UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM NEUROCIÊNCIAS E BIOLOGIA CELULAR

LOUISE BOGÉA RIBEIRO

COMPLEXIDADE SEMÂNTICA E HABILIDADE DE DECODIFICAÇÃO: UM MODELO QUANTITATIVO DA COMPREENSÃO DE TEXTOS DENOTATIVOS EM LÍNGUA PORTUGUESA BASEADO NA TEORIA DA INFORMAÇÃO

Belém
2018

LOUISE BOGÉA RIBEIRO

COMPLEXIDADE SEMÂNTICA E HABILIDADE DE DECODIFICAÇÃO: UM MODELO QUANTITATIVO DA COMPREENSÃO DE TEXTOS DENOTATIVOS EM LÍNGUA PORTUGUESA BASEADO NA TEORIA DA INFORMAÇÃO

Dissertação apresentada como requisito final para a obtenção do grau de Mestre em Neurociências pelo Programa de Pós-Graduação em Neurociências e Biologia Celular do Instituto de Ciências Biológicas da Universidade Federal do Pará.
Área de concentração: Neurociências.
Orientador: Prof. Dr. Manoel da Silva Filho.
Coorientador: Prof. Dr. Anderson Raiol Rodrigues.

Belém
2018

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

RIBEIRO, LOUISE
COMPLEXIDADE SEMÂNTICA E HABILIDADE DE DECODIFICAÇÃO: UM MODELO QUANTITATIVO DA COMPREENSÃO DE TEXTOS DENOTATIVOS EM LÍNGUA PORTUGUESA BASEADO NA TEORIA DA INFORMAÇÃO / LOUISE RIBEIRO. — 2018
74 f. : il. color

Dissertação (Mestrado) - Programa de Pós-graduação em Neurociências e Biologia celular (PPGNBC), Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, 2018.
Orientação: Prof. Dr. MANOEL SILVA FILHO
Coorientação: Prof. Dr. ANDERSON RAIOL RODRIGUES.

1. NEUROLINGUÍSTICA. I. SILVA FILHO, MANOEL, *orient.* II. Título

---

CDD 612.82336

LOUISE BOGÉA RIBEIRO

COMPLEXIDADE SEMÂNTICA E HABILIDADE DE DECODIFICAÇÃO: UM MODELO QUANTITATIVO DA COMPREENSÃO DE TEXTOS DENOTATIVOS EM LÍNGUA PORTUGUESA BASEADO NA TEORIA DA INFORMAÇÃO

Dissertação apresentada como requisito final para a obtenção do grau de Mestre em Neurociências pelo Programa de Pós-Graduação em Neurociências e Biologia Celular do Instituto de Ciências Biológicas da Universidade Federal do Pará.

Área de concentração: Neurociências.

Orientador: Prof. Dr. Manoel da Silva Filho.
Coorientador: Prof. Dr. Anderson Raiol Rodrigues.

Data de aprovação: 26 de fevereiro de 2018.

Banca examinadora:

_____
Prof. Dr. Manoel da Silva Filho (orientador)
Universidade Federal do Pará

_____
Prof. Dr. Carlomagno Pacheco Bahia
Universidade Federal do Pará

_____
Profa. Dra. Maria Elena Crespo López
Universidade Federal do Pará

_____
Prof. Dr. Daniel Valle Vasconcelos Santos (suplente)
Universidade Federal do Pará

Ao Z.

AGRADECIMENTOS

Agradeço sempre à leitura a ter me ensinado a superar os momentos de quaisquer tipos de falta de comunicação comigo mesma. Não poderia esquecer da música. Nunca conseguirei agradecer suficientemente ao meu orientador, Prof. Dr. Manoel da Silva Filho, a confiança e oportunidade de continuar meus estudos, abrindo-me a porta do seu Laboratório de Neuroengenharia, o que transformou a minha vida.

Agradeço ao meu coorientador, Prof. Dr. Anderson Raiol Rodrigues, em todo o processo, mas darei ênfase à data da minha qualificação, um dia marcante na minha vida, e que ele se fez presente, apoiando-me.

Agradeço a todos os professores que contribuíram com a minha formação acadêmica, desde a minha formação infantil, básica, técnica e superior, com resultados significativos na minha vida profissional e pessoal.

Agradeço àqueles que conheci durante as aulas e eventos da pós-graduação, principalmente ao amigo Jonabeto Vasconcelos e Arthur Costa, e à secretaria e direção do Programa de Pós-graduação em Neurociências e Biologia Celular pelo auxílio e incentivo.

Agradeço a todos do Museu da UFPA, sobretudo à Jussara Derenji que como ela mesma disse uma vez: "Finalmente nos entendemos".

Agradeço aos professores que constituíram as bancas examinadoras da qualificação e defesa do projeto, com seus comentários e retificações.

Tenho demasiada gratidão ao Kauê Machado Costa, pela solicitude e críticas construtivas.

Um obrigado especial ao Sr. Ismael, quem me indicou a porta do laboratório pela primeira vez.

Ainda que ela não leia este documento, agradeço à minha mãe a todos os minutos comigo que valorizo cada segundo.

E desculpo aqueles que já partiram da minha vida, restando apenas eternos agradecimentos.

Sumário

RESUMO

Com base em princípios da neurociência cognitiva e da teoria da informação, com ênfase no trabalho de Claude Shannon, realizou-se uma análise estatística de 33.101 palavras a partir da coleta de textos científicos da Biblioteca Digital Brasileira de Teses e Dissertações e da Biblioteca Digital da USP, mediante a utilização da linguagem de programação *C#* e do *Microsoft Visual Studio* 2012 enquanto complemento do código, incluindo o *SQL Server Management Studio* 2012 para o gerenciamento do banco de dados, em prol do desenvolvimento do programa de processamento de informação intitulado de CalcuLetra, com o objetivo de mensurar a dificuldade de compreensão textual em Língua Portuguesa. A partir das premissas de que o aprendizado dos significados das letras, palavras e outros símbolos favorece o desenvolvimento do sistema nervoso central de humanos; que o comportamento metacognitivo do leitor permite a resposta a estímulos advindos do processo de leitura; e que as palavras de maior ocorrência no banco representam as mais conhecidas pelos seus autores, o algoritmo determina, assim, o grau de familiaridade das palavras conforme os parâmetros matemáticos e estatísticos do banco. Ao comparar textos não literários ou denotativos com os valores probabilísticos encontrados, revela-se quão compreensivo é o texto inserido no programa, considerando leitores neurotípicos e que o conteúdo possua os devidos elementos de coesão textual, conforme as regras gramaticais da língua. Nossos resultados revelam grupos de palavras que causam a incompreensão ou facilitam a leitura. Adicionalmente, mostramos lacunas de vocabulário e na utilização do dicionário. Apesar dos resultados preliminares, este estudo foi mais uma prova de conceito para o método empregado e demonstrou seu potencial para futuras pesquisas.
A metodologia do modelo de quantificação pode ser adaptada a outras línguas, e espera-se que a pesquisa possa contribuir em prol da elaboração de diagnóstico objetivo de transtornos do comportamento (ex. dislexia), mediante classificação quantitativa da incompreensão escrita; e ter a sua aplicabilidade enquanto instrumento auxiliar na análise de exames dissertativos de vestibulares, do Enem e de concursos públicos, cuja avaliação é ainda de forma subjetiva.

Palavras-chave: neurolinguística; estudos cognitivos; análise de frequência.

ABSTRACT

Based on the principles of cognitive neuroscience and information theory, with emphasis on the work of Claude Shannon, a statistical analysis of 33,101 words was done from the collection of scientific texts of the Brazilian Digital Library Of Thesis And Dissertation and the Digital Library of USP, using the C # programming language and Microsoft Visual Studio 2012 as a code complement, including SQL Server Management Studio 2012 for database management, for the development of the information processing program titled CalcuLetra, with the purpose of measuring the difficulty of textual comprehension in Portuguese Language. From the premises that the learning of the meanings of letters, words and other symbols provides the development of the central nervous system of humans; that the reader's metacognitive behavior allows the response to stimuli coming from the reading process; And that the words of greatest occurrence in the bank represent those best known by their authors, the algorithm thus determines the degree of familiarity of the words according to the mathematical and statistical parameters of the bank. Therefore, when comparing non-literary or denotative texts with the probabilistic values found, it shows how comprehensible is the text inserted in the program, considering neurotypical readers and that its content has the necessary elements of textual cohesion, according to the grammatical rules of the language. Our results reveal groups of words that cause misunderstanding or make reading easier. Additionally, we show gaps in vocabulary and dictionary usage. Despite the preliminary findings, this study was more a proof of concept for the method employed, and demonstrated its potential for future research. The quantification model's methodology presented in the present study can be adapted to other languages, and we hope it contributes to the development of objective evaluation of behavioral disorders (e.g., dyslexia), by quantitative classification of written incomprehension; and helps in the analysis of dissertations of vestibular, ENEM and other public examinations, whose evaluation is still of a subjective form.

Keywords: neurolinguistics; cognitive studies; frequency analysis.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

GLOSSÁRIO

SNC – Sistema Nervoso Central

OCDE - Organisation for Economic Co-operation and Development

PISA – Programme for International Student Assessment

ENEM – Exame Nacional do Ensino Médio

CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

USP – Universidade de São Paulo

SELMA – Semantic Linguistic Maturity

VSM - Vetores de Valores Derivados de Frequências de Eventos

P-valor: É uma estatística utilizada para sintetizar o resultado de um teste de hipóteses. Formalmente, o p-valor é definido como a Probabilidade do se obter uma estatística de teste igual ou mais extrema que aquela observada em uma amostra, assumindo como verdadeira a hipótese nula. Como geralmente define-se o nível de significância em 5%, uma p-valor menor que 0,05, gera evidências para rejeição da hipótese nula do teste.

D.P. – Desvio Padrão. É uma das principais medidas de dispersão dos dados. Pode ser definida como a raiz quadrada da variância. Sua medida representa o quanto os dados se afastam da média.

1a Q – 1a Quartil: O primeiro quartil é uma medida de posição que representa que pelo menos 25% das respostas são menores que ele.

2a Q – 2a Quartil: O segundo quartil, também conhecido como mediana é uma medida de posição que representa que pelo menos 50% das respostas são menores que ele.

3a Q – 3a Quartil: O terceiro quartil é uma medida de posição que representa que pelo menos 75% das respostas são menores que ele.

I.C – 95% = Intervalo de 95% de confiança: É um intervalo estimado para um parâmetro estatístico. Em vez de estimar o parâmetro por um único valor é dado um intervalo de estimativas prováveis. Um intervalo de 95% de confiança garante que o parâmetro pontual estimado com 95% de confiança estará dentro do intervalo estimado em outras amostras da mesma população.

1 INTRODUÇÃO

O Sistema Nervoso Central (SNC) é capaz de armazenar a memória e consolida a aprendizagem. A linguagem é a principal mediadora da formação e do desenvolvimento das funções psicológicas superiores, seja ela oral, gestual, escrita, artística, musical ou analítica (LOPES, 2012). Para o desenvolvimento cognitivo, a aquisição da linguagem e de outros símbolos é um marco, por meio de relações sociais e culturais.

A organização combinatória da comunicação humana é única (NOWAK; PLOTKIN; JANSEN, 2000). A partir das habilidades de codificação e decodificação, o ato de comunicação incorpora representações mentais do ambiente (RAMOS, 2014). A comunicação estimula a neuroplasticidade, o que leva ao desenvolvimento cognitivo complexo, relacionado ao valor associado com letras e imagens e suas correlações para o meio externo (SIGMAN *et al*., 2014). Ao formar conexões sinápticas oriundas do processo de leitura, da habilidade de decodificação de objetos e do processamento da informação, a educação também estimula o fenômeno da plasticidade, ocasionando a aprendizagem.

O conhecimento tem várias realidades relativas que dependem do ambiente evolutivo que o indivíduo esteja inserido e a linguagem escrita tenta transmitir a informação de uma forma mais permanente. Com um papel cultural que muitas vezes incorpora vieses implícitos positivos ou negativos, o ensino da língua tem a responsabilidade de adequar o encéfalo primitivo a viver na sociedade

moderna, provendo adaptação em vários níveis no ambiente que o indivíduo esteja inserido.

Na última edição do Programa Internacional de Avaliação de Estudantes (PISA), em 2015, o Brasil ficou em 59º lugar entre 70 países no *ranking* de leitura, revelando que muitos estudantes têm habilidades insuficientes de leitura e escrita, o que pode ter sido devido a diagnósticos pobres ou métodos de instrução inadequados. Em geral, o desenvolvimento de competências de leitura e de escrita é avaliado usando critérios subjetivos, a exemplo de exames dissertativos de vestibulares, do Exame Nacional do Ensino Médio (ENEM), ou de concursos públicos ao redor do país são analisados de maneira subjetiva, cujos resultados são frequentemente irrecorríveis. Portanto, a avaliação da leitura e da produção escrita atual carece de um método objetivo ou quantitativo. Faz-se necessária a realização de pesquisas na tentativa de elucidar tais questões, a fim de auxiliar tanto a produção de textos dissertativos quanto a sua avaliação. Ou ainda para a elaboração de redes neurais em Língua Portuguesa.

Também não existe um método preciso para diagnosticar deficiências de leitura, sendo o principal indicador a incompreensão escrita, como a dislexia. A compreensão escrita foi definida como as representações mentais feitas pelas conexões semânticas em um texto à medida que os códigos são traduzidos e combinados (KENDEOU *et al.,* 2014), o que revela o desempenho das operações internas realizadas pelo leitor ou o entendimento do texto (PEREIRA, 2012).

O processo de aprendizagem pode ser impulsionado a partir de estudos acerca da metacognição. Salienta-se que os fatores relativos à idade, formação e ensino têm ligação direta com o desenvolvimento de habilidades metacognitivas em prol da compreensão (RIBEIRO, 2003); já quaisquer alterações na formação e/ ou no desenvolvimento do SNC ainda não estão bem esclarecidas (ex., afasia, dislexia), bem como seu diagnóstico e tratamento.

É imprescindível o desenvolvimento de estudos interdisciplinares, e a neurolinguística, apesar de ser ainda uma ciência recente (JEFFERIES, 2013; KEMMERER, 2014), vem com essa proposta, ao unir pesquisadores de várias áreas do conhecimento (ex., biologia computacional, educação) em busca de respostas a fenômenos específicos (PURCELL, 2011; INDEFREY, 2011). Faz-se necessário o estímulo da aprendizagem da língua padrão pelos seus usuários – considerando os atuais métodos avaliativos de dissertação do Brasil –, bem como a formação de estudos interdisciplinares, a pesquisa em neurolinguística e análise de frequências e inteligibilidade textual em português.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Determinar uma medida probabilística que expresse a dificuldade da compreensão textual em Língua Portuguesa, com base em princípios da neurociência cognitiva e da teoria da informação.

2.2 OBJETIVOS ESPECÍFICOS

- Construir um banco de palavras com sentido denotativo, por meio da coleta de textos da Biblioteca Digital Brasileira de Teses e Dissertações e da Biblioteca Digital da USP, para servirem de base ao estudo;

- Estabelecer fórmulas probabilísticas do conhecimento para cada palavra dos textos coletados, considerando princípios da teoria da informação, com as devidas adaptações;

- Desenvolver o *software* CalcuLetra, utilizando a linguagem de programação C# (C Sharp) e o *Microsoft Visual Studio,* juntamente com o *SQL Server Management*, para o armazenamento do banco de palavras, das equações e fórmulas;

- Avaliar a complexidade de textos denotativos sob o olhar da compreensão escrita, considerando que o texto siga a norma culta, e que o leitor neurotípico tenha ciência dos aspectos sintáticos e estruturais da língua;

- Inserir para o teste de textos não-literários ou denotativos no *software,* comparando sua dificuldade de compreensão de acordo com os parâmetros objetivos do banco armazenado;

3 REFERENCIAL TEÓRICO

3.1 O DESENVOLVIMENTO DE ATIVIDADES COGNITIVAS AO LONGO DOS ANOS

O signo linguístico é a combinação de um significante – a palavra – e de um significado – conceito –, sendo este arbitrário até quando representar aquele, para compor um código linguístico social que não poderá ser alterado facilmente (NETTO, 2010), apesar de a língua ser dinâmica e ter variações linguísticas em sua forma de comunicação falada. Na analogia inconsciente e significante, este é superposto às representações de palavras, com a tendência do significado se identificar às representações das coisas, podendo a palavra ser considerada uma expressão do inconsciente, ao ser selecionada pela consciência (DOR, 1996).

Os processos da consciência (NUNES, 2002) dividem-se em cognição, ou seja, operações inconscientes sem um objetivo mais amplo, relacionadas a atividades de apreensão, processamento, recuperação e memória (MATURANA; MPODOZIS; LETELIER, 1995; SPERBER; WILSON, 1989); e metacognição, responsável por realizar operações conscientes com um objetivo definido, envolvendo atividades de reflexão e de monitoramento da própria compreensão (JOLY; SANTOS; MARINI, 2006; FLAVELL, 1987; NELSON; NARENS, 1996).

Apesar de não existir ainda um paradigma experimental que mostre precisamente quais áreas do encéfalo são ativadas em um dado pensamento,

as regiões de ativação relacionadas à compreensão escrita já foram identificadas em muitos estudos (BOOTH *et al*., 2002a; COHEN *et al*., 2000; YARKONI *et al*., 2008; VIRTUE *et al*., 2006), incluindo a frontal inferior e a temporal, estas associadas à compreensão e à produção textual, e ainda outras áreas corticais cerebrais distribuídas nos lobos temporal, parietal e frontal ligadas à construção coerente de representações do texto, entretanto, o processamento semântico não está totalmente claro (VISSER; LAMBON, 2011). É possível quantificar a semântica de eventos, independentemente da sintaxe (CHAMPOLLION, 2011), considerando que a maioria das linguagens escritas tendem a ter regras pré-estabelecidas, cuja aprendizagem sintática se consolida aos 12 anos de idade (BERNS, 2002). Recentemente, demonstrou-se que decodificamos uma frase da mesma forma em inglês e português (YANG *et al*., 2016).

Definida como o significado de uma palavra, o uso da semântica induz a cognição e as associações linguísticas através das correlações entre o vocabulário e seus conceitos associados (BRENT; SISKIND, 2001), promovendo o desenvolvimento cognitivo.

Já definiram a maturidade semântica como a capacidade de derivar o significado correto de uma coocorrência de palavras (LANDAUER; DUMAIS, 1997). Em um estudo semelhante, as representações semânticas foram encontradas ao se associarem a distribuições de palavras à medida que as crianças as liam ou ouviam (HANSSON *et al*., 2015). As palavras abstratas são mais desafiadoras para aprender do que palavras com significado concreto (BERGELSON; SWINGLEY, 2013). Há estudos que mostram o aprendizado de palavras

concretas (ex., laranja, caderno) por crianças de 6 meses de idade (BERGELSON; SWINGLEY, 2012), já palavras abstratas (ex., comer, ler) são adquiridas posteriormente entre 10 e 13 meses (BERGELSON; SWINGLEY, 2013), cuja diferença nos testes relaciona-se diretamente ao desenvolvimento das capacidades sociocognitivas básicas.

O teste *Semantic Linguistic Maturity* (SELMA) analisa o desenvolvimento da linguagem em crianças. Considerando que a aquisição de significados das palavras é uma das questões fundamentais no estudo da mente, estudos concluíram que os significados das palavras estavam fortemente relacionados às estatísticas de uso de palavras, ao utilizarem os Vetores de Valores Derivados de Frequências de Eventos (VSMs); no entanto, foi enfatizado que dados empíricos adicionais eram necessários para melhorar a análise de frequência (TURNEY; PANTEL, 2010; FOURTASSI; DUPOUX, 2016). O processo de aquisição da língua escrita baseia-se, portanto, em valores de letras, imagens, que se definem por sua correspondência com as coisas (BORGES, 2010), em meio a um processo de formação de conceitos mediante a linguagem, com influência da escolarização, o que define nosso ser, pensamento e atitudes.

A construção de conceito, sentido, ou ainda da compreensão textual dá-se em um conjunto de tarefas complexas que envolvem a apropriação de informação, representações mentais, generalização e recriação de ideias, mediante habilidades de codificação e decodificação (DIAS, 2014). Ainda que o leitor não seja totalmente passivo durante a leitura, precisará de informação advinda do texto para então buscar seus conhecimentos prévios e realizar inferências; caso

contrário, ocorrerá uma falha na interação. Em outras palavras, apesar de a compreensão textual não se esgotar na habilidade de decodificação, esta é essencial e passível de quantificação em prol do entendimento do texto.

O Indicador Nacional de Alfabetização Funcional (INAF) relata que 68% dos 30, 6 milhões de brasileiros entre 15 e 64 anos que estudaram até 4 anos; e que 75% dos 31,1 milhões que estudaram até 8 anos permanecem nos níveis básicos de alfabetização. Para a simplicidade textual, com foco em aspectos estruturais e sintáticos, algumas ferramentas do inglês, como o nível do índice Flesch-Kincaid e as métricas do Coh-Metrix, foram adaptadas para avaliar a inteligibilidade e a complexidade textual em português usando o número de palavras, parágrafos, frases e outros parâmetros (GASPERIN *et al*., 2009; ALUÍSIO *et al*., 2008A; CASELI *et al*., 2009; CANDIDO JUNIOR *et al*., 2009; WATANABE *et al*., 2009; GASPERIN; MASIERO; ALUISIO, 2010; SCARTON; ALUÍSIO, 2010).

Embora o inglês tenha construções computacionais complexas, observou-se que estas são usadas com pouca frequência (THORNE; SZYMANIK, 2015), o que ocorre também na língua portuguesa, a ser analisada no presente estudo. Muitos bancos de textos e palavras, livros e contos literários em português estão disponíveis (BIBLIOTECA VIRTUAL DE LITERATURA, c2017; FUNDAÇÃO BIBLIOTECA NACIONAL, c2017; PÁGINA SOBRE EÇA DE QUEIRÓS, c2017; CORPUS DO PORTUGUÊS, c2017; CORPUS BRASILEIRO, c2017), como as listas das 1.000 palavras mais frequentes no português falado (PORTUGUESE 101, c2017), as 3.000 palavras com ocorrências mais altas (FERREIRA, 2014),

as 5.000 palavras mais recorrentes (WIKTIONARY; c2017; MARK , 2011), e recentemente, as 50 mil palavras portuguesas mais comuns (HERMIT, c2017). A disponibilidade desses repositórios levou a estudos associados a ocorrências de letras, estruturas silábicas e sentenças em português (HARTMANN *et al*., 2014; SOARES *et al*., 2017; QUARESMA, 2008; VIARO; GUIMARÃES-FILHO, 2007).

A partir da teoria matemática da informação (SHANNON, 1948; D'ALFONSO, 2010), percebe-se que em inglês e português a letra *a* é utilizada com alta frequência; no entanto, contém pouca informação e alta entropia informacional, considerando que, quanto mais semelhantes as ocorrências de palavras, maior a entropia (cf. ANEXO A). Estudos baseados nessa teoria já buscaram medir a quantidade de informação, a entropia informacional (MONTEMURRO; ZANETTE, 2002; MONTEMURRO; ZANETTE, 2011; DEBOWSKI, 2011; KALIMERI *et al*., 2015) e a informação semântica em uma mensagem (BAR-HILLEL; CARNAP, 1952; D'ALFONSO, 2011; BAO, 2011; MARCELO; DAMIAN, 2010; MATTHEWSON, 2014), mostrando, a partir de bancos de textos, uma universalidade na entropia entre idiomas e uma ligação entre a informação semântica e o linguístico estatístico (MONTEMURRO; ZANETTE, 2016; MONTEMURRO, 2014).

Apesar de muito trabalho já conduzido no campo para alcançar uma verdadeira teoria do significado, há mais para melhoria e aperfeiçoamento, especialmente em português, considerando-se que poucos estudos usaram abordagens baseadas na teoria da informação para examinar probabilidades de letras e

sentenças em português para determinar a entropia do texto (RABÊLO; MORAES, 2008; VILLAVICENCIO, 1995).

Nosso estudo utilizou princípios da teoria da informação para determinar a dificuldade média da compreensão escrita mediante a decodificação semântica, sob as premissas de que o aprendizado dos símbolos linguísticos favorece o desenvolvimento do SNC de humanos; que o comportamento metacognitivo permite aos leitores responderem a estímulos advindos do processo de leitura; e que as palavras mais utilizadas na produção textual são aquelas mais familiares aos usuários da língua. Apresentamos um modelo quantitativo da compreensão de textos não-literários ou denotativos em língua portuguesa, ao testarmos a hipótese de que a compreensão de textos dependa de quanto maior for a probabilidade de conhecimento de suas palavras que está inversamente proporcional a sua entropia informacional. Testamos nossa hipótese usando parâmetros objetivos da teoria da informação para ocorrências de palavras em um banco de textos não-literários. Amostras de textos denotativos foram coletadas para serem inseridas no *software* desenvolvido nesta pesquisa e comparar os seus níveis de complexidade semântica. O modelo quantitativo de compreensão escrita desenvolvido neste estudo pode ser universalmente adaptado a outras línguas para avaliação de leitura e escrita.

Isto exposto, pergunta-se qual a mínima previsibilidade de conhecimento necessária para um texto ser compreensível a maioria dos usuários da língua?

4 METODOLOGIA

4.1 TIPO DE ESTUDO

A pesquisa é de abordagem quantitativa, a partir da coleta randômica de textos científicos, organização e classificação da amostra, com análise estatística dos resultados obtidos, para a formação de inferências e discussão.

4.2 LOCAL DE ESTUDO

Laboratório de Neuroengenharia (LANEGE) do Instituto de Ciências Biológicas (ICB) da Universidade Federal do Pará (UFPA), pioneiro no centro, para a realização de pesquisas interdisciplinares em Neurociências, com a produção de protótipos e dispositivos auxiliares que visem a avaliação e diagnóstico de acometidos pelas mais diversas limitações mentais e/ ou físicas. Concomitantemente, o presente estudo foi realizado no Laboratório de Bioestatística e Matemática Computacional, no Núcleo de Medicina Tropical (NMT), unidade de integração da UFPA.

4.3 PERÍODO DE ESTUDO

Período de 2016 e 2017.

4.4 AMOSTRAGEM DA PESQUISA

A coleta do universo de estudo se deu mediante seleção randômica de textos científicos, disponíveis na Biblioteca Digital Brasileira de Teses e Dissertações e

na Biblioteca Digital da USP, correspondendo o total da amostra igual a 1.032 textos, entre dissertações de mestrado e teses de doutorado, de diferentes áreas do conhecimento.

4.5 CRITÉRIOS DE INCLUSÃO E EXCLUSÃO

O formato científico, denotativo ou não-literário dos textos em forma de dissertação ou tese escritos em Língua Portuguesa foram critérios de inclusão considerados, bem como a sua data de publicação ser de até 10 anos atrás. Portanto, aqueles textos com conteúdo conotativo, figurado ou literário, ou ainda escritos em sua maior parte em língua diferente do português brasileiro, foram excluídos da análise. As palavras não pertencentes aos mais recentes dicionários da Língua Portuguesa (ex., Aurélio, *Houiass*) foram igualmente removidas. Figuras, nomes próprios, palavras com erros ortográficos ou estrangeiras, números e outros símbolos (ex., ( ) [ ] { } _ ? + \ / = * < > : ; ‘ % @ ! # - & |) também foram desconsiderados, para evitar quaisquer discrepâncias nos cálculos estatísticos da amostra.

4.6 PROCEDIMENTOS INICIAIS

Os textos científicos foram baixados da Biblioteca Digital Brasileira de Teses e Dissertações e da Biblioteca Digital da USP, considerando os critérios de inclusão e exclusão expostos, e convertidos para serem armazenados no *software* de processamento de informação “CalcuLetra”, desenvolvido durante o presente estudo, a partir da utilização da linguagem de programação *C#* (C Sharp) e do *Microsoft Visual Studio* 2012 enquanto complemento do código,

incluindo o *SQL Server Management Studio* 2012, para o gerenciamento do banco de dados.

4.7 PROCEDIMENTOS DE COLETA DE DADOS

4.7.1 Tratamento do banco

Após a armazenagem dos textos no CalcuLetra, foi realizada uma filtragem no SQL dos caracteres constituídos por um conjunto superior a 14 letras por serem erros oriundos da conversão dos textos de PDF a DOC, e/ ou com ocorrência (*f*) ≤ 3 no universo ou probabilidade de ocorrência (P) ≤ ,0001. Prefixos e sufixos também foram filtrados no SQL (ex., *pós-*, *sub-*, *vice-*, *intra-*, *inter-*, *pro-*, *pre-*, *anti-*, *ante-*, *-mente*, *-issimo*, *-oso*, *-osa*) por terem probabilidades similares aos seus radicais sem alteração no significado, bem como palavras duplicadas. Ao final, obteve-se uma amostra de 33.101 palavras para a análise.

4.7.2 Parâmetros objetivos de análise

Para cada palavra armazenada no CalcuLetra, o algoritmo determina o seu respectivo grau de familiaridade, conforme a sua *f* e *P* no universo de textos, mediante a utilização de fórmulas matemáticas, para a constituição das intituladas probabilidades do conhecimento (P_c) e do desconhecimento (P_d), ambas obtidas a partir de interpolação linear entre os extremos dos valores

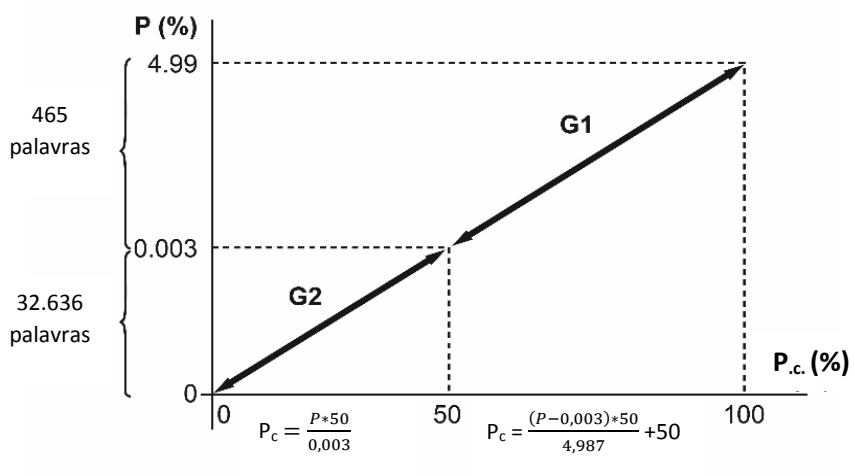
probabilísticos do banco. A quantidade de informação (I) = log (1/ P_d) e a entropia informacional (h) = I*P_d também foram calculadas, gerando representações mentais a partir da ocorrência de palavras (SAHLGREN, 2008).

Tabela 1. Exibição aleatória do banco de dados com cada parâmetro armazenado. Ao inserir um texto qualquer no CalcuLetra, é possível obter o número total de palavras, de letras e de sentenças do texto, bem como os valores da *f* e *P* de cada letra e palavra. Para a medição da quantidade de compreensão escrita e de sua consequente complexidade para pelo menos a maioria dos usuários (autores/ leitores) da língua, a P_c média do texto inserido é obtida a partir da soma das P_c de cada uma de suas palavras de acordo com os valores dos parâmetros do banco armazenado. Desta forma, os valores individuais da P_c de cada palavra armazenada no banco não se alteram ao inserir um novo texto no CalcuLetra.

ID	Palavra	*f*	*P (%)*	*Pc (%)*	*Pd (%)*	*I (bits)*	*h (bits)*
33426	Baixa	2704	0,0030827	50,008337	49,991663	1,000241	0,5000371
124058	Seção	2699	0,0030770	50,007762	49,992238	1,000224	0,50003436
60415	Concentração	2693	0,0030701	50,007067	49,992933	1,000204	0,50003131
103391	Massa	2683	0,0030587	50,005917	49,994083	1,000171	0,50002631
60366	Conceitos	2682	0,0030576	50,005806	49,994194	1,000168	0,50002593
64579	Critérios	2677	0,0030519	50,005232	49,994768	1,000151	0,50002317
129308	Tradução	2676	0,0030508	50,005121	49,994879	1,000148	0,50002278
81713	Direção	2664	0,0030371	50,003740	49,996260	1,000108	0,50001659
128044	Temos	2638	0,0030074	50,000746	49,999254	1,000022	0,50000353
117790	Próprio	2633	0,0030017	50,000171	49,999829	1,000005	0,50000078
127813	Taxa	2631	0,0029995	49,991667	50,008333	0,999760	0,49996331

No presente estudo, a P_c e a P_d tiveram de ser estabelecidas para a determinação da entropia em termos semânticos ou comunicacionais, tendo em vista que nos estudos da teoria da informação tanto a letra *c* quanto a *m,* por exemplo, têm entropia informacional similares, porém, há palavras que são mais conhecidas com a letra *c,* e, seguindo o mesmo raciocínio, tanto a quantidade de informação quanto a entropia podem ser iguais para um texto de fácil compreensão ou não, razão pela qual as fórmulas foram estipuladas e armazenadas também no CalcuLetra juntamente com as demais. Desta maneira, esperou-se corrigir possíveis deturpações advindas de palavras com maiores probabilidades de surgirem em textos de níveis de instrução inferior.

Figura 1. Fórmulas de equivalência para os grupos de palavras pertencentes ao banco de dados armazenado no CalcuLetra. Ao dividir o banco em dois grupos de palavras (G1 e G2), foram estabelecidas suas P_c e P_d. Para o G1, a $Pc = \frac{(P-0{,}003)*50}{4{,}987} + 50$; e para o G2, a $Pc = \frac{P*50}{0{,}003}$. Por outro lado, $Pd = 100 - Pc$. As inferências têm sua veracidade confirmada a partir da análise do amplo universo, levando-se em consideração que o presente estudo se trata de médias probabilísticas para a quantificação de textos, e não de palavras específicas ou isoladas. Probabilidade de ocorrência média (P_m) do banco = ,003.



Nota: $P_d = 100-P_c$

4.8 INSTRUMENTOS DE ANÁLISE E DE INTERPRETAÇÃO DOS RESULTADOS

Para a análise estatística, operamos o BioEstat 5.3, para análise das variáveis, com foco nas variações da P_c e suas correlações entre os outros parâmetros objetivos do banco. Para estabelecer as faixas de Probabilidade do Conhecimento e Probabilidade do Desconhecimento em relação ao nível de Entropia foi utilizada uma Árvore de Decisão (BREIMAN *et al*.,1984).

As árvores de decisão surgem como uma alternativa preditiva/exploratória de grande valia para problemas de regressão e classificação, sendo possível tratar tanto variáveis categóricas como numéricas, sendo que quando a variável resposta é do tipo categórica, tem-se a árvore de classificação e quando a

variável resposta é do tipo contínua o tipo da árvore é de regressão. O algoritmo de árvores de decisão auxilia na predição e classificação dos dados, com base nos valores da variável resposta. Este método consiste na execução de partições binárias sucessivas, com o intuito de obter subconjuntos cada vez mais homogêneos em relação à variável resposta, e tem como resultado uma árvore hierárquica de regras de decisão utilizadas para prever ou classificar, sendo que cada ponto em que há partição dos dados denomina-se nó. Vale ressaltar que há nós internos e terminais (folhas). O algoritmo tem como principais passos: (1) geração do nó raiz (contém todo o banco de dados); (2) encontrar a melhor decisão para dividir os dados em dois grupos (encontrar nós a serem divididos); (3) escolher um atributo que melhor classifica os dados, levando em consideração que existe um único caminho entre o nó raiz e cada nó (divisão de nó) e; (4) criação e desenho do nó e as suas ramificações. O algoritmo retorna para o passo 2.

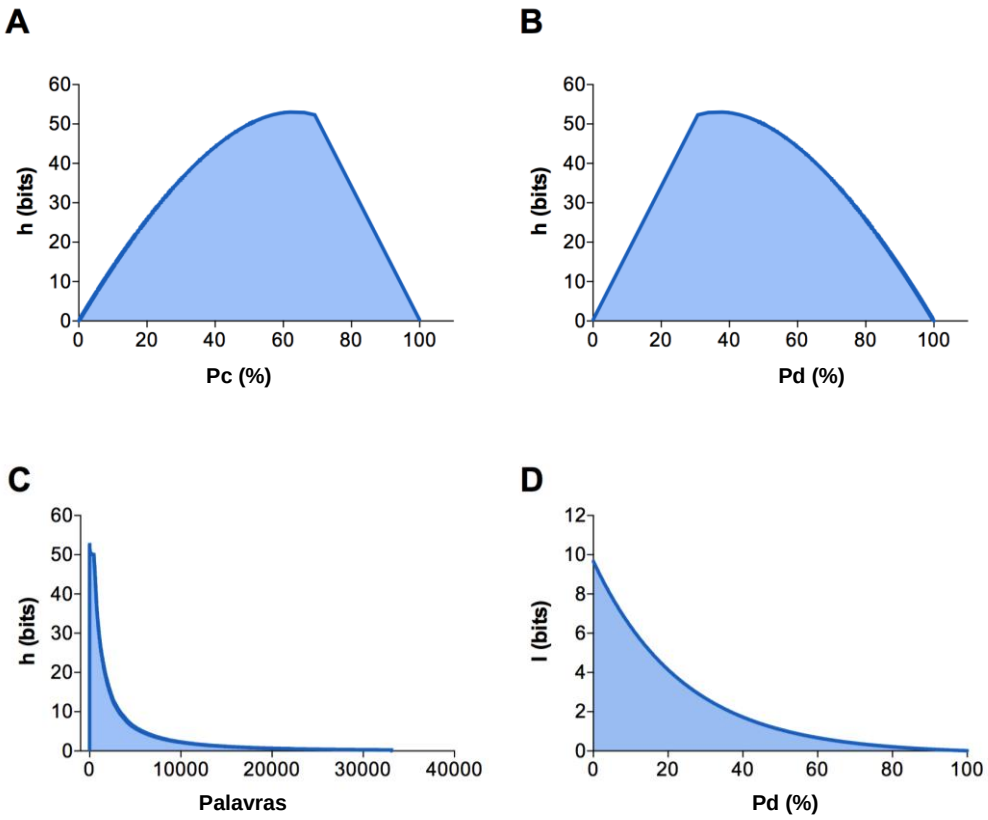
Como vantagem, pode-se ressaltar que, por ser um método não paramétrico, não assume nenhuma distribuição; além de apresentar uma excelente inspeção inicial dos dados, o que fornece uma imagem bastante clara de sua estrutura, garantindo resultados de fácil interpretação. Entre as suas desvantagens, podemos ressaltar que existe uma limitação a uma linguagem descritiva, baseada em atributos-valores; e que o mesmo conceito pode ser representado por diferentes árvores de decisão. Ainda, modelou-se a relação quadrática entre Entropia e a Probabilidade do Conhecimento por meio de um Modelo de Regressão Gaussiano (MONTGOMERY *et al*., 2006).

O *software* utilizado nas análises foi o R (versão 3.2.4). E utilizamos o CalcuLetra para a inserção e comparação aleatória de textos denotativos e suas respectivas Pc médias. Para identificar os pontos aproximados onde existe mudança na inclinação entre a quantidade de textos denotativos inseridos no CalcuLetra e as suas probabilidades de conhecimento médias, utilizamos um método de Segmentação Binária (SCOTT; KNOTT, 2006) que identifica mudanças na média e variância das probabilidades de conhecimento média, a medida em que o número de textos crescia. Para cada intervalo dos pontos de mudança, a relação entre as probabilidades de conhecimento e os textos foram descritas por regressões lineares.

5 RESULTADOS

Os limites do banco de palavras obtido são os resultados mais relevantes da pesquisa, já que um extremo representa um grupo de palavras com alta probabilidade de formar um texto de fácil compreensão ou ainda de contribuir neste sentido; enquanto que no outro grupo há alta probabilidade para a formação de um texto com ruídos no processamento mental, com palavras com menor utilização nos textos e, portanto, com maiores chances de não serem conhecidas a possíveis leitores em geral.

Figura 2. Distribuição das probabilidades, quantidade de informação e entropia informacional de cada palavra do banco armazenado no CalcuLetra. A, B: Desvio para as palavras pertencentes ao G1, com $P_c \geq ,7$. Resultado inverso ocorre em B, ao serem eventos exclusivos. **C**: Distribuição harmônica da *h*, exceto para o G1, envolvendo conjunções, preposições e artigos. Aproximadamente 5 mil palavras foram classificadas como as mais familiares ou conhecidas pela maioria dos usuários, sendo que há uma concentração das probabilidades entre 6 a 9 mil palavras. **D:** Relação monótona decrescente e não linear entre as variáveis. Quando a P_d é máxima, a I é mínima.



Em 2C, percebe-se que a maioria dos textos contém um vocabulário restrito, com uma quantidade relativamente pequena de palavras. Cerca de 17,33% dos textos apresentaram entropia maior que 10 e, estes, estão concentrados nas contagens mais baixas de número de palavras. Por exemplo, o ponto aproximado do número de palavras onde a entropia é igual a 10 é de 3259 e para entropia igual a 5 o número de palavras é aproximadamente igual a 5724. A contagem de 4 palavras está associada a entropia máxima (53,05).

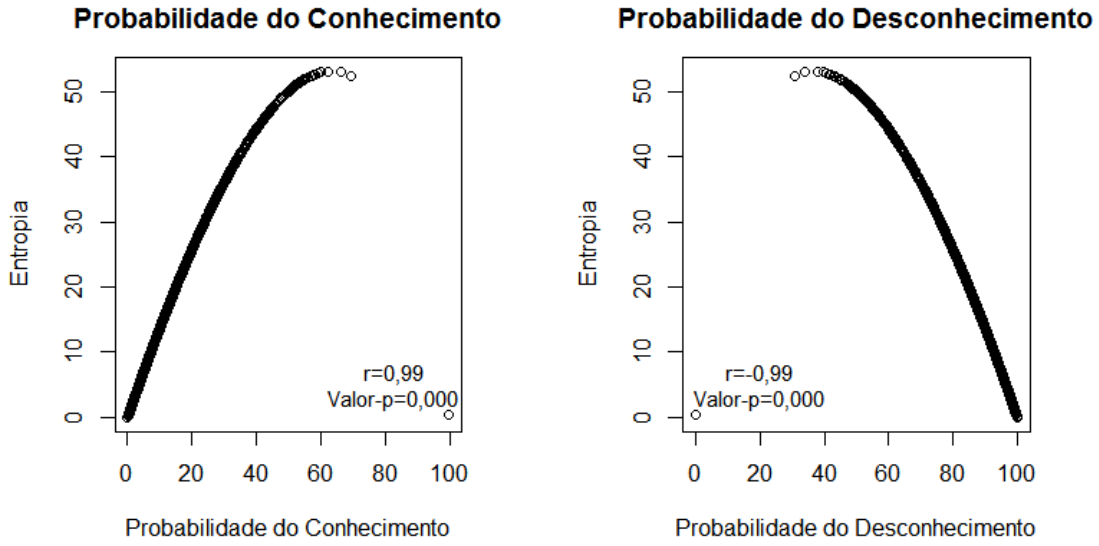
Nossos resultados mostraram um grupo de palavras com maior probabilidade de ocasionar a incompreensão e, consequente, falha na comunicação, durante a interação texto-leitor, bastando apenas uma ou duas palavras deste grupo estarem presentes em meio a outras redundantes. A Tabela 2 e o Gráfico 1 apresentam a correlação de Spearman da entropia com a probabilidade do conhecimento e a probabilidade do desconhecimento.

Tabela 2. Correlação de Spearman da entropia com as probabilidades do banco. Tem-se correlação positiva e significativa (r = 0,99, valor-p = 0,000) da entropia com a probabilidade do conhecimento, dessa maneira, quanto maior for a entropia, maior será a probabilidade do conhecimento e vice-versa; e correlação negativa e significativa (r = -0,99, valor-p = 0,000) da entropia com a probabilidade do desconhecimento, logo, quanto maior for a entropia, menor será a probabilidade do desconhecimento e vice-versa.

Variáveis	Probabilidade do Conhecimento	Probabilidade do Desconhecimento
Entropia	0,99 (0,000)	-0,99 (0,000)

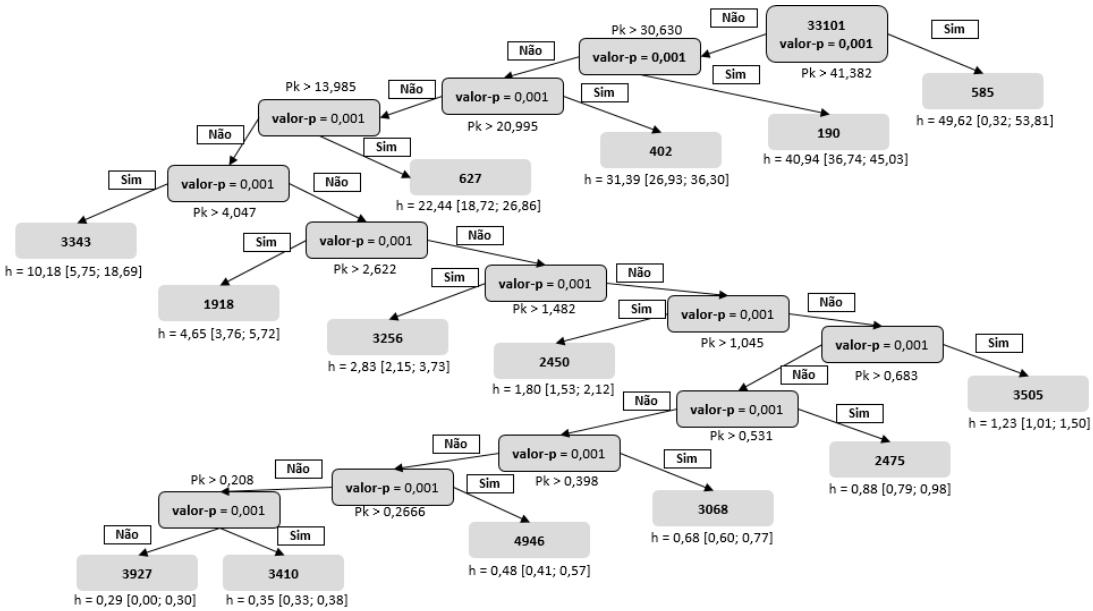
*Correlação de Spearman e valor-p – r (valor-p).

Gráfico 1. **Correlação de Spearman das variáveis com a entropia.** As distribuições possuem dois pontos atípicos: há uma observação que apresenta alto valor de Probabilidade do Conhecimento (99,97) e nível de entropia baixo (0,31) e há uma observação com alta Probabilidade do Desconhecimento (100,00) e nível entropia nulo (0,00). Palavras isoladas têm, em média, alta entropia informacional, ou seja, alto grau de incerteza de surgir ou não em um texto qualquer, entretanto, para um pequeno grupo de palavras responsáveis pela composição estrutural da linguagem, a entropia se mostrou baixa e as palavras técnicas com alta P_d têm entropia igualmente baixa por ser mais provável não as encontrarmos na leitura.



Para avaliar os pontos de corte da Probabilidade do Conhecimento e da Probabilidade do Desconhecimento em que a distribuição de Entropia fosse homogênea, usou-se a técnica de Aprendizado de Máquina "Árvore de Regressão". Como método de decisão foi estipulado que o número mínimo para divisão dos grupos seria 5000.

Figura 3. Árvore de Regressão para entropia (variável resposta) e Probabilidade do Conhecimento (explicativa). Em cada folha, a entropia está representada da seguinte forma: h = média da folha [valor mínimo da folha; valor máximo da folha].



A partir dos pontos de corte identificados pela Árvore de Regressão para a Probabilidade do Conhecimento, fez-se uma descritiva da Entropia a partir das faixas e os resultados são apresentados na Tabela 3 e no Gráfico 2. Tem-se que na faixa mais baixa de Probabilidade do Conhecimento ([0,000; 0,20833]) a média de Entropia foi igual a 0,286 [0,286; 0,287], sendo que os valores mínimo e máximo foram, respectivamente, 0,000 e 0,200; na faixa mais alta de Probabilidade do Conhecimento ([41,3823; 99,9730]) a média de Entropia foi igual a 49,615 [49,377; 49,780], sendo que os valores mínimos e máximos foram, respectivamente, 0,315 e 53,049. A árvore de decisão não foi capaz de diferenciar o ponto atípico devido falta de dados com altas probabilidades do conhecimento.

Tabela 3. Descritiva da variável Entropia por faixa de Probabilidade do Conhecimento.

Faixa de Pc	Entropia								
	N	Média	D.P.	I.C - 95%[1]	Mín.	1º Q	2º Q	3º Q	Máx.
(41,3823; 99,9730]	585	49,615	2,419	[49,377; 49,780]	0,315	50,008	50,057	50,179	53,049
(30,6300; 41,3823]	189	40,941	2,469	[40,599; 41,270]	36,739	38,780	40,784	43,269	45,030
(20,9950; 30,6300]	402	31,388	2,797	[31,124; 31,655]	26,925	28,924	31,221	33,705	36,601
(13,9980; 20,9950]	627	22,436	2,335	[22,254; 22,615]	18,717	20,366	22,318	24,387	26,860
(4,0466; 13,9980]	3343	10,183	3,541	[10,070; 10,302]	5,746	7,143	9,319	12,514	18,694
(2,2616;4, 0466]	1918	4,652	0,565	[4,627; 4,676]	3,760	4,160	4,609	5,113	5,718
(1,4820; 2,6116]	3256	2,829	0,465	[2,813; 2,845]	2,150	2,420	2,769	3,226	3,732
(1,0450, 1,4820]	2450	1,799	0,178	[1,793; 1,807]	1,526	1,635	1,797	1,961	2,122
(0,6833; 1,0450]	3505	1,228	0,147	[1,224; 1,233]	1,011	1,092	1,228	1,364	1,500
(0,5316; 0,6833]	2475	0,882	0,062	[0,879; 0,884]	0,794	0,820	0,875	0,930	0,982
(0,3983; 0,5316]	3068	0,677	0,055	[0,675; 0,679]	0,602	0,629	0,684	0,739	0,765
(0,2666; 0,3983]	4946	0,484	0,054	[0,482; 0,485]	0,411	0,437	0,492	0,521	0,574
(0,2083; 0,2666]	3410	0,354	0,023	[0,353; 0,355]	0,329	0,329	0,355	0,384	0,384
[0,000; 0,20833]	2927	0,286	0,014	[0,286; 0,287]	0,000	0,274	0,274	0,300	0,300

[1]Intervalo de Confiança Bootstrap para a média

Gráfico 2. **Descritiva da variável Entropia por faixa de Probabilidade do Conhecimento.**

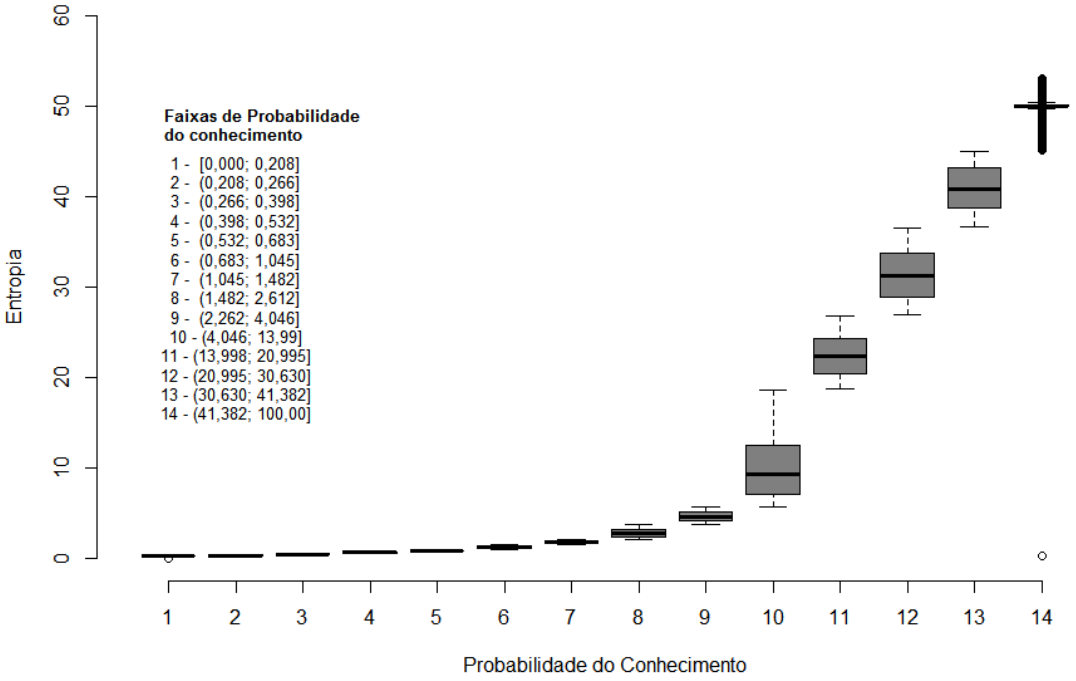
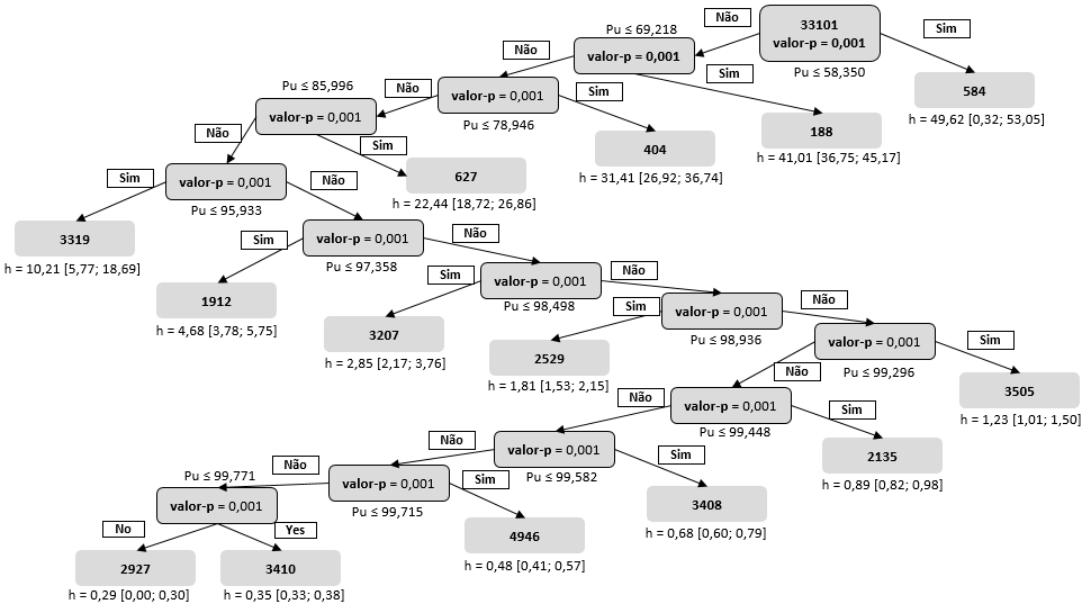


Figura 4. Árvore de regressão para Entropia (variável resposta) e Probabilidade do Desconhecimento (explicativa). Em cada folha, a entropia está representada da seguinte forma: h = média da folha [valor mínimo da folha; valor máximo da folha].



A partir dos pontos de corte identificados pela árvore de regressão para a Probabilidade do Desconhecimento, fez-se uma descritiva da Entropia a partir das faixas e os resultados são apresentados na Tabela 4 e no Gráfico 3. Tem-se que a faixa mais baixa de Probabilidade do Desconhecimento ([0,0265;

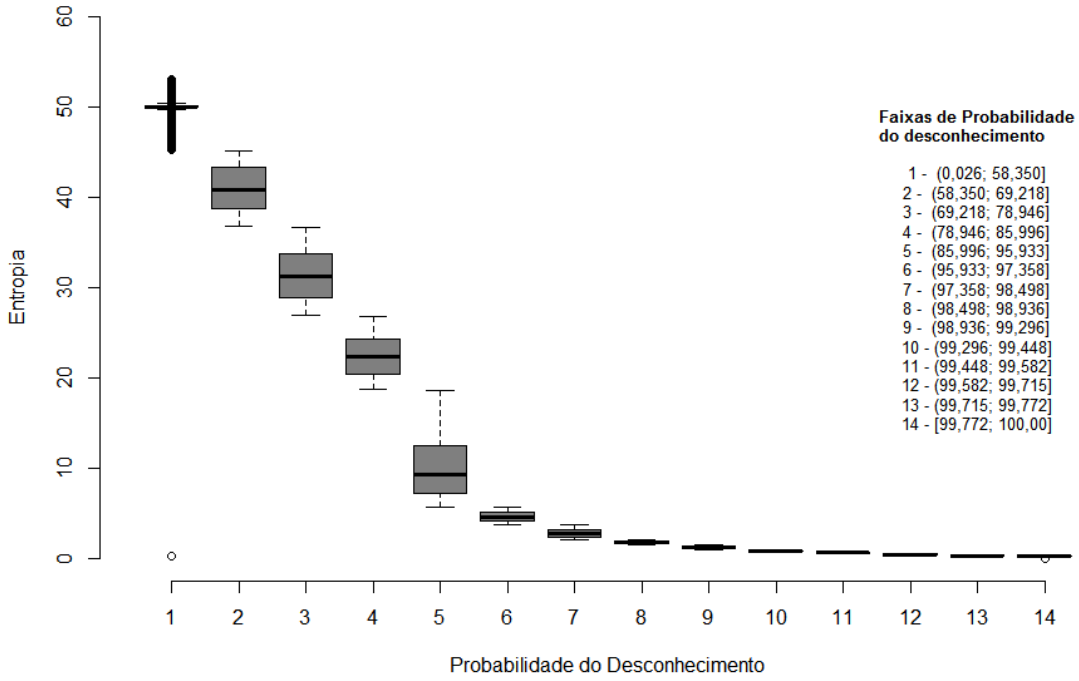
58,3500]) apresentou Entropia média de 49,622 [49,407; 49,783], sendo que os valores mínimo e máximo foram 0,315 e 53,049. A árvore de decisão não foi capaz de diferenciar o ponto atípico (como descrito anteriormente) devido a falta de dados com baixas probabilidades do conhecimento; e a faixa mais alta de Probabilidade do Desconhecimento ((99,7716; 100,00]) apresentou Entropia média de 0,286 [0,286; 0,287], sendo que os valores mínimo e máximo foram respectivamente, 0,000 e 0,300.

Tabela 4. Descritiva da variável Entropia por faixa de Probabilidade do Desconhecimento.

Faixa de Pd	Entropia								
	N	Média	D.P.	I.C - 95%[1]	Mín.	1º Q	2º Q	3º Q	Máx.
[0,0265; 58,3500]	584	49,622	2,414	[49,407; 49,783]	0,315	50,008	50,057	50,180	53,049
(58,3500; 69,2183]	188	41,009	2,456	[40,638; 41,354]	36,756	38,813	40,830	43,305	45,171
(69,21833; 78,9466]	404	31,414	2,815	[31,151; 31,690]	26,925	28,924	31,230	33,742	36,739
(78,9466; 85,9966]	627	22,436	2,335	[22,262; 22,617]	18,717	20,366	22,318	24,387	26,860
(85,9966; 95,9333]	3319	10,215	3,533	[10,102; 10,331]	5,771	7,183	9,371	12,564	18,694
(95,9333; 97,3583]	1912	4,679	0,568	[4,655; 4,704]	3,786	4,186	4,637	5,166	5,746
(97,3583; 98,4983]	3207	2,854	0,464	[2,838; 2,870]	2,176	2,446	2,795	3,252	3,760
(98,4983; 98,9366]	2529	1,810	0,186	[1,803; 1,817]	1,526	1,661	1,797	1,961	2,150
(98,9366; 99,2966]	3505	1,228	0,147	[1,223; 1,233]	1,011	1,092	1,228	1,364	1,500
(99,2966; 99,4483]	2135	0,896	0,055	[0,893; 0,898]	0,820	0,846	0,901	0,930	0,982
(99,4483; 99,5816]	3408	0,689	0,062	[0,687; 0,691]	0,602	0,629	0,684	0,739	0,794
(99,5816; 99,7150]	4946	0,484	0,054	[0,482; 0,485]	0,411	0,437	0,492	0,521	0,574
(99,7150; 99,7716]	3410	0,354	0,023	[0,353; 0,355]	0,329	0,329	0,355	0,384	0,384
(99,7716; 100,00]	2927	0,286	0,014	[0,286; 0,287]	0,000	0,274	0,274	0,300	0,300

[1]Intervalo de Confiança Bootstrap para a média

Gráfico 3. **Descritiva da variável Entropia por faixa de Probabilidade do Desconhecimento.**



5.1 REGRESSÃO QUADRÁTICA

Foram ajustados diversos modelos de Regressão Gaussiano na tentativa de modelar a relação das Probabilidades de Conhecimento e Desconhecimento com a Entropia. Percebe-se que a relação entre as variáveis não é linear, dessa forma foram ajustados vários modelos de regressão polinomiais na tentativa de encontrar uma função que descrevesse a relação entre as variáveis.

Ao final, foi preferível adotar a função com um termo quadrático (MONTGOMERY *et al*., 2006), já que a inclusão de mais termos polinomiais dificultava a interpretação da função e causava multicolinearidade do modelo, o que resultava em modelos sobre ajustados.

5.1.1 Probabilidade do conhecimento

Para descrever a relação entre Entropia e a Probabilidade do Conhecimento, utilizou-se um modelo de regressão quadrático, permitindo assim encontrar o valor da probabilidade do conhecimento que maximiza a entropia. Para um melhor ajuste, a variável Probabilidade do Conhecimento foi centrada na média. O modelo ajustado apresentou Fator de Inflação da Variância (VIF) de 6,24, evidenciando ausência de multicolinearidade.

A multicolinearidade acontece quando duas ou mais variáveis explicativas apresentam forte relação linear entre si, o que prejudica a qualidade do ajuste de regressão. A multicolinearidade é medida pelo VIF, que é um indicador que mede o quanto a variância dos coeficientes de regressão é influenciada por outras variáveis presentes. Segundo Fox (2008), variáveis que apresentam valor de VIF maior que 10 influenciam desproporcionalmente na estimação dos coeficientes de regressão e devem ser retirados do modelo. O modelo com termo quadrático apresentou VIF de 6,24, logo o mesmo pode ser considerado bem ajustado.

Logo, a equação do modelo para Entropia (h) e Probabilidade do Conhecimento (Pc) pode ser definida da seguinte forma:

$$\overline{Pc} = \text{Pc} - 3{,}210436$$

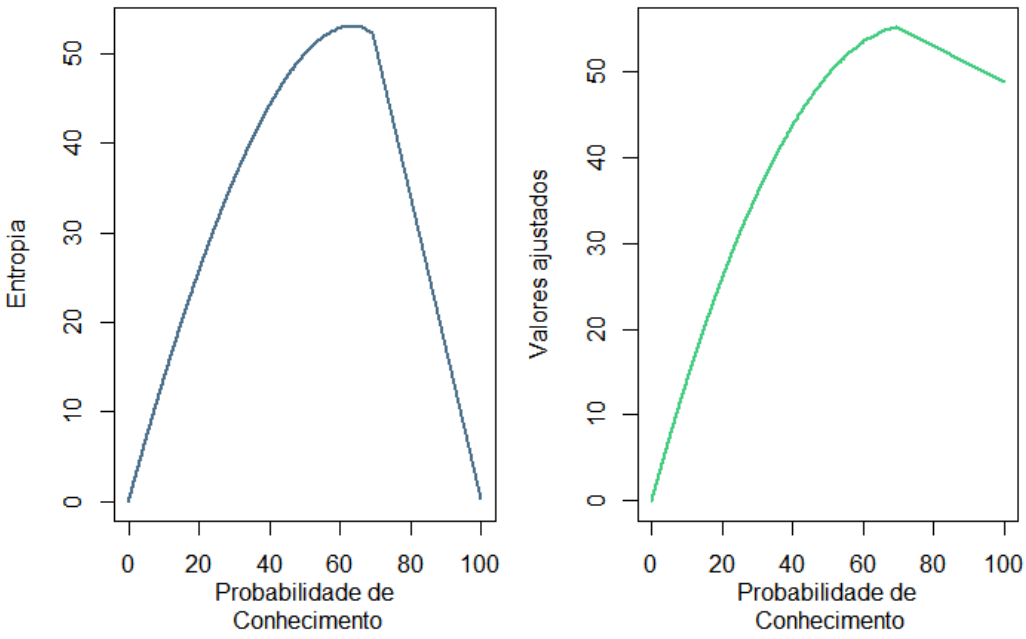
$$\hat{h} = 4{,}6457866 + 1{,}4350649 * \overline{Pc} - 0{,}0101171 * \overline{Pc}^2$$

Neste sentido, quando a Probabilidade do Conhecimento aumenta em 10

unidades a Entropia sofre um aumento médio de aproximadamente 13,92 unidades. A função nos permite encontrar pontos de máximo e estimar a Entropia a partir da Probabilidade de Conhecimento. Com o modelo de regressão quadrática ajustada pode-se verificar que o ponto que maximiza a Entropia é a probabilidade de Conhecimento de aproximadamente 70,92%, após esse valor à medida que se aumenta a probabilidade de conhecimento o valor da entropia tende a diminuir.

No Gráfico 4 é possível comparar os valores reais e os valores ajustados. Nota-se que após o ponto de inflexão, a curva ajustada não consegue acompanhar exatamente a real. Isso ocorre devido a pequena quantidade de pontos examinados nessa faixa, o que é uma limitação do banco, mas podemos inferir que as faixas de entropia tendem a diminuir para as palavras muito conhecidas.

Gráfico 4. **Modelo de Regressão Ajustado - Probabilidade do Conhecimento.** O modelo ajustado consegue modelar de forma razoável as características dos dados até próximo do ponto 60 de Probabilidade do Conhecimento. A falta de ajuste para valores superiores de Pc são um reflexo da falta de amostras com probabilidades superiores a 60 e também da presença de uma observação atípica (Pc = 99,97 e h = 0,31).



A Tabela 5 apresenta algumas medidas de erro, sendo elas: Desvio Absoluto Médio (MAD), que é a média simples das diferenças absolutas entre valores reais e ajustados; Desvio Quadrático Médio (MSD), que é a média simples do quadrado das diferenças entre valores reais e ajustados; Erro percentual absoluto médio (MAPE), que expressa o percentual de erros entre os valores reais e ajustados.

Tabela 5. **Medidas de desvio.** As estimativas de MAD e MSD foram baixas, sugerindo a proximidade dos valores ajustados com os dados reais. Além disso, a estimativa do MAPE foi de que, em média, a previsão está incorreta em 3,009%.

Desvios	
MAD	0,069
MSD	0,081
MAPE	3,009%

A Tabela 6 apresenta uma descritiva de Entropia e dos Valores ajustados. É possível observar que as médias são iguais e as estimativas dos quartis são muito próximas, evidenciando a qualidade de previsão do modelo.

Tabela 6. Comparação de Entropia e os Valores Ajustados. A diferença absoluta mediana foi de 0,047 e a máxima de 48,465, que é proveniente da observação atípica supracitada.

Variável	**N**	**Média**	**D.P.**	**Mín.**	**1º Q**	**2º Q**	**3º Q**	**Máx.**
Entropia	33101	4,019	8,552	0,000	0,466	0,982	2,984	53,049
Valores Estimados	33101	4,019	8,547	-0,066	0,418	0,955	3,025	55,295
|Entropia - Valores Ajustados|	33101	0,069	0,276	0,000	0,031	0,047	0,055	48,465

5.1.2 Probabilidade do desconhecimento

Para um melhor ajuste a variável Probabilidade do Desconhecimento foi centrada na média. A equação do modelo para Entropia (h) e Probabilidade do

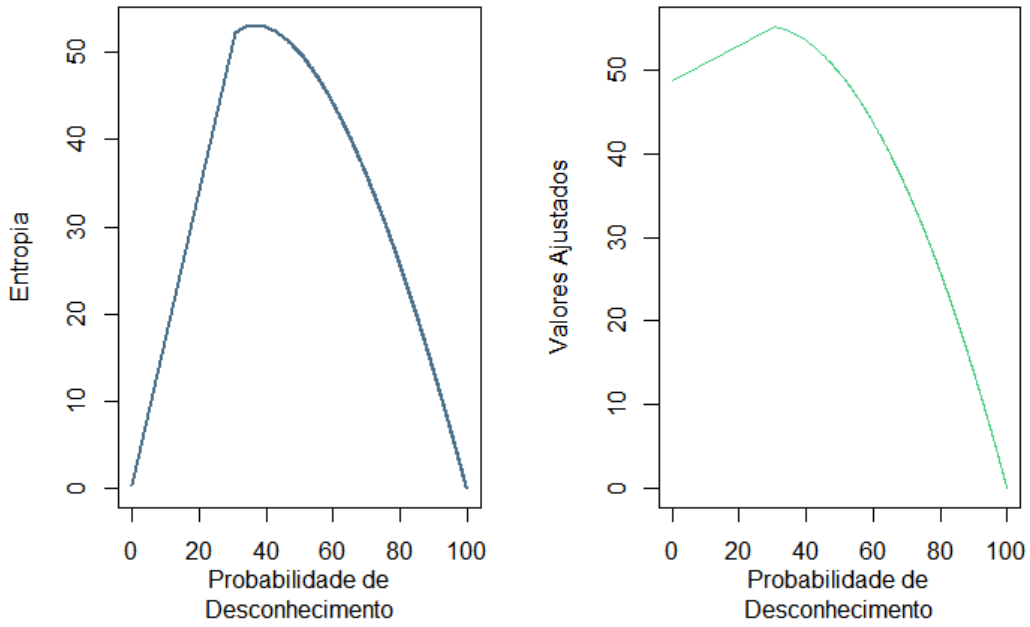
Desconhecimento (Pd) é definida da seguinte forma:

$$\overline{Pd} = \text{Pd} - 96{,}78956$$

$$\hat{h} = 4{,}6457866 - 1{,}4350649 * \overline{Pd} - 0{,}0101171 * \overline{Pd}^2$$

Com o modelo de regressão quadrática ajustada, pode-se verificar que o ponto que maximiza a Entropia é a probabilidade de desconhecimento de aproximadamente 30,74%, após esse valor à medida que se aumenta a probabilidade de desconhecimento o valor da entropia tende a diminuir. As estimativas de MAD, MSD e MAPE do modelo para Probabilidade de Desconhecimento foram idênticas às do modelo para Probabilidade do Conhecimento. Ainda, nota-se que as funções quadráticas para os dois modelos são muito parecidas e só diferem no sinal do coeficiente referente a probabilidade de conhecimento/desconhecimento.

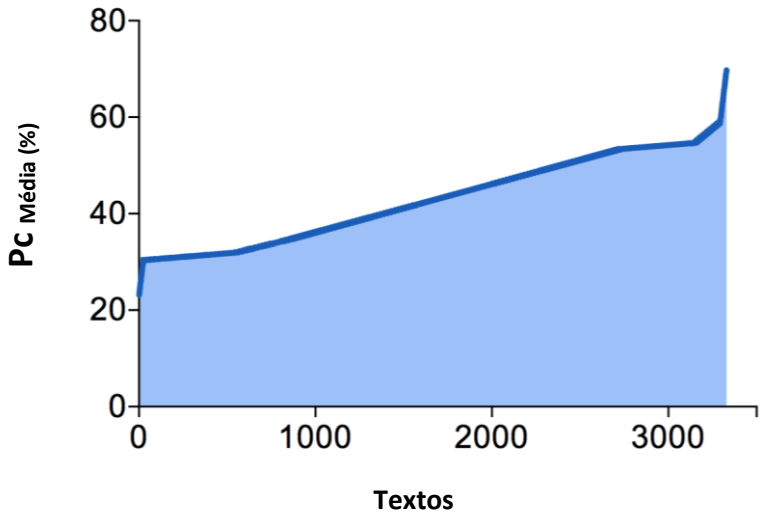
Gráfico 5. **Modelo de Regressão Ajustado** - Probabilidade de Desconhecimento.



5.2 QUANTIFICAÇÃO DE TEXTOS DENOTATIVOS NO CALCULETRA

A partir da inserção de textos não-literários ou denotativos no CalcuLetra (n=3.330), comparando-os com os valores matemáticos e estatísticos obtidos no banco de palavras obtido, revela-se quão compreensivo é o texto inserido, considerando que o seu conteúdo possua os devidos elementos de coesão textual, conforme as regras gramaticais da língua, já que o aprendizado sintático se consolida por volta dos 12 anos de idade (BERNS, 2002). Verificamos o valor de Pc média inferior a ,3 para textos com maiores chances de causar incompreensão.

Figura 5. Registro randômico da inserção de textos não-literários no CalcuLetra e as suas respectivas Pc médias. Os textos inseridos (ex., artigos, redações, notícias, resumos, materiais didáticos) foram delimitados conforme modelo similar ao mostrado no Apêndice 2, o que corresponde a aproximadamente 30 linhas no MS Word, com margens padrão de 3 cm. As palavras não encontradas no banco do Calculetra são automaticamente desconsideradas. Pc = ,3 ≥ ,5 para a maioria dos textos analisados, podendo os textos com Pc média ≤ ,3 serem considerados como de difícil compreensão. Significância estatística Pc ≥ ,21. Variação total da Pc média = ,23 ≥ ,69. Pc média = ,42.



Utilizamos o método de Segmentação Binária (SCOTT; KNOTT, 2006) que

identifica mudanças na média e variância de P_c a medida em que o número de textos crescia. Dessa forma, o método identificou os pontos: 24, 551, 2724, 3152 e 3295.

Nota-se que, dentro de cada intervalo dos pontos de mudança, a relação entre Pc e os textos é estritamente linear. Dessa forma, 6 regressões lineares foram ajustadas considerando os intervalos de tempo identificados nos pontos de mudança e os resultados são apresentados na Tabela 7 e no gráfico 6. Logo, vale ressaltar que:

· No intervalo [1; 24], quando o número de textos aumenta em 10 unidade, a P_c aumenta, em média, em 3,00 unidades;

· No intervalo (25; 551), quando o número de textos aumenta em 10 unidades, a P_c aumenta, em média, em 0,03 unidades;

· No intervalo (552; 2724], quando o número de textos aumenta em 10 unidades, a P_c aumenta, em média, em 0,10 unidades;

· No intervalo (2725; 3152], quando o número de textos aumenta em 10 unidades, a P_c aumenta, em média, em 0,03 unidades;

· No intervalo (3153; 3295], quando o número de textos aumenta em 10 unidades, a P_c aumenta, em média, em 0,30 unidades;

· No intervalo (3296; 3330], quando o número de textos aumenta em 10 unidades, a P_c aumenta, em média, em 3,00 unidades.

· Em todos os modelos o valor de R^2 foi praticamente perfeito.

Tabela 7. Regressão linear gaussiana para os textos denotativos e a Pc.

Amostra de textos	β_o			β_1			R^2
	Coef.	E.P	Valor-p	Coef.	E.P	Valor-p	
[1; 24]	22,881	0,000	0,000	0,300	0,000	0,000	1,000
(25; 551)	30,378	0,000	0,000	0,003	0,000	0,000	1,000
(552; 2724]	31,737	0,002	0,000	0,010	0,000	0,000	0,999
(2725; 3152]	53,400	0,000	0,000	0,003	0,000	0,000	0,999
(3153; 3295]	54,734	0,004	0,000	0,030	0,000	0,000	0,999
(3296; 3330]	59,298	0,000	0,000	0,300	0,000	0,000	0,999

O Gráfico 6 apresenta a comparação entre os valores reais e os valores ajustados. É possível perceber que os modelos conseguem captar bem as características dos dados.

Gráfico 6. **Modelo de Regressão Ajustado para os textos denotativos e as suas respectivas Pc médias.**

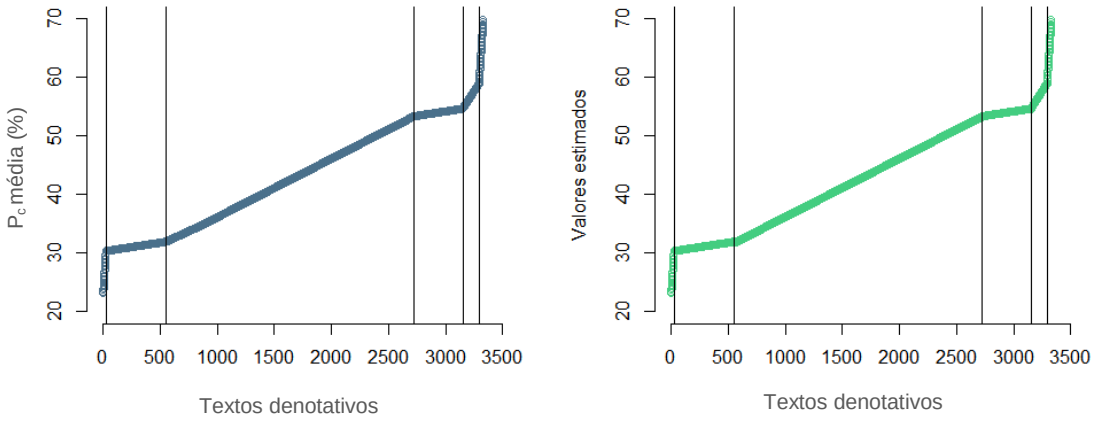


Tabela 8. Medidas de desvio. As estimativas de MAD (0,024) e MSD (0,002) foram baixas, sugerindo a proximidade dos valores ajustados com os dados reais. Além disso, a estimativa do MAPE foi de que, em média, a previsão está incorreta em 0,064%.

Desvios	
MAD	0,024
MSD	0,002
MAPE	0,064%

Para a classificação dos textos, a P_c se mostrou mais fidedigna em relação à entropia informacional, já que esta indica apenas a caoticidade da distribuição probabilística, podendo, porém, ser utilizada para a otimização da aprendizagem quanto menor for a sua presença em um texto, este sendo necessariamente composto em grande parte por palavras comuns ou redundantes. Desta maneira, um texto considerado complexo a partir do método exposto deve ser incompreensível aos usuários da língua em geral, contendo necessariamente em sua maior parte palavras de baixa familiaridade.

6 DISCUSSÃO

As análises de complexidade textual geralmente incluem vários domínios (CROSSLEY *et al*., 2017; GRAESSER; MCNAMARA; KULIKOWICH, 2011). Nesta pesquisa, a dificuldade de compreensão escrita foi quantificada para avaliar objetivamente a complexidade de textos de caráter predominantemente denotativo em português e já organizados sintaticamente na norma padrão da língua, com base no vocabulário usado por um determinado autor. A compreensão do texto foi estimada a partir das correlações entre as representações semânticas das palavras e suas respectivas probabilidades para assegurar que a complexidade textual estava profundamente relacionada com o seu grau de familiaridade e a entropia semântica das palavras.

A compreensão textual é o objetivo final do processo de decodificação exigido na atividade de leitura, considerando que os processos de decodificação (reconhecimento de palavras e entendimento do texto) podem dar-se independentemente, contudo, a sua colaboração é necessária para que possa atingir a compreensão (CITOLER; SANZ, 1997) e construir um conhecimento semântico a partir das informações obtidas pela interação com o material lexical já organizado sintaticamente.

Na literatura, o presente estudo pode ser considerado inovador no campo, sendo a primeira investigação sobre dificuldade de compreensão escrita em português utilizando métodos quantitativos e parâmetros objetivos, com contribuições relevantes para pesquisas futuras. Nossos resultados mostraram que houve déficit no processamento de saída (escrita), o que aumenta chances de ocorrer

limitações no processamento de entrada (leitura), considerando que os dicionários de português dispõem de mais de 400 mil verbetes e nossos resultados apontam para a utilização média de 6 a 9 mil palavras, há, portanto, ainda lacunas a serem ultrapassadas. Entendemos que a assimilação e a utilização de palavras funcionam no processo evolutivo como um catalisador cognitivo em prol da transmissão cultural, esta capaz de modificar vieses.

Ao apresentarmos uma análise probabilística de palavras em língua portuguesa moderna, explicitamos todos os dados relevantes obtidos. Mostramos que o texto constituído em sua maior parte por palavras redundantes tem alta quantidade de informação, com maior probabilidade de ser compreendido pela maioria dos leitores, o que por outro lado faz com que mesmo um usuário proficiente na língua tenha ruídos no seu processamento mental ao se deparar com palavras incomuns.

Embora a heterogeneidade semântica tenha sido frequentemente considerada enquanto obstáculo à interoperabilidade dos conjuntos de dados, encontramos semelhanças entre grupos de palavras onde a entropia informacional revelou-se baixa, conforme mostrado na Tabela 9 e na Tabela 10.

Tabela 9. Exemplos aleatórios de palavras com Pc ≥ ,7. Este grupo de palavras foi predominante por elementos conectivos para coesão textual (ex., conjunções e preposições).

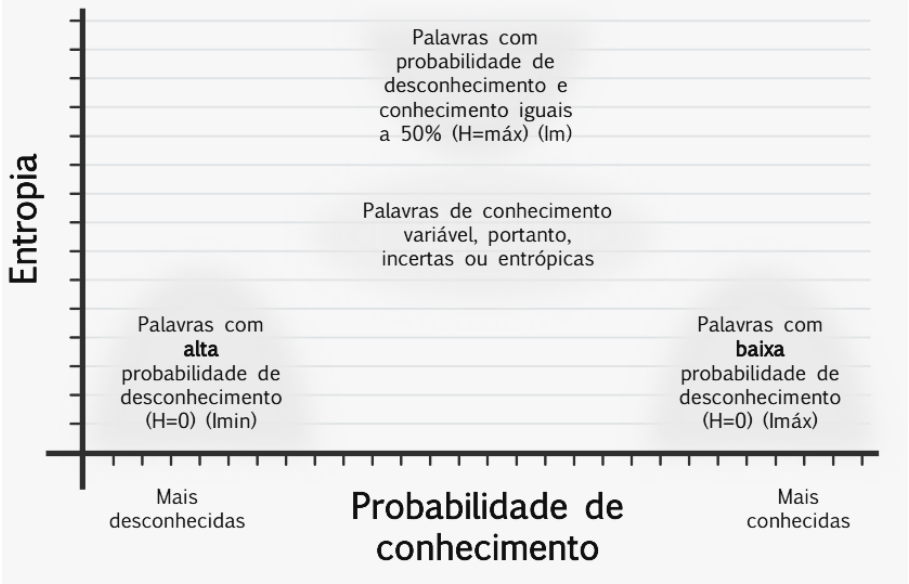
ID	Palavra	*F*	*P (%)*	*Pc (%)*	*Pd (%)*	*I (bits)*	*h (bits)*
107504	Com	240327	2,739839	97,843672	2,156328	5,535279	0,119358
110952	Para	232542	2,651089	94,674283	5,325717	4,230880	0,225324
118876	Que	227632	2,595119	92,675482	7,324518	3,771122	0,276216

Tabela 10. Exemplos aleatórios de palavras com Pd ≥ ,3. Este grupo de palavras foi predominante por elementos semanticamente significativos para a coerência textual (ex., substantivos e adjetivos).

ID	Palavra	*f*	*P (%)*	*Pc (%)*	*Pd (%)*	*I (bits)*	*h (bits)*
112540	Percolação	40	0,000456	0,760000	99,240000	0,011005	0,010921
227	Alvarenga	64	0,000730	1,216667	98,783333	0,017660	0,017445
159	Albumina	289	0,003295	5,491667	94,508333	0,081487	0,077012

Em média, textos voltados ao público infantil contêm palavras familiares, comuns ou redundantes a maioria dos usuários da língua. O rebuscamento textual mostra-se passível de quantificação, ao ser capaz de gerar a incompreensão por sua imprevisibilidade ou baixa familiaridade ao leitor, pouca quantidade de informação e baixa entropia informacional, conforme a Figura 6 sumariza. A partir da análise, foi possível identificar tais ruídos no processamento cognitivo.

Figura 6. Visão geral dos parâmetros probabilísticos do banco de dados.



O modelo apresentado contribui para uma avaliação objetiva da compreensão textual, ao relacionar as representações semânticas de palavras e sua ocorrência, para quantificar a compreensão a partir dos mais prováveis ruídos semânticos da linguagem denotativa em português. Neste sentido, podemos sugerir a dificuldade da compreensão escrita para uma avaliação da

complexidade textual, que está mais frequentemente relacionada ao grau de entropia semântica embutida no significado de suas palavras, ou seja, no seu grau de familiaridade. As nuances do discurso podem até passar despercebidas pelo leitor ou mesmo se diferirem entre um grupo de leitores, mas, se as palavras forem redundantes, a compreensão nunca será nula.

Embora o entendimento do funcionamento do processamento e armazenamento de informações requerer cada vez mais metodologias inovadoras, já se mostrou que as palavras relacionadas são armazenadas em conjunto, já verbos e substantivos permanecem alocados em diferentes áreas (VIGLIOCCO *et al*., 2011). Este também foi o caso em nosso banco de dados, considerando que a maioria das conjugações verbais apresentou alta entropia e que, apesar da baixa ocorrência, seus radicais implicavam o significado ou a decodificação semântica mais provável, o que as tornaram irrelevantes no presente estudo.

Não é possível ter ciência de quais palavras virão em um texto, porém, o alto grau de imprevisibilidade indica maior probabilidade de seu conteúdo não ser compreendido, pois se previsível, é redundante. Em um estudo de frequência de palavras em livros, a quantidade de informação em 300 a 3000 palavras imprevisíveis foi relatada como sendo de 0,2 bits por palavra (MARCELO; DAMIAN, 2010), o que foi congruente com os achados em nosso conjunto de dados.

As equações da teoria da informação e a confiabilidade dos seus princípios foram amplamente validados, confirmando que aspectos complexos da

informação, organizados em sequências simbólicas, eram susceptíveis de quantificação e análise usando essa teoria (MONTEMURRO; ZANETTE, 2002; MONTEMURRO; ZANETTE, 2011; DEBOWSKI, 2011; KALIMERI *et al*., 2015). No entanto, neste trabalho, classificamos a dificuldade de decodificação semântica usando fórmulas de equivalência pré-estabelecidas, de modo que um índice médio para a compreensão de leitura de texto denotativo em português fosse definido. Este estudo também diferiu de outras pesquisas que buscavam quantificar a entropia da ordem das palavras em frases, cujos resultados mostraram que houve entropia relativamente constante, apesar da diversidade de vocabulário (MONTEMURRO, 2014).

Avaliamos, portanto, a compreensão textual com a premissa de que quanto mais compreensível, menos complexo, sem dissociar compreensão e complexidade, tendo em vista a improbabilidade de um texto de alta complexidade em termos semânticos ser compreendido por indivíduos de baixa instrução. Ao ter ciência de que cada texto armazenado no CalcuLetra é um produto do repertório linguístico do seu autor, consideramos a medida da complexidade e consequentemente a compreensão como sendo o grau de familiaridade das palavras, ao estabelecermos que textos complexos são aqueles com palavras incomuns de baixa frequência, passíveis de identificação e quantificação como demonstrado nos nossos resultados.

Nossos resultados sugerem que usuários da língua alfabetizados e neurotípicos podem não compreender textos inseridos no CalcuLetra classificados como de difícil compreensão, apesar da sua visão de mundo, cultura, memórias,

experiências e percepções semânticas individuais. Acreditamos, contudo, que experimentos mentais concretos também precisam ser realizados em avaliações de escrita e leitura quando comparados com a média estatística de nível universitário encontrada no CalcuLetra, para confirmar a eficácia deste modelo e assim encorajar o uso de palavras incomuns por indivíduos, o que aumentaria sua capacidade de compreensão em si, com a possibilidade de alterar a distribuição global da capacidade de compreensão como um todo entre os estudantes. À medida que identificamos níveis de compreensão textual, nossos resultados podem ser aplicados também para a elaboração ou seleção de livros didáticos e avaliações escritas ou redações.

Pesquisas adicionais são necessárias para a quantificação da compreensão escrita em português, principalmente em redes neurais artificiais e na identificação de leitores com distúrbios de aprendizagem, que ainda carece de um método de diagnóstico objetivo, bem como estudos sobre redes cognitivas para elucidar a quantificação da memória semântica, análise de frequência e representações mentais (FEINSTEIN, 2011), para um melhor entendimento das relações entre a compreensão escrita em nível semântico, a frequência de palavras na língua e o processamento mental de signos de forma quantitativa e estatística.

Planejamos expandir o banco de dados do CalcuLetra com o mínimo de erros possível e disponibilizá-lo para futuras pesquisas. Atualmente, estamos criando um dicionário digital para as palavras desconhecidas mais usadas identificadas neste estudo para estimular o uso mais eficaz da ferramenta de aprendizado e

pretendemos desenvolver um aplicativo educativo auxiliar na aquisição de vocabulário dirigido tanto para usuários iniciantes quanto fluentes de língua portuguesa. É nossa meta também testar indivíduos alfabetizados (tanto neurotípicos como não-neurotípicos), com idade superior a 11 anos, uma vez que esta é a idade em que as pessoas iniciam a realização de operações cognitivas complexas (ILIAS; ESA, 2017), usando os textos denotativos inseridos no CalcuLetra, com base nas suas probabilidades de conhecimento média e grau de familiaridade, para confirmar nossos resultados e verificar quão distante os indivíduos se encontram da média universitária obtida. Nosso estudo certamente abre vias para outras aplicações e deve ser amplamente lido por profissionais interdisciplinares em semântica, linguística computacional e aprendizagem de línguas.

Apesar dos resultados preliminares, este estudo foi mais uma prova de conceito para o método empregado, diferentemente das outras metodologias de abordagem sintática existentes na literatura (GASPERIN *et al*., 2009; ALUÍSIO *et al*., 2008A; CASELI *et al*., 2009; CANDIDO JUNIOR *et al*., 2009; WATANABE *et al*., 2009; GASPERIN; MASIERO; ALUISIO, 2010; SCARTON; ALUÍSIO, 2010), demonstrando o seu potencial para futuras pesquisas. O modelo aqui introduzido pode ser universalmente adaptado a outras línguas para determinar a dificuldade de compreensão textual, no sentido de estendê-la ou aprimorá-la para alcançar diferentes objetivos e metas.

7 CONCLUSÃO

A partir do banco de palavras de sentido denotativo obtido e de suas respectivas fórmulas probabilísticas de conhecimento ou grau de familiaridade armazenados no *software* CalcuLetra, inserimos para o teste e quantificação de textos não-literários, encontrando um índice de dificuldade de compreensão textual em português. É possível, assim, obter níveis médios estatísticos de compreensão escrita a partir dos parâmetros objetivos pertencentes ao modelo quantitativo apresentado, o que pode ser o início para a elaboração de uma relevante ferramenta de avaliação de padrões comportamentais no processo de escrita e leitura, bem como de transtornos nesse comportamento, cujo principal indicador é a incompreensão textual. Nossos resultados revelam grupos de palavras que causam a incompreensão ou facilitam a leitura. Adicionalmente, apontamos lacunas em média enfrentadas, no que tange à aquisição e utilização de vocabulário da língua, ainda distante de ser usufruído em sua totalidade. Nosso modelo pode ser aplicado como instrumento de avaliação auxiliar em exames dissertativos de concursos públicos e/ ou de vestibulares em prol de uma avaliação objetiva da produção textual, ou ainda para a elaboração e análise de materiais didáticos.

REFERÊNCIAS

Arnaldo Candido Junior, Erick Maziero, Caroline Gasperin, Thiago Pardo, Lucia Specia and Sandra M. Aluisio. 2009. Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese. In the Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications, pages 34–42, Boulder, Colorado, June 2009.

Bao J. Towards a theory of semantic communication. Technical report, Rensselaer Polytechnic Institute. 2011.

Bar-Hillel Y, Carnap R. An outline of a theory of semantic information. Research Laboratory of Electronics Technical Report 247, Massachusetts Institute of Technology. 1952.

Bergelson E, Swingley D. The acquisition of abstract words by young infants. Cognition, 2013;127.

Bergelson, E., & Swingley, D. At 6–9 months, human infants know the meanings of many common nouns. Proceedings of the National Academy of Sciences, 109(9), 3253–3258, 2012.

Berns RM. O Desenvolvimento da Criança. Tradução por: Cecilia Camargo Bartalotti e Marcos Bagno. Edições Loyola, 2002; 800.

biblio.com.br [Internet]. Biblioteca Virtual de Literatura; c2017 [cited 2017 Mar 20]. Available from: http://www.biblio.com.br

Booth JR, Burman DD, Meyer JR, Gitelman DR, Parrish TB, Mesulam MM. Modality independence of word comprehension. Hum Brain Mapp. 2002a; 16: 251-61.

BORGES, S. Psicanálise, linguística, linguisteria. São Paulo: Escuta, 2010.

Breiman L, Friedman J, Olshen R, Stone C. Classification of regression trees. Wadsworth Books. 1984; 358.

Brent M., Siskind JM. The role of exposure to isolated words in early vocabulary development. Cognition, 2001; 81: B33-B44.

Caroline Gasperin, Erick Masiero and Sandra M. Aluisio. Challenging choices for text simplification, 2010.

cervantesvirtual.com [Internet]. Fundação Biblioteca Nacional; c2017 [cited 2017 Mar 20]. Available

from:http://www.cervantesvirtual.com/portal/fbn/cat_titulos.shtml

Champollion L. Quantification and negation in event semantics. *Baltic International Yearbook of Cognition, Logic and Communication*: 2011; 6: 1-23.

Citoler, D. C. & Sanz, O. R. A Leitura e a Escrita: Processos e Dificuldades na sua Aquisição. In: R. Bautista, Necessidades Educativas Especiais,(pp. 127-129). Lisboa: Dinalivro, 1997.

Cohen L, Dehaene S, Naccache L, Lehéricy S, DehaeneLambertz G, Hénaff M, et al. The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. Brain. 2000; 123: 291-307.

corpusbrasileiro.pucsp.br [Internet]. Pontifícia Universidade Católica de São Paulo (PUC-SP): Corpus Brasileiro; c2017 [cited 2017 Mar 20]. Available from: http://corpusbrasileiro.pucsp.br/cb/Inicial.html

corpusdoportugues.org [Internet]. Mark Davies: Corpus do Português; c2017 [cited 2017 Mar 20]. Available from: http://www.corpusdoportugues.org/

Crossley SA, Skalicky S, Dascalu M, McNamara D, Kyle K. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. Discourse Processes. 2017; 54: 340.

D'Alfonso S. On quantifying semantic information. Information. 2011; 2: 61-101.

D'Alfonso S. Review of information: A very short introduction. Essays in Philosophy. 2010; 11: 2-10.

Debowski T. On the vocabulary of grammar-based codes and the logical consistency of texts. IEEE Trans Inf Theory. 2011; 57: 4589-4599.

DIAS, Maria Sara de Lima; KAFROUNI, Roberta; BALTAZAR, Camilla Silva and STOCKI, Juliana. A formação dos conceitos em Vigotski: replicando um experimento. Psicol. Esc. Educ, vol.18, n.3, p.493-500, 2014.

DOR, J. Inconsciente. In: KAUFMANN, P. Dicionário Enciclopédico de Psicanálise: o legado de Freud e Lacan. Rio de Janeiro: Jorge Zahar, 1996.

en.wiktionary.org [Internet]. Wiktionary; c2017 [cited 2017 Mar 20]. Available from: https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/BrazilianPortuguese_wordlist

Feinstein, S. A Aprendizagem e o Cérebro. Lisboa: Horizontes Pedagógicos, 2011.

Ferreira, A. B. H.. Dicionário Aurélio. Editora positivo, 5º Edição, 208 p., 2014.

figaro.fis.uc.pt [Internet]. Página sobre Eça de Queirós; c2017 [cited 2017 Mar 20]. Available from: http://figaro.fis.uc.pt/queiros/eca_intro.html
FINATTO, M. J. B. Complexidade Textual Em Artigos Científicos: Contribuições Para O Estudo Do Texto Científico Em Português. Organon (UFRGS), v. 50, p. 30-45, 2011.

Flavell, J. Speculations about the nature and development of metacognition. In: F. Weinert & R. Kluwe (Ed.), Metacognition, motivation, and understanding. Hillsdale, NJ: Lawrence Erlbaum, p. 21-29, 1987.

Fourtassi A, Dupoux E. The role of word-word co-occurrence in word meaning learning. In proceedings of the 38th Annual Meeting of the Cognitive Science Society, Poster; 2016.

Fox, J. Applied Regression Analysis and Generalized Linear Models, Second Edition. Sage, 2008.

Gasperin C, Specia L, Pereira T, Aluísio S. Learning when to simplify sentences for natural text simplification. In: Proceedings of ENIA; 2009. 809-818.

github.com [Internet]. Hermit D; c2017 [cited 2017 Mar 20]. Available from: https://github.com/hermitdave/FrequencyWords/blob/master/content/2016/pt_br/pt_br_50k.txt

Graesser, A., McNamara, D., & Kulikowich, J. Coh-Metrix: Providing multilevel analysis of text characteristics. *Educational Researcher*, *40*, 223–234, 2011.

Hansson K, Baath R, Löhndorf S, Sahlén B, Sikström S. Quantifying Semantic Linguistic Maturity in Children. *Journal of Psycholinguistic Research*, 2015.

Hartmann N, Avanço L, Balage P, Magali D, Booth MGV, Pardo T, Aluísio S. A large corpus of product reviews in Portuguese: Tackling out-of-vocabulary words. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC' 14), Reykjavik, Iceland. European Language Resources Association (ELRA); 2014.

Helena Caseli, Tiago Pereira, Lucia Specia, Thiago Pardo, Caroline Gasperin and Sandra Aluísio. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In: Alexander Gelbukh (ed). Advances in Computational Linguistics, Research in Computer Science, vol 41, p. 59-70, 2009.

Ilias MR, Esa A. Children's psychology development. Psychology, 2017; 6924-46925. ISSN 2229-712X.

Indefrey, P. The spatial and temporal signatures of word production components: A critical update. Frontiers in Psychology, Volume 2, 255, 2011.

Jefferies, E. The neural basis of semantic cognition: Converging evidence from neuropsychology, neuroimaging, and TMS. Cortex, 49, 611-625, 2013.

JOLY, M. C., SANTOS, L. M., & MARINI, J. A. Uso de estratégias de leitura por alunos do Ensino Médio. Paidéia: Revista Cadernos de Psicologia e Educação, Ribeirão Preto, v. 34, p. 204-214, 2006.

Kalimeri M, Constantoudis V, Papadimitriou C, Karamanos K, Diakonos FK, Papageorgiou H.Word-length entropies and correlations of natural language written texts. J Quant Linguist. 2015; 22(2): 101-118.

Kemmerer, D. The cognitive neuroscience of language: An introduction. New York: Psychology Press, 2014.

Kendeou P, van den BROEK P, Helder A, Karlsson J. A cognitive view of reading comprehension: Implications for reading difficulties. Learn Disabil Res Pract. 2014; 29(1): 10-16.

Landauer T, Dumais S. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev. 1997; 104: 211-240.

LOPES, E. J. (Org.). Temas em Ciências Cognitivas & Representação Mental. Porto Alegre: Sinopsys, 2012.

Marcelo AM, Damian HZ. Towards the quantification of the semantic information encoded in written language. Adv Compl Sys. 2010; 13(02):135-153.

Mário Eduardo Viaro, Zwinglio O. Guimarães-Filho. Análise quantitativa da frequência dos fonemas e estruturas silábicas portuguesas. USP, São Paulo, Estudos Linguísticos XXXVI (1), p.27 / 36, janeiro-abril, 2007.

Mark D. A Frequency Dictionary of Portuguese. 2011: Routledge.

Matthewson, L. The measurement of semantic complexity: How to get by if your language lacks generalized quantifiers. In: Frederick J. Newmeyer & Laurel B. Preston (eds.), Measuring grammatical complexity. Oxford: Oxford University Press, 241—263, 2014.

MATURANA, Humberto, MPODOZIS, Jorge, LETELIER, Juan Carlos, Brain, Language and the Origin of Human Mental Functions, Biological Research, 28: 15-26, 1995.

Montemurro M, Zanette DH. Universal entropy of word ordering across linguistic families. PLOS ONE. 2011; 6(5): e19875.

Montemurro MA, Zanette DH. Complexity and universality in the long-range order of words. In: M. Degli Esposti et al., eds., Creativity and Universality in Language, Springer, Berlin. 2016.

Montemurro MA, Zanette DH. Entropic analysis of the role words in literary texts. Adv Compl Sys. 2002; 5(1): 7-17.

Montemurro MA. Quantifying the Information in the Long-Range Order of Words: Semantic Structures and Universal Linguistic Constraints. Cortex, in press (http://dx.doi.org/10.1016/j.cortex.2013.08.008); 2014.

Montgomery D, Peck A, Viving G. Introduction to Linear Regression Analysis, 4th ed. Hoboken: John Wiley. 2006.

Nelson, T., & Narens, L. Why investigate Metacognition?. In: J. Metcalfe & A. P. Shimamura (Ed.), Metacognition. Knowing about knowing. Cambridge, MA: MIT Press, p. 1-27, 1996.

Netto, G. A. F. Doze Lições Sobre Freud e Lacan. Ed. Pontes, Campinas, p.69), 2010.

Nowak MA, Plotkin JB, Jansen VA. The evolution of syntactic communication. Nature. 2000; 404: 495-498.

NUNES, B. R. S. S. Leitura em língua inglesa: a resolução colaborativa de exercícios de compreensão textual. 2002. Dissertação (Mestrado em Letras e Lingüística) – Faculdade de Letras, Universidade Federal de Goiás, Goiânia, 2002.

Pedro Quaresma. Frequency Analysis of the Portuguese Language. University of Coimbra, 2008.

Pereira., L. N. Fatores compartilhados no processamento de leitura em L1 e L2. Porto Alegre: Edipucrs, 2012.

Purcell, J.J., Turkeltaub, P.E., Eden, G.F., & Rapp, B. Examining the central and peripheral processes of written word production through meta-analysis. Frontiers in Psychology, 2, Article 239, 2011.

Rabêlo LGN, Moraes, RM. Entropia e geração de séries de aproximação utilizando uma ferramenta JAVA. In: Anais do XXVI Simpósio Brasileiro de Telecomunicações (SBrT), Rio de Janeiro. 2008; 1-6.

Ramos RT. The concepts of representation and information in explanatory theories of human behavior. Front Psychol. 2014; 5: 1034.

Ribeiro., C. Metacognição: Um Apoio ao Processo de Aprendizagem. Universidade Católica portuguesa, 2003.

Sahlgren, M. The distributional hypothesis. *Italian Journal of Linguistics*, *20*, 33–53, 2008.

Sandra Aluísio, Lucia Specia, Thiago Pardo, Erick Maziero and Renata Fortes. Towards Brazilian Portuguese Automatic Text Simplification Systems. In: Proceedings of The Eight ACM Symposium on Document Engineering. São Paulo, Brazil, 240-248, 2008a.

SCARTON, Carolina, e Sandra Maria ALUÍSIO. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adapatando as métricas do CohMetrix para o português. LinguaMática 2, 2010.

Scott, AJ, Knott M. A Cluster analyses method for grouping means in the analysis variance. Biometrics. 2006; 30(3): 507-512.

Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948. 27: 379-423, 623-656.

Sigman M, Peña M, Goldin AP, Ribeiro S. Neuroscience and education: Prime time to build the bridge. Nat Neurosci. 2014; 17(4): 497-502.

Soares AP, Costa AS, Machado J, Comesana M, Oliveira HM. The Minho Word Pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. Behav Res Meth. 2017; 49(3), 1065-1081.

SPERBER, D., WILSON, D. La Pertinence. Communication et Cognition. Paris, Les Éditions de Minuit, 1989.

Thorne C, Szymanik J. Semantic complexity of quantifiers and their distribution in corpora. In Proceeding of the International Conference on Computational Semantics. 2015.

Turney PD, Pantel P. From frequency to meaning: vector space models of semantics. J Artif Intel Res. 2010; 37: 141-188.

Vigliocco G, Vinson DP, Druks J, Barber H, Cappa, SF. Nouns and verbs in the brain: a review of behavioural, electrophysiological, neuropsychological and imaging studies. Neurosci Biobehav Rev. 2011; 35(3): 407-426.

Villavicencio A. Avaliando um rotulador estatístico de categorias morfo-sintáticas para a língua portuguesa. Dissertação de mestrado. 1995; Porto Alegre: CPGCC-UFRGS. 154 p.

Virtue S, Haberman J, Clancy Z, Parrish T, Jung Beeman M. Neural activity of inferences during story comprehension. Brain Res. 2006; 1084(1): 104-114.

Visser M, Lambon Ralph, MA. Differential contributions of bilateral ventral anterior temporal lobe and left anterior superior temporal gyrus to semantic processes. J Cogn Neurosci. 2011; 23: 3121-3131.

Willian Watanabe, Arnaldo Candido Junior, Vinícius Uzêda, Renata Fortes, Tiago Pardo and Sandra Aluísio. Facilita: reading assistance for lowliteracy readers. In: Proceedings of the 27th ACM International Conference on Design of Communication. SIGDOC '09. ACM, New York, NY, 29-36, 2009.

www.101languages.net [Internet]. Portuguese 101; c2017 [cited 2017 Mar 20]. Available from: http://www.101languages.net/portuguese/most-common-portuguese-words/

Yang Y, Wang J, Bailer C, Cherkassky V, Just MA. Commonality of neural representations of sentences across languages: Predicting brain activation during Portuguese sentence comprehension using an English-based model of brain function. NeuroImage. 2016. Referenced in doi: 10.1016/j.neuroimage.2016.10.029.

Yarkoni T, Speer NK, Balota DA, McAvoy MP, Zacks JM. Pictures of a thousand words: Investigating the neural mechanisms of reading with extremely rapid event-related fMRI. NeuroImage. 2008; 42(2): 973-987.

APÊNDICE 1 – EXEMPLOS ALEATÓRIOS DOS TEXTOS ARMAZENADOS NO CALCULETRA PARA A CRIAÇÃO DO BANCO DE DADOS

APÊNDICE 1.1 – EXEMPLO A.

Tiago Mendes de Almeida

A ORIGEM DO CARBONO NO UNIVERSO - INSIGHTS A PARTIR DE OBSERVAÇÕES DE ESTRELAS POBRES EM METAIS NAS NUVENS DE MAGALHÃES

Dissertação apresentada ao Departamento de Astronomia do Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo como parte dos requisitos para a obtenção do título de Mestre em Ciências

Área de Concentração: Astronomia
Orientadora: Prof.ª Dr.ª Silvia Rossi

São Paulo

2009

APÊNDICE 1.2 – EXEMPLO B.

VALDIANA DO BOMFIM ALVES

O trabalho docente em uma turma de alfabetização na rede municipal de ensino de São Bernardo do Campo: entre objetos ensinados e dispositivos didáticos

Dissertação apresentada à Faculdade de Educação da Universidade de São Paulo para obtenção do título de Mestre em Educação

Área de concentração: Linguagem e Educação

Orientador: Prof. Dr. Sandoval Nonato Gomes-Santos

São Paulo

2013

APÊNDICE 1.3 – EXEMPLO C.

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE EDUCAÇÃO

DANIELE KOHMOTO AMARAL

Histórias de (re)provação escolar:
vinte e cinco anos depois

SÃO PAULO
2010

APÊNDICE 1.4 – EXEMPLO D.

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS
DEPARTAMENTO DE LETRAS CLÁSSICAS E VERNÁCULAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LITERATURA BRASILEIRA

Daniela Utescher Alves

A crônica de Cecília Meireles:
uma viagem pela ponte de vidro do arco-íris

São Paulo
2012

APÊNDICE 2 – AMOSTRAS DOS TEXTOS DENOTATIVOS INSERIDOS NO CALCULETRA PARA A QUANTIFICAÇÃO DA COMPREENSÃO ESCRITA

https://vestibular.uol.com.br/redacao/
http://www.fuvest.br/vest2013/bestred/bestred.html
http://edilaine-bancoderedacoesblogspotcom.blogspot.com.br/
http://vestibular.brasilescola.uol.com.br/banco-de-redacoes/tema-caminhos-para-combater-intolerancia-religiosa-no-brasil.htm
http://www.igualdadedegenero.cnpq.br/igualdade.html
https://www.educacao.mg.gov.br/cidadao/banco-de-noticias
http://clipping.radiobras.gov.br/
http://www.acrobatadasletras.com.br/2017/02/inventario-de-redacoes-nota-1000-enem-2001-2016.html
https://www.yumpu.com/pt/document/view/12780608/banco-de-redacoes-colegio-anglo-brasileiro
https://educacao.uol.com.br/bancoderedacoes/temas.jhtm
http://www.mundotexto.com.br/redacao.html
http://www.vestibular1.com.br/redacao/modelos-de-redacao/
http://www.fnde.gov.br/programas/livro-didatico/guias-do-pnld

APÊNDICE 2.1 – EXEMPLO A.

Atenção: Leia atentamente as instruções do caderno de questões antes de preencher essa folha.

Consumismo e felicidade.

O mundo vive uma época de consumismo e individualismo cada vez mais exacerbados. Os homens são constantemente incentivados a produzir mais para que possam ganhar mais e, consequentemente, consumir mais. Apregoa-se a ideia de que quanto maior o sucesso financeiro do indivíduo e maior a quantidade de bens e serviços que este possa adquirir, mais próximo ele estará da felicidade. Mas será que isto é verdade? Ou melhor, será que é toda a verdade?

A florescência de "shopings centers" cada vez mais opulentos e destinados às camadas mais ricas da população faz parecer que a realidade desta sociedade de consumo seja uma experiência muito bem sucedida e que responda plenamente a um dos supostos anseios mais primevos do ser humano: o de ter para ser. A facilidade da obtenção de financiamento e crédito junto a instituições bancárias e a procura cada vez maior da população por estes recursos parecem sinalizar na mesma direção.

Na contramão desta realidade, porém, diversos pesquisadores e ativistas sociais e ambientais sinalizam que seriam necessárias várias "Terras" para suprir esta ânsia consumista desenfreada caso a totalidade da população global tivesse a mesma taxa de produção e consumo dos países mais ricos. Isto mostra que se todos os cidadãos do planeta assumissem esta mesma postura dos "mais bem sucedidos" a Terra se tornaria um lugar insustentável, para todos.

Enganamo-nos se pensamos que esta crítica à posse desmedida seja exclusividade da sociedade contemporânea. Já no século XVIII, o filósofo francês Rousseau aludia ao fato de que nenhum homem fez tanto mal à humanidade quanto aquele que pela primeira vez cercou suas terras e as declarou como propriedade sua. A diferença é que ~~a~~ o planeta nunca teve uma população tão grande como a atual e este desejo de possuir é exponencializado cotidianamente. Vivenciamos uma quantidade crescente de bens de consumo e serviços "imprescindíveis" que, até ontem, nem existiam.

Faz-se necessário, então, que a sociedade global repense seus valores e assuma um modelo de produção e consumo mais humanizado e responsável e que seja capaz de abranger a totalidade da população mundial de forma sustentável, afinal, todos os indivíduos devem ter direito ao acesso aos mesmos bens, à mesma dignidade e à mesma felicidade. Somente assim cada indivíduo poderá desfrutar do melhor que o mundo tem a oferecer, de maneira mais comedida, coletiva e verdadeira.

APÊNDICE 2.2 – EXEMPLO B.

INCLUSÃO E APRENDIZADO

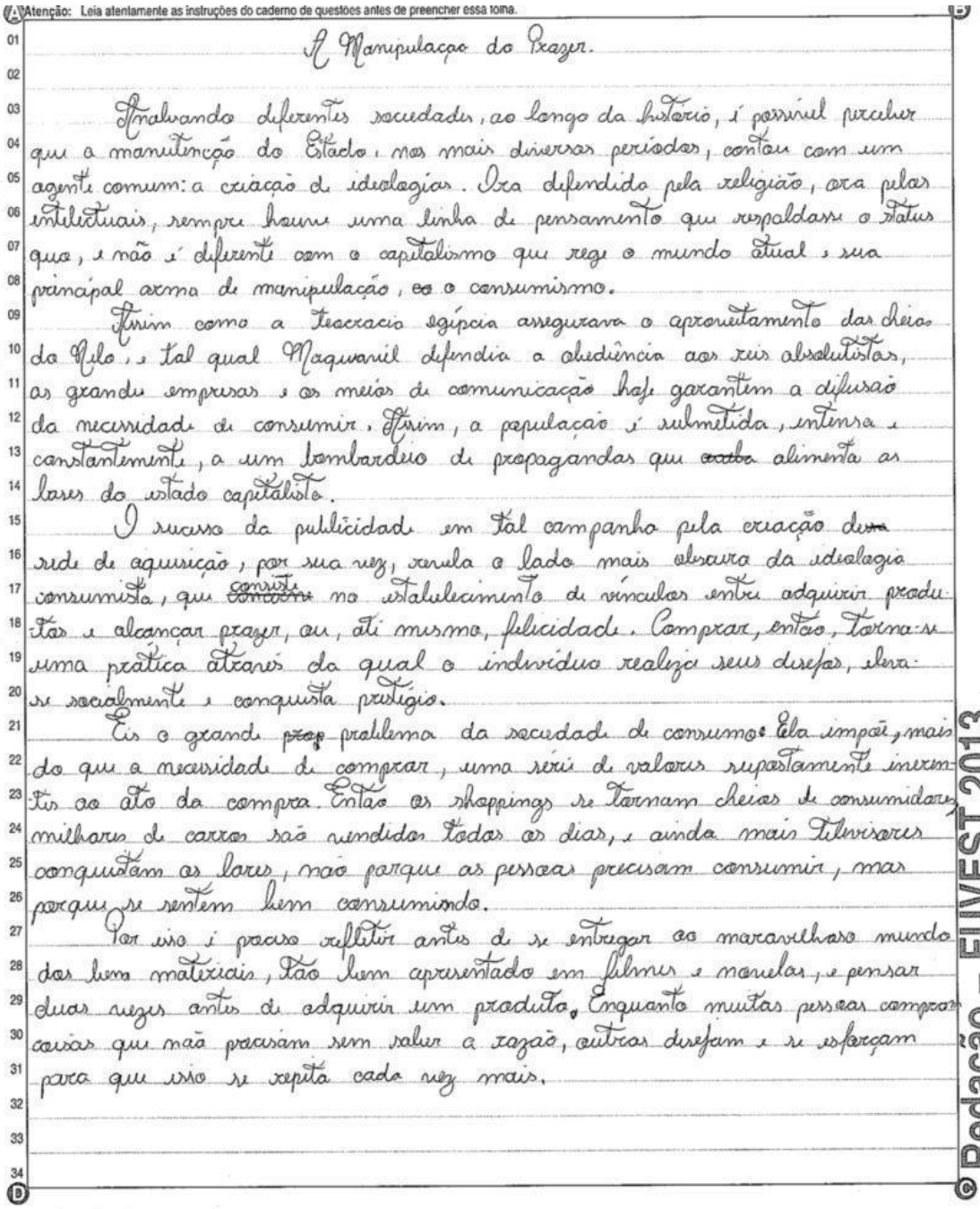
Decisivo para melhorar a inclusão e a permanência na escola dos filhos de seus beneficiários, o programa Bolsa Família continua devendo mecanismos capazes de demonstrar claramente, 10 anos depois de sua implantação, resultados concretos sob o ponto de vista da qualidade do ensino. A particularidade de exigir precondições dos participantes, como matricular os filhos na escola e se comprometerem com a frequência de 85% das aulas, permitiu um salto que até mesmo os críticos dessa alternativa de complementação de renda reconhecem. Falta, porém, uma avaliação técnica e mais aprofundada sobre o que vem ocorrendo sob o ponto de vista do aprendizado em si e da qualidade do ensino dos contemplados pela iniciativa. Afinal, nada menos de 50 milhões de brasileiros, mais do que a população da Argentina e do Uruguai somadas, recebem o Bolsa Família.

Na falta de estudos concretos que possam ir um pouco além de casos bem-sucedidos tomados como exemplo, incluindo brasileiros para os quais o ingresso na universidade seria impensável até há alguns anos, o jornal Valor Econômico analisou 26 cidades de todos os Estados, entre mais de 2 mil consideradas prioritárias pelo Ministério da Educação. Uma das conclusões, até certo ponto previsível, foi a de uma melhora generalizada nos indicadores de aprovação e de distorção idade-série, o que já é suficiente para reafirmar a importância do programa. Sob o ponto de vista da avaliação do Índice de Desenvolvimento da Educação Básica (Ideb) entre 2005 e 2011, porém, o resultado é bem menos animador. O principal instrumento de aferição da qualidade do ensino no país mostra a dificuldade de serem alcançadas as metas ou mesmo uma piora na maioria dos municípios analisados. É o que ocorre inclusive nas duas cidades sulinas incluídas na mostra, Redentora, no Rio Grande do Sul, e Entre Rios, em Santa Catarina.

Certamente, o impacto de programas como o Bolsa Família não se dá apenas sobre o ensino, pois se estende a diferentes áreas, a começar por um inevitável aumento da cidadania. Mas, por sua importância para o país e pelos elevados montantes de recursos orçamentários que mobiliza, os resultados não podem ser avaliados apenas pelas propagandas oficiais.

No caso específico do ensino, o trabalho seria facilitado pelo fato de o governo federal já contar hoje com dados de 75 mil escolas públicas do país, de um total de 200 mil, nas quais a maioria dos alunos se beneficia do programa. O país, porém, não pode se contentar com quantidade: precisa dispor de instrumentos eficazes para aferir também a qualidade da educação dos beneficiários do Bolsa Família.

APÊNDICE 2.3 – EXEMPLO C.



Atenção: Leia atentamente as instruções do caderno de questões antes de preencher essa folha.

A Manipulação do Prazer.

Analisando diferentes sociedades, ao longo da história, é possível perceber que a manutenção do Estado, nas mais diversas períodos, contou com um agente comum: a criação de ideologias. Ora defendido pela religião, ora pelas intelectuais, sempre houve uma linha de pensamento que respaldasse o status quo, e não é diferente com o capitalismo que rege o mundo atual, sua principal arma de manipulação, eo o consumismo.

Assim como a teocracia egípcia assegurava o aproveitamento das cheias do Nilo, e tal qual Maquiavel defendia a obediência aos reis absolutistas, as grandes empresas e os meios de comunicação hoje garantem a difusão da necessidade de consumir. Assim, a população é submetida, intensa e constantemente, a um bombardeio de propagandas que ~~cada~~ alimenta as bases do estado capitalista.

O sucesso da publicidade em tal campanha pela criação de ~~uma~~ rede de aquisição, por sua vez, vela o lado mais obscuro da ideologia consumista, que ~~consiste~~ no estabelecimento de vínculos entre adquirir produtos e alcançar prazer, ou, até mesmo, felicidade. Comprar, então, torna-se uma prática através da qual o indivíduo realiza seus desejos, eleva-se socialmente e conquista prestígio.

Eis o grande ~~prog~~ problema da sociedade de consumo: Ela impõe, mais do que a necessidade de comprar, uma série de valores supostamente inerentes ao ato da compra. Então os shoppings se tornam cheios de consumidores, milhares de carros são vendidos todos os dias, e ainda mais televisores conquistam os lares, não porque as pessoas precisam consumir, mas porque se sentem bem consumindo.

Por isso é preciso refletir antes de se entregar ao maravilhoso mundo das bem matizadas, tão bem apresentado em filmes e novelas, e pensar duas vezes antes de adquirir um produto. Enquanto muitas pessoas compram coisas que não precisam sem saber a razão, outras desejam e se esforçam para que isso se repita cada vez mais.

APÊNDICE 2.4 – EXEMPLO D.

Projeto de escola sobre relação do homem com a fauna é premiado

A diversidade da fauna brasileira foi tema de projeto desenvolvido por educadores e alunos da Escola Estadual São Francisco de Assis. A unidade escolar que funciona em um Centro Socioeducativo no município de Governador Valadares, no Vale do Rio Doce, usou a arte para estudar os animais domésticos e selvagens com os mais de 100 estudantes dos ensinos Fundamental e Médio que cumprem medidas socioeducativas.

O trabalho foi idealizado pela professora de Uso da Biblioteca, Mariana Oliveira da Rocha, mas contou com a interdisciplinaridade das áreas de Ciências, Geografia, História, Língua Portuguesa e Matemática. “Em um primeiro momento exibimos o filme Rio e realizamos uma discussão sobre assuntos, como os animais da fauna brasileira e o tráfico de animais”, lembra Mariana Oliveira.

A etapa seguinte do projeto contou com a leitura do livro Marley e Eu e o filme de mesmo nome que contam a história da relação afetiva de uma família com um cachorro. “Com essas obras, nós procuramos discutir o amor e o carinho com os animais, como uma família acolhe o cachorro, porque o ser humano tem esse carinho e dedicação com os animais”, ressalta a educadora.

Reconhecimento

A iniciativa em trabalhar a conscientização ambiental sobre os cuidados com a fauna doméstica foi inscrita no concurso ‘Guarda Responsável: que bicho é esse?’, realizado pelo Ministério Público de Minas Gerais em parceria com a Secretaria de Estado de Educação e a Associação regional de Proteção Ambiental de Divinópolis. O projeto foi premiado e a escola foi contemplada com um notebook que será utilizado para fins pedagógicos.

Para a idealizadora e coordenadora da ação na escola: “Fiquei muito surpresa, porque não tinha muita dimensão, eu trabalho projeto porque é a melhor maneira de se trabalhar os temas. Procuramos trabalhar de uma forma mais divertida, suave. Eles não podem sair a campo por cumprirem medidas socioeducativas, então buscamos ações que podem ser feitas na sala de aula. Ficamos satisfeitos como resultado, o ego da escola fica muito bom”, finaliza Mariana Oliveira.

ANEXO 1 – ESTATÍSTICAS DE METADADOS DOS BANCOS DE TEXTOS EM PORTUGUÊS

ANEXO 1.1 – BIBLIOTECA DIGITAL BRASILEIRA DE TESES E DISSERTAÇÕES

Fonte: http://bdtd.ibict.br/vufind/

DOCUMENTS	
By type	
Master's dissertations	371,202
Doctoral thesis	172,636
Total	543,838

ANEXO 1.2 – BIBLIOTECA DIGITAL DE TESES E DISSERTAÇÕES DA USP

Fonte: http://www.teses.usp.br/

DIGITAL LIBRARY		2014	2013	2012	2011
DOCUMENTS		50.564	43.752	37.161	31.618
By year		6.813	6.591	5.543	5.195
By type	Master's dissertations	29.998	26.209	22.277	18.974
	Doctoral theses	20.200	17.250	14.660	12.460
	Habilitation theses	366	292	223	183
By access	Public	48.884	42.209	36.193	31.111
	Restricted	320	322	322	322
	Withheld	1.360	1.221	646	185
STORAGE (MB)		406.345	311.145	228.209	160.236
By year		88.039	82.936	67.973	33.153
# of files		13.265	13.712	11.834	6.763
Avg. size		6,64	6,44	5,98	4,91
DOWNLOADS		19.694.828	13.113.110	9.883.923	7.312.018
Visitors		14.716.790	9.344.477	7.433.992	5.536.999
Robots		4.978.038	3.768.633	2.449.931	1.775.019
USING					
Searches		100.592	708.055	782.493	889.994
Terms of searches		295.609	329.162	357.514	397.673
GOOGLE ANALYTICS		**2014**	**2013**	**2012**	**2011**
Pagesviews		6.458.894	7.578.309	8.653.179	13.924.876
Visitors		979.289	1.097.241	1.145.340	3.144.719
	Unique	688.440	749.185	775.321	2.355.910
	New	670.351	727.924	747.214	2.309.926
	Returning	308.938	369.317	398.126	834.786
	Desktop	921.700	1.058.158	1.126.519	3.123.433
	New	631.840	701.975	736.407	2.280.282
	Returning	289.860	356.183	390.112	843.151
	Mobile	57.589	39.083	18.820	21.287
	New	40.075	26.302	13.548	18.252
	Returning	17.514	12.781	5.272	3.035
Avg. Visit Duration		04:58	05:17	05:33	02:52
	New	04:24	04:49	05:05	02:19
	Returning	06:12	06:13	06:26	04:24
Pages/session		6,60	6,91	7,56	4,43
	New	6,26	6,78	7,44	3,86
	Returning	7,31	7,16	7,77	6,00
Location	Brazil	921.252	1.021.577	1.071.002	2.929.174
	Other countries/territories	58.037	75.664	74.338	215.538
	# of countries/territories	160	165	151	178

ANEXO 2 - DISTRIBUIÇÃO DE PROBABILIDADES TÍPICA (LÍNGUA INGLESA)

Fonte: Claudio Gomes Mello. Codificação Livre de Prefixo para Cripto-Compressão. 2006. Tese (Doutorado em Engenharia da Computação) - Pontifícia Universidade Católica do Rio de Janeiro.

Letra	Probabilidade (%)
A	8,04
B	1,54
C	3,06
D	3,99
E	12,51
F	2,30
G	1,96
H	5,49
I	7,26
J	0,16
K	0,67
L	4,14
M	2,53
N	7,09
O	7,60
P	2,00
Q	0,11
R	6,12
S	6,54
T	9,25
U	2,71
V	0,99
W	1,92
X	0,19
Y	1,73
Z	0,09