 desordens 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 3.6.

3.2.5 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

Tabela 3.6: 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%

mestrado. Nele, os autores analisaram textos produzidos por sujeitos de três grupos: controles sadios 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 morfossintático[10];
- 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.

Com base nessas métricas, foram treinados três classificadores de paradigmas diferentes: Naïve Bayes, Regressão Logística e SVM. Os resultados obtidos estão reproduzidos na tabela 3.7. 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.

[10] 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 3.7: 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
Baseline	61,5	53,3	58,3

4. Proposta de pesquisa: Coh-Metrix-Dementia

Neste capítulo, é apresentada a proposta de pesquisa deste trabalho. Na seção 4.1, é descrito o córpus que será coletado para uso durante a pesquisa; na seção 4.2, é apresentada a arquitetura planejada para a ferramenta, bem como as métricas que serão adaptadas para o portugês do Brasil e as técnicas e desafios de aprendizado de máquina que a tarefa impõe; na seção 4.3, é descrita a forma como a ferramenta será avaliada; finalmente, na seção 4.4, são apresentados dois experimentos piloto, um envolvendo sujeitos sadios em uma tarefa de descrição de figuras, e outro apresentando as dificuldades que os textos com que o Coh-Metrix-Dementia lidará impõem para ferramentas básicas de PLN como parsers e taggers (analisadores morfossintáticos).

4.1 Córpus de pesquisa

Os pacientes que produzirão os textos serão provenientes do ambulatório de Psicogeriatria do Laboratório de Neurociências do IPq - HCFMUSP. Os dados serão coletados pela Dra. Márcia Radanovic, neurologista com formação em linguagem, e pela Dra. Ariella Fornachari Belan, fonoaudióloga com especialização em neurolinguística, pertencentes ao grupo coordenado pelo Prof. Dr. Orestes Vicente Forlenza. Os sujeitos controles serão recrutados entre acompanhantes não-consanguíneos dos pacientes, pareados de acordo com idade e escolaridade. Serão, no total, 20 sujeitos com DA, 20 com CCL e 20 controles sadios.

Os critérios para o diagnóstico de CCL são os de Petersen [2004] e, para a DA, os do NINCDS-ADRDA [McKhann et al., 1984]. O diagnóstico é feito por uma equipe multidisciplinar com psiquiatras, geriatras, neurologista, neuropsicólogos e terapeuta ocupacional, por um critério de consenso. O critério de inclusão no grupo de pacientes é que estes devem ser idosos com queixa de memória ou deficit cognitivo, seguidos no ambulatório. Os critérios de exclusão são: doenças clínicas mal controladas, déficits sensitivos que não possam ser compensados e interfiram na execução dos testes, outros diagnósticos neurológicos ou psiquiátricos que possam cursar com demência ou déficit cognitivo, e uso de medicações em doses que afetem a cognição. Os teste a serem aplicados sobre os participantes são:

- Mini-exame do Estado Mental [Folstein et al., 1975].
- CAMCOG [Roth et al., 1986].
- *Rivermead Behavioral Memory Test* [Wilson et al., 2008].
- *Short Cognitive Test* [Lehfeld and Erzigkeit, 1997].
- Figura complexa de Rey.
- *Fuld Object Memory Evaluation* [Fuld et al., 1990].
- EXIT 25.

- *Direct Assessment of Functional Status Revised*.
- Escala de Hamilton para depressão.
- Escala de Young para mania.

Os sujeitos normais serão submetidos ainda, além dos testes acima, aos seguintes:

1. Tarefas de fluência verbal semântica e fonêmica.
2. Entrevista do MAYO OLDER AMERICAN NORMATIVE STUDIES: MOANS, para garantir os critérios para classificação como normal (para indivíduos de 55 anos ou mais):

    (a) Ausência de doença psiquiátrica ou neurológica em atividade.

    (b) Ausência de queixa de dificuldade cognitiva durante a anamnese e interrogatório sobre os diferentes aparelhos, e ausência, ao exame físico, de achado sugestivo de transtornos com potencial para afetar a cognição.

    (c) Ausência de uso de medicação psicotrópica em quantidades que possam comprometer a cognição ou sugerir transtorno neuropsiquiátrico.

    (d) Status de vida independente na comunidade.

    (e) Histórias pregressas de transtornos (p.ex: alcoolismo) com potencial para afetar a cognição não são excluídos automaticamente desde que os transtornos não estejam em atividade e tenha havido recuperação sem sequela cognitiva aparente.

    (f) Doenças médicas crônicas não são critério para exclusão, desde que o médico responsável relate que a condição não comprometa a cognição.

Parte dos textos analisados será composta 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. 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 intervem 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 teste é registrado com câmera de vídeo e com gravador de áudio. O áudio é, então, transcrito, gerando os textos a serem utilizados neste trabalho.

Serão utilizados também textos de uma entrevista autobiográfica com o participante. Os participantes serão convidados a escolher os eventos a partir de cinco períodos de vida: início da infância (primeira infância

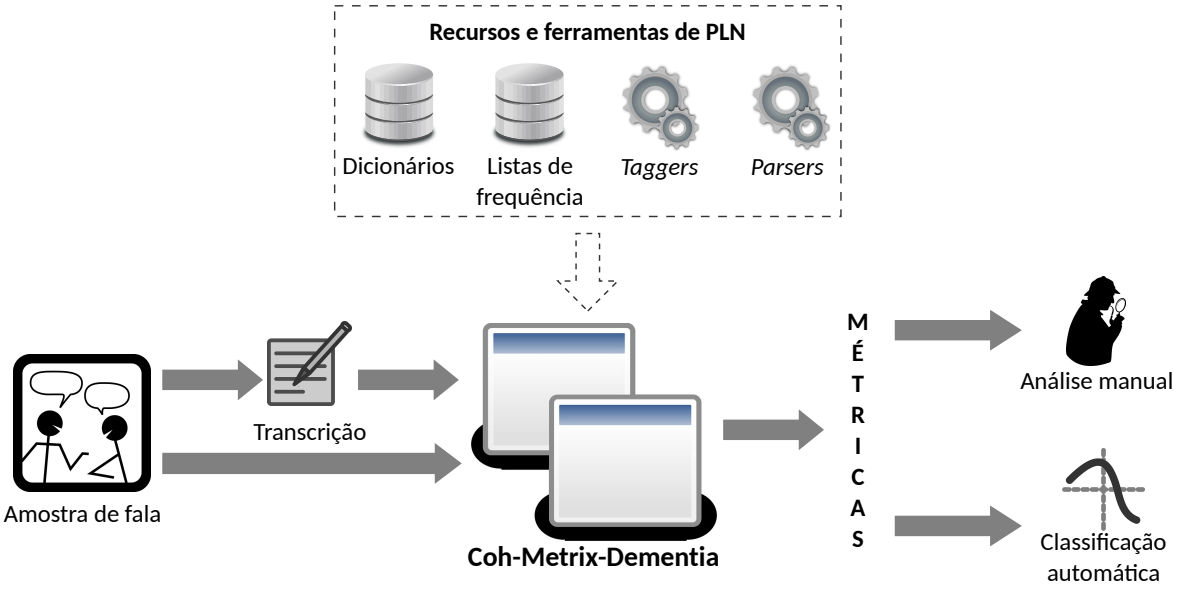


Figura 4.1: Arquitetura geral do Coh-Metrix-Dementia.

aos 11 anos); adolescência-adolescente (idades 11-17), início da idade adulta (idades 18-35) , a meia-idade (35-55), e eventos que ocorreram no ano passado (2013). Todos os participantes selecionarão um evento de cada período de vida.

As instruções a seguir serão dadas: "*Vou pedir para me contar sobre o evento a partir de cada um desses períodos de tempo de sua vida (é dada a lista de períodos de vida) . Você pode escolher eventos que você deseja. Vou pedir para descrever os eventos, então eu vou fazer algumas perguntas sobre eles. O evento deve ser um que você estava pessoalmente envolvido, e você deve ter uma lembrança de estar pessoalmente envolvido. Não escolher eventos sobre os quais você ouviu falar sobre a partir de outros. Eles devem ser os eventos de um tempo e lugares específicos. Eu quero que você saiba que eu vou pedir para você dar alguns detalhes sobre esses eventos mais tarde, por isso escolha apenas os eventos em que você se sente vontade para discutir em detalhe.*"

4.2 Arquitetura do Coh-Metrix-Dementia

A arquitetura geral do Coh-Metrix-Dementia está representada na Figura 4.1. A partir dela, podemos ver que uma amostra de fala é transcrita, e tanto a transcrição quanto o áudio são fornecidos como entrada ao Coh-Metrix-Dementia. Na versão do Coh-Metrix-Dementia que se espera produzir durante este projeto de mestrado, apenas a transcrição será analisada, mas a ferramenta será projetada de maneira a suportar a extensão por meio de ferramentas de análise de fala. O Coh-Metrix-Dementia, então, dará como saída um conjunto de métricas, que podem ser utilizadas tanto para análises manuais quanto para classificação automática.

Inicialmente, ocorrerá a reescrita da versão atual do Coh-Metrix-Port, da linguagem Ruby para a linguagem Python. Essa reescrita acontecerá porque as bibliotecas utilizadas na codificação da versão atual não se mostraram de fácil atualização e integração com outras ferramentas. Essa escrita se dará já com vistas

à inserção das novas métricas, e ocorrerá concomitantemente à uma pesquisa e avaliação de bibliotecas, de preferência na linguagem Python, que ofereçam funções voltadas ao PLN (como o `NLTK`[1], ao AM (como o `scikit-learn`[2]), à visualização de dados (como o `matplotlib`[3]), e à Web (como o `Django`[4]). Com isso, todas as métricas atualmente disponíveis no Coh-Metrix-Port (tabela 2.5) estarão disponíveis no Coh-Metrix-Dementia. É importante notar que alguns recursos lexicais, como listas de frequências, serão atualizadas, pois após 3 anos e meio de uso da ferramenta, novos grandes córpus foram compilados como o Corpus Brasileiro[5], com 1 bilhão de palavras do português do Brasil.

Em seguida, serão adicionadas novas métricas ao sistema. As dezoito métricas que se deseja adicionar nessa primeira versão são:

1. LSA (seção 3.1.4), com a contribuição de 8 novas métricas semelhantes ao Coh-Metrix, pois é de fácil codificação na linguagem Python, a ser adotada no projeto (cf. `http://blog.josephwilk.net/projects/latent-semantic-analysis-in-python.html`).

2. Medidas de diversidade lexical: MLU, Índice de Brunét e Estatística de Honoré (seção 3.1.1).

3. Medidas de complexidade sintática: complexidades de Yngve e Frazier, distância de dependência, nível de desenvolvimento (possivelmente) e entropia cruzada (seção 3.1.2).

4. Medidas de densidade semântica: Análise de Densidade de Ideias e densidade de CSU's (seção 3.1.2).

A criação do Coh-Metrix-Dementia, tal como proposta aqui, apresenta diversas dificuldades. As medidas de complexidade sintática dependem de um analisador sintático pleno; para o português, existe o PALAVRAS [Bick, 2000], utilizado pelo AIC (confira na seção 2.2.2), mas o PALAVRAS não é um software livre e não é personalizável ou treinável (pois utiliza uma abordagem baseada em conhecimento, não em aprendizado de máquina), o que distoa dos propósitos desejados para o Coh-Metrix-Dementia, que deverá ser um software flexível, livre e acessível a todos. Por isso, será feita uma pesquisa sobre outros analisadores disponíveis publicamente para o português, a fim de se evitar utilizar o PALAVRAS.

A fim de inserir a medida de distância de dependência no sistema, é preciso utilizar um analisador sintático que forneça as relações de dependência existentes na sentença. Para tanto, serão analisadas alternativas de analisadores para essa tarefa, com base em *reviews* como os de Buchholz and Marsi [2006] e de Hall and Nilsson [2006], bem como nos dois analisadores avaliados no projeto PorSimples, o MaltParser [Nivre et al., 2006] e o MSTParser McDonald et al. [2006]. Vale notar que o PALAVRAS é um analisador de dependências para o qual um processo de levantamento da árvore sintática é utilizado na sua saída.

Além disso, medidas como a de densidade de ideias e a densidade de CSU's recaem sobre conjuntos de regras, alguns deles bastante extensos (como é o caso de manuais de análise de densidade de ideias, como Chand et al. [2010]) e todos criados para o inglês. A tarefa de adaptá-los para o português não é trivial, pois a

[1] `http://nltk.org/`
[2] `http://scikit-learn.org/stable/`
[3] `http://matplotlib.org/`
[4] `https://www.djangoproject.com/`
[5] `http://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS`

análise minuciosa necessária para adequar cada uma das regras, possivelmente excluindo algumas regras e criando outras, demanda conhecimento linguístico relativamente profundo. O conhecimento de linguistas que trabalham no NILC pode ser utilizado para a avaliação das regras para o português.

A inserção do nível de desenvolvimento no Coh-Metrix-Dementia será estudada, uma vez que já existem softwares para o inglês que a calculam [Lu, 2009]. Essa métrica também recai sobre um conjunto de regras, e pode ser calculada com base nas informações fornecidas por analisadores sintáticos.

Deve-se notar também que, apesar de ser balanceado, o córpus possui um tamanho reduzido (60 textos, 20 textos por classe), e o Coh-Metrix-Dementia produzirá uma grande quantidade de métricas para cada texto, levando a dados esparsos. Por isso, serão investigadas técnicas de redução de dimensionalidade para tentar aliviar o problema e melhorar o desempenho dos classificadores, bem como serão pesquisados algoritmos de classificação que lidem melhor com conjuntos de dados pequenos.

4.3 Avaliação

A avaliação do Coh-Metrix-Dementia se dará com base em dois critérios complementares. O primeiro deles é a **correção** do classificador, medida conforme critérios tradicionais da área de Aprendizado de Máquina (precisão, revocação, medida F). O segundo é relacionado com a qualidade do ambiente Coh-metrix-Dementia como um produto de *software*. Para tanto, será utilizada uma das seis características apresentadas pela norma ISO/IEC-9126-1[6] para avaliar a qualidade de um produto de *software:* a **usabilidade**.

A razão de se enfatizar essa característica - embora o Coh-Metrix-Dementia também será construído visando as outras cinco (funcionalidade, confiabilidade, eficiência, manutenibilidade e portabilidade) - é o fato de a ferramenta ser construída para ser utilizada no futuro por profissionais da área médica que trabalham com textos transcritos do português. As três subcaracterísticas da usabilidade serão consideradas: (i) a Inteligibilidade, que evidencia o esforço do usuário para reconhecer o conceito lógico e sua aplicabilidade; (ii) a Facilidade de aprender e de usar, que evidência o esforço do usuário para aprender sua aplicação, ligada a facilidade de aprendizado do *software*; e (iii) Operacionalidade, que evidencia o esforço do usuário para sua operação e controle.

4.4 Experimentos piloto

Nesta seção, encontram-se descritos dois experimentos piloto. O primeiro deles trata textos de sujeitos sadios em uma tarefa de descrição de figuras, e mostra a utilidade de se utilizar ferramentas de PLN na automatização da análise de textos com objetivos clínicos.

4.4.1 Experimento 1

Grande parte das pesquisas com lesados cerebrais compara o desempenho destes com o de indivíduos sadios [Smith and Ivnik, 2003]. Apesar disso, pouca ênfase tem sido dada à produção discursiva de indivíduos

[6] ISO/IEC 9126-1 (2001) Software engineering — Product quality — Part 1: Quality model. Disponível em `http://www.iso.org/iso/`

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 et al., 2012; Wright et al., 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 and 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 and 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-Metrix-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-Metrix-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 analise 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: Naïve 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 CFS (*Correlation-based Feature Selection* [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-Metrix-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 4.1. Com base nela, pode-se perceber que não foi possível obter bom desempenho, e que os algoritmos de seleção implicaram ganho pequeno.

Tabela 4.1: Medida F (%) dos métodos para o primeiro conjunto de experimentos (3-8 versus 9+).

Algoritmo	**Todos**	**IG ∩ SVM**	**Manual**	**CFS**
Naïve Bayes	66,3	**71,8**	68,0	68,2
SVM	41,8	41,8	41,6	64,1
MLP	64,4	66,0	69,7	62,8
SimpleLogistic	**69,8**	69,4	**71,3**	**71,2**
JRip	65,7	69,7	67,2	65,4
J48	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 4.2, 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).

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 4.3. Primeiramente, removemos a classe 9-15, ficando com 3-8 e 15+, e não obtendo bom desempenho;

Tabela 4.2: Medida F (%) dos métodos com classes extremas e intermediárias.

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

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 4.3: Medida F (%) dos métodos de classificação com remoção de classes intermediárias.

Algoritmo	3-8 x 15+	3-4 x 9+	3-4r x 9+
Naïve Bayes	70,9	74,2	75,7
SVM	43,9	68,0	**97,7**
MLP	69,0	81,7	93,5
SimpleLogistic	73,4	86,7	86,3
JRip	**73,6**	**87,9**	88,6
J48	66,2	79,3	93,2
Baseline	***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 4.4, e mostram que, pelo menos para o melhor algoritmo (SVM), os métodos de seleção não alteraram o desempenho do classificador.

Com isso, concluimos 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.

Tabela 4.4: Medida F (%) dos métodos de classificação e seleção de atributos com remoção de classes intermediárias.

Algoritmo	**Todos**	**IG ∩ SVM**	**Manual**	**CFS**
Naïve Bayes	75,7	81,3	74,2	80,4
SVM	**97,7**	**97,7**	**97,7**	**97,7**
MLP	93,5	93,2	91,6	93,5
SimpleLogistic	86,3	85,2	81,8	84,8
JRip	88,6	86,3	87,8	88,3
J48	93,2	91,2	90,5	92,8
Baseline	***51,1***	***51,1***	***51,1***	***51,1***

Tabela 4.5: Estatísticas básicas das entrevistas.

Indivíduo	CAP	LMVG	RCLO	OB	EFA	JAM
Estágio da DA	moderada	moderada	moderada	leve	leve	leve
N. de caracteres	16.630	11.548	14.617	26.401	24.920	25.911
N. médio de caracteres por pal.	3.82	3.79	3.61	4.15	3.89	3.75
N. de palavras	4.351	3.040	4.047	6.360	6.391	6.893

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 inicial, realizado para nos familiarizarmos com a área e trocarmos experiências com a coorientadora deste trabalho, a Prof.ª Letícia Mansur, da USP. 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, que se constituirão referência para o estudo de lesados cerebrais no futuro desenvolvimento deste projeto.

4.4.2 Experimento 2: desempenho das ferramentas de PLN no corpus DA-PLN-EVAL

O córpus DA-PLN-EVAL é composto das entrevistas transcritas que foram utilizadas na tese de Mansur [1996]. Como comentado na seção 2.1.4, o córpus é composto das entrevistas dos seis pacientes com DA, 3 com DA leve e 3 com DA moderada, transcrito seguindo o padrão de anotação do Projeto NURC (veja detalhes na Seção 2.1.4). A tabela 4.5 apresenta estatísticas de cada uma das seis entrevistas, com anotação da transcrição, mantendo as 2 vozes, mas sem inclusão das siglas indicativas dos falantes.

Para avaliar o impacto das ferramentas de etiquetação morfossintática (*tagger*) e da sintática (*parser*), que são utilizadas em várias métricas de análise automatizada de alterações de linguagem em doenças neu-

rológicas degenerativas, escolhemos um trecho das entrevistas, com um paciente com DA leve (o sujeito EFA), do córpus DA-PLN-EVAL. O trecho relativo à DA leve foi classificado com pertencente à análise do problema de repetição de informações.

Foi apresentado, na seção 4.4.2.1, para o trecho de entrevista escolhido, o impacto de 2 ferramentas: o *tagger* MXPOST treinado no córpus MAC-MORPHO, um córpus de 1.2 milhões de palavras de notícias da Folha de São Paulo [Aluísio et al., 2003] e que é utilizado na ferramenta Coh-Metrix-Port (apresentada no capítulo 2); e do *parser* Palavras Bick [2000], utilizado na ferramenta AIC (também apresentada no Capítulo 2). A análise aqui apresentada é ilustrativa apenas, pois na pesquisa a ser desenvolvida neste mestrado um novo córpus de pesquisa será compilado (detalhes no Capítulo 4) e suas características devem ser analisadas. Mostramos três versões dos trechos de entrevistas escolhidos:

1. a primeira versão é a original, com turnos de entrevistador e entrevistado, transcrita seguindo o padrão de anotação do Projeto NURC;

2. na segunda, removemos as falas do entrevistador e as siglas indicativas dos falantes; e

3. na terceira os trechos foram editados com base no pós-processamento sugerido em Fraser et al. [2012].

Fraser et al. [2012] realiza 5 modificações para utilizar *taggers* e *parsers* não adaptados á fala de afásicos. As modificações são: (i) as marcas de pausas preenchidas transcritas foram eliminadas da análise automática, mas foram computadas; (ii) tokens que não são considerados palavras da língua foram removidos; (iii) falsos começos e repetições foram eliminados da análise automática, mas foram computados; (iv) houve inclusão de pontuações e capitalizações, que foram incluídas com ajuda de *features* semânticas, sintáticas e prosódicas; (v) neologismos e fala incompreensível não foram incluídos na análise automática nem computados, pois não havia a certeza de quantas palavras a fala era composta.

As modificações realizadas no córpus DA-PLN-EVAL, dando origem ao córpus DA-PLN-EVAL-pós_processado foram:

1. Foram mantidas somente as falas do paciente.

2. Foram retiradas as marcas de prolongamento de vogais (dois pontos). Por exemplo, “agora” foi gerado de “a::gora”.

3. Foram substituídas as marcas de pausa “...” por ponto final, quando se encontrava no final de uma elocução;

4. Foram mantidas as marcas “...” quando a pausa estava diante de substantivos ou entre pronome e verbo; em resumo, em lugares “inadequados” e que poderiam sinalizar “problemas” ou no fim de frases “abandonadas”, incompletas.

5. Foi usada a vírgula quando o “timing” da micropausa era admitido como “normal” (menos de 3 segundos).

Tabela 4.6: Trecho da transcrição original do sujeito com DA leve - EFA, simplificado por E.

Linha	Pessoa	Turno de fala
1	O-	antigamente eu tomava muito café... porque eu trabalhava eu trabalhava em vendas... em
2	E-	{hum}
3	O-	vendas pra iniciar a venda e ser bem sucedido você precisa convidar o o (companheiro) prum café... né?... (porque bom)... deixa... a:... as infelicidades pra trás... e vamos comprar... e do café sai aquela
4	E-	((riso de E))
5	O-	(conversa) gostosa e tal... cê tá vendendo tal...eu fui:... gerente da
8		{hum hum} {hum hum}
6	O-	Kelsons... gerente de vendas... aqui em São Paulo porque a fábrica é no Rio...
7	E-	{hum}
8	O-	eu era gerente de São Paulo... cheguei a ter até sessenta pessoas...
9	E-	bastante né?
10	O-	{para o Brasil... (venda) para o Brasil...

6. Foram harmonizadas as emissões, retirando marcas culturais por exemplo, “genti”, “né” e substituindo por “gente” e “não é”.

7. Foram utilizadas letras maiúsculas em inícios de orações.

8. Foram mantidas as palavras truncadas, repetição de fonemas e sílabas, repetição de segmentos, etc.

Da mesma forma que o realizado em Fraser et al. [2012], pretendemos contabilizar as características das falas de pacientes com demência, antes de pós-processá-las para serem usadas pelas ferramentas de PLN, no novo córpus de trabalho.

4.4.2.1 Análise computacional do trecho transcrito do paciente com DA leve - o sujeito EFA

A tabela 4.6 mostra em 10 linhas a transcrição original de um trecho da entrevista do paciente EFA. A tabela 4.7 mostra, do lado esquerdo, o trecho original sem a fala do entrevistador e, do lado direito, o mesmo trecho da esquerda, com o pós-processamento descrito acima. Ambos os trechos estão tokenizados, para que o *tagger* MXPOST possa realizar as anotações. A tabela de etiquetas principais do *tagger* é apresentada na tabela 4.8.

A saída da etiquetação morfossintática com o *tagger* MXPOST do texto do lado esquerda da tabela 4.7 é mostrada abaixo, separada em 5 sentenças, via marcas de pausas no final de uma elocução. Foram sublinhados pontos de destaque do desempenho, discutidos mais abaixo.

Tabela 4.7: Segunda (esquerdo) e terceira (direito) versões do trecho apresentado na tabela 4.6.

Trecho sem as falas do entrevistador	Trecho pós-processado
antigamente eu tomava muito café ... porque eu trabalhava eu trabalhava em vendas ... em vendas pra iniciar a venda e ser bem sucedido você precisa convidar o o (companheiro) prum café ... né ?... (porque bom)... deixa ... a :... as infelicidades pra trás ... e vamos comprar ... e do café sai aquela (conversa) gostosa e tal ... cê tá vendendo tal ... eu fui :... gerente da Kelsons ... gerente de vendas ... aqui em São Paulo porque a fábrica é no Rio ... eu era gerente de São Paulo ... cheguei a ter até sessenta pessoas ... para o Brasil ... (venda) para o Brasil ...	Antigamente eu tomava muito café , porque eu trabalhava eu trabalhava em vendas . Em vendas , para iniciar a venda e ser bem sucedido , você precisa convidar o o companheiro para um café , porque bom , deixa a as infelicidades para trás e vamos comprar . E do café sai aquela conversa gostosa e tal , você está vendendo tal ... Eu fui gerente da Kelsons , gerente de vendas , aqui em São Paulo ; porque a fábrica é no Rio , eu era gerente de São Paulo . Cheguei a ter até sessenta pessoas , para o Brasil , venda para o Brasil .

Tabela 4.8: Conjunto de Etiquetas do *tagger* MXPOST treinado com o Mac-Morpho.

Classe gramatical	Etiqueta
ADJETIVO	ADJ
ADVÉRBIO	ADV
ADVÉRBIO CONECTIVO SUBORDINATIVO	ADV-KS
ADVÉRBIO RELATIVO SUBORDINATIVO	ADV-KS-REL
ARTIGO (def. ou indef.)	ART
CONJUNÇÃO COORDENATIVA	KC
CONJUNÇÃO SUBORDINATIVA	KS
INTERJEIÇÃO	IN
NOME	N
NOME PRÓPRIO	NPROP
NUMERAL	NUM
PARTICÍPIO	PCP
PALAVRA DENOTATIVA	PDEN
PREPOSIÇÃO	PREP
PRONOME ADJETIVO	PROADJ
PRONOME CONECTIVO SUBORDINATIVO	PRO-KS
PRONOME PESSOAL	PROPESS
PRONOME RELATIVO CONECTIVO SUBORDINATIVO	PRO-KS-REL
PRONOME SUBSTANTIVO	PROSUB
VERBO	V
VERBO AUXILIAR	VAUX
SÍMBOLO DE MOEDA CORRENTE	CUR
CONTRAÇÕES e ÊNCLISES	|+
MESÓCLISES	|!

antigamente_ADV eu_PROPESS tomava_V muito_ADV café_ADJ ..._... porque_KS eu_PROPESS trabalhava_V eu_PROPESS trabalhava_V em_PREP vendas_N ..._...

em_PREP vendas_N pra_PREP iniciar_V a_ART venda_N e_KC ser_V bem_ADV sucedido_PCP você_PROPESS precisa_V convidar_V o_PRO-KS o_PRO-KS (_(companheiro_N)_) prum_Vcafé_N ..._... né_PROPESS?_? ..._... (_(porque_KS bom_ADJ)_) ..._... deixa_V ..._... a_ART :_: ..._... as_ART infelicidades_N pra_PREP trás_N ..._... e_KC vamos_VAUX comprar_V ..._...

e_KC do_KS café_KS sai_V aquela_PROADJ (_(conversa_N)_) gostosa_ADJ e_KC tal_PROSUB ..._... cê_ADVtá_VAUX vendendo_V tal_PROSUB ..._...

eu_PROPESS fui_V :_: ..._... gerente_N da_N Kelsons_NPROP ..._... gerente_N de_PREP vendas_N ..._... aqui_ADV em_PREP São_NPROP Paulo_NPROP porque_KS a_ART fábrica_N é_NPROP no_NPROP Rio_NPROP ..._... eu_PROPESS era_V gerente_N de_PREP São_NPROP Paulo_NPROP ..._...

cheguei_Va_PREP ter_Vaté_Nsessenta_NUM pessoas_N ..._... para_PREP o_ART Brasil_NPROP ..._... (_(venda_N)_) para_PREP o_ART Brasil_NPROP ..._...

A saída da etiquetação morfossintática com o *tagger* MXPOST do texto do lado direito da tabela 4.7 é mostrada abaixo, separada por sentenças terminadas por ponto final, totalizando 5 sentenças.

Antigamente_ADV eu_PROPESS tomava_V muito_ADV café_ADJ ,_, porque_KS eu_PROPESS trabalhava_V eu_PROPESS trabalhava_V em_PREP vendas_N ._.

Em_PREP vendas_N ,_, para_PREP iniciar_V a_ART venda_N e_KC ser_V bem_ADV sucedido_PCP ,_, você_PROPESS precisa_V convidar_V o_PRO-KS o_PRO-KS companheiro_N para_PREP um_ART café_N ,_, porque_KS bom_ADJ ,_, deixa_V a_PREP|+ as_ARTinfelicidades_N para_PREP trás_N e_KC vamos_VAUX comprar_V ._.

E_NPROP do_NPROP café_NPROP sai_V aquela_PROADJ conversa_N gostosa_ADJ e_KC tal_PROSUB ,_, você_PROPESS está_VAUX vendendo_V tal_PROSUB ..._...

Eu_PROPESS fui_V gerente_ADJda_N Kelsons_NPROP ,_, gerente_N de_PREP vendas_N ,_, aqui_ADV em_PREP São_NPROP Paulo_NPROP ;_; porque_KS a_ART fábrica_N é_NPROP no_NPROP Rio_NPROP ,_, eu_PROPESS era_V gerente_N de_PREP São_NPROP Paulo_NPROP ._.

Cheguei_VAUX a_PREP ter_VAUX até_PCPsessenta_NUM pessoas_N ,_, para_PREP o_ART Brasil_NPROP ,_, venda_N para_PREP o_ART Brasil_NPROP ._.

Uma análise comparativa dos dois trechos acima, visando analisar os acertos que o pós-processamento aplicado na versão 3 ocasiona, sem analisar erros possíveis do *tagger* (dado que ele apresenta 96,98% de precisão em textos do mesmo gênero para o qual foi treinado) mostra que:

1. a primeira sentença não sofreu diferenças na etiquetação; a marca de pausa para separar a oração subordinada não interferiu na anotação do *tagger*;

2. a segunda sentença se beneficiou do pós-processamento na troca dos coloquialismos "prum" e "né", que foram etiquetados corretamente;

3. a terceira sentença se beneficiou do pós-processamento na troca do coloquialismo "cê" que foi etiquetado corretamente, porém ao ficarem próximos as palavras "a as", o *tagger* etiquetou-as incorretamente, no trecho processado;

4. na quarta sentença, o fato de haver um erro de etiquetação na palavra "gerente" e acerto na sentença não processa parece ser erro comum do *tagger*; e

5. a capitalização do início da sentença parece beneficiar a correta etiquetação. Em ambos os trechos a palavra "até" é etiquetada erroneamente.

Mais testes precisam ser realizados em termos de pósprocessamento da transcrição para que o *tagger* faça sua tarefa com maior precisão. Esta é uma das metas do mestrado. Quanto a erros comuns e previsíveis da saída do *tagger*, o Coh-Metrix-Port, que dará a base ao Coh-Metrix-Dementia, já possui um pósprocessamento para revisar os erros frequentes.

Quanto à saída do *parser*, mostramos abaixo um trecho da saída para o texto à esquerda da tabela 4.7. É possível ver que a cada pausa (...) o *parser* reconhece uma sentença (anota como </s>) e assim prejudica qualquer análise de complexidade sintática que se faça com esta saída.

```
antigamente  [antigamente] ADV @ADVL>
eu  [eu] PERS M/F 1S NOM @SUBJ>
tomava  [tomar] <fmc> <mv> V IMPF 1S IND VFIN @FS-QUE
muito  [muito] <quant> DET M S @>N
café  [café] <drink> N M S @<ACC
$...
</s>
porque  [porque] <clb> KS @SUB
eu  [eu] PERS M 1S NOM @SUBJ>
trabalhava  [trabalhar] <mv> V IMPF 1S IND VFIN @FS-<ADVL
em  [em] PRP @<PIV
vendas  [venda] <act> N F P @P<
$...
</s>
em  [em] PRP @<ADVL
vendas  [venda] <act> N F P @P<
pra  [pra] PRP @<ADVL
iniciar  [iniciar] <mv> V INF @ICL-P<
```

```
a  [o] <clb> <artd> DET F S @>N
venda  [venda] <act> N F S @SUBJ>
e  [e] <co-fmc> <co-inf> KC @CO
ser  [ser] <mv> V INF 0/1/3S @ICL-P<
bem  [bem] <quant> ADV @>A
sucedido  [suceder] V PCP M S @<SC
você  [você] PERS M/F 3S NOM @SUBJ>
precisa  [precisar] <fmc> <mv> V PR 3S IND VFIN @FS-QUE
convidar  [convidar] <mv> V INF @ICL-<ACC
o  [o] <artd> DET M S @>N
$(
companheiro  [companheiro] <Hfam> N M S @<PRED
$)
prum  [prum] N F S @<ACC
café  [café] <drink> N M S @<ACC
$...
</s>
né  [né] ADV @<ADVL
$?
$...
$(
porque  [porque] <clb> KS @SUB
bom  [bom] ADJ M S @SUBJ>
$)
$...
</s>
```

Para o texto á direita na tabela 4.7, há o correto processamento das sentenças, permitindo assim cálculo corretos da complexidade sintática. Mostramos abaixo a saída completa para o texto.

```
Antigamente  [antigamente] ADV @ADVL>
eu  [eu] PERS M/F 1S NOM @SUBJ>
tomava  [tomar] <fmc> <mv> V IMPF 1S IND VFIN @FS-STA
muito  [muito] <quant> DET M S @>N
café  [café] <drink> N M S @<ACC
$,
porque  [porque] <clb> KS @SUB
eu  [eu] PERS M/F 1S NOM @SUBJ>
```

```
trabalhava  [trabalhar] <mv> V IMPF 1S IND VFIN @FS-<ADVL
em  [em] PRP @<PIV
vendas  [venda] <act> N F P @P<
$.
</s>
Em  [em] PRP @ADVL>
vendas  [venda] <act> N F P @P<
$,
para  [para] PRP @ADVL>
iniciar  [iniciar] <mv> V INF @ICL-P<
a  [o] <artd> DET F S @>N
venda  [venda] <act> N F S @<ACC
e  [e] <co-inf> KC @CO
ser  [ser] <mv> V INF @ICL-P<
bem  [bem] <quant> ADV @>A
sucedido  [suceder] V PCP M S @<SC
$,
você  [você] PERS M/F 3S NOM @SUBJ>
precisa  [precisar] <fmc> <mv> V PR 3S IND VFIN @FS-STA
convidar  [convidar] <mv> V INF @ICL-<ACC
o  [o] <artd> DET M S @>N
companheiro  [companheiro] <Hfam> N M S @<ACC
para  [para] PRP @<ADVL
um  [um] <arti> DET M S @>N
café  [café] <build> N M S @P<
$,
porque  [porque] <clb> KS @SUB
bom  [bom] ADJ M S @PRED>
$,
deixa  [deixar] <mv> V PR 3S IND VFIN @FS-<ADVL
a  [a] PRP @<PIV
as  [o] <artd> DET F P @>N
infelicidades  [infelicidade] <ac> N F P @P<
para  [para] PRP @<ADVL
trás  [trás] ADV @<ADVL
e  [e] <co-fin> <co-fmc> <co-fin> KC @CO
vamos  [ir] <fmc> <aux> V PR 1P IND VFIN @FS-STA
comprar  [comprar] <mv> V INF @ICL-AUX<
$.
```

```
</s>
E  [e] KC @CO
de  [de] <sam-> PRP @ADVL>
o  [o] <artd> <-sam> DET M S @>N
café  [café] <build> N M S @P<
sai  [sair] <fmc> <mv> V PR 3S IND VFIN @FS-STA
aquela  [aquele] <dem> DET F S @>N
conversa  [conversa] <talk> N F S @<SUBJ
gostosa  [gostoso] <np-close> ADJ F S @N<
e=tal  [e=tal] ADV @<ADVL
$,  [$,] <co-fmc> <co-fin> PU @CO
você  [você] PERS M/F 3S NOM @SUBJ>
está  [estar] <fmc> <aux> V PR 3S IND VFIN @FS-STA
vendendo  [vender] <clb> <mv> V GER @ICL-AUX<
tal  [tal] <KOMP> <diff> DET M/F S @<ACC
$...
</s>
Eu  [eu] PERS M/F 1S NOM @SUBJ>
fui  [ser] <fmc> <mv> V PS 1S IND VFIN @FS-STA
gerente  [gerente] <Hprof> <NER:official> N M/F S @<SC
de  [de] <sam-> <NER2> <np-close> PRP @N<
a  [o] <NER2> <artd> <-sam> DET F S @>N
Kelsons  [Kelsons] <org> PROP F S @P<
$,
gerente  [gerente] <Hprof> <np-long> N M/F S @N<PRED
de  [de] <np-close> PRP @N<
vendas  [venda] <act> N F P @P<
$,
aqui  [aqui] ADV @<ADVL
em  [em] PRP @A<
São=Paulo  [São=Paulo] <civ> PROP M S @P<
$;
</s>
porque  [porque] <clb> KS @SUB
a  [o] <artd> DET F S @>N
fábrica  [fábrica] <build> N F S @SUBJ>
é  [ser] <mv> V PR 3S IND VFIN @FS-ADVL>
em  [em] <sam-> PRP @<SC
o  [o] <artd> <-sam> DET M S @>N
```

```
Rio  [Rio] <civ> PROP M S @P<
$,
eu  [eu] PERS M/F 1S NOM @SUBJ>
era  [ser] <fmc> <mv> V IMPF 1S IND VFIN @FS-STA
gerente  [gerente] <Hprof> <NER:official> N M/F S @<SC
de  [de] <NER2> <np-close> PRP @N<
São=Paulo  [São=Paulo] <civ> PROP M S @P<
$.
</s>
Cheguei  [chegar] <fmc> <aux> V PS 1S IND VFIN @FS-STA
a  [a] PRP @PRT-AUX<
ter  [ter] <mv> V INF @ICL-AUX<
até  [até] PRP @<ADVL
sessenta  [sessenta] <card> NUM F P @>N
pessoas  [pessoa] <H> N F P @P<
$,
para  [para] PRP @<ADVL
o  [o] <artd> DET M S @>N
Brasil  [Brasil] <civ> PROP M S @P<
$,
venda  [venda] <act> N F S @<ACC
para  [para] PRP @<ADVL
o  [o] <artd> DET M S @>N
Brasil  [Brasil] <civ> PROP M S @P<
$.
</s>
```

5. Plano de Trabalho e Cronograma

Com base no que foi exposto na seção anterior, o desenvolvimento da pesquisa de mestrado será dividido em 15 etapas, apresentadas abaixo:

1. Revisão bibliográfica de trabalhos relacionados ao tema do mestrado.
2. Levantamento das características potenciais a serem extraídas das transcrições dos sujeitos.
3. Levantamento dos requisitos de anotação e pré-processamento das transcrições.
4. Levantamento e seleção de ferramentas e recursos de PLN para o português do Brasil que possam ser disponibilizados gratuitamente para a comunidade.
5. Reescrita da versão atual do Coh-Metrix-Port. Pesquisa e avaliação de bibliotecas.
6. Codificação das novas métricas e do ambiente de trabalho a ser usado pelos pesquisadores médicos.
7. Acompanhamento da coleta dos dados, junto à Prof.ª Letícia.
8. Processamento dos dados textos coletados; treinamento e teste dos classificadores e eventuais ajustes às métricas e às formas de anotação e pré-processamento.
9. Avaliação do ambiente de trabalho por parte de uma equipe de pesquisadores, designados pela Prof.ª Letícia.
10. Cumprimento de créditos de disciplinas, exigidos pelo ICMC/USP.
11. Realização do exame de proficiência em inglês, exigido pelo ICMC/USP.
12. Entrega da qualificação de mestrado.
13. Escrita da dissertação de mestrado.
14. Escrita de relatórios e artigos científicos.
15. Defesa da dissertação de mestrado.

A tabela 5.1 mostra o cronograma previsto para o desenvolvimento do projeto.

Tabela 5.1: Cronograma previsto para o projeto.

At.	2013										2014												2015	
	mar	abr	mai	jun	jul	ago	set	out	nov	dez	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez	jan	fev
1																								
2																								
3																								
4																								
5																								
6																								
7																								
8																								
9																								
10																								
11																								
12																								
13																								
14																								
15																								

6. Considerações finais

Este projeto de mestrado está sendo desenvolvido no âmbito de uma cooperação binacional entre a Universidade de São Paulo (USP) e a Universidade de Toronto (UoT), com um projeto intitulado *Análise de distúrbios de linguagem nas demências: perspectiva translinguística*, sob a coordenação da Prof.ª Letícia Mansur (pela USP) e de Jed A. Meltzer (pela UoT), com colaboração da Prof.ª Sandra Aluísio e do Prof. Graeme Hirst, dentre outros. As atividades aqui descritas serão realizadas, em sua maior parte, no Brasil, sendo que existe a possibilidade de o aluno em questão realizar um estágio na UoT, sob supervisão do Prof. Graeme Hirst e de Jed A. Meltzer. Com a bolsa FAPESP aprovada em janeiro de 2014, o aluno poderá solicitar uma bolsa BEPE para a realização deste estágio. Além da possível ida do aluno para a UoT, o NILC receberá uma aluna de doutorado, Katy C. Fraser, em abril de 2014, para trabalho conjunto. Com certeza, a visita da aluna da UoT trará benefícios ao trabalho aqui descrito.

Referências Bibliográficas

Ahmed, S., de Jager, C. A., Haigh, A.-M., and Garrard, P. (2013). Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. *Neuropsychology*, 27(1):79–85.

Albert, M. L., Spiro, A., Sayers, K. J., Cohen, J. A., Brady, C. B., Goral, M., and Obler, L. K. (2009). Effects of Health Status on Word Finding in Aging. *Journal of the American Geriatrics Society*, 57(12):2300–2305.

Almor, A., Kempler, D., MacDonald, M. C., Andersen, E. S., and Tyler, L. K. (1999). Why Do Alzheimer Patients Have Difficulty with Pronouns? Working Memory, Semantics, and Reference in Comprehension and Production in Alzheimer's Disease. *Brain and Language*, 67(3):202–227.

Altmann, L. J. and McClung, J. S. (2008). Effects of semantic impairment on language use in Alzheimer's disease. *Semin Speech Lang*, 29(1):18–31.

Aluísio, S., Pelizzoni, J., Marchi, A., Oliveira, L., Manenti, R., and Marquiafável, V. (2003). An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese. In Mamede, N., Trancoso, I., Baptista, J., and Graças Volpe Nunes, M., editors, *Computational Processing of the Portuguese Language*, volume 2721 of *Lecture Notes in Computer Science*, pages 110–117. Springer Berlin Heidelberg.

Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, IUNLPBEA '10, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aluísio, S. M. and Gasperin, C. (2010). 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*, YIWCALA '10, pages 46–53, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aluísio, S. M., Specia, L., Pardo, T. A. S., Maziero, E. G., Caseli, H. M., and Fortes, R. P. M. (2008). 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*, SIGDOC '08, pages 15–22, New York, NY, USA. ACM.

Amaral-Carvalho, V. and Caramelli, P. (2012). Normative Data for Healthy Middle-Aged and Elderly Performance on the Addenbrooke Cognitive Examination-Revised. *Cognitive and Behavioral Neurology*, 25(2):72–76.

Anderson, S., White-Schwoch, T., Parbery-Clark, A., and Kraus, N. (2013). A dynamic auditory-cognitive system supports speech-in-noise perception in older adults. *Hearing Research*, 300(0):18–32.

Andrade, C. R. F. and Martins, V. O. (2010). Variação da fluência da fala em idosos. *Pro-Fono Revista de Atualização Científica*, 22(1):13–18.

Andreetta, S., Cantagallo, A., and Marini, A. (2012). Narrative discourse in anomic aphasia. *Neuropsychologia*, 50(8):1787–1793.

Ardila, A., Bertolucci, P. H., Braga, L. W., Castro-Caldas, A., Judd, T., Kosmidis, M. H., Matute, E., Nitrini, R., Ostrosky-Solis, F., and Rosselli, M. (2010). Illiteracy: The Neuropsychology of Cognition Without Reading. *Archives of Clinical Neuropsychology*, 25(8):689–712.

Aronoff, J. M., Gonnerman, L. M., Almor, A., Arunachalam, S., Kempler, D., and Andersen, E. S. (2006). Information content versus relational knowledge: Semantic deficits in patients with Alzheimer's disease. *Neuropsychologia*, 44(1):21–35.

Ash, S., Moore, P., Antani, S., McCawley, G., Work, M., and Grossman, M. (2006). Trying to tell a tale: Discourse impairments in progressive aphasia and frontotemporal dementia. *Neurology*, 66(9):1405–1413.

Balage Filho, P., Pardo, T., and Aluísio, S. (2013). An Evaluation of the Brazilian Portuguese LIWC Dictionary. *To be published in the Proceedings of STIL 2013*, page 5.

Bayles, K. A., Tomoeda, C. K., and Trosset, M. W. (1992). Relation of linguistic communication abilities of Alzheimer's patients to stage of disease. *Brain and Language*, 42(4):454–472.

Berndt, R. S., Wayland, S., Rochon, E., Saffran, E., and Schwartz, M. (2000). *Quantitative Production Analysis: A Training Manual for the Analysis of Aphasic Sentence Production*. Psychology Press, Hove, UK.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Bick, E. (2000). *The Parsing System"Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Bickel, C., Pantel, J., Eysenbach, K., and Schröder, J. (2000). Syntactic Comprehension Deficits in Alzheimer's Disease. *Brain and Language*, 71(3):432–448.

Boeuf, C. (1971). *Raconte...: 55 historiettes en images*. Ecole.

Brown, C., Snodgrass, T., Kemper, S., Herman, R., and Covington, M. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2):540–545.

Brown, R. (1973). *A first language*. Harvard University Press, Cambridge, MA.

Brucki, S. and Rocha, M. (2004). Category fluency test: effects of age, gender and education on total scores, clustering an switching in Brazilian Portuguese-speaking subjects. *Brazilian Journal of Medical and Biological Research*, 37(12):1771–1777.

Brucki, S. M. D., Malheiros, S. M. F., Okamoto, I. H., and Bertolucci, P. H. F. (1997). Dados normativos para o teste de fluência verbal categoria animais em nosso meio. *Arq Neuropsiquiatr*, 55(1):56–61.

Brunet, É. (1978). *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*. Travaux de linguistique quantitative. Slatkine.

Bryant, L., Spencer, E., Ferguson, A., Craig, H., Colyvas, K., and Worrall, L. (2013). Propositional Idea Density in aphasic discourse. *Aphasiology*, 27(8):992–1009.

Bschor, T., Kühl, K.-P., and Reischies, F. M. (2001). Spontaneous Speech of Patients With Dementia of the Alzheimer Type and Mild Cognitive Impairment. *International Psychogeriatrics*, 13:289–298.

Buchholz, S. and Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 149–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.

Cannizzaro, M. S. and Coelho, C. A. (2012). Analysis of Narrative Discourse Structure as an Ecologically Relevant Measure of Executive Function in Adults. *Journal of Psycholinguistic Research*, pages 1–23.

Caramelli, P., Mansur, L. L., and Nitrini, R. (1998). Language and Communication Disorders in Dementia of the Alzheimer Type. In Stemmer, B. and Whitaker, H., editors, *Handbook of Neurolinguistics*. Academic Press.

Carlomagno, S., Santoro, A., Menditti, A., Pandolfi, M., and Marini, A. (2005). Referential Communication in Alzheimer's Type Dementia. *Cortex*, 41(4):520–534.

Cera, M. L., Ortiz, K. Z., Bertolucci, P. H., and Minett, T. S. (2013). Speech and orofacial apraxias in Alzheimer's disease. *Int Psychogeriatr*, 25(10):1679–1685.

Cerhan, J. H., Ivnik, R. J., Smith, G. E., Tangalos, E. C., Petersen, R. C., and Boeve, B. F. (2002). Diagnostic Utility of Letter Fluency, Category Fluency, and Fluency Difference Scores in Alzheimer's Disease. *The Clinical Neuropsychologist*, 16(1):35–42. PMID: 11992224.

Chand, V., Baynes, K., Bonnici, L., and Farias, S. T. (2010). *Analysis of Idea Density (AID): A Manual*. University of California at Davis.

Cheung, H. and Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, 13:53–76.

Cook, C., Fay, S., and Rockwood, K. (2009). Verbal Repetition in People With Mild-to-Moderate Alzheimer Disease: A Descriptive Analysis From the VISTA Clinical Trial. *Alzheimer Disease & Associated Disorders*, 23(2).

Cooper, P. V. (1990). Discourse Production and Normal Aging: Performance on Oral Picture Description Tasks. *Journal of Gerontology*, 45(5):210–214.

Creamer, S. and Schmitter-Edgecombe, M. (2010). Narrative comprehension in Alzheimer's disease: assessing inferences and memory operations with a think-aloud procedure. *Neuropsychology*, 24(3):279–290.

Crook, T., Bartus, R. T., Ferris, S. H., Whitehouse, P., Cohen, G. D., and Gershon, S. (1986). Age-associated memory impairment: Proposed diagnostic criteria and measures of clinical change — report of a national institute of mental health work group. *Developmental Neuropsychology*, 2(4):261–276.

Cuetos, F., Rodríguez-Ferreiro, J., and Menéndez, M. (2009). Semantic markers in the diagnosis of neurodegenerative dementias. *Dementia and Geriatric Cognitive Disorders*, 28(3):267–274.

de Carvalho, I. A. and Mansur, L. L. (2008). Validation of ASHA FACS-functional assessment of communication skills for Alzheimer disease population. *Alzheimer disease and associated disorders*, 22(4):375–381. Validation Studies,Research Support, Non-U.S. Gov't,.

de Lira, J. O., Ortiz, K. Z., Campanha, A. C., Bertolucci, P. H. F., and Minett, T. S. C. (2011). Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics*, 23:404–412.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407.

Druks, J., Masterson, J., Kopelman, M., Clare, L., Rose, A., and Rai, G. (2006). Is action naming better preserved (than object naming) in Alzheimer's disease and why should we ask? *Brain and Language*, 98(3):332–340.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using Latent Semantic Analysis to Improve Access to Textual Information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '88, pages 281–285, New York, NY, USA. ACM.

Dunn, J. C., Almeida, O. P., Barclay, L., Waterreus, A., and Flicker, L. (2002). Latent Semantic Analysis: A New Method to Measure Prose Recall. *Journal of Clinical and Experimental Neuropsychology*, 24(1):26–35. PMID: 11935421.

Facal, D., Juncos-Rabadán, O., Rodríguez, M., and Pereiro, A. X. (2012). Tip-of-the-tongue in aging: influence of vocabulary, working memory and processing speed. *Aging Clin Exp Res*, 24(6):647–656.

Facal-Mayo, D., Juncos-Rabadán, O., Álvarez, M., Pereiro-Rozas, A. X., and Díaz Fernández, F. (2006). Efectos del envejecimiento en el acceso al léxico: el fenómeno de la punta de la lengua ante los nombres propios. *Rev Neurol*, 43:719–723.

Fergadiotis, G., Wright, H. H., and Capilouto, G. J. (2011). Productive vocabulary across discourse types. *Aphasiology*, 25(10):1261–1278.

Ferguson, A., Spencer, E., Craig, H., and Colyvas, K. (2013). Propositional Idea Density in women's written language over the lifespan: Computerized analysis.

Fillmore, C. J. (1968). The Case for Case. In Bach, E. and Harms, R. T., editors, *Universals in Linguistic Theory*, pages 0–88. Holt, Rinehart and Winston, New York.

Fillmore, C. J. (1969). Toward a modern theory of case. In Reibel, D. A. and Schane, S. A., editors, *Modern Studies in English*, pages 361–375. Prentice Hall.

Finatto, M. J. B., Scarton, C. E., Rocha, A., and Aluísio, S. M. (2011). Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. In *The 8th Brazilian Symposium in Information and Human Language Technology*, volume 1.

Fischman, H. C., Fernandes, C. S., Lourenço, R. A., Paradela, E. M. P., Carthery-Goulart, M. T., and Caramelli, P. (2009). Age and educational level effects on the performance of normal elderly on category fluency tasks. *Dement Neuropsychol*, 3(1):49–54.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3):285–307.

Foltz, P. W., Laham, D., and Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Education Journal of Computer enhanced learning On-line journal*, 1(2).

Forbes-McKay, K. and Venneri, A. (2005). Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological Sciences*, 26(4):243–254.

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., and Rochon, E. (2012). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*.

Frazier, L. (1985). Syntactic Complexity. In Dowty, D. R., Karttunen, L., and Zwicky, A. M., editors, *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, pages 129–189. Cambridge University Press, Cambridge.

Fried, L. P. (2000). Epidemiology of Aging. *Epidemiologic Reviews*, 22(1):95–106.

Frota, N. A. F., Nitrini, R., Damasceno, B. P., Forlenza, O., Dias-Tosta, E., da Silva, A. B., Herrera Junior, E., and Magaldi, R. M. (2011). Critérios para o diagnóstico de doença de Alzheimer. *Dement Neuropsychol*, 5(1):5–10.

Fuld, P. A., Masur, D. M., Blau, A. D., Crystal, H., and Aronson, M. K. (1990). Object-memory evaluation for prospective detection of dementia in normal functioning elderly: Predictive and normative data. *Journal of Clinical and Experimental Neuropsychology*, 12(4):520–528. PMID: 2211974.

Garcia, F. H. A. and Mansur, L. L. (2006). Habilidades funcionais de comunicação: idoso saudável. *Acta fisiatr*, 13(2):87–89.

Garrard, P., Ralph, M. A. L., Patterson, K., Pratt, K. H., and Hodges, J. R. (2005). Semantic feature knowledge and picture naming in dementia of Alzheimer's type: A new approach. *Brain and Language*, 93(1):79–94.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *The Assessment of Aphasia and Related Disorders*. Lippincott Williams & Wilkins, 3 edition.

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., and Grossman, M. (2011). Classification of primary progressive aphasia and its variants. *Neurology*, 76(11):1006–1014.

Graesser, A., McNamara, D., Louwerse, M., and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Graesser, A. C., Cai, Z., Louwerse, M. M., and Daniel, F. (2006). Question Understanding Aid (QUAID) A Web Facility that Tests Question Comprehensibility. *Public Opinion Quarterly*, 70(1):3–22.

Graesser, A. C. and McNamara, D. S. (2011). Computational Analyses of Multilevel Discourse Comprehension. *Topics in Cognitive Science*, 3(2):371–398.

Graesser, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, 40(5):223–234.

Graesser, A. C., Singer, M., and Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101:371–395.

Grossman, M., Koenig, P., DeVita, C., Glosser, G., Moore, P., Gee, J., Detre, J., and Alsop, D. (2003). Neural basis for verb processing in Alzheimer's disease: an fMRI study. *Neuropsychology*, 17(4):658–74.

Grossman, M. and Rhee, J. (2001). Cognitive resources during sentence processing in Alzheimer's disease. *Neuropsychologia*, 39(13):1419–1431.

Haberlandt, K. F. and Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General*, 114(3):357–374.

Hall, J. and Nilsson, J. (2006). CoNLL-X Shared Task: Multi-lingual Dependency Parsing. Technical report, Växjö University.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. English language series. Longman, London.

Henry, J. D., Crawford, J. R., and Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. *Neuropsychologia*, 42(9):1212–1222.

Henry, M. L. and Gorno-Tempini, M. L. (2010). The logopenic variant of primary progressive aphasia. *Current Opinion in Neurology*, 23(6):633–637.

Hodges, J. R., Patterson, K., Graham, N., and Dawson, K. (1996). Naming and Knowing in Dementia of Alzheimer's Type. *Brain and Language*, 54(2):302–325.

Holmes, D. I. and Singh, S. (1996). A Stylometric Analysis of Conversational Speech of Aphasic Patients. *Literary and Linguistic Computing*, 11(3):133—140.

Honoré, A. (1979). Some Simple Measures of Richness of Vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7:172–177.

Hooper, T. and Bayles, K. A. (2007). Management of neurogenic communication disorders associated with dementia. In Chapey, R., editor, *Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders*, pages 988–1008. Wolters Kluwer, Lippincott Williams & Wilkins, Philadelphia, 4 edition.

Horton, W. S. and Spieler, D. H. (2007). Age-related effects in communication and audience design. *Psychol Aging*, 22:281–290.

Howard, D., Patterson, K., and Company, T. V. T. (1992). *The Pyramids and Palm Trees Test: A Test of Semantic Access from Words and Pictures : [manual]*. Thames Valley Test Company.

Jacob Filho, W. (2000). Envelhecimento e atendimento domiciliário. In Duarte, Y. A. O. and Diogo, M. J. D., editors, *Atendimento domiciliar: um enfoque gerontológico*, pages 19–25. Atheneu, São Paulo.

Jarrold, W., Peintner, B., Yeh, E., Krasnow, R., Javitz, H., and Swan, G. (2010). 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., and Huang, J., editors, *Brain Informatics*, volume 6334 of *Lecture Notes in Computer Science*, pages 299–307. Springer Berlin Heidelberg.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Just, M. and Carpenter, P. (1987). *The Psychology of Reading and Language Comprehension*. Allyn and Bacon, Boston.

Kalpouzos, G., Eustache, F., Sayette, V., Viader, F., Chételat, G., and Desgranges, B. (2005). Working memory and FDG–PET dissociate early and late onset Alzheimer disease patients. *Journal of Neurology*, 252(5):548–558.

Kaplan, E., Goodglass, H., and Weintraub, S. (2001). *Boston Naming Test*. Lippincott Williams & Wilkins, Philadelphia, 2 edition.

Kemper, S., Greiner, L. H., Marquis, J. G., Prenovost, K., and Mitzner, T. L. (2001a). Language decline across the life span: findings from the Nun Study. *Psychol Aging*, 16(2):227–239.

Kemper, S., Herman, R., and Lian, C. (2003). Age Differences in Sentence Production. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 58(5):260–268.

Kemper, S., Kynette, D., Rash, S., O'Brien, K., and Sprott, R. (1989). Life-span changes to adults' language: Effects of memory and genre. *Applied Psycholinguistics*, 10:49–66.

Kemper, S. and Sumner, A. (2001). The structure of verbal abilities in young and older adults. *Psychology and Aging*, 16(2):312–322.

Kemper, S., Thompson, M., and Marquis, J. (2001b). Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging*, 16(4):600–614.

Kempler, D. (1995). Language changes in dementia of the Alzheimer's type. In Lubinsky, editor, *Dementia and communication: Research and clinical implications*, pages 98–114. Singular, San Diego.

Kempler, D., Almor, A., Tyler, L. K., Andersen, E. S., and MacDonald, M. C. (1998). Sentence Comprehension Deficits in Alzheimer's Disease: A Comparison of Off-Line vs. On-Line Sentence Processing. *Brain and Language*, 64(3):297–316.

Kendall, K. (2007). Presbyphonia: a review. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 15(3).

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95:163–182.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition.* Cambridge University Press, Cambridge.

Kintsch, W. and Keenan, J. (1973). Reading Rate and Retention as a Function of the Number of Propositions in the Base Structure of Sentences. *Cognitive Psychology*, 5:257–274.

Lam, B., Masellis, M., Freedman, M., Stuss, D. T., and Black, S. E. (2013). Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome. *Alzheimers Res Ther*, 5(1):1.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Landauer, T. K., Laham, D., Rehder, B., and Schreiner, M. E. (1997). How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. *Cognitive Science*.

Le Dorze, G. and Bédard, C. (1998). Effects of age and education on the lexico-semantic content of connected speech in adults. *Journal of Communication Disorders*, 31(1):53–71.

Lehfeld, H. and Erzigkeit, H. (1997). The SKT–a short cognitive performance test for assessing deficits of memory and attention. *Int Psychogeriatr*, 9(1):115–121.

Lin, D. (1996). On the Structural Complexity of Natural Language Sentences. In *COLING*, pages 729–733.

Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1):3–28.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

MacDonald, M. C., Almor, A., Henderson, V. W., Kempler, D., and Andersen, E. S. (2001). Assessing Working Memory and Language Comprehension in Alzheimer's Disease. *Brain and Language*, 78(1):17–42.

MacKay, A., Connor, L., Albert, M., and Obler, L. K. (2002). Noun and verb retrieval in healthy aging. *J Int Neuropsychol Soc*, 8(6):764–770.

Mackenzie, C. (2000). Adult spoken discourse: the influences of age and education. *Int. J. Lang. Comm. Dis.*, 35(2):269–285.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition.

MacWhinney, B., Fromm, D., Holland, A., Forbes, M., and Wright, H. (2010). Automated analysis of the Cinderella story. *Aphasiology*, 24(6-8):856–868.

Manning, W. H. and Shirkey, E. A. (1981). Fluency and the Aging Process. In Beasley, D. S. and Davis, G. A., editors, *Aging Communication Processes and Disorders*. Grune & Straton, Nova Iorque.

Mansur, L. L. (1996). *Formulação e reformulação: contribuição ao estudo da linguagem oral de indivíduos com demência do Tipo Alzheimer*. PhD thesis, FFLCH.

Mansur, L. L., Carthery, M. T., Caramelli, P., and Nitrini, R. (2005). Linguagem e Cognição na doença de Alzheimer. *Psicologia, Reflexão e Crítica*, 18(3):300–307.

Mar, R. A. (2004). The neuropsychology of narrative: story comprehension, story production and their interrelation. *Neuropsychologia*, 42(10):1414–1434.

Marini, A., Boewe, A., Caltagirone, C., and Carlomagno, S. (2005). Age-related Differences in the Production of Textual Descriptions. *Journal of Psycholinguistic Research*, 34(5):439–463.

Marques, J. F., Cappa, S. F., and Sartori, G. (2011). Naming from definition, semantic relevance and feature type: the effects of aging and Alzheimer's disease. *Neuropsychology*, 25(1):105–113.

Mattis, S. (1988). *Dementia rating scale: professional manual*. Psychological Assessment Resources, Inc.

Maziero, E., Pardo, T., and Aluísio, S. (2008). Ferramenta de análise automática de inteligibilidade de córpus (AIC). Technical report, NILC - ICMC - USP.

McDonald, R., Lerman, K., and Pereira, F. (2006). Multilingual Dependency Analysis with a Two-stage Discriminative Parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 216–220, Stroudsburg, PA, USA. Association for Computational Linguistics.

McKenna, P. and Warrington, E. (1983). *Graded naming test*. NFER-Nelson.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34(7):939–944.

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., and Phelps, C. H. (2011). 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*, 7(3):263–269.

McNamara, D., Kintsch, E., Songer, N. B., and Kintsch, W. (1996). Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning from Text. *Cognition and Instruction*, 14(1):1–43.

McNamara, D., Louwerse, M., and Graesser, A. (2002). Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Grant Proposal. http://cohmetrix.memphis.edu/cohmetrixpr/publications.html.

McNamara, D. S. and Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, 22(3):247–288.

Mendez, M. F., Lee, A. S., Joshi, A., and Shapira, J. S. (2012). Nonamnestic Presentations of Early-Onset Alzheimer's Disease. *American Journal of Alzheimer's Disease and Other Dementias*, 27(6):413–420.

Meyer, B. J., Wijekumar, K., Middlemiss, W., Higley, K., Lei, P.-W., Meier, C., and Spielvogel, J. (2010). Web-Based Tutoring of the Structure Strategy With or Without Elaborated Feedback or Choice for Fifth- and Seventh-Grade Readers. *Reading Research Quarterly*, 45(1):62–92.

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology*, 43(11):2412–2414.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., and Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

Nippold, M. A., Cramond, P. M., and Hayward-Mayhew, C. (2013). Spoken language production in adults: Examining age-related differences in syntactic complexity. PMID: 24093162.

Nivre, J., Hall, J., and Nilsson, J. (2006). MaltParser: A data-driven parser-generator for dependency parsing. In *In Proc. of LREC-2006*, pages 2216–2219.

O'Brien, E. J., Rizzella, M. L., Albrecht, J. E., and Halleran, J. G. (1998). Updating a situation model: a memory-based text processing view. *J Exp Psychol Learn Mem Cogn*, 24(5):1200–10.

O'reilly, T. and McNamara, D. S. (2007). Reversing the Reverse Cohesion Effect: Good Texts Can Be Better for Strategic, High-Knowledge Readers. *Discourse Processes*, 43(2):121–152.

Ozuru, Y., Dempsey, K., and McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, 19(3):228–242.

Pakhomov, S., Chacon, D., Wicklund, M., and Gundel, J. (2011). Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing. *Behavior Research Methods*, 43(1):136–144.

Park, D. C. and Bischof, G. N. (2013). The aging mind: neuroplasticity in response to cognitive training. *Dialogues Clin Neurosci*, 15(1):109–119.

Paschoal, S. M. P. (1996). Epidemiologia do Envelhecimento. In Papaléo Netto, M., editor, *Gerontologia*, pages 26–43. Atheneu, São Paulo.

Pasqualini, B., Scarton, C. E., and Finatto, M. J. B. (2011). Comparando Avaliações de Inteligibilidade Textual entre Originais e Traduções de Textos Literários. In *The 8th Brazilian Symposium in Information and Human Language Technology*, volume 1.

Peintner, B., Jarrold, W., Vergyri, D., Richey, C., Tempini, M. L. G., and Ogar, J. (2008). 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*, pages 4648–4651.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. (2007). The development and psychometric properties of LIWC2007. http://homepage.psy.utexas.edu/homepage/faculty/Pennebaker/reprints/LIWC2007_LanguageManual.pdf, Austin, TX.

Pennebaker, J. W., Francis, M. E., and J., B. R. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Lawrence Erlbaum Associates, Mahwah.

Pentimonti, J. M., Zucker, T. A., Justice, L. M., and Kaderavek, J. N. (2010). Informational Text Use in Preschool Classroom Read-Alouds. *The Reading Teacher*, 63(8):656–665.

Perfetti, C. (2007). Reading Ability: Lexical Quality to Comprehension. *Scientific Studies of Reading*, 11(4):357–383.

Peters, F., Majerus, S., Baerdemaeker, J. D., Salmon, E., and Collette, F. (2009). Impaired semantic knowledge underlies the reduced verbal short-term storage capacity in Alzheimer's disease. *Neuropsychologia*, 47(14):3067–3073.

Petersen, R. C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, 256(3):183–194.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., and Kokmen, E. (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology*, 56(3):303–308.

Pichora-Fuller, M. K. and Levitt, H. (2012). Speech Comprehension Training and Auditory and Cognitive Processing in Older Adults. *American Journal of Audiology*, 21(2):351–357.

Pinheiro, M. M. C. and Desgualdo, P. L. (2004). Processamento auditivo em idosos: estudo da interação por meio de testes com estímulos verbais e não-verbais. *Revista Brasileira de Otorrinolaringologia*, 70(2):209–214.

Rapp, D. N., van den Broek, P., McMaster, K. L., Kendeou, P., and Espin, C. A. (2007). Higher-Order Comprehension Processes in Struggling Readers: A Perspective for Research and Intervention. *Scientific Studies of Reading*, 11(4):289–312.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., and Seidenberg, M. S. (2001). How Psychological Science Informs the Teaching of Reading. *Psychological Science in the Public Interest*, 2(2):31–74.

Resnik, P. (1992). Left-Corner Parsing And Psychological Plausibility. In *COLING*, pages 191–197.

Riverol, M. and López, O. (2011). Biomarkers in Alzheimer's disease. *Frontiers in Neurology*, 2(46).

Roark, B., Hosom, J.-P., Mitchell, M., and Kaye, J. A. (2007a). Automatically derived spoken language markers for detecting mild cognitive impairment. In *2nd International Conference on Technology and Aging (ICTA)*.

Roark, B., Mitchell, M., and Hollingshead, K. (2007b). Syntactic complexity measures for detecting Mild Cognitive Impairment. In *Biological, translational, and clinical language processing*, pages 1–8, Prague, Czech Republic. Association for Computational Linguistics.

Roark, B., Mitchell, M., Hosom, J., Hollingshead, K., and Kaye, J. (2011). Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2081–2090.

Rohrer, J. D., Rossor, M. N., and Warren, J. D. (2012). Alzheimer's pathology in primary progressive aphasia. *Neurobiology of aging*, 33(4):744–752.

Rosenberg, S. and Abbeduto, L. (1987). Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, 8:19–32.

Roth, M., Tym, E., Mountjoy, C. Q., Huppert, F. A., Hendrie, H., Verma, S., and Goddard, R. (1986). CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. *The British Journal of Psychiatry*, 149(6):698–709.

Saling, L. L., Laroo, N., and Saling, M. M. (2012). When more is less: Failure to compress discourse with retelling in normal ageing. *Acta Psychologica*, 139(1):220–224.

Salmon, D. P., Thomas, R. G., Pay, M. M., Booth, A., Hofstetter, C. R., Thal, L. J., and Katzman, R. (2002). Alzheimer's disease can be accurately diagnosed in very mildly impaired individuals. *Neurology*, 59(7):1022–1028.

Salthouse, T. (1991). *Theoretical Perspectives on Cognitive Aging*. Erlbaum.

Salthouse, T. A. and Mandell, A. R. (2013). Do Age-Related Increases in Tip-of-the-Tongue Experiences Signify Episodic Memory Impairments? *Psychological Science*, 24(12):2489–2497.

Saur, D., Schelter, B., Schnell, S., Kratochvil, D., Küpper, H., Kellmeyer, P., Kümmerer, D., Klöppel, S., Glauche, V., Lange, R., Mader, W., Feess, D., Timmer, J., and Weiller, C. (2010). Combining functional and anatomical connectivity reveals brain networks for auditory language comprehension. *NeuroImage*, 49(4):3187–3197.

Scarton, C. and Aluísio, S. (2010). Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamática*, 2(1):45–62.

Scarton, C., Gasperin, C., and Aluisio, S. (2010). Revisiting the Readability Assessment of Texts in Portuguese. In Kuri-Morales, A. and Simari, G., editors, *Advances in Artificial Intelligence – IBERAMIA 2010*, volume 6433 of *Lecture Notes in Computer Science*, pages 306–315. Springer Berlin Heidelberg.

Silva, T. B. L., Yassuda, M. S., Guimarães, V. V., and Florindo, A. A. (2011). Fluência verbal e variáveis sociodemográficas no processo de envelhecimento: um estudo epidemiológico. *Psicol Reflex Crit*, 24(4):739–746.

Ska, B. and Duong, A. (2005). Communication, discours et démence. *Psychol NeuroPsychiatr Vieil*, 3(2):125–133.

Ska, B. and Joanette, Y. (2006). Vieillissement normal et cognition. *Médecine sciences*, 22(3):284–287.

Smith, E. and Ivnik, R. J. (2003). Normative neuropsychology. In Petersen, R. C., editor, *Mild cognitive impairment*, pages 63–88. Oxford University Press, New York.

Snowdon, D., Kemper, S., Mortimer, J., Greiner, L., Wekstein, D., and Markesbery, W. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *JAMA*, 275(7):528–532.

Snowdon, D. A., Greiner, L. H., and Markesbery, W. R. (2000). 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*, 903(1):34–38.

Soares, L. M., Cachioni, M., Falcão, D. V. d. S., Batistoni, S. S. T., Lopes, A., Neri, A. L., and Yassuda, M. S. (2012). Determinants of cognitive performance among community dwelling older adults in an impoverished subdistrict of São Paulo in Brazil. *Archives of Gerontology and Geriatrics*, 54(2):187–192.

Sperling, R. A., Karlawish, J., and Johnson, K. A. (2013). Preclinical Alzheimer disease - the challenges ahead. *Nat Rev Neurol*, 9(1):54–58.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21(4):360–406+.

Steiner, V. A. G. (2012). *Fluência semântica, fonêmica e de verbos em indivíduos com comprometimento cognitivo leve*. PhD thesis, Faculdade de Medicina da Universidade de São Paulo.

Steiner, V. A. G., Mansur, L. L., Brucki, S. M., and Nitrini, R. (2008). Phonemic verbal fluency and age: a preliminary study. *Dement Neuropsychol*, 2(4):328–332.

Stopford, C. L., Snowden, J. S., Thompson, J. C., and Neary, D. (2008). Variability in cognitive presentation of Alzheimer's disease. *Cortex*, 44(2):185–195.

Takao, A. Y., Prothero, W. A., and Kelly, G. J. (2002). Applying Argumentation Analysis To Assess the Quality of University Oceanography Students' Scientific Writing. *Journal of Geoscience Education*, 50(1):40–48.

Taler, V. and Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5):501–556. PMID: 18569251.

Tannen, D. (1982). *Spoken and written language : exploring orality and literacy / Deborah Tannen, editor*. ABLEX Pub. Corp Norwood, N.J.

Thomas, C., Keselj, V., Cercone, N., Rockwood, K., and Asp, E. (2005). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *Mechatronics and Automation, 2005 IEEE International Conference*, volume 3, pages 1569–1574 Vol. 3.

Thorson, E. and Snyder, R. (1984). Viewer recall of television commercials: Prediction from the propositional structure of commercial scripts. *Journal of Marketing Research*, 21:127–136.

Togher, L. (2001). Discourse sampling in the 21st century. *Journal of Communication Disorders*, 34(1–2):131–150.

Toledo, C. M. (2011). Variáveis sociodemográficas na produção do discurso em adultos sadios. Master's thesis, Faculdade de Medicina da Universidade de São Paulo, São Paulo.

Traykov, L., Rigaud, A.-S., Cesaro, P., and Boller, F. (2007). Le déficit neuropsychologique dans la maladie d'Alzheimer débutante. *L'Encéphale*, 33(3):310–316.

Turner, A. and Greene, E. (1977). The Construction and Use of a Propositional Text Base. Technical report, Institute for the Study of Intellectual Behavior, University of Colorado.

van Dijk, T. and Kintsch, W. (1983). *Strategies of discourse comprehension*. Monograph Series. Academic Press.

Verhaegen, C. and Poncelet, M. (2013). Changes in Naming and Semantic Abilities With Aging From 50 to 90 years. *Journal of the International Neuropsychological Society*, 19:119–126.

Vliet, E. C.-V., Manly, J., Tang, M.-X., Marder, K., Bell, K., and Stern, Y. (2003). The neuropsychological profiles of mild Alzheimer's disease and questionable dementia as compared to age-related cognitive decline. *Journal of the International Neuropsychological Society*, 9:720–732.

Williams, L. J., Abdi, H., French, R., and Orange, J. B. (2010). 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*, 53(5):1372–1393.

Wills, C., Capilouto, G., and Wright, H. (2012). Attention and Off-Topic Speech in the Recounts of Middle-Age and Elderly Adults: A Pilot Investigation. *Contemporary issues in communication science and disorders*, 39:105–112.

Wilson, B., Greenfield, E., Clare, L., Baddeley, A., Cockburn, J., Watson, P., Tate, R., Sopena, S., and Nannery, R. (2008). *Rivermead Behavioural Memory Test-Third Edition. (RBMT-3)*. Pearson Assessment, Santo Antonio. TX.

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., and Petersen, R. (2004). Mild cognitive impairment – beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *Journal of Internal Medicine*, 256(3):240–246.

Wingfield, A. and Grossman, M. (2006). Language and the Aging Brain: Patterns of Neural Compensation Revealed by Functional Brain Imaging. *Journal of Neurophysiology*, 96(6):2830–2839.

Wright, H. H., Capilouto, G. J., and Koutsoftas, A. (2013). Evaluating measures of global coherence ability in stories in adults. *International Journal of Language & Communication Disorders*, 48(3):249–256.

Wright, H. H., Koutsoftas, A. D., Capilouto, G. J., and Fergadiotis, G. (2014). Global coherence in younger and older adults: Influence of cognitive processes and discourse type. *Aging, Neuropsychology, and Cognition*, 21(2):174–196. PMID: 23656430.

Yassuda, M. S., Diniz, B. S., Flaks, M. K., Viola, L. F., Pereira, F. S., Nunes, P. V., and Forlenza, O. V. (2009). Neuropsychological Profile of Brazilian Older Adults with Heterogeneous Educational Backgrounds. *Archives of Clinical Neuropsychology*, 24(1):71–79.

Yngve, V. H. (1960). A Model and an Hypothesis for Language Structure. *Proceedings of the American Philosophical Society*, 104:444–466.

Zwaan, R. A. and Radvansky, G. A. (1998). Situation Models in Language Comprehension and Memory. *Psychological Bulletin*, 123(2):162–185.

A. Extratos de entrevistas compilados do Anexo IV da tese de Mansur [1996]

Os extratos estão separados por sujeitos da pesquisa (6 sujeitos) e por problemas que os extratos ilustram. Dentro de cada problema há um conjunto de exemplos.

1) SUJEITO 1 (DA leve) OB

1. REPETIÇÕES DE INFORMAÇÃO

EXEMPLO 1

E- o senhor correu muito pra chegar até aqui seu Osvaldo?

O- não... viemos tranquilos

E- não correu não?

O- {nós moramos em Mairiporã né?

E- ah em Mairiporã?

O- é... tomamos o: ônibus pra podê saltar no metrô aqui... tranquilão/

{hum}

E- e o senhor anda bastante aqui em são Paulo seu seu Osvaldo?

O- aqui não porque eu não moro aqui... só quando eu estou aqui... nós moramos em Mairiporã né? ... lá é minha vida né?

EXEMPLO 2

O- eu gostava muito de vendas sempre fui homem de vendas ... já ouviu falá em Kelsons?

Kelsons é a maior fábrica de bolsas de senhoras do Brasil ... no Rio de Janeiro...

então eles tinham uma: ... várias lojas deles aí também ... trabalhei vinte e cinco anos como gerente de vendas... tinha sessenta pessoas... sessenta vendedores...

1ª REPETIÇÃO DAS INFORMAÇÕES FORNECIDAS NO TRECHO 2 (ACIMA EM ITÁLICO)

O- antigamente eu tomava muito café.. porque eu trabalhava eu trabalhava em vendas... em vendas pra iniciar a venda e ser bem sucedido você precisa convidar o o (companheiro) prum café... né?... (porque bom)... deixa... a:... as infelicidades pra trás... e vamos comprar... e do café sai aquela (conversa) gostosa e tal... cê tá vendendo tal...eu fui:... gerente da Kelsons... gerente de vendas... aqui em São Paulo porque a fábrica é no Rio... eu era gerente de São Paulo... cheguei a ter até sessenta pessoas...

2ª REPETIÇÃO DA INFORMAÇÃO FORNECIDA NO TRECHO 2 (ACIMA EM ITÁLICO)

em São Paulo que eu era... gerente de vendas da Kelsons... que era a maior fábrica de plástico do Brasil... no Rio de Janeiro

EXEMPLO 3

O- principalmente a partir de determinada idade né? ...eu vivi passei toda a minha vida na cidade de São Paulo ... sempre trabalhei na capital vivi... passava as noites e tal... tudo no centro de São Paulo ... trabalhei na praça do Patriarca... pra você te uma ideia...

E- bem no centro

REPETIÇÃO DA INFORMAÇÃO FORNECIDA NO TRECHO 3 (ACIMA, EM ITÁLICO)

O- {e olha que eu fui e olha que eu nasci em São Paulo ... na capital ... sempre trabalhei no centro da cidade ... na Barão de Itapetininga por ali Praça do Patriarca Dom José de Barros... tive escritório por ali...

EXEMPLO 4.

O- é: eu disse:... adotei:... o: sobrinho da minha mulher né?... que foi pra casa com uns... três quatro anos e já está com quarenta e um... e ele ca ele é casado... e tem um filho... então ele me arrumou um neto...

REPETIÇÃO DA INFORMAÇÃO FORNECIDA NO TRECHO 4 (ACIMA, EM ITÁLICO)

O- é cê vê ele é filho... do meu sobrinho... que eu adotei... o sobrinho da minha mulher ... que nós adotamos...

E- ah que jóia... o orgulho da família

O- {é... ele já está com quarenta e um anos... foi pra casa com: oito anos... sete anos... está com quarenta e um

2. FALHAS DE ACESSO LEXICAL

SUBSTANTIVOS COMUNS

EXEMPLO 1.

Descrição: dificuldade em acessar o termo “prédios”.

Pistas para análise: pausa de hesitação diante de substantivo [F0E0?] emprego de termo correlato (apartamentos) [F0E0?] substituição brusca pelo termo desejado (prédios)

O- cê nunca ouviu fala?... clube de Mairiporã?

E- eu num num conheço

O- é uma beleza

E- é?

O- tem/ é um é privilegiado ... tem tudo tem três piscinas... mais o terreno é que é próprio é adequado ... uma sede maravilhosa tal **e: existe um:...um:...ô meu Deus deixa pra lá...** apartamentos dentro do próprio

terreno... o terreno tem uma divisão...o clube é uma depois lá eles fizeram... três... **apartamentos mui/ três prédios** muito bons... nós já moramos lá: quando eu deixei de trabalhar

EXEMPLO 2.

Pistas para a análise:

Prolongamento, pausa de hesitação, comentário sobre a dificuldade de resgate lexical (“solúvel”), pista por meio de descrição [F0E0?] intervenção do entrevistador [F0E0?] confirmação do paciente

E- mas me dá sua receita aí do cafezinho vamo ve se é tão boa quanto da macarronada... como é que o senhor faz?

O-{não café é simples... geralmente quem faz café é a Haide... ela esquenta a água... bota o adoçante na xícara... bocadinho de pó... e: bota a água direto... (na realidade) nós usamos o:...((estala os dedos para tentar se lembrar do nome)) aquele pó como é que chama que vem na na na na... no vidrinho... não precisa coar... (então cê vê) como que está minha cabeça...

E- não... eu também não estou

O-{imagina se antigamente eu esquecia alguma coisa que se come ou que se bebe... nunca na minha vida... é aquele pó que vem em vidro...

E- aquele solúvel?

O- é aquele solúvel... eu uso aquilo... não dá muito trabalho... não precisa fazer... negócio de... de coador: lava coador

NOMES PRÓPRIOS

EXEMPLO 1

Pistas para a análise:

Prolongamento e pausa de hesitação diante de nome próprio (nome da rua)

E- então o senhor acha que o o progresso atrapalha um pouco a vida das pessoas?

O- principalmente a partir de determinada idade né? ...eu vivi passei toda a minha vida na cidade de São Paulo ... sempre trabalhei na capital vivi... passava as noites e tal... tudo no centro de São Paulo ... trabalhei na praça do Patriarca... pra você te uma idéia...

E- bem no centro

O- ti tive escritório na:... Barão de Itapetininga na praça da República na Dom José de Barros... sempre por ali

EXEMPLO**2**

Pistas para a análise:

Prolongamento, pausa de hesitação e repetição, diante de nome próprio (nome da cidade do interior)

E- o senhor te o senhor teve filhos?

O- nós adotamos... nós não tivemos filho lamentavelmente e:... a família dela mora em:... em: morava... no interior... sempre moraram no interior...

EXEMPLO**3**

Pistas para a análise:

Pausa de hesitação, preenchimento (“e”) diante de nome próprio

O- eu frequentei o jóquei clube quarenta e cinco anos... não (é que eu tava indo toda semana ... então aparecia corrida eu ía) todo domingo chegava no primeiro páreo... e no último ía saber se eu não tinha mais se podia ir embora... é muito tempo passei ir até aos sábados sábado e domingo ía todo sábado e domingo ... nos grandes prêmios ía pro Rio de Janeiro passava uma semana lá que... o CHIQUE do negócio é: na Gávea no Rio de Janeiro... fui pra e ... Argentina... assim que um cavalo brasileiro foi correr

EXEMPLO **4**

Pistas para a análise:

Pausa de hesitação diante de predicativo (“apelidado”) do verbo de ligação, repetição (foi – foi), pausa de hesitação e repetição diante de nome próprio (Diamante Negro)

E- e quem eram os grandes jogadores?

O- bom como o grande Leônidas né?... o: inclusive na na Europa no campeonato:... mundial ele foi... foi apelidado de o:... o:... de Diamante Negro...

EXEMPLO **5**

Pistas para a análise:

Prolongamento (amigo:), pausa de hesitação, uso de expressão descritiva para ganhar tempo (o famoso aí) repetição + prolongamento (o:), repetição (de, de), prolongamento + pausa de hesitação (o:) diante do nome próprio (Pelé)

E- ah é?... que bom... a gente/falando dos jogadores de futebol... eles acabam virando estrelas né? assim como as...ahm... pessoas da televisão

O- é... o caso maior foi o nosso amigo:... o:... famoso aí de de de Santos o:... Pelé

3. OUTRAS HESITAÇÕES

Pausas, prolongamentos e repetições

EXEMPLO 1

E- como é que está o futebol agora que eu não tô muito:

O {é está: está bem... está no:... no auge do do do do do do do campeonato... sei lá... agora termino aqui em São Paulo... o Corinthians foi campeão eu sou:... corinthiano... então

EXEMPLO 2

O- é... mas tem uma coisa... ele telefonou lá da Argen da da Argentina... VÔ:... quanto cê vai me dar?... ele é LOUCO por dinheiro

2) SUJEITO 2 (DA leve) EFA

1. USO DE TERMOS VAGOS

(NA LITERATURA APARECE COMO “LINGUAGEM VAZIA”)

TERMOS VAZIOS EMPREGADOS: “COISA” , “COISINHA”, “NEGÓCIO”, “TUDO”, “TAL”, “VÁRIOS”

EXEMPLO 1

EFA- ... bom ... eu já morei aqui em São Paulo quando era minina ... depois meus pais mudaram pra Santos ... i: adoro Santos ... num tem ... pode se dizê qui me criei ... mesmo lá ... eu estudei ... fiz tudo lá em Santos ... eu jamais quiria voltá morá aqui em São paulo ... por causa dessa correria ... **essa coisa assim** ... ADORO São Paulo ... de vir pra São Paulo pra passeá assim ... tá tudo tá ... tá comigo ... mas : morá não ... de jeito nenhum ...

EXEMPLO 2

EFA- não ... eu sempre venho né cu meu marido ... a gente vem de carro ... é: ... faz as compra **uma coisinha e otra** né ... assim ... eu gosto de vir pra São Paulo ... adoro vir pra São paulo passeá ... ... assim ... mas morá num queria morá aqui não

EXEMPLO 3

EFA- eu venho muito à Sào Paulo ... sempre venho ... gosto muito de passeá **todas essas coisas** ... mas situação desagradável eu num tive ...

EXEMPLO 4

EFA- ... i: aí: num sei ah: ... bom si foi uma coisa que chegô queimá tudo ... num deu tempo nem di você podê desligá a força ... qui a primeira coisa que a gente faz numa hora dessa ... começa um trovão ... **um negócio assim** eu corro no relógio ... desligo ... né?

EXEMPLO 5

EFA- i ... desci ... lá pra baixo ... i tem o quintal ... tem piscininha também ... fiz mais uma piscina ... **uma coisinha** qui a gente andô aprontando lá ...

EXEMPLO 6

E- então me dá sua receita de macarronada

EFA- ah ... eu faço um BELO ... molho né ?... com carne ... uma carne assada bem temperada bem gostosa bem (assim) ... aí ponho cebola ... o tomate ... ponho u: ... molho ... põe ... us temp/ us cheiros verdes **aquela coisa toda**

EXEMPLO 7

E- mas como é que é o café? ... como que é a receita?... se a senhora tiver que ensinar prum estrangeiro que chega

EFA- ah ... como é que eu faço café bom ... eu fervo água ... quando a água já tá fervendo (depois) aí eu ponho o pó ... mexo fê/ mexo bem ... deixo ... quando começa a querê subi eu desligo ... passo pro coador ... né? ... côo o café ... fica u cafezinho bom ... aí pego depois u coadinho ... ponho em banho maria ... o bule pra ficá quentinho ... né? ... pego leite ... pãozinho biscoito ... **essas coisas mais**

EXEMPLO 8

EFA- ... aí a tarde faço um lanche ... **uma coisa assim** mais ligeira né ?... faço um bolo pro sábado ... pu dia seguinte ... domingo sábado ... **qualquer outra coisa assim de manhã** a gente faz eu já faço já deixo tudo pronto pro domingo ... também posso comer fora pizza **essas coisas**

EXEMPLO 9

E- hum . a sua diversão é essa?

EFA- é ... i a tarde a gente dor ... descansa um pouco depois a gente pega um cine:ma ... u: ... ah inventa **qualqué coisa** ... vai fazê uma visi:ta né? ... prus amigos

EXEMPLO 10

EFA- ... i: ... dá um tempo pra empregada ela tem um menininho né? ... então tem um molequinho pra bagunçá lá dentro ... mas ela é boazinha coitada ... i: ... ela mi limpa a casa ... dá um jeito na casa **... coisa** i eu tenho faxineira uma vez por semana ... vem fazê limpeza ... pesada né ... ela conserva a casa bonitinha ... i quando ... **qualqué coisa** ... ela ... o dia que ela tem folga então eu vô ... passo né ... a gente vai ... levando ...

EXEMPLO 11

EFA- ... então ... fico cum ela ... lá na casa dela ... no apartamento dela que eu num quero ... não ... né ... ela tem a parte dela ... eu já fiz separado ... eu tranco a porta no meio ... ela fica pra lá ... cus cachorro dela ... e ela fica cus cachorro dentro de ca:sa ... **tudo** ... ela se vira lá cus cachorro ... agora a minha não ... a minha dorme fo:ra **... tudo direitinho**

EXEMPLO 12

EFA- ... então ela mora em casa ... então tô tranquila ... que eu saio eu sei que ela cuida direitinho ... faz a comida dela ... eu já faço a comida pra dois três dias deixo dentro do freezer né? ... de noite ela / de manhã ela já tira deixa as ... amorná a comida ... **i tal** pra dá à noite ... que ela come uma vez por dia né?

EXEMPLO 13

EFA- ah: ... quando eu viajo assim ... primeiro: ... nós: ... geralmente : ... antigamente tinha um sítio **tudo** né? ... nós iamos pra lá domingo sábado domingo ia pro sítio **aquela coisa**... gra:ças a deus vendemos

EXEMPLO 14

E- i: ... com relação ao carro não precisa tomar nenhuma providência?

EFA- não ... aí sim ... bom isso aí é a parte dele né ?((ri))... ele qui leva pra ofic/ manda vê examiná tá tudo bem ... **aquela coisa** ... enche o tanque de gasolina ... aquela história toda é ele que

faz ... num quero nem sabe ... agora na viagem sim ... muitas vezes ele pára ... ele cansa então ele ... fica ... então eu pego a direção eu ando ... vô embora ... depois ele ...

EXEMPLO 15

EFA- já sabe já ta acostumado com a gente ... ...então vem ... fica em casa ... eu deixo ... tenho um quarto la ... um quarto pra hóspede ... **tudo direitinho** arrumadinho ... **tudo em ordem** ... vem ... vem então ... quando vem uma pessoa assim ... que seja amigo ou ... coleguinha da minha filha ... qui fique **qualquer coisa** ... porque ela também tem o apartamento dela ela pode receber lá ... mas ... si quizé qui ... recebê em casa ... eu tenho um quarto todo arrumadinho ... ((inint.)) ... entende? ... em ordem **tudo direitinho** ... i : ... a nossa vida continua apenas no momento ... uma questão di co/ di comi: da di refeição **essas coisa** ... a gente só aumenta e faz **umas coisinha** ... faz uns docinhos mas ...

EXEMPLO 16

EFA- { um bolo ... faz um doce de abóbora ... um doce di ... mamão ... **essas coisinhas assim** qui a genti faz né ... tem que ver na hora ... entende?

EXEMPLO 17

EFA- pois é : ... ma num qué istudá ... num qué: di jeito nenhum ... e é isso qui mi faz muito mal agora cheguei ao ponto de dizer ... porque que eu tenho varios problemas assim ... de nervoso ... **de coisa**? ... é por causa dele ... tanto meu marido também ... coitado ... porque o pai num aguenta ...

EXEMPLO 18

EFA- o meu é só o pobre desse esquecimento...o esquecimento das coisas né? ...depois de um certo tempo pra cá ... num era ass assim mas acho que tanta coisa ...sabe? tanto aborrecimento que eu já te ... que eu tenho tido ... esse negócio de neto ... i filho também ... o casamento num foi lá aquelas coisa sabe? ... tudo isso: ... mete mexe cum a cabeça da gente mesmo que você num queira mexe ... de vê um rapaz ... formado ... aqui pela Politécnica ... engenheiro ... largô ... bandonô tudo ... si mandô pra Santos atrás dela ... nu fim acabaram casando ... um casamento péssimo ... agora vive lá né?... de de ... de vendê **coisinhas e ...** bobagem na casa dele ... contrabando essas bobagens que ele traz do Paraguai **... essas coisas assim** ... é dureza né filha ... fazê um sacrifício danado ... dá tudo pra um filho se formá ... engen engenheiro ... afinal de contas ele é formado por uma das melhores escolas de São Paulo ... (i tá numa situação dessa)

EXEMPLO 19

EFA- ah olha ... eu gosto tudo quanto é tipo de programas assim ... políticos ... esses programas assim eu gosto ... sabe? ... num gosto desses programas de ... muita bobagem **coisa assim** eu num assisto não ...

quando tem algum qui mi pega ... qui mi interessa assim ... qui mi prende ... aí eu eu assisto ... mas **essas coisas** de :

EXEMPLO 20

E- {teve algum que a senhora acho que fosse mais marcante ... do ponto de vista de interesse

EFA- é tem ... tem vários que eu gosto de assisti

E- então conta um que a senhora gostou

EFA- {**vários ai** : ... i agora peraí ... (tem qui) qui dizê assim de cabeça cê vê como é?((inint)) escapa viu? ... tem **vários** que eu gosto de assisti viu? ... do: ... eu ouço muito lá em casa a gente vê muito o oito né? ...

EXEMPLO 21

E- que que a senhora achou do impeachement do Collor?

EFA- ... do impeachement do Collor? ... bom ... eu achei que tinha qui sê né? ... qui do jeito que o negócio estava ...n num tinha condições ... foi **a melhor coisa** qui podia ter acontecido pro Brasil tê si livrado desse homem né ?

EXEMPLO 22

EFA- eu ....sabe ... qui uma coisa que eu vô te dizer ... num sô MUI:to lá desse negócio de política não viu ... eu acompanho assim então... meu marido que go:sta tal ... mas mesmo assim eu me desinteresso um poco ... num me pego muito sabe? dessas ((inint)) nu Collor a gente tem ... porque foi um **negócio tão assim** ((inint.)) que todo mundo não se falava em outra coisa né? ...a a: situação ... mas agora acho que tá bem né ... o Fernando Collor tá bom ... tô gostando dele ... parece que ele tá indo bem né ?

EXEMPLO 23

EFA- fui mi virei ... depois de casada ... fu/quis ... melhorá ... fui fazê facul/ ... vinha pra São paulo ... levantava a hora que levantava ... cinco hora da manhã porque eu tinha qui tá aqui ... no ... na na na faculdade ... aqui na na ... como é que é o nome dessa avenida qui tem aqui na ... **coisa aqui** ?... ah ... onde tem u: ... perto da Luz ali ... como é o nome daquela avenida? já até esqueci ... tem a faculdade de so/ Sociologia e Política de São Paulo ... foi lá que eu vim fazê a faculdade ...

EXEMPLO 24

E- e a senhora acha que as ... que as novelas tem um papel na educação das pessoas?

EFA- olha ... mais ou menos né? ... às vezes ela só dá maus exemplos eu acho viu? ... num é uma coisa assim que seja ... qui si diga ... qui seja: ... BOA pra pra ... tem muita bobagem que a gente vê qui num ... né? ... agora tem coisas interessantes também ... é: é o **tal negócio** né ... o (que vale) é a cabeça né?

2. USO DE ONOMATOPÉIAS EM SUBSTITUIÇÃO A ITENS LEXICAIS

EXEMPLO 1

E- e dus outros jogadores que penduraram a chuteira ... a senhora lembra quem eram

EFA- {não porque eu num sô muito di marca sabe? ... nome de de futebol ... eu vejo futebol assim ... um tá assistindo futebol ... eu sento assisto futebol ... torço ... vai ganhá **vá vá va** ... mas eu num fixo assim nus ... nus jogadores

EXEMPLO 2

E- i na segunda feira? ... qual é sua rotina di segunda feira?

EFA- bom segunda feira eu tô com a empregada em casa ... então eu já tô di beleza

E- ((inint.)) aí é que a senhora se diverte né?

EFA- {aí ... é que eu me divirto... eu só saio pra fazê compra ... já vejo o que que ela vai fazê de almoço ... **(pé pé pé pé pé pé)** dô as ord as ordens lá ... ela se vira

EXEMPLO 3

EFA-... aquele belo molho junto com a carne ... aquele molhinho a gente cozinha o macarrão separado ((faz o barulho do cozimento)) **ch ch...** queijo bastante

EXEMPLO 4

EFA- tanto qui eu vô ao futebol eu vô assisti ... (assim) ... fico um pouco ... faço companhia ... ((inint)) um pouquinho **... já tô:: s s s** ... vô fazê outra coisa ((ri)) ... num me prendo muito não

EXEMPLO 5

EFA- ... bom ... quando é um amigo mesmo ... que é m m ... amigo (antigo) gente num tem muito **lé ... nhe nhé nhé** ... cê recebe é amigo de casa ...

3. FALHAS DE ACESSO LEXICAL

EXEMPLO 1

Pistas para a análise:

Pausas de hesitação, uso de preenchedores, comentários sobre o problema.

EFA- tem us cachorro da minha fi:lha ... tem dois ... desse tamanhinho ... eu tenho a minha ... pastora

E- hum hum ... como é o nome dos cachorros?

EFA**- ... ah é o: o: ... um é Pirata** ... o otro **como é que é como é o cachorro dela?** ... eu sei que um é Pirata ... e o **outro é: ...** ai **tem uns nome isquisito que ela põe nus cachorro lá** ... **sabe qui eu nem mi lembro** ... (que eu num lido quase) cum elis

EXEMPLO 2

Pistas para a análise:

Pausas de hesitação, repetição, apoio no interlocutor.

EFA- ... eu moro no canal cinco mesmo ... **sabe ... no ... cê sabe o SESC**? ... sabe o SESC ... conhece o SESC ? né?

EXEMPLO 3

Pistas para a análise:

Pausas de hesitação, comentários sobre o problema, pistas vagas para o interlocutor.

EFA- i : ... a minha filha tem dois ... tem dois não ... tem TRES: ... é o Pirata ... que é o pai mais dois ... filhos dele ... são **uns** cachorrinho ... **daquele**: ... como é o nome ?... num é poodle não ... são um pelinho baixo assim ... mas num é ((inint.)) ... são dessi tamanho ... mas são uma gracinha também

EXEMPLO 4

Pistas para a análise:

Repetições, prolongamentos, pausas de hesitação, preenchimentos para ganhar tempo, busca explícita do item lexical.

EFA- a mãe é muito ... dispersiva dessas (com cabeça) então ... a mãe é dessas que num tá nem aí ... sabe assim ... (esse tipo de pessoa assim) eu já tenho um gênio assim que eu sempre fui muito ... seve:ra ... sempre fiz as coisas muito certa ... meus filho graças a deus ... meu filho é engenheiro ... formado pela Politécnica de São paulo ... minha filha é professora ... é professora **di di ... foi ... foi a: ... como é ?...** educação física em Santos... é professora de normal ... e tem um curso num sei do que que ela fez aqui em São Paulo também ...

EXEMPLO 5

Pistas para a análise:

Pausas de hesitação, repetições.

EFA- então eu f fui fazê a Biblioteconomia aqui em São Paulo ... mi formei aqui em São Paulo **... na na** Biblioteconomia aqui de São Paulo ... i: ... mas ... tamos cum esse problema do neto né?

EXEMPLO 6

Pistas para a análise:

Pausa, comentário sobre o problema, pistas vagas ao interlocutor.

EFA- eu tomo **um** remédio que o médico já me deu **lá** ... mas nem sei o nome do remed/ uns comprimido **lá** que ele mando fazê

EXEMPLO 7

Pistas para a análise:

Pausas de hesitação, repetições, comentários sobre o problema, busca ativa do nome, emprego de termos vazios, pistas ao interlocutor.

EFA- fui mi virei ... depois de casada ... fu/quis ... melhorá ... fui fazê facul/ ... vinha pra São paulo ... levantava a hora que levantava ... cinco hora da manhã porque eu tinha qui tá aqui **... no ... na na na** faculdade ... aqui **na na ... como é que é o nome dessa avenida qui tem aqui na ... coisa aqui** ?... ah ... onde tem **u: ... perto da Luz ali ... como é o nome daquela avenida? já até esqueci** ... tem a faculdade de so/ Sociologia e Política de São Paulo ... foi lá que eu vim fazê a faculdade ...

EXEMPLO 8

Pistas para a análise:

Repetições da formulação, apoio no interlocutor, busca ativa da forma fonológica.

E- {a senhora não lembra de nenhum ... dus que penduraram a chuteira?

EFA- ah ... Leônidas né? ... teve o Leonidas ... teve **aquele** ... como é que é o nome dele? Baltazar como é que é? num era baltazar? ... num era Baltazar ? não/ num é Baltaz ? ... é **um** nome assim parecido ... num me lembro... **tinha um** que eu gostava dele ...

EXEMPLO 9

Pistas para a análise:

Indicações vagas ao interlocutor, preenchedores, hesitações, comentários evasivos, apoio no interlocutor.

E- e qual a novela que deixou as pessoas mais ligadas nos últimos tempos?

EFA- **essa última** da Globo né? ... **qui tá levando agora** ... como é o nome? **a ...i** eu nem sei porque num presto muito atenção geralmente novela eu saio ... vô fazê **qualquer coisa** num mi pego muito ... essa qui tá levando agora essa aí às vezes eu vejo um pedacinho ... **essa qui tá levando na Globo agora ... como é que é? ... como que é o nome dela?** ... (minha cabeça num) ... também num marco né ... eu num mi ligo muito em nome de novela

EXEMPLO 10

Pistas para a análise:

Preenchedores, pausas de hesitação, indicações vagas ao interlocutor, comentários sobre o problema.

E- que que a senhora gostaria de ver nos últimos tempos? ... o que que marco a senhora?

EFA- {eu adoro cinema ... **olha ... deixo ver ...** qual foi o filme que nós vimo **bom agora ultimamente ... ai ai ai o pior é pra eu guardá us nomes viu? ... peraí** ... qui tá levando em Santos ... (cê num lembra de) nunhum nome assim? ... agora também ... num tô lembrando

4. FALHAS DE FORMULAÇÃO – FRASES ABANDONADAS NAS QUAIS A PACIENTE SE APOIA NO CONTEXTO OU NO SABER COMPARTILHADO COM O ENTREVISTADOR

EXEMPLO 1

EFA- Santos é mais calmo né ...

E- é mais calmo

EFA- gente sai : ... é mais tranqui:lo ... condução é mais fácil ... num tem **esse** problema di ... como aqui em são Paulo **essa** loucura né?

EXEMPLO 2

E- e o que mais que a senhora faz aos domingos? ... o que a senhora faz pra se divertir além de ir ao Internacional?

EFA- ... bom eu geralmente de manhã vou pra praia ... como eu num almoço em casa né? ... então ((inintel.)) passamo a manhã toda passeando ... eu gosto muito de andá ... apesar que eu ando cum problema no joelho aqui dana:do ... ando cumo dificuldade danada ... mas eu ando com dor i tudo lá vou eu ... vô andá vamos pela praia andando dize qui ... um quilometro ... do dois canais pra lá ... a gente vai e volta ...eu e ele ... minha filha já num gosta ... minha filha já gosta di ficá mais é mais molengo:na ... gosta di ficá deita:da ... vendo televisão: ... **num é num é di** ... mas: ... então saio eu e ele ((inint))

EXEMPLO 3

E- a senhora gosta de sair sozinha?

EFA- eu?... não eu geralmente saio assim co meu marido né? ... assim sozinha eu já num ... num aprecio muito não ... mas saio às vezes quando tem necessidade eu saio mas num ... eu gosto di sai em companhia eu ... ou com a minha filha ... ou co meu marido ... a gente sempre sai muito passeia muito ... viaja muito ... gosto muito de viajá

3) SUJEITO 3 (DA leve) JAM

1. FALHAS DE FORMULAÇÃO

Pistas para análise: número excessivo de pausas de hesitação, preenchedores, prolongamentos durante toda a formulação da elocução.

Pigarreia e tosse durante toda a entrevista.

EXEMPLO 1

J- ah é u shopping aí ... todo SÁBADO E DOMINGO ... é um inferno isso aqui né? às vezes sexta feira à

E- {férias né?} {ahm ahm}

noite já começa a enchê de gente aqui qui num tem tamanho... então ... é :: essa é a:: ... é a:: dificuldade

J- que a gente encontra aqui ... ((pigarreia))

EXEMPLO 2

E-... o senhor já tinha morado aqui?

J- não/...eu ((pigarreia e tosse)) ...minha mulher é que é daqui de Santos ...

E- ah sei ...

J- é:: ela sempre::... nasceu i: i:: i:: si ... i:: viveu bastante tempo aqui ...

EXEMPLO 3

E- ah sei sei ... daí o senhor morô no interior?

J- ah morei... moramos...moramos quatro ou cinco anos nu interior depois eu vim prá S... prá SãoPaulo

E- hum hum ... i o senhor morô bastante tempo im São Paulo? ...

J- ah morei ... quase toda a minha vida lá ne? ...((pigarreia)) que eu sô:: cirurgião dentista ... tenho consultório lá junto com o meu filho né?

E- {ahm ahm}

J-meu filho também é...da::...éé dentista ...i:: ((pigarreia)) i ele:: ... ((pigarreia)) i ele tem ...

J-consultório lá((inint) ... temos consultório juntos né? ... ele tem uma sala ... i tenho na outra ...i:: ... graças a Deus ... deu tudo muito bem não é? ...

EXEMPLO 4

E- morreram alguns mendigos qui tavam (morando)

J- {é qui tavam **lá ... é:: ... como se diz? tavam deitados lá** ... pra pra::: é num sei si...o que ... tavam fazendo lá né? ... então ... i eu vi isso aí ...

EXEMPLO 5

J- eu num sei purquê qui eu **nu:m ... num** tive essa: curiosidade ainda **di: di di** levá-los pra vê u u portu ... mas geralmente eles conhecem através di jorna:l ... através di: ... ((pigarreia)) ou di fotografia vista **im im im...** televisão ... ou memo im revista ...

EXEMPLO 6

J- qué dizê/ num tá bem no centro da cidade ... mais tá: ... um ou dois quarteirões lá du centro da cidade num é... ... praticamente é na cidade mesmo né? ... qui a

E- {hum hum ... hum hum}

J- cidade tá si ispalhando pro ... pra mais pro pra praa periferias né? ...

EXEMPLO 7

J- ((pigarreia)) olha:: ((pigarreia)) bom a única coisa qui: qui é: ... qui eu posso:: qui eu posso proporcioná aos meus ...familiares seria a minha parte como dentista né? qué dizê/ ... há muitos anos né? ... de forma qui a... quando eu eu eu adoecí ...é é:: então eles procuraram ...por exemplo depois u meu filho também é dentista ... u:: ((pigarreia)) u me:u ... u meu sobrinho também é dentista ... di forma qui: ... é:: ... ((pigarreia)) quando precisa essas coisas eles procuram logo:u parente mais próximo i não pra mim qui ta aqui em Santos né? ... seria mais difícil... mas quando tava im São Paulo: i t i/ i ti i tenho consultório lá im São Paulo ainda ...eles procuravam ... mi procuravam lá em São Paulo ... agora aqui ... é que eu vim pra cá: ... então num mi procuraram mais ...

EXEMPLO 8

J- u meu papel é:: si pricisá alguma coisa assim di di di

E- {retaguarda ...

J- é retaguarda ... entã:o me consulta mi:

EXEMPLO 9

E- {faz de conta qui eu num sei nada di café ...

J- ((pigarreia)) bom eu teria qui esquentá a água ... fervê a água ... i depois pricisa vê quantas ... quantas xícras eu vô ... vô usá' ... pra por di água pra podê fazê dipois u café ... pras pessoas qui quiistão ... qui istão mi visitando né? ...

EXEMPLO 10

J- eu sigo a minha mulher também ... si ela põe uus café/ ... éé t vamos supor ... três quatro xícaras... põe tanto di água ...tantas...culher di di di: di café i depois

u açucar faz na hora ... qui é ao gostu da da: ... vamos dizê assim ... ao gosto du freguês né?... cada um aç adoça como qué ...

EXEMPLO 11

E- i si u senhor tivesse qui si virá sozinho fazendo uma macarronada como é qui o senhor faria?

J- bom ai tem qui isque tem qui isque fervê bem a água ...((pigarreia)) lava u u u: macarrão ... i depois põe na água fervendo pra pra ... pra: praamolecê ... ((pigarreia)) ai a a parte di di do molho eu num mi lembro porque poco eu vô à cozinha pra fazê mas em todo o caso tem o o molho pra macarrão né?

EXEMPLO 12

E- como que é o seu dia?

J- eu gosto di vê muita vitrine né? vitrine a:: ((pigarreia)) a::enfim a: ... s si tem uma mudança na na na: nu vestiário na:: né? ... isso eu gosto di vê ... di resto num tem muita assim ... muita:é::... ((tosse)) num gosto de saí da m da da rotina ne? ... eu acho qui aquilo ... pra mim tudo dia é quase igual ... então só nu sá/ nu domingo sábado i domingo qui eu gosto di saí um pouco da rotina ...

EXEMPLO 13

J- {é ... purque ele ele ele ele é ...ele quando fala ...

E- {malabarismos

J-é ... quando ele fala ele fala cu:m ... cum convicção das coisas né? então... i:sso impressiona muito u u u povo né? principalmente u povo ... u povo assim di di di: qui num tem muita instrução... eu acho qui isso aí é ruim pru país né?...

EXEMPLO 14

J- muita gente pricisô:a: apelá purquê: quando ele istava nu comando quando ele saiu di: ... é:... quando ele veio praqui pra pra: pra São Paulo teve um negócio qui piorô né? São Paulo ... u correio ... então ele se tornô um ditá ... como se dissesse

E- {hum }

J- ... quase qui um ditador né ... qui ele queria impor a ... as coisas dele né?... as idéias dele ... acho qui isso num é ... num foi bom não ...

LACUNAS NA FORMULAÇÃO

J- bom então é esse qui esse é esse esse qui istá indo em cima do quiis [??????????] do do branco né? ((pigarreia))

INSERÇÃO DE PREENCHEDORES

"vamos dizê assim", "qué dizê"

EXEMPLO 1

J- eu tenho:: ... ((pigarreia)) eu tenho muita confiança nu:: ... aqui no prédio i i na: ... **vamos dizê assim** ... na administração du prédio ... né? **qué dizê**...tomam todos os cuidados possíveis i... i até: as vezes hum é é: impossível né? i:: i nu resto tá tudo muito bem né? ... porque num ... eu num eu num corro nenhum ...((pigarreia)) num fico cum medo di nada ... nunca sofri:: uma uma uma ((pigarreia)) **vamos dizê assim** uma:: ... ô entrá ladrão ou assaltá qualqué cois(a) nunca sofremos nada aqui em Santos ... graças a Deus não ...

EXEMPLO 2

J- ((pigarreia)) olha:: ((pigarreia)) bom a única coisa qui: qui é: ... qui eu posso:: qui eu posso proporcioná aos meus ...familiares seria a minha parte como dentista né? **qué dizê**/ ... há muitos anos né? ... **de forma qui** a... quando eu eu eu adoecí ...é é:: então eles procuraram ...por exemplo depois u meu filho também é dentista ... u:: ((pigarreia)) u me:u ... u meu sobrinho também é dentista ... **di forma qui:** ... é:: ...

EXEMPLO 3

J- eu sigo a minha mulher também ... si ela põe uus café/ ... éé t vamos supor ... três quatro xícaras... põe tanto di água ...tantas...culher di di di: di café i depois u açucar faz na hora ... qui é ao gostu da da: ... **vamos dizê assim** ... ao gosto du freguês né?... cada um aç adoça como qué ...

TERMOS VAZIOS

EXEMPLO 1

eu num eu num corro **nenhum** ... ((pigarreia)) num fico cum medo di **nada** ... nunca sofri:: uma uma uma ((pigarreia)) vamos dizê assim uma:: ... ô entrá ladrão ou assaltá qualqué cois(a) **nunca** sofremos **nada** aqui em Santos ... graças a Deus não ...

EXEMPLO 2

E- que que o senhor ... porque que o senhor mudô? só curiosidade minha...

J- eu acho qui o o estadão é mais completo no noticiário né?

E- noticiário de...

J- **tudo ... de tudo de tudo** ... ((enfático))

EXEMPLO 3

J- mas ... essa daí ele ... si ele si ele pudé: voltá **no coisa** ... ele volta pru governo i di distruição mesmo ... eu acho isso aí um perigo pra: pra aqui pru país né?...

REPETIÇÃO DE SEGMENTOS

COM APOIO NA FALA DO INTERLOCUTOR

Pistas para análise:

O paciente repete o que o entrevistador pergunta como forma de apoio e também repete as frases emitidas, como forma de preenchimento.

Além disso, mantém as hesitações e prolongamentos.

EXEMPLO 1

E- Santos agora já é cidade grande né?

J- ah é ... isso aqui ... isso aqui ... já é:: muito grande ... i já:: é: tem uma vida::

E- {já é uma cidade grande ...

J- ... vamos dizê assim é:: independente né? ... muito bom ...

EXEMPLO 2

E- {hum hum} {i o senhor acha qui o progresso ... atrapalha a vida das pessoas? ... o progresso de uma cidade grande? assim ... como Santos?...

J- {ahm:::::eu acho qui não ... eu acho qui não ...

EXEMPLO 3

E- {é} { qui coisa né? ... qui coisa mais ... terrível né? ...a gente pensa qui é só em outros lugares...na Tailândia qui caem muitos prédios ...

J- é ... na Tailândia sim ... lá ... ...

E- diz qui é mal qui são mal construídos né?

J- lá eu tenho a impressão qui:: qui...

E- {não tem fiscalização ...

J- é eu acho qui num tem fiscalização ... lá lá é Deus ... Deus é qui salva a situação

EXEMPLO 4

J- u meu papel é:: si pricisá alguma coisa assim di di di

E- {retaguarda ...

J- é retaguarda ... entã:o me consulta mi:

2. RESPOSTAS EVASIVAS e VAGAS

EXEMPLO 1

Pista para análise:

"NÃO SEI" , "NÃO GOSTO..." "ACHO QUE FOI BOM" "NÃO SENTI NADA" , "EU NUM PRECISO FAZÊ:: NADA MAIS DO QUE AQUILO QUI EU QUE EU TÔ ACOSTUMADO FAZÊ

E- i o outro lado do progresso? violência ... essas coisas? ...

J- ah eu acho qui isso aí eu:: ... nu:m sei ... eu num tenho ... convivido assim muito não com essa orla aqui ... eu fico aqui em casa num ... num ... num gosto muito de saí ...minha mulher também num gosta muito de saí ... de formas qui o qui a gente ouve é qui:: ... eu acho qui foi bom viu? ...

EXEMPLO 2

E- i:: já aconteceu algum proble::ma com o senhor? relativo a esse ... esse aumento de população ... alguma coisa assim? ... não?

J-**não ...eu não senti::...absolutamente nada** ...nem em São Paulo...quando eu estava lá

EXEMPLO 3

E- nem em prestação de serviços o senhor num sente problemas ...assim... serviços municiPAis ... serviços ...

J- não eu (tô) ... num num atuo nessa nessa área ... então ... ah ... eu ...

E- {na sua casa num aconteceu di acabá a lu::z? ... ou di:: ...

J- **{não ... é difícil isso aqui acontecê** di di apagá a luz ... isso é coisa rápida e volta logo ... num tem ... num tem problema assim ... é maiores prá ... é perturbá: a vida::da gente é: aqui na cidade ... né?... a cho qui é isso ...

EXEMPLO 4

E- morreram alguns mendigos qui tavam (morando)

J- {é qui tavam lá ... é:: ... como se diz? tavam deitados lá ... pra pra::: é **num sei si...o que ... tavam fazendo lá** né?

EXEMPLO 5

J- i nu resto tá tudo muito bem né?

EXEMPLO 6

E- quando o senhor recebe as pessoas de Jaú pra cá? ... quais as providências qui o senhor toma aqui pra recebê-las?

J-a:h e:eu... geralmente o meu pessoal me conhece i já sabe como é qui é a minha vida aqui ... então ... eu **num preciso fazê:: nada mais do que aquilo qui eu que eu tô acostumado fazê** ... recebo da mesma maneira i::como se eu estivesse lá em Jaú ... eu vindo pra cá ... **mesma** ... i a recepção é é recíproca né?

EXEMPLO 7

intão: **num si interessa pur nada ... num me pergunta nada também...então eu também num vô ...tocá nu assunto né? ...**

EXEMPLO 8

E- i esse técnico aí o que o senhor acha?

J- {((ri)) esse técnico ((rindo)) aí eu num sei ... ele tá meio:: ele tá meio:: ... acho qui ele tá é meio confuso ainda viu? ...

E- {é?}

J-num sei se é porque ele pegô agora assim di sopetão ... mas ele tá meio confuso ... ele num tá:: ele **num tá muito: seguro da da da da: situação não né?...**

E- o senhor acha qui ele escalô algum perna di pau?

J- ((pigarreia)) ... olha num num ... num diria ... é ele ele:: ((tosse)) **ele num tá muito assim seguro da das funções dele** ...porque **eu acho qui u camarada tem qui fazê ... principalmente técnico de de esporte ... ele tem qui abrangê uma uma uma: toda a extensão da do esporte** num é isso?

EXEMPLO 9

E- i: dos jogadores do passado? ... tem algum qui era seu preferido?

J- nã:o eu eu era era ... **todos** pra mim eram bons viu?... **TODOS** por aí

3. FALHAS DE ACESSO LEXICAL

EXEMPLO 1

J- não/ saiu otra vez... saiu novamente uma uma reportagem a esse respeito né? qui us prédios istão inclina:dos qui... i qui si... num tem probabilidade ...

agora i aaliás ATÉ ... us engenheiros já falaram a esse respeito ... num tem porque ...

u u aqui ... Santos **é um ...t.. é é::... é sobre uma água ... um ...** aliás uma uma camada de água né?

EXEMPLO 2

J- bom...quando tem oportunidade deu ir prá São Paulo eu vô pra lá...eu vô pra pra casa du meu filho né?... i às vezes eu vô pra pro interior **pra:: pra Jaú** qui é onde ...eu tenho...ãh ãh...meus pais tão enterrados lá né?...i tem...i eu tenho uma irmã qui mora lá...então eu faço visita pra ela lá...u resto é só São Paulo

EXEMPLO 3

i:: onde mais poderia i? ((pigarreia)) depois da Ponte Pensil lá **lado di di como é qui chama aquela Grande né?** ... num tem muita coisa pra si fazê aqui ... ii geralmente eles conhecem já a vida de Santos né? ... us qui moram lá em Jaú i que me visitam né?

EXEMPLO 4

E- ham ham'... como é o seu ... u seu domingo aqui em Santos? macarronada lembra um pouco domingo né? como é que é o seu domingo aqui em Santos? ... como é a sua rotina?

J- {((pigarreia)) {((pigarreia))} {geralmente ... geralmente eu eu vô:: eu almoço fora né? ... qué dizê...ou eu vô lá nu ban ...nu: nu restaurante du Banco du Brasil ou então nessas ... nessas ah::m casas qui servem refeição assim por quilo né? ... intão eu vô lá i almoço como o que::... o que for ... **u qui tivér di:** oferecendo né? ... i:: eu almoço assim ... num sô muito **di: di: comê assim comida muito ... comida muito a a:: como é**? ... temperada muito: ... eeu espero qui a minha mulher faça pra mim ... né? eu espero ...

EXEMPLO 5

E- qui qui o senhor vê na Globo?

J- ((pigarreia)) ah todo noticiário qui tem da Globo ... TODO o noticiário da Globo eu

E- {noticiário?}

J- ouço ... i da Manchete também a parte qui qui ... o esporte a parte di di ... eu ouço noticiários (qui num tá) a Globo as vezes ...falha eu pego na na na na:: ((pigarreia)) ...no canal 13 né?

EXEMPLO 6

E- hum hum ... hum hum ... o senhor algum outro tipo de programação na televisão? ... o senhor chega a vê alguma novela ... algum seria:do? ...

J- ... bom agora ... agora tá surgindo **aí a ...a a como é que chama**? (a inint)...a Babalu né? tá aparecendo agora ... o companheiro dela **que é o o... num sei eu num guardo totalmente bem o nome dos dos artistas né**? ... ((pigarreia)) o Raí né?

EXEMPLO 7

E- tem algum bandido nessa novela?

J- ... aquele qui é:: bigodinho né? ... **o nome dele? ... é u: é u: ... eu num sei u nome é u:m eu num lembro...totalmente**

EXEMPLO 8

E- que que o senhor achô du impeachment do Collor?

J- bom ... chegô em boa hora viu? ...

E- é?

J- é ... ele era: ... ele queria... ele queria transformá **isso aqui numa: ...** ele queria se ditador ... eu acho qui ele queria sê ditador ...i ele foi ... ((pigarreia)) iele fo:i ...barrado na hora certa viu? ... pra mim... ele ia ... **ele ia torna isso**

aqui uma: uma: ... como é que se diz? um segundo:((tosse)) **onde tá o Fidel Castro né**? a::((pigarreia)) eu acho qui ele teve ... ele foi barrado na hora certa viu?

4. FRASES ABANDONADAS

J- não/ saiu otra vez... saiu novamente uma uma reportagem a esse respeito né?

qui us prédios istão inclina:dos

qui... i qui si... num tem probabilidade ...

agora i aaliás ATÉ ... us engenheiros já falaram a esse respeito ...

num tem porque ...

u u aqui ...

Santos é um ...t.. é é::... é sobre uma água ... um ... aliás uma uma camada de água né?

... então ... essa camada pode está(r) influ ... tendo influência du du du:: ... ou da elevação da da:: água ou da diminuição ...

ah u prédio deve sofrê alguma coisa né? ...

isso que eu tenho lido também né? **... não sei se...**

5. MUDANÇAS DE DIREÇÃO

J- eu vô cum ... eu vô di ônibus pra lá ... ((pigarreia))

E- i onde o senhor leva pra passeá? aqui em Santos? ... quais são os seus

J- não não eu ... aqui só ... eu só as visitas assim... ... qué dizê assim nu Shopping né? ... vê u Shopping qui tem aqui ...

6. MUDANÇAS DE DIREÇÃO COM REFORMULAÇÃO

J- às vezes eu vô pra pro interior pra:: pra Jaú qui é onde ... **eu tenho** ... ahm ahm ... meus pais tão enterrados lá né? ... i tem **... i eu tenho uma irmã qui mora lá** ... então eu faço visita pra ela lá ... u resto é só São Paulo ou...ou pra Jaú

REFORMULAÇÃO COM MÚLTIPLOS ENSAIOS

EXEMPLO 1

J- AH é fácil ... tem ônibus toda ... num sei di... acho qui di...di tr... duas ou três horas ... tem ônibus pra lá ...

EXEMPLO 2

E-quantos anos ele tem?

J- ele tá cum ...cato:rze ..... treze catorze anos..... catorze quinze anos ... então tudo:: pra ele é é ... ele é MAIS ...mais vivo assim do que a minha filha...do que minha fi/ ah minha filha ((fala rápido))a minha neta ...((pigarreia))a minha neta tá cum dezessete mas num é assim tão ligada nu computador não ... agora ele não ...

4) SUJEITO 4 (DA moderada) CAP

1. REPETIÇÃO DA QUESTÃO FORMULADA – DIFICULDADE DE COMPREENSÃO

EXEMPLO 1

E- acha difícil? como é que é a sua cidade?

C- a minha cidade ? ãhm: agora eu moro no interior né?

EXEMPLO 2

E- ih:... a senhora sai muito?

C- se eu saio? eu saio sim... um pouco...

EXEMPLO 3

E- que que a senhora faz no domingo?

C- nu domingo? agora a gente... come muito pouco... a a tudo que queé cum carne com isso... com aquilo...

EXEMPLO 4

E- a senhora correu muito pra chegá até aqui?

C- hoje?

E- é

EXEMPLO 5

E- i:: como é... que é sua casa Dona dona C.?

C- tem mais? a minha casa?

E- {é} {descreva pra mim a sua casa?}

C- é::... no no momento eu... estou sem casa... mas tem casa...

E- no momento a senhora?

C- no momento eu já destrincho logo:... i acabo...

E- mas a senhora tava falando da sua casa?

C- é da casa... também se alguem me chama assim se eu saio correndo pra i atendê a porta ... tem coisas que sim né? sse eu to isperando uma ... u um resultado importante:... na hora que toca a capainha ou telefone a gente corre vai lá lê...vê o que que é... o que que não é né?

E- hum hum...

E- e a sua casa? descreva pra mim a sua casa/? como que ela é?

C- minha casa? minha casa... é uma casa di::... uma casa que tem madeira no chão ... tem m tem madeira também assim nas portas... i que mais? é muito simples ...tem... tem faxineira que vem ... mais ou menos assim di di... duas a três semanas vem porque se não fica perdido né? e aí num... trabalha mais direito né? si mandá vim todo dia ... i ... de vez em quando eu dô umas ...olhada cum ela... pur tudo aquilo pra vê se... ficô limpo si num ficô... a gente muda né ?... depois... e (assim se é) a vida né? di... di dona de casa né? qui num é muito bom não...

E- {hum}

E- quantos quartos tem a sua casa?

C- minha casa? tem... acho que é cinco ou seis... porque temum tem um... a gente fez uma reforma...i: como é que se diz? aumentavam acho que dois... três quartos... então aumentô... mais um pouco...

E- hum...

C- jáagora já terminô... já tá funcionando...

E- i como é que é u resto da casa?

C- u resto da casa? é ...ssão... mmas umas três... como é que se diz? uns três quartos... uns dois quartos por aí... o que num ta: do jeito que a gente gosta... qui acha qui num tá bem porque num ficô bem... a gente já ... in indireita antes di entrá ... i:: intão é isso ..a gente mesmo é que resolve o problema...

E- hum hum

C- porque tem que sê assim né? se não tá perdido né? ((ri))

2. DIFICULDADE DE COMPREENSÃO DA QUESTÃO – ORGANIZAÇÃO DA RESPOSTA A PARTIR DE FRAGMENTOS E RESPOSTAS VAGAS E EVAZIVAS

EXEMPLO 1

E- que que a senhora achô da saída do Collor?

C- do Collor? eu achei é boa... porqueeu achei que fosse melhor... mas num ...num é não...

E- {hum... achô que era isso mesmo...}

C- eu achei qui podia tê sido melhor...

E- é? como poderia ter sido Dona C.?

C- ah... ele poderia te:r... ficado com um cargo de mais responsabilidade né?

E- {hum}

EXEMPLO 2

E- i:: i o resto da política? das eleições ? que a senhora... a senhora diz... tem acompanhado?

C- hum hum

E- acompanha?

C- {acompanho sim...

E- e qual é sua posição? que que a senhora acha di tudo isso qui tá acontecendo?

C- eu acho qui melhorô... mas também qui num foi muito legal não...

E- não? ((inint))

C- {(num foi não) eles demoraram muito pra decidí...

E- decidí? a parte... qual parte?

C- ué? a parte deles... né ?queeles... cada um recebeu... uma um um xis né? i eles é que tinham que... estudá entre eles né? mas eu acho que saiu bem sim...

E- {hum hum}

C- acho que saiu bem sim...

EXEMPLO 3

E- mas então ... Dona C.? mas eu queria pensá assim mais na sua parte? e senhora pra podê ficá lá cum ele? a senhora tem que tô/ com essa sua amiga... a senhora tem que tomá providências... em relação às suas coisas...

C- {lógico porque eu tomei também pra num deixá tudo de qualquer jeito?

E- {e pra/ em relação à sua saí:da? é:: pra chegá até o loca:l? pra podê ficá/? como é que é essa organização da viagem ... passo a passo?

C- cada um ((ininteligível)) de um jeito... muita gente qui num *anda sozinha* né? e eu acho certo... você num pode *andá sozinha*... São Paulo é uma cidade... uma cidade muito difícil... né? num pode... agora você tem tem sempre qui tá *dando uma arranjadinha* porque... pra... ssabê si si tem que fazê isso ou assado... si a faxineira fez ou si num fez...se ela/ né? n˜fao num é muito legal não... tomá conta da casa viu?

EXEMPLO 4

E- ih::... o que que a senhora... que que a senhora vê? televisão? mais jornal? rádio?

C- o jornal... às vezes tem (nacional) a gente tem umas... tem uns trechos assim...bons queeles passam na televisão é interessante... bem... a gente varia... mas num gosto também di tá andando pela cá/... pela rua ... assim... feito gente abandonada não... num gosto...

E- {hum hum} {hum hum}

C- eu gosto de andá um uma pessoa junto de mim... em caso de perigo tê uma pessoa né?

E- {sei}

E- e na televisão? a senhora assiste televisão?

C- hu::m nada...

E- nada?

C- não

EXEMPLO 5

E- Dona C. e futebol?

C- futebol?

E- a senhora...

C- eu assisto quandué uma ma ma ma ma uma... como é que se diz? um dia... qui tá bonito... ensolarado... eu gosto...

EXEMPLO 6

E- ih? pra se diverti que que a senhora faz no domingo?

C- no domingo? fico mais em casa do que saio...

E- é? mas o que que a senhora faz em casa?

C- {é }

C- em casa? eu mexo u dia inteirinho... vejo se as gavetas tão em or:di... vejo si

C- aquela COisa que eu vi na rua... queeu achei muito bonita... vale a pena comprá ou num vale ... sabe? essas coisas de mulher...

EXEMPLO 7

E- sei... ih ...i na segunda feira?como que é a sua segunda feira? ((ininteligível))

C- {segunda feira é:...

C-levanta até mais cedo... na segunda feira ... que é por isso... tá também se já deixa tudo pronto né?

C- {a roupa...

E- mas u que que a senhora faz? qual/ i depois? u resto do dia? como que é?

C- u resto du dia? ah::? u resto du dia mexe... mexe a ca/ a ca/... cê já viu

C- alguém ficá parado tem/ tem casa... né? a gente mexe... i aí num tem muito

E- {hum}

C- companhia d du marido... que o marido tá trabalhando... então a gente ...fica sozi:nha

3. DIFICULDADE NA APREENSÃO DO SENTIDO METAFÓRICO

EXEMPLO 1

E- ih:: ((pigarreia)) ahm quando o motorista perguntô:... a que altura i... o que ele queria sabê? quando perguntô?

C- ele ele num teve chance porque... eu quando ele falô... ele pode errá pode errá u que vai fazê eu já tinha começado e já destrinchava... não esperava ele chegá lá ...

EXEMPLO 2

E- Dona C.? que que significa apertá o cinto pra senhora?

C- apertá o cinto?

E- é...

C- é apertá o cinto né ?

E- como assim?

C- si tá fazendo muita arte com o cinto... deixa ele um pouco de lado i i discansá um pouco...

4. IMPRECISÃO DE TERMOS – TERMOS VAGOS – FALHAS LEXICAIS

EXEMPLO 1

E- uma vez na minha cas u cabo de eletricidade em frente... entrô em curto i ... houve uma sobrecarga i ... todos os meus aparelhos eletrodomésticos se queimaram ... como é que a senhora resolveria essa situação?

C- ah/ eu pegaria u telefone da du du... como é que chama da lista... u telefone... i telefonava né? se ninguém vai... se ninguém me a me atendesse eu ficaria insistindo

E- pra onde a senhora telefonava?

C- pra pra *pra onde eu to/ estô* né?

E- ahm... i quem é que iria resolvê u problema?

C- quem ia resolvê ? *é um funcionário de alta categoria... dentro da:... dentro do ambiente* né?

EXEMPLO 2

E- mas então ... Dona C.? mas eu queria pensá assim mais no no s s na na sua parte? e senhora pra podê ficá lá com essa sua amiga... a senhora tem que tomá providências... em relação às suas coisas

C- {lógico porque eu tomei também pra num deixá **tudo** de **qualquer jeito**...

E- {e pra/ em relação à sua saí:da? é:: pra chegá até o loca:l? pra podê ficá/? como é que é essa organização da viagem ... passo a passo?

C- cada um ((ininteligível)) **de um jeito**... muita gente qui num anda sozinha né? e eu acho certo... você num pode andá sozinha... São Paulo é uma cidade... uma cidade muito difícil... né? num pode... agora você tem tem sempre qui tá dando uma **arranjadinha** porque... pra... ssabê si si tem que fazê **isso ou assado**... si a faxineira fez ou si num fez...se ela/ né? não num é muito legal não... tomá conta da casa viu?

E- ((ri)) não é?

C- prefiro comprá comida pronta

EXEMPLO 3

E- ih... pra se diverti que que a senhora faz no domingo?

C- no domingo? fico mais em casa do que saio...

E- é? mas o que que a senhora faz em casa?

C- {é }

C- em casa? **eu mexo** u dia inteirinho... vejo se as gavetas tão em or:di... vejo si

C- **aquela COisa** que eu vi na rua... queeu achei muito bonita... vale a pena comprá ou num vale ... sabe? essas coisas de mulher...

EXEMPLO 4

E- i qual... i qual que é a sua atividade principal atualmente?

C- atualmente... é num sei viu? porque é **tanta coisa**... eu **tive tanta gente** nas férias ... cê funciona... cê **mexe** né? mas correu tudo bem... graças a Deus... deu pra gente conversá um pouco... como é qui ... o que queeles andam fazendo de melhor né? porque a gente tem sempre... oportunidade de falá... i ocês já tão... indo mais na igreja? o num tão? que que é ? i que que padre... que ocê acha melhor? qui agora tem uns padres muito bons... te::m... tem mui::to...

E- {é?}

E- qual é o padre que a senhora acha bom?

C- eu acho (queeu tenho) uns quatro ou cinco...

E- que tipo... que que é um padre bom pra senhora?

C- um padre bom é aquele queeu/...por exemplo... si prici... si eu tenho necessidade de **alguma coisa**... eu chego até ele (ininteligível) tá me acontecendo... tá me acontecendo **isso**... o que que eu acho/ o senhor acha que é o melhor? ....

EXEMPLO 5

E- i quais são seus projetos?

C- meus projetos?

E- é

C- são modestos... porque... eu num gosto de nada extravagante... di levá... desse tamanho *assim*... numa viagem queeu vô fazê... na nada *disso*... eu quero levá *assim*... u u poco procê num ficá preocupada com *aquilo*... i aquela *coisa*... porque *isso que aquilo*... u que eles fazem... na na nas viagens... é é um horror né? eles arrancam

tudo... depois *assim* (não acha) u otro tira... ((inint))

EXEMPLO 6

E- nada? a senhora prefere o seu tempo de mocidade?

C- não... não é que a gente... queeu prefira... mas... *uma coisa* qui:... qui já

C-tivesse bem estado melhor né? tá sempre com aquela *mesma coisa... nu sai*...

E- {hum}

E- e quando o técnico escala um perna de pau?

C- bom ... esses tempos ocê num qué acabá de *enxergá ...* cê só sai i vai embora ...

EXEMPLO 7

E- como é que é? mas eu quero sua receita... passo a passo...

C- a primeira coisa que tem que fazê é:... vê se você tem ingredientes... pra pra começá a fazê né?

porque ((inint)) *como é que chama*? u pacote né? qui tá duro... cê tem que tomá u expediente né? pra vê o que que vai nu tempero... e isso e aquilo... si vai mantega si vai queijo si num vai né? cê tem que... é *sei lá ... uma coisa* porque se não... a comida fica muito esquisita né? ainda mais pra pessoa qui trabalhô u dia inteirinho né? a gente ajuda sim...

E- hum hum

EXEMPLO 8

C- café? café: eu tomo: muito pouco café... num tomo mais muito café não...

E- mas e se chegá alguém?

C- aí u: a a... faz... GRAças a Deus nós temos *aqueles que cê faz assim tim tim tim* ... i já tá tudo pronto o café... num tem problema

E- mas como é que é a receita? fala?

C- a gente né já... u sSEMpre cê tem que tê uma ma vasilhinha já prontinha pro cê ... botá u pozinho dentro du café...né? cê num pode i ainda lá... mexê nu café... pra vê si i... si precisa *ralá mais... si num rala... essas besteira né*?

E- { hum hum}

C- então cê já vai ca pronto... ch quando cê sabe que a família... que vai chegá a família cê já deixa até dentro du:... *como é que chama*? d dentro da xícara ... mas aí ... já é é um problema... porque... um gosta disso... um gosta daquilo cê num pode botá u igualzinho em todo lugar né?

EXEMPLO 9

E- acha difícil? como é que é a sua cidade?

C- a minha cidade ? ãhm: agora eu moro no interior né? então a vida é mais...

E- {isso?}

C- tranqüila... mas quando a gente tem algum problema que tem que i prá São Paulo... já é um problema maior ainda né? fica difícil**... intercalá** tudo que tem pra fazê né ? ((inint))

EXEMPLO 10

E- ahm ahM... e a senhora acha que o progresso atrapalha a vida das pessoas ?

C- tem umas coisas... que atrapalha né?

E- hum... o que que atrapalha?

C- que a pessoa as vezes num num... como é que se diz ? num tem confiança cum u que qui tá falando... num fala o suficiente né?

E- hum?

C- então pode prejudicá um pouco né?

E- hum... certo...

C- mas eu acho que **frequentando** um pouco antes um pouco... ele já vai... vai como é que se di:z ? vai trabalhando normalmente e vence...

E- vai trabalhando normalmente? e vence

C- { é...é ...i vence}

5. HESITAÇÕES

EXEMPLO

E- ham ham... pra que time que a senhora torce?

C- eeu sempre torci... pru *como é que chama*? mas ele agora n num tá... trabalhan/ num ta::... num tá muito bem bem de saúde...então num tá funcionando bem... sabe? mas... ele é bom...

5) SUJEITO 5 (DA moderada) RLCO

1. DIFICULDADES DE COMPREENSÃO E PRODUÇÃO DO TEXTO ORAL.

DIFICULDADE PARA RETER A PERGUNTA DIFICULDADE EM COMPREENDER E ORGANIZAR A ENUNCIAÇÃO SEGUNDO EIXO SEMÂNTICO

EXEMPLO 1

E- qual o ... qual o ... que tipo de artista a sra prefere ... os que fazem papel de ...

E- mocinho ... ou os que fazem papel de bandido? ...

R- depende de cada um né?... **fala otra vez a pergunta ...**

E- é:: quê ... que tipo de artista a sra gosta? ... artista de novela? ... por exemplo ... os que

{não tenho paciência ...eu não tenho...}

E- fazem...

E- não? ...

EXEMPLO 2

E- comigo aconteceu assim ... uma vez u cabo de eletricidade da rua da minha casa entrou em curto ... daí todos os aparelhos da minha casa qui funcionavam

E- em 110V queimaram ...

R- é? ...

E- daí eu liguei pro corpo de Bombeiros mas eles num quiseram atender u meu pedido ... daí fiquei sem sabê u qui fazê ... u que a sra faria nesse caso? ...

R- eu pe... pegava um ... pur exemplo na na na is na istrada ... ((tosse e funga)) a gente pega i pede pru:: ... u: ... alguma um ... alguma coisa qui tem ... qui sabe ... como é que chama? ... é :: ... porque tem gente qui fica na istrada né?

R- ... é gente dio sitio ... então nós ficamos lá um tempão ...

E- naquele lugar? ...

R- é ruim purque ...a gente tem medo né?...eu tenho medo... passei muito medo ...

EXEMPLO 3

E- já aconteceu alguma coisa desagradável pra sra aqui na cidade grande? ...

R- a quando na na :: ... ((ininteligível)) onde nós fomos ... no iscã nu iscãd qual que foi?... o que é que foi mesmo? u que queu ia falá ? ... pera aí fal... fala otra vez...

EXEMPLO 4

E- e segunda-feira o que que a sra faz ? ... conta um pouco ... assim ...

R- bom ... i na escola todo dia ... né? ... depois ... depois ((inint)) o Ignácio é meio doido ... as vezes ele ... chega lá ... fala assim vô combiná/ ... combina cum num sei quem ... i vai embora ... i gente vai

nu nus parente...pegá dus otro...((ri)) não/... por por isso a gente...trabalha {((ri))} nu...falá cum eles né? ... a:: u meu marido tem uma :: ... uma mulher dela ... agora (a coitada) tá doente ... (inint) ela já é velha e ela ... tá doente ... num sei ... preciso vê o que que é ...preciso vê ... diz que ela qui num estava bem ... ... acho

E- {hum hum}

R-agora nós vamos dá uma chegá ... um chego lá ((initeligível))...

E- {hum hum}

EXEMPLO 5

R- agora eu esqueci ... mas não é ... minha cabeça é assim mesmo ...

{esqueceu?}

R-eu falo aQUI ... quan ... seu faço uma coisa ... eu falo com uma pessoa uma coisa

R-quando eu chego aqui eu num tenh ... eu já esqueci ...

E- por que a sra acha que acontece isso? ...

R- ... é que :: ... eu per ... eu perco ... num lembro num lembro da coisa ...

E- é?

R- é ... a cabeça num ....u meu pai ...que tinha isso... meu p meu p meu pai ... ele ele eu f

R- ... eu num sei se eu falei procê do meu pai ... faz muito tempo né? ... ((inint))

E- {ahm}

R- mas ele no fim ... ele trabalhava ... num ... engenho ... engenho de a-açucar ... na QUE le

{a;hm}

R-tempo ... nós tínhamos um ... nós tínhamos um muito bom na nossa casa ... gran ... linda tinha

E- {puxa}

EXEMPLO 6

E- e assim ... qual a novela qui deixô ... qui mais deixô as pessoas ligadas nos últimos tempos? ...

R- us tempos ? ... Jesus Cristo ...

E- é? ... por quê?

R- porque a gen a gente tem que tê **alguém** ... que a gente vê: ... que Je ... que ele que a gente

R-fala ... (**ó i... começô a cabeça**) ... eu acho que Jesus é no melhor ... que deve (inint) que: que:

R-... que ajuda a gente ... que ele dá ... eu acho que é um: ... que a gente precisa de alguma pessoa ... uma f de uma figura ... que a gente precisa ter um ponto de ... pra melhorar né?...

E- hum hum

R- cê vê ... quem ... quem num tem nada nada ... que que tem? ... não tem nada de si? ... (então

R- nunca vai sê melhor ...

E- hum hum

R- vamos? (vamo andando)

2. DIFICULDADES NA ORGANIZAÇÃO DE SCRIPTS (ESTABELECIMENTO DA SEQUENCIA DE PASSOS, ESTABELECIMENTO DO EIXO SEMÂNTICO BÁSICO PARA ORGANIZAÇÃO - MACROESTRUTURA)

EXEMPLO 1

E- i cafezinho? ... como que a sra. ... gosta du café? ...

R- {ele ele ele que ele gosta eu também gosto ...

E- é?

R- mas eu faço pouco né porque num é bom pra mim... (logo)

E- {mas como que a sra gosta du café? ...

R- mais forte ...

E- forte? ... hum hum ... eu também prefiro forte ...

R- é ...

E- igual o SEU jeito de fazer café? ...

R- eu pego a água ... pego o café (vai) **vai fica fica a** água né ? ... depois eu pego ... um pouco do pó.... **daí fica meio assim assim** ... **depois põe na na coisa**...

DESRESPEITO A PRINCIPIOS DE ORGANIZAÇÃO DA ENUNCIAÇÃO – ESTABELECIMENTO DE REFERENTES E SEQUENCIAMENTO DA INFORMAÇÃO

EXEMPLO 2

E- hum hum... e tinha m/ animais na chácara?...

R- lá te:m ... cachorro ... eu **tive tudo queu tinha** ... quando ma mo morria um ... **eu ia** na escola ... a pé né?... porque num ... então quando mo morria um ... o primeiro **que ela via** pega um cachorrinho i levava pra casa ... nunca fiquei sem um cachorro

R- ... eu ainda tenho dois ...

EXEMPLO 3

E- elas num entram em casa? ...

R- ela fica só no ... quarto da menina ... qui **ela fica com ela** ...

E-ah:: a gata fica só no quarto?

R- { qui é .../

EXEMPLO 4

E- eles num resolvem né?

R- num é ... nem ... num qué sabê ...

R- ele ele ... ele ... ele namorô ... namorô uma moça ... depois ele num quis ela ... **num qué ela** ... **num quis sabê dele** ... não ... não ... é/ num quis sabê...

E- aí tão solteiros os dois ...

R- os dois ...

E- nem nem namorado eles num tem? ...

R- não ... ele ele namorô ... namorô mas ... mas ...

E- desistiu ...

R- não ... ele ele foi ... ele foi pra:: namorá ... chegô lá ... pa pra pega a namorada ... ele

R- ... ela falô cum ele ... ele chegô lá ... ele chegô pra ela ela ... (pra pra namorá)chegô lá

R- **a avó** ... ele largô dela e foi lá vê a: **lá vê ... (na vê da moça)** ...

E- o quê ?...

EXEMPLO 5

E- sra viesse pass ... pra cá ... pra sua casa ... é :: como a sra se organizaria pra

E- recebê-la? ...

R- ah: ... eu sabia ... a gente geralmente ... se a gente é :: a gente s s se (sua) com **as coisas** ... com **as duas** né ... então agente já sabe o que que eles podem trazê ... então ... que que eu ... a gente podia comê que **elas** sabia ... também que**ela** gosta...

3. A ORGANIZAÇÃO DAS FRASES: SEGMENTOS CONFUSOS

a) a informação é genérica demais;

b) são omitidos itens, em construção elíptica.

c) sequência de unidades confusa pelas várias direções de reformulações

INFORMAÇÃO GENÉRICA – falhas lexicais, termos vagos, imprecisões

EXEMPLO 1

E- que que a sra acha de vivê numa cidade grande? ...

R- eu acho qui ... é um pouco de ... *melhoassi*... acho que é muito ... muito: ... todo mundo mata ... nessa nesta ... nessa ... nessa cidade aqui ...

E- hum hum ... parece que o progresso atrapalha ... um pouco a vida das pessoas

R- {atrapalha um poco a gente ...

R- agente para aqui o cê vê uma pessoa aí de de *caindo* ... num sei todo mundo ... esses que vem lá de cima ... vem lá de de de lá de do do ((ininteligível)) eles vem de lá prá cá *e não fica ... não faz* o que eles fazem ... porque aqui eles têm lá u jeitinho deles lá nu nu (hum) i chega aqui eles fic ficando nu meio da rua ...

EXEMPLO 2

E- e se a sra fosse visitá um uma amiga que mora no interior ... pra passá uns dias lá

E- ... qui providências a sra precisaria tomá ... pra í ... viajá ?

R- eu eu tenho o essa normalmente elas ... sabem ... o que ela vai fazê né? ... então a gente **pega as coisas** que a gente precisa pra ... **ficarem nu pedaço** que a gente ... nu que a gente ... **no que a gente usa** né? ...**fazê uma coisa pra levá** prá lá ... **levá uma coisa pra fazê** um grato ... um agrado né? ...

E- {hum hum ... i si ... i si a ... uma amiga da

EXEMPLO 3

R- eu pe... pegava um ... pur exemplo na na na is na istrada ... ((tosse)) a gente pega i **pede pru:: ... u: ... alguma um ... alguma coisa** qui tem ... qui sabe ... como é que chama? ... é :: ... porque tem gente qui fica na istrada né?

R- ... é gente di o sitio ... então nós ficamos lá um tempão ...

E- naquele lugar? ...

R- é ruim purque ...a gente tem medo né?...eu tenho medo... passei muito medo ...

EXEMPLO 4

R- {eu cheguei ... eu cheguei a tê ... dezoito cachorros ... porque a gente ... a

R- gente *fazia* us cachorrinhos né? ... a gente (vendia) vendia ... ...

E- ah cês vendiam filhotinhos? ...

R- é: por ... é porque :: roubavam (inint.)

E- daí cês vendiam us cachorros ...

R- daí a gente *fazia* us cachorros ... *fazia* fera na fe:ra ... feira di cachorrinho ...

E- ah é? ... ah qui jóia ...

R- a gente *fazia* ... ...

EXEMPLO 5

E- a sra acha que as novelas atuais têm algum papel na educação das pessoas? ...

R- eu acho que ... eu acho que:: ... num fi... não é bom pra criança ...

E- hum ... por quê?

R- a porque a : a: ... (inint) tudo né... enfia todo mundo lá dentro né ... a gente fica lá ... fica

R- tudo *fazendo* ... vendo novela ... novela ... novela ((fala rápido)) e eu às vezes ... as vezes

R- tem novela ... qui num é tão boa ... né? ... então *fazê uma coisa* qui num ... *faz* errado ...

E- {hum}

E- hum hum ... a sra se sente ligada nos tempos modernos?

R- tem tem que sê né? ... vai pará por quê? ... tem que ficá sim ...

EXEMPLO 6

E- a sra prefere artistas que fazem papel de mocinho ou de bandido? ...

R- bom isso eu acho que não há ... num há num t num tem ... num dá... depende do: ...

R- depende do que se ... do que é ... (do jeito do jeito) do que *vai fazendo*

R- né? ... *vai fazendo* assim ...

E- {hum hum}

R- então a gente: ... as vezes o: ... o bom ... as vezes enquanto ele fica o:fica ... ruim ...

E- {hum hum} {hum}

R- então ele que fica assim ... essa cabeça luta comigo ... ((ri))

(32)

E- é importante estar informada né? ... a sra assiste televisão ... lê jornal? ...

R- {é...} {televisão eu já num sô muito ligada

R- ... assisto só quando é uma coisa que a gente ... que eu que o queu gosto ... mas ficá

R-assim mais *nada* na na *coisa* ... num fico não ... ficá ... num fico...

EXEMPLO 7

E- igual o SEU jeito de fazer café? ...

R- eu pego a água ... pego o café (vai) vai fica fica a água né ? ... depois eu pego ...

R- um pouco do pó.... daí fica meio assim assim ... depois põe na na *coisa*...

CONSTRUÇÕES ELÍPTICAS

EXEMPLO 8

E- tem dois cachorros?...

R- {fiquei du ... fiquei fiquei du ... eu eu chorei porque a minha otra ... a minha que tem em casa ...q estava ... estava ... i ela morreu ... ela morreu ...

HESITAÇÕES

MÚLTIPLOS ENSAIOS DE REFORMULAÇÃO

EXEMPLO 9

E- ah: sra sai muito? ...

R-não ... só pra cha ... chacra (fui) pro ... eu t... **fiz algumas ... algumas** ((tosse)) (eu j) **eu fui a ... lá nu** ... como é que chama?... eu **fu:i ...fui lá** ... **minha sobrinha ... ela** hum ... ela ... eu fui lá ... quando qui ela saiu da ... ela casô i ... **foi lá ... em cima** da ... do ... do: ... como é que fala ?... ((tosse)) ahm lá em cima do: do do nosso: ... do do Brasil aqui... mas lá em cima né?...

E- hum

R- tem água né?... tem mata lá em cima ... agora mi ... trap ... su ... atrapalha tudo as minhas coisas ... ((ri)) falha ... ((ininteligível))

EXEMPLO 10

E- a sra lê jornal ... escuta rádio ?...

R- eu t/ eu eu e::u eu fazia ... eu eu como é que chama? ... tem um jornal né? ... eu t tinha ... eu fazia ... eu comprava ... mas faz tempo ... agora ... depois ... parei ...

E- que que a sra assistiu ... leu de importante nos últimos tempos?

4. DIFICULDADES NA COESÃO DE TEMPOS VERBAIS E ESTABELECIMENTO DE REFERÊNCIAS TEMPORAIS

EXEMPLO 1

E- eu perguntei ... si já aconteceu alguma coisa é:...desagradável pra senhora ...

R- {ah sim ... ah sim então ... então nós fomos na is na ish na ist ... na istrada ...na istrada ((tosse)) dois ... dois motoris ... uma menina/ ... uma moça i dois ...homens ... nós ficamos mais de três ... seis horas ... pra eles saí da gente ... ainda tivemos que (pe)di dinheiro ... dá dinheiro pra ele i imbora

E- nossa u que qui a sra fez?

R- é :: eu dei u dinheiro ... daí ele vai embora ... vô fazê u quê? ... por dinheiro vô pega minha minha minha vida ...

E- nossa ...

R-fica um temPÃO ...

DIFICULDADE NA CONSTRUÇÃO DE SCRIPTS

E- e se a sra fosse visitá um uma amiga que mora no interior ... pra passá uns dias lá

E- ... qui providências a sra precisaria tomá ... pra í ... viajá ?

R- eu eu tenho o essa normalmente elas ... sabem ... o que ela vai fazê né? ... então a gente pega as coisas que a gente precisa pra ... ficarem nu pedaço que a gente ... nu que a gente ... no que a gente usa né? ...fazê uma coisa pra levá prá lá ... levá uma coisa pra fazê um grato ... um agrado né? ...

E- {hum hum ... i si ... i si a ... uma amiga da

5. DIFICULDADE DE ACESSO À FORMA FONOLÓGICA

EXEMPLO 1

E- a sra costuma sair sozinha?...

R- não ... muito pel... muito p.... ... muito perto ...

E- hum hum ... a sra gosta mais de ficar em casa?...

R- {é}

R- ah eu gostu... purque é é isso eu... nã:o ... é que: é por causa da gente sê

R- assim mesmo que a gente fica muito *segundo* ... *segura* né?...

EXEMPLO 2

E- e se a sra fosse visitá um uma amiga que mora no interior ... pra passá uns dias lá

E- ... qui providências a sra precisaria tomá ... pra í ... viajá ?

R- eu eu tenho o essa normalmente elas ... sabem ... o que ela vai fazê né? ... então a

R- gente pega as coisas que a gente precisa pra ... ficarem nu pedaço que a gente ... nu

R- que a gente ... no que a gente usa né? ...fazê uma coisa pra levá prá lá ...

R- levá uma coisa pra fazê um *grato* ... um *agrado* né? ...

E- {hum hum ... i si ... i si a ... uma amiga da

EXEMPLO 3

E- e qual que a sra gosta mais?

R- a branquinha é mais brava ... a pretinha é mansa ... mas ela já tá velhinha ... ela já tá

R- cum us ... cum us *óc* ... cum us *óculos* ... *óculos* ... já tá ficando: ... ficando ff ...

R- ficando ce cega ... essa qui ca/ ... essa qui é agora ... que eu mandei mandei

E- {ahm}

R- pra i pra ... (murmúrio) chorei chorei qui nem cão ...

E- qui nem cão ((ri))

R- seu marido gosta de bicho ... né?

{go:sta ...}

EXEMPLO 4

E- que que a sra assistiu ... leu de importante nos últimos tempos? ...

R- {ah: foi foi o *elipse* ... foi o *elipse*:

E- ah:: eclipse? ...

R- é ...eclipse ...

EXEMPLO 5

R- só no quarto ... porque eu tenho as duas *ca* ... duas *cas* ... duas cadelinhas que eu

R- tenho se põe lá ela faz um ... ela mata ele ... ela é brava ... a branquinha é brava

R- ... é brava ...

EXEMPLO 6

R- não ... ela tem ... ela *méd* ... ela tem medo ...

E- ahm...

R- ela tem medo porque eu tenho duas duas cachorrinhas ...ela num sai ... porque o gato

R- vai embora ... ela num ... ela num deixa ...

E- e as cachorras ficam só no quintal? ...

6) SUJEITO 6 (DA moderada) LMVG

1. COMPREENSÃO

Dificuldades para reter os elementos da questão ou elaborar seu conteúdo.

EXEMPLO 1

L- vamo falá de novo?... pode?...

E- que que o senhor quer queeu fale de novo... ((riem juntos)) que qui o senhor quer queeu faça? ... queeu faça as perguntas ou queeu conte a piada?

L- é então vamos ...

EXEMPLO 2

E- parece que o progresso atrapalha a vida das pessoas... que que o senhor acha disso?

L- {ah isto é}

L- eu ... como que cê falou?

E- o progresso atrapalha a vida das pessoas ...

L- ah isto é ... eu acho também ... isso acho ...

EXEMPLO 3

E- em que época da sua vida o senhor teve que apertar o cinto?

L- quando:: ... eu fui lá ... casa ...

E- ah é... o que que aconteceu na época? como foi?

L- foi muito bom né ...

E- ahm?

L- nós ... nós ficamos é ... todo dia chegando lá e ... as as pessoas ... pediam que ... manter um traba:lho ... e a gente ficava lá:: ... parava corrá ... corria ... então foi muito bom ...

2. DIFICULDADES PARA FORMULAR

DIFICULDADES NA ORGANIZAÇÃO DE FRAMES E SCRIPTS (ESTABELECIMENTO DA SEQUENCIA DE PASSOS, ESTABELECIMENTO DO EIXO SEMÂNTICO BÁSICO PARA ORGANIZAÇÃO - MACROESTRUTURA)

EXEMPLO 1

E- me diz uma coisa/... onde u senhor mora?

L- moru a:: ... bom um ... é essas coisas assim é qui tudo (parque)... na

L-Lapa ...

E- na Lapa... i como é sua casa...

L- minha casa é boa ...

E- descreve pra mim como é ...

L- é um lugar bo-nito ... com casas todas juntas i ... i também ... tirando mais ... i:

E- {ahm} {hum hum}

L- ... i é um lugar espetacular ... bom...com as meninos co:rrem ... meninos ... i a gente vai atrás... lugar muito bom

Dificuldades discursivas com a extensão e esquematização da formulação:

A dependência do apoio do interlocutor se dá tanto para construir segmentos longos como para encadeá-los.

EXEMPLO 2

E- já aconteceu alguma coisa desagraDÁvel aqui com o senhor.?

L- aqui não ... aqui não::

E- {na cidade de São Paulo ... é}

L- { na cidade ... na cidade:: e:u ... estou aqui ... im São Paulo ... i gostu da cidade

E- {ahm} {ahm}

... i gostu de tudu ... i às vezes a gente d

E- {ahm}

L- ((ininteligível)) interior né?... qui é menor também né... ((ri)) mas ahm é melhor aqui ...

Referências construídas de tal forma que os papéis desempenhados pelos personagens de uma narrativa ficam ambíguos:

EXEMPLO 3

E- o senhor vota?

L- claro ... o meu pau ... o meu pai ... o meu pau ... não/ ... calma ... calma ...

E- {ahm}

L-u meu pai ... é velho né... eu num queria ... i ele fala/ não cê tem que i lá ((imita a voz do pai agravada))

E- {ahm}

E-L- e E ((riem))

L- saiu di si sis uma cidade longe ... pra vim aqui ... imagina ... ce vê que coisa?

E- {ahm} {nossa}

E- com essa idade... o seu pai ainda te obriga a fazê as coisas...

L- {com essa idade ...}

L- é ...

E- puxa vida... ((ri)) o senhor foi votá então...

L- ele foi votá ... velhinho...

E- ah ELE foi votá ...

L- Ele foi votá ... ((ininteligível))

E-{ e o senhor.?

L- eu também ...

A dificuldade para organizar o tema soma-se à dificuldade para formular sentenças extensas, o que é exemplificado nas duas amostras apresentadas a seguir:

EXEMPLO 4

L- eu gosto porque aqui eu estive a ... aqui ... eu ESTOU aqui ... com ... os meus filhos ... esse que ce encontrô ...

L-entendeu.?...ih ::: a gente fica:: junto com i com isso né?...

EXEMPLO 5

L- é um lugar bo-nito ... com casas todas juntas i ... i também ... tirando mais ... i:

E- {ahm} {hum hum}

L- ... i é um lugar espetacular ... bom...com as meninos co:rrem ... meninos ... i a gente vai atrás... lugar muito bom

LMVG responde em estilo sintético e necessita da intervenção constante do interlocutor para expandir a resposta.

EXEMPLO 6

E- u senhor sai muito?

L- eu saio ...

E- é... ih ... u senhor costuma sair sozinho?

L- sozinho ...

E- é ... que que o senhor faz?

L- ando ...

E- anda?

L- ando ando ando ... ando ando ... ando ...

E- todos os dias?

L- todos os dias...

E- é.?

L- é ...

E- qui bom ... faz bem né...

EXEMPLO 7

E- bom ... que que o senhor faz domingo?

L- futebol ...

E- todo domingo?

L- todo domingo ... mas em casa ... ((ri))

E- {hum} {ah em casa ... não no Tietê ?

L- {não no Tietê ...}

E- hum hum...

L- nu Tietê também mas não chega ... em casa ... aí é futebol ... futebol du São Paulo ...

Organização sintática das frases tende ao estilo telegráfico ou é confusa.

EXEMPLO 8

E- bom ... que que o senhor faz domingo?

L- futebol ...

E- todo domingo?

L- todo domingo ... mas em casa ... ((ri))

E- {hum} {ah em casa ... não no Tietê ?

L- {não no Tietê ...}

E- hum hum...

L- nu Tietê também mas não chega ... em casa ... aí é futebol ... futebol du São Paulo ...

E- futebol du São Paulo? ... u senhor é são-paulino?

L- são-paulino ...(aliás tamu) numa fase ... uma num **negócio a**í ...

EXEMPLO 9

L- nós ... nós ficamos é ... todo dia chegando lá e ... as as pessoas ... pediam que ... manter um traba:lho ... e a gente ficava lá:: ... parava corrá ... corria ... então foi muito bom ...

3. DIFICULDADES LEXICAIS

As dificuldades em relação ao léxico evidenciam-se pelo emprego de termos vagos e frases incompletas.

EXEMPLO 1

E- e o que o senhor acha de morá ...em São Paulo...que é uma cidade muito grande ?

L- eu gosto ...

E- gosta ... por quê ?

L- eu gosto porque aqui eu estive a ... aqui ... eu ESTOU aqui ... com ... os meus filhos ... esse que ce encontrô ...

L-entendeu.?...ih ::: a gente fica:: junto com i com isso né?...

EXEMPLO 2

E- me diz uma coisa/... onde u senhor mora?

L- moru a:: ... bom um ... é essas coisas assim é qui tudo (parque)... na

L-Lapa ...

EXEMPLO 3

E- o senhor acha que foi culpa de quem?... quem que é o responsável?

L- um:: jogador lá do ... Palmeiras ... que é terrível ... então ele que ... levantô u

E- {quem ?}

L- negócio aí ... depois ((ininteligível))

E- quem do ... quem do Palmeiras é terrível ?...

L- sei lá ... (meu Deus...) é: ..... cê sabe... cê num sabe né...

E- eu... num sô du Palmeiras ... pergunta pra mim... eu num vô sabê...

L- cê é do Palmeiras?

E- eu não ... eu sô são-paulina como o senhor hein...

L- ah bom ...

E e L riem

L- é:: ... u ... nome dele ... puxa vida... **isso aí** ... tá na cara ... nos jornais ...

E- acho queeu ouvi falá ... num é o Edmundo?... Evair?...

L- {é: é esse}

E- Edmundo... fazê o que né?... fazê o quê?...

EXEMPLO 4

L- i **essa siituação** me deixa chateado .... muito chateado né?...

porqueu eu fui **homem di rádio ... i:: futebol** ...

EXEMPLO 5

E- ahm ... pra quem que o senhor votô pra presidente?

L- é:: ... presidente: **é esse das ... bolinha** aqui né ... como é que chama...

E- bolinha?

L- bolinha não ... bolinha não... é da:: daqui queu tenho ó ... ((ri))

E- AH... das MOEDAS... do REAL...

L- ((ri)) REAL ...

E- ahm Fernando Henrique? ... é...?

L- é ...

EXEMPLO 6

L- i essa siituação me deixa chateado .... muito chateado né?...porqueu eu fui homem di rádio ... i:: futebol ...

E- {hum hum} {hum hum}

fazia tudo isso né? ... i: vivi muito ... saí pro pra o: mundo todo: ... mais chegando:**nessa ... nessa nesse**

E- {hum hum}

L- .**lugarzinho** agora ... num dá mais né? ... num dá ... i: eu fico chateado ... bastante sim

E- {hum hum}

L- purque:... a gente qué falá:: ... num pode... i a coisa MAIS...QUI EU TINHA...era **isso**..

E- {hum hum} {era a fala?}

L- futebol ainda né?... rápido **zzzzzzzuzu** né... ih: mas ... Deus ... sabe o que faz ...

4. ALTERAÇÕES LEXICAIS – FORMA FONOLÓGICA

EXEMPLO 1

E- o senhor vota?

L- claro ... *o meu pau* ... o meu pai ... *o meu pau* ... não/ ... calma ... calma ...

E- {ahm}

L-u meu pai ... é velho né... eu num queria ... i ele fala/ não cê tem que i lá ((imita a voz do pai agravada))

EXEMPLO 2

E- uma vez estourou um cabo de eletricidade perto da minha casa ... que que o senhor faria se acontecesse isso com o senhor ?...

L- aconteceu ...

E- ah aconteceu.?.. não me diga... ((os dois riem)) i aí o que que o senhor fez ...

L- sabe o que que aconteceu.?...

E- ahm

L- é:: faz uns dois anos ... mais ou menos ... eu inda tava bem ... i:: um ... ma

E- {ahm}

L- **réss réss** ... um ((sinal de transtorno))

E- { um raio...}

L- raio ... um raio ... i qui passô ... passô mas num passô ... juntinho ((ri)) di

E- { hum}

L- mim i eu ... caí né ...