UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular

Sidney Evaldo Leal

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

ICMC USP
SÃO CARLOS

Sidney Evaldo Leal

Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular

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

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Sandra Maria Aluísio

USP – São Carlos
Junho de 2021

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

Leal, Sidney Evaldo
L435p Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular / Sidney Evaldo Leal; orientadora Sandra Maria Aluísio. -- São Carlos, 2021.
208 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Predição de Complexidade de Sentenças. 2. Inteligibilidade. 3. Simplificação de Textos. 4. Rastreamento Ocular. 5. Transfer Learning. I. Aluísio, Sandra Maria, orient. II. Título.

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

Sidney Evaldo Leal

Sentence-based readability prediction in Brazilian Portuguese, using linguistic, psycholinguistic and eye tracking metrics

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Sandra Maria Aluísio

USP – São Carlos
June 2021

AGRADECIMENTOS

Em primeiro lugar agradeço à melhor orientadora que um estudante poderia almejar, professora Sandra, se algum dia eu vier a orientar alguém, quero tentar ser pelo menos metade do que você foi pra mim.

À minha co-orientadora não oficial, a incrível linguista e pesquisadora Magali Duran, nem consigo listar o quanto aprendi com você.

À pesquisadora Vanessa Magalhães e à Embrapa, pelo apoio, motivação e acompanhamento em todas as fases deste trabalho.

Aos demais co-autores dos artigos desta tese. Se tive algum mérito no resultado final, certamente foi o de servir de motivo para reunir tanta gente boa em torno do objetivo de me ajudar. Em especial às professoras Carol, Érica, Elis, Katerina e Teresa; aos professores Gustavo, Denis e Renê; e aos colegas João, Nathan e Edresson.

À minha esposa, bailarina e professora, que há 16 anos vem me apoiando nessa jornada.

Ao meu finado pai, que silenciosamente me ensinou que ser honesto é mais importante que saber ler e escrever.

À minha mãe, que despertou em mim o gosto pela leitura por meio do exemplo, e sempre me incentivou a continuar estudando.

Ao meu irmão e guia, que me trouxe para a área da computação e agora tenta me levar para a cafeicultura nas montanhas mineiras.

Ao professor Thiago Pardo, por ensinar PLN com um entusiasmo contagiante.

Ao professor Thiago e às professoras Graça Nunes, Carol Scarton, Maria José Finatto e Lilian Hubner pelas preciosas dicas nas bancas de qualificação de mestrado e doutorado.

Às professoras Lilian Hubner e Renata Vieira e ao professor Marcelo Finger pela cuidadosa revisão e avaliação da tese na banca final.

Aos demais colegas do NILC: Roney, Murilo, Marcelo, Henrico, Rogério, Erick, Leandro, Marcio, Ana, Carol, João, Laura, Renata, Marco(s), Rafaéis, ... pelas dicas, ideias e risadas.

Ao Leo Comelli, que me convenceu a tentar a pós na USP.

Aos mentores Alfredo Deak e André Silva, pela confiança no início com as cartas de indicação.

Ao time da Konfido: JP, Ricardo, Barizon, Romancini e Addo, que apagaram incêndios

em vários sistemas em produção, enquanto eu estudava.

Aos companheiros de banda: Marcão, Lau e Barna, pelos momentos de musicoterapia durante todo esse tempo.

Às forças invisíveis que estão por aí nos ajudando, sem se importar com que nome recebam.

Aos mais de 450 participantes do RastrOS; tanto aos que responderam ao teste Cloze quanto aos que leram os parágrafos monitorados pelo Eye-tracker.

À FAPESP, que contribuiu com o financiamento do projeto RastrOS (processo número 2019/09807-0).

E finalmente à USP e ao ICMC, que me aceitaram como aluno especial do mestrado cinco anos atrás, quando a única coisa que eu podia oferecer era a disposição em aprender. Reproduzo abaixo por pura nostalgia a mesma carta que apresentei ao programa da pós. Muita coisa mudou desde então, mas o brilho nos olhos e a sede por conhecimento só aumentaram.

Carta de intenção - 16.02.2016

Vi um computador pela primeira vez aos quinze anos de idade, o que não está muito fora da curva se considerar que a primeira vez que assisti televisão foi aos dez.

Nasci e cresci no interior, no norte de Minas Gerais, às margens do rio Jequitinhonha. Só pisei em uma sala de aula aos oito, e para chegar nela andava quarenta minutos na ida e mais quarenta na volta. Como geralmente acontece, em vez de atrapalhar, essas dificuldades só aumentaram meu empenho na busca por conhecimento, o que acabou me impulsionando a mudar para cidades cada vez maiores até chegar em São Paulo.

Foi quando vi o computador, um 386 que meus tios compraram. Aprendi o MS-DOS em dois dias e comecei a programar em Basic nas madrugadas, após todo mundo ir dormir, com o brilho do monitor praticamente zerado, pois na época era ponto indiscutível na casa que o computador "queimava" se ficasse ligado muito tempo.

Semanas mais tarde encontrei meu primeiro emprego na área e não parei mais. Fiz o colegial técnico em Processamento de Dados e bacharelado em Ciência da Computação.

Terminei a faculdade empolgado e pronto para começar um mestrado, mas aí veio uma grande decepção: os cursos eram oferecidos durante o dia, e minhas condições financeiras na época não permitiam muita margem para negociação.

Mergulhei chateado no mundo corporativo e venho evoluindo satisfatoriamente.

Nesses vinte anos de área, trabalhei em empresas pequenas, médias, grandes e multinacionais, de capital público e privado. Liderei e empreendi algumas vezes, fiz projetos para o governo, bancos e empresas de telecom, atuando como desenvolvedor, analista, arquiteto, administrador de dados e gerente de projetos. Também me tornei profissional certificado nas plataformas Java e Microsoft, e fiz MBA em gestão da tecnologia da informação.

Hoje aos trinta e seis, comecei a rever algumas escolhas. A maturidade chegou e juntamente com ela vieram melhores condições financeiras e um emprego com flexibilidade de horário. Outra vantagem que acredito dispor é de um vasto campo de pesquisa dentro do ambiente corporativo onde estou inserido, que torna relativamente simples colher métricas e testar soluções.

É nesse ponto que me apresento diante do ICMC, com saudades do mundo acadêmico, recuperando o brilho dos olhos da época da faculdade e ainda sedento por conhecimento.

“Aprender a ler é aprender a ser livre.”
(Paulo Freire)

RESUMO

LEAL, S. E. **Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular**. 2021. 208 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

O INAF (Indicador de Alfabetismo Funcional), criado para mensurar o nível de letramento da população brasileira entre 15 e 64 anos, apontou, em seu relatório de 2018, que somente um em cada dez brasileiros adultos é considerado letrado de forma proficiente. No setor da agropecuária, apenas uma em cada cem pessoas consegue ler e compreender textos sem enfrentar dificuldades. Isso significa que a grande maioria dos produtores rurais pode não se beneficiar das tecnologias publicadas por entidades de pesquisa como a Embrapa (Empresa Brasileira de Pesquisa Agropecuária). Uma solução bastante viável para esse problema é simplificar essas publicações para torná-las mais acessíveis para público alvo. Uma das etapas da simplificação é a tarefa conhecida como "predição da complexidade sentencial", responsável por identificar as sentenças mais complexas de um texto, as quais serão alvo das operações de simplificação subsequentes. Para o português brasileiro, antes do presente trabalho, a tarefa de predição de complexidade sentencial ainda não havia sido avaliada e nem havia córpus criados para o aprendizado da tarefa. Outra lacuna observada foi a falta de um córpus com métricas de rastreamento ocular, semelhante aos disponíveis em inglês e utilizados pelos trabalhos internacionais mais recentes sobre predição de complexidade. O objetivo principal desta pesquisa é avaliar métodos de predição de complexidade sentencial para o português brasileiro escrito, a fim de criar um método no estado da arte para a tarefa. Para implementar esse método, projetou-se um ambiente denominado Simpligo, que tem por objetivo auxiliar na simplificação de textos, especialmente os produzidos pela Embrapa para o domínio rural. Para atingir esses objetivos, foram criados dois córpus: um com as sentenças alinhadas do PorSimples (CASELI *et al.*, 2009), e um com métricas de rastreamento ocular e normas de previsibilidade de estudantes do ensino superior. Também disponibilizou-se a versão de 2021 da ferramenta NILC-Metrix, de código-fonte aberto, com 200 métricas linguísticas e psicolinguísticas, as quais são utilizadas nas avaliações dos métodos de predição de complexidade sentencial . Por fim, nesta pesquisa foram avaliadas abordagens de *ranking* e *transfer learning*, sendo que esta última, com a adição das métricas de rastreamento ocular, atingiu o estado da arte para a tarefa de predição da complexidade sentencial na língua portuguesa, com 97,5% de acurácia. Este trabalho contribui com novos córpus, métodos e aplicações, voltados à tarefa de avaliação da complexidade sentencial. Além disso, ao serem disponibilizados publicamente todos os recursos desenvolvidos, torna-se possível sua utilização em outras tarefas e investigações.

Palavras-chave: Predição de Complexidade de Sentenças, Inteligibilidade, Simplificação de Textos, Rastreamento Ocular, *Transfer Learning*.

ABSTRACT

LEAL, S. E. **Sentence-based readability prediction in Brazilian Portuguese, using linguistic, psycholinguistic and eye tracking metrics**. 2021. 208 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

The INAF (Functional Literacy Indicator), created to measure the literacy level of the Brazilian population between 15 and 64 years old, pointed out in its 2018 report that only one in ten adult Brazilians is considered proficiently literate. In the farming sector, only one in a hundred people can read and understand texts without facing difficulties. This means that the vast majority of rural producers may not benefit from the technologies published by research entities such as Embrapa (Brazilian Agricultural Research Corporation). A very viable solution to this problem is to simplify these publications to make them more accessible to the target audience. One of the simplification steps is to assess the sentence complexity, a task known as "sentence complexity prediction", responsible for identifying the most complex sentences in a text, which will be the target of subsequent simplification operations. For Brazilian Portuguese, before the present work, the task of sentence complexity prediction had not been evaluated and there was no corpus available for learning the task. Another gap observed was the lack of a corpus with eye-tracking metrics, similar to those available in English and used by the most recent international studies on complexity prediction. The main goal of this research is to evaluate methods for predicting sentential complexity for written Brazilian Portuguese in order to create a state-of-the-art method for the task. To implement this method, we designed a computational environment called Simpligo to support texts simplification, especially those produced by Embrapa for the rural domain. To achieve these goals, we created two corpora: one with PorSimples (CASELI *et al.*, 2009) aligned sentences, and one with eye-tracking metrics and predictability norms for higher education students. In addition, we released the 2021 version of the open-source NILC-Metrix tool with 200 linguistic and psycholinguistic metrics, which we use in our evaluations of sentence complexity prediction methods. Finally, this research evaluated ranking and transfer learning approaches, and the latter, with the addition of eye-tracking metrics, reached the state-of-the-art for the task of predicting sentential complexity in the Portuguese language, with 97.5% accuracy. This work contributes with new corpora, methods and applications focused on the task of evaluating sentential complexity. Additionally, by making all the resources developed publicly available, we enable them to be used in other tasks and investigations.

Keywords: Sentence-based Readability Prediction, Readability, Text Simplification, Eye-tracking, Transfer Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizagem de Máquina
ARA	*Automatic Readability Assessment*
BERT	*Bidirectional Encoder Representations from Transformers*
IA	Inteligência Artificial
INAF	Indicador de Alfabetismo Funcional
LSA	*Latent Semantic Analysis*
LSTM	*Long Short Term Memory*
MLP	*Multi Layer Perceptron*
MSE	*Mean Squared Error*
NILC	Núcleo Interinstitucional de Linguística Computacional
PB	Português Brasileiro
PCA	*Principal Component Analysis*
PLN	Processamento de Línguas Naturais
RMSE	*Root Mean Squared Error*
RNN	*Recurrent Neural Network*
SVD	Singular Value Decomposition
SVM	*Support Vector Machines*
TF-IDF	*Term Frequency - Inverse Document Frequency*

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 O cenário e os problemas de pesquisa

Segundo o Indicador de Alfabetismo Funcional (INAF), apenas um em cada dez brasileiros adultos é considerado letrado de forma proficiente (IPM, 2018). Esse indicador explicita um dos grandes desafios brasileiros: o acesso à evolução econômica e tecnológica pela população. Percebe-se que a situação é ainda mais assustadora quando isolamos certos setores da economia, como o da agropecuária, em que apenas 1% dos entrevistados foram considerados proficientes (IPM, 2016).

Na prática, isso significa que a quase totalidade dos produtores rurais pode não ser capaz de usufruir das novas tecnologias desenvolvidas e publicadas por entidades de pesquisa. A falta de acesso ao conhecimento prejudica bastante este setor, um dos mais importantes do Brasil, que é responsável por 23% do Produto Interno Bruto (PIB) (MIN.AGRICULTURA, 2017) e 40% da renda da população economicamente ativa (SEAD, 2018).

Uma alternativa bastante viável na atualidade é simplificar essas publicações para permitir maior acesso e utilização pelo público alvo. Um bom exemplo disso é a abordagem da Embrapa Gado de Leite, no projeto APP@Rural (MAGALHÃES *et al.*, 2017). A Embrapa decidiu transformar este cenário simplificando os informativos e textos técnicos publicados e tornando-os mais acessíveis aos produtores, estudantes e extensionistas[1]. Outro exemplo é o movimento Linguagem Simples (*Plain Language*), que vem ganhando força em diversos países, pricipalmente na comunicação entre governos e cidadãos (FISCHER, 2020).

Uma das tarefas propostas para apoiar essa simplificação foi a evolução dos métodos de classificação textual do projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) (ALUÍSIO; GASPERIN, 2010). O PorSimples teve como objetivo

[1] Profissionais de iniciativa pública ou privada, que atuam como braço de divulgação das tecnologias e boas práticas desenvolvidas pelas entidades de pesquisas no meio rural.

promover o acesso a textos em Português Brasileiro (PB) por analfabetos funcionais e crianças ou adultos em fase de alfabetização e criou os modelos automáticos com base em textos jornalísticos com dois níveis de simplificação. No PorSimples, houve a proposta de indicar automaticamente as sentenças alvos de simplificação em um texto, entretanto essa tarefa não foi implementada durante aquele projeto.

Para suprir essa lacuna, esta pesquisa teve como principal objetivo avaliar métodos de predição da complexidade sentencial para o PB, pois, até onde foi possível verificar, embora a tarefa tenha sido explorada para outras línguas como o inglês, ela ainda não havia sido explorada, de forma automática, para o português.

Do ponto de vista de domínio de aplicação, o objetivo foi desenvolver um conjunto de soluções apoiadas em métodos do Processamento de Línguas Naturais (PLN) e Aprendizagem de Máquina (AM), que permita a classificação de textos e sentenças nos quatro níveis indicados pelo INAF 2018 mapeados para o nível de letramento do público alvo desse projeto, conforme Tabela 1. Esse conjunto de soluções servirá de apoio para os diversos canais de comunicação desenvolvidos no projeto APP@Rural da Embrapa, provendo o texto adequado para os diferentes públicos e aumentando a compreensão e uso das tecnologias para produtores de leite.

Tabela 1 – Níveis de letramento do público alvo

Proficiente	Pesquisadores
Intermediário	Extensionistas
Elementar	Produtores rurais mais avançados
Rudimentar	Produtores rurais do nível mais básico

Fonte: Elaborada pelo autor.

Além dos benefícios para o domínio rural, os métodos e ferramentas desenvolvidos neste trabalho também estão disponíveis publicamente para livre utilização em qualquer outra área ou domínio que possa se beneficiar deles.

1.2 A tarefa: predição da complexidade sentencial

O principal foco do presente trabalho é a tarefa de **predição da complexidade sentencial**, também referenciada na literatura como avaliação automática da inteligibilidade de sentenças, ou *sentence-based* ARA (*Automatic Readability Assessment*) em inglês, que consiste em prever o nível de complexidade de uma determinada sentença por meio da análise de suas características.

A predição da complexidade não se preocupa com a simplificação em si, apenas em medir o quão fácil ou difícil é um texto ou sentença para um determinado público alvo, geralmente colocando a resposta em uma escala numérica (cf. o Índice Flesch na Seção 2.2.1).

O que é complexidade sentencial?

O desenvolvimento da habilidade da leitura exige esforço e tempo de dedicação, assim, teremos leitores distribuídos num amplo espectro com vários níveis de proficiência, seja por estarem em processo de aprendizado (incluindo os falantes não nativos de uma determinada língua), seja por terem estacionado em determinado nível de letramento, ou ainda por algum distúrbio cognitivo como afasia ou dislexia.

> "Inteligibilidade é a facilidade de leitura de um texto criada pela escolha de conteúdo, estilo, estruturação e organização que atende ao conhecimento prévio, habilidade de leitura, interesse e motivação da audiência"(DUBAY, 2007), pg.6

A inteligibilidade[2] de um texto, do inglês *text readability* é uma tarefa bem antiga e bastante estudada por várias áreas de pesquisa. As primeiras fórmulas de inteligibilidade foram criadas há quase um século, na década de 1920 nos Estados Unidos, com o objetivo de ajudar a selecionar material de leitura para aulas por professores e administradores do sistema educacional (DAVISON; GREEN, 1988 apud BOHN, 1990).

Estas fórmulas consideravam que a complexidade poderia ser inferida por métricas de palavras e sentenças, baseadas na frequência e tamanho (quantidade de letras ou sílabas) das palavras e na média da quantidade de palavras por sentença.

Desde então, a Inteligibilidade Textual se tornou uma grande área de pesquisa multidisciplinar, com uma vasta bibliografia, e ganhou novas abordagens neste século, com o uso de métodos de PLN e AM.

Complexo para quem?

A complexidade só existe a partir do ponto de vista específico de quem está lendo, não é possível estudá-la sem o sujeito envolvido no processo da leitura. Mesmo em níveis de letramento próximos, pessoas diferentes podem achar o mesmo texto complexo e simples. Isso varia de acordo com o conhecimento de mundo adquirido e armazenado em cada um, experiência, habilidade de leitura e grau de interesse no texto (DUBAY, 2007).

O INAF é um ótimo retrato geral dos potenciais leitores adultos do Brasil (IPM, 2018). O levantamento é feito em média a cada dois anos, desde 2001 e classifica a população nos seguintes níveis de letramento:

[2] Por vezes também traduzida como leiturabilidade. O autor optou por inteligibilidade pela relação óbvia com "entender"e dominar a língua no cotidiano (isto é, ser letrado) em contraste com a habilidade de decodificar o sistema de escrita, ou seja, ser alfabetizado. O mais importante aqui é evitar o termo legibilidade, relacionado com as características que tornam um texto fácil de ser lido, como tamanho de fonte, cor, estruturação em itens, paragrafação, etc.

- **Analfabeto:** Não consegue ler;
- **Rudimentar:** Localiza informações explícitas e literais;
- **Elementar:** Realiza pequenas inferências em textos de tamanho médio;
- **Intermediário:** Consegue interpretar textos e confrontar a moral da história com sua própria opinião ou senso comum; e
- **Proficiente:** Interpreta e elabora textos de maior complexidade sem dificuldades.

É importante frisar que essa classificação em cinco níveis é arbitrária e agrupa internamente diversos níveis de letramento. O próprio INAF até 2011 utilizava apenas quatro níveis (Analfabeto, Rudimentar, Básico e Pleno). Mais um nível foi adicionado ao identificar que, após as ações do governo de combate ao analfabetismo a maioria das pessoas subiu para o nível básico, porém os níveis superiores permaneceram estáveis. De 2001 a 2018 o nível analfabeto caiu de 12% para 8%, enquanto o nível proficiente se manteve em 12%.

Outro ponto importante é que apesar de ter alfabetismo explícito no nome, o INAF avalia o nível de letramento (ou literacia) da população. Alfabetização está relacionada ao processo mecânico de reconhecer os grafemas, ligando-os aos fonemas, enquanto letramento é o uso social desse processo.

> "Letramento é o resultado da ação de ensinar ou de aprender a ler e escrever: o estado ou a condição que adquire um grupo social ou um indivíduo como consequência de ter-se apropriado da escrita."(SOARES, 1996)

Por fim, um conceito ligado à habilidade de leitura versus motivação é o estado de fluxo (CSIKSZENTMIHALYI, 2008). Aplicando ao contexto da leitura, se o texto for demasiado simples e a habilidade do leitor for alta, a experiência vai se tornar enfadonha. Por outro lado se a habilidade do leitor for pequena demais para o nível de complexidade ou desafio apresentado, o esforço exigido vai ser bastante desmotivador. O estado de fluxo seria o casamento do nível de dificuldade adequado para o nível de proficiência do leitor.

A maior motivação deste trabalho é apoiar a criação e adaptação de conteúdos para apresentar o nível ideal de complexidade para o nível de habilidade do leitor.

Por que não ensinar a ler em vez de simplificar?

Esta é uma crítica recorrente e muito importante, logo é bom abordá-la bem no início. O autor concorda plenamente que é sempre melhor ensinar a ler do que simplificar. Dito isso, são citadas a seguir duas grandes exceções para se utilizar a simplificação:

1. **Tempo x Acesso:** Ensinar a ler exige tempo, enquanto simplificar pode permitir o acesso à informação no momento presente. Isso é uma verdade para a população adulta que possui dificuldades na leitura e, por diversos motivos, menos tempo para investir na própria educação. Além disso, para obter um resultado abrangente o suficiente, o investimento necessário na educação precisa partir do governo. Iniciativas isoladas conseguem bons resultados, mas quantitativamente o acesso à informação é maior simplificando os conteúdos publicados.

2. **Nível ideal de complexidade:** Conforme mencionado no tópico anterior, para um estudante em processo de aprendizagem, ser exposto a um texto demasiadamente difícil pode trazer mais prejuízos do que benefícios. A evolução do processo de ensino-aprendizagem pode ser muito mais eficiente se os textos fornecidos aos estudantes apresentarem um nível de desafio adequado.

Por que o foco nas sentenças?

Tradicionalmente, a tarefa tem sido aplicada no nível textual, estimando uma nota (ou nível de *ranking* de proficiência) para um documento inteiro. Porém, em um documento classificado como simples podem ocorrer sentenças complexas, assim como existem sentenças simples em um documento complexo.

Uma sentença é uma unidade importante que traz contida em si, na maioria das vezes, informação suficiente para inferência e análise da complexidade. Um exemplo disso pode ser visto na Tabela 2 em uma sentença simplificada em dois níveis, extraída do córpus PorSimples[3].

Tabela 2 – Exemplo de sentença simplificada em dois níveis

Original	Uma parcela critica o uniforme, porque acredita que ele ameaçaria a individualidade de cada um.
Simplificação Natural	Alguns criticam o uniforme, porque acreditam que ele ameaça a individualidade de cada um.
Simplificação Forte	Alguns acreditam que o uniforme ameaça a individualidade de cada um.

Fonte: Elaborada pelo autor.

A avaliação do nível de inteligibilidade de sentenças é uma tarefa de pesquisa recente. Os primeiros artigos sobre o tema começaram a surgir na última década (2010). A ideia principal é analisar e avaliar individualmente as sentenças de um texto, permitindo uma informação mais acurada dos pontos complexos para auxiliar na tarefa simplificação do texto ou auxiliar estudantes a compreenderem fenômenos linguísticos mais complexos para uma dada série.

[3] O córpus PorSimples é detalhado na Seção 2.3.2

Dell'Orletta, Montemagni e Venturi (2014a) reforçam isso ao afirmarem que a maioria das abordagens de classificação de inteligibilidade levam em consideração os textos inteiros, o que não traz grandes vantagens para a utilização em métodos automáticos de simplificação, aplicados posteriormente. O fato de um texto ser classificado como de difícil leitura não significa que todas as suas sentenças sejam complexas. Esse mal entendimento pode prejudicar o treinamento dos métodos, principalmente quando são utilizadas as sentenças de um córpus com anotação no nível dos textos para avaliar a tarefa de predição da complexidade sentencial. Isso foi comprovado por Vajjala e Meurers (2014a) durante a investigação dos motivos da baixa acurácia que obtiveram utilizando o córpus Wikipedia-SimpleWikipedia, sem alinhamento na época da pesquisa.

Os primeiros trabalhos a considerarem a tarefa de complexidade especificamente para o nível sentencial foram Dell'Orletta, Montemagni e Venturi (2011), comparando a sua dificuldade em relação ao nível textual. Sjöholm (2012) abordou o tema em seu mestrado no ano seguinte. Porém, a definição da forma de avaliação da tarefa só foi consolidada por Vajjala e Meurers (2016) e permitiu que os trabalhos posteriores aperfeiçoassem os resultados comparativamente. Ambati, Reddy e Steedman (2016) conseguiram melhorar significativamente os resultados utilizando um parser do tipo *Combinatory Categorial Grammar* (CCG), e Gonzalez-Garduño e Søgaard (2018) chegaram no estado da arte atual para o inglês, utilizando métricas de rastreamento ocular aliadas com as linguísticas e psicolinguísticas.

Mais recentemente Stajner, Ponzetto e Stuckenschmidt (2017) e Scarton, Paetzold e Specia (2018b) contribuíram para a tarefa avaliando a complexidade com o apoio do córpus Newsela e Bosco, Pilato e Schicchia (2018a) obtiveram bons resultados para o italiano utilizando Redes Neurais Recorrentes do tipo *Long Short Term Memory* (LSTM)[4]. Finalmente, Brunato *et al.* (2018) contribuíram com um trabalho sobre a percepção da complexidade e concordância entre anotadores, enquanto Timm (2018) investigou simplificações sentenciais automáticas utilizando rastreamento ocular.

Esses trabalhos compõem a base para a definição do tema principal, das questões, e dos objetivos desta pesquisa.

1.3 Objetivos, questões de pesquisa e lacunas

Delimitação do objeto

O objeto deste trabalho consiste na pesquisa e desenvolvimento de um modelo computacional que permita a classificação da complexidade das sentenças de um texto, de acordo com o público alvo e níveis de interesse.

Este modelo poderá ser aplicado em um ambiente de apoio à simplificação, tendo como

[4] Mais detalhes na Seção 2.4

base o Simplifica (CANDIDO-JUNIOR; OLIVEIRA; ALUÍSIO, 2009; SCARTON *et al.*, 2010) e também o NILC-Metrix (Capítulo 3), que reúne mais de uma década de métricas desenvolvidas no Núcleo Interinstitucional de Linguística Computacional (NILC), iniciadas com o Coh-Metrix-Port (SCARTON; ALUÍSIO, 2010).

Esse objeto cumpre uma das atividades previstas no projeto da Embrapa: APP@Rural - Desenvolvimento de conteúdos e soluções computacionais para dispositivos móveis, dentro da linha: Transferência de tecnologias para a sustentabilidade da cadeia produtiva do leite no Brasil. A criação dos classificadores/regressores poderá viabilizar a simplificação sentencial específica para o público alvo em mídias eletrônicas e impressas, além de também permitir a predição de conteúdo a ser recomendado com base no histórico do usuário e em suas preferências de navegação.

Justificativa

O setor agropecuário possui evidente importância para o Brasil, pois ele é responsável por 23% do PIB (MIN.AGRICULTURA, 2017). A agricultura familiar é base da economia de 90% dos municípios com até 20 mil habitantes e responsável pela renda de 40% da população economicamente ativa (SEAD, 2018).

Em contraste, o nível de letramento dos trabalhadores rurais está concentrado nos níveis analfabeto (39%), rudimentar (18%) e elementar (5%) segundo o IPM (2016). Apenas 1% dos entrevistados no levantamento do indicador foi considerado proficiente, e 2% no nível intermediário.

As publicações existentes tratando das tecnologias e contendo informações para melhoria dos processos rurais estão disponíveis em grande volume no PB, entretanto continuam longe do alcance dos pequenos produtores, em especial dos com problemas de leitura devido à baixa ou nenhuma escolaridade. Esse problema é relativamente antigo, iniciativas como a Extensão Rural, implantada no Brasil desde a década de 1940, tropeçaram em barreiras como a distância entre o conhecimento científico e o conhecimento tradicional do agricultor (FETTER, 2017).

Como a alfabetização de todos é uma meta relativamente ousada na conjuntura política atual, e afirmadamente de longo prazo, é necessário criar alternativas para levar esse conhecimento disponível a quem mais precisa dele. Uma solução viável é a utilização de métodos de Processamento de Língua Natural (PLN) e Aprendizagem de Máquina (AM) para a avaliação e simplificação desses textos, para os níveis adequados que permitam leitura e compreensão das informações.

Lacunas

Até onde foi possível investigar, não foi encontrado nenhum trabalho que avalie a tarefa de predição de complexidade sentencial para o PB. Portanto, também não existia córpus com sentenças alinhadas especificamente para treinamento e avaliação da tarefa.

Outra lacuna importante foi a inexistência de um córpus com métricas de rastreamento ocular para o PB, semelhante ao utilizado no trabalho que atingiu o estado da arte para o inglês (cf. Seção 2.3.3).

Objetivos

Desenvolver um método computacional de classificação da complexidade sentencial para avaliação de inteligibilidade e simplificação, utilizando técnicas de PLN e AM. O foco inicial é no domínio rural, com o objetivo de apoiar a criação de textos informativos, técnicos e procedimentais para leitores com níveis de letramento variados.

Objetivos específicos

- Pesquisar e desenvolver modelos de classificação de complexidade sentencial;
- Disponibilizar publicamente o NILC-Metrix, com a reunião das métricas linguísticas e psicolinguísticas disponibilizadas pelo NILC na última década, além da criação de novas medidas para auxiliar na tarefa principal deste trabalho;
- Criar o córpus RastrOS[5], o primeiro grande córpus com métricas de rastreamento ocular para o PB (vide Subseção 2.2.4 e Subseção 2.3.3), para permitir a exploração dessas medidas na tarefa de avaliação da complexidade sentencial;
- Compilar o córpus PorSimplesSent, com as sentenças alinhadas do PorSimples (CASELI *et al.*, 2009), para o treinamento e avaliação dos modelos desenvolvidos.

Questões

Como a abordagem da tarefa de avaliação da complexidade no nível sentencial é relativamente nova, principalmente para o PB, existem muitas questões a serem exploradas. Este trabalho focou nas questões relacionadas às métricas (*features*) e córpus adequados à tarefa, além de algumas questões mais teóricas sobre complexidade. Os conjuntos de *features*, métodos e córpus disponíveis serão aprofundados no Capítulo 2.

1. Qual o impacto de grupos de *features* linguísticas (morfossintáticas, lexicais, sintáticas, semânticas), clássicas, psicolinguísticas e do rastreamento ocular para a tarefa de predição de complexidade sentencial?

[5] <http://www.nilc.icmc.usp.br/nilc/index.php/rastros>

2. Qual o impacto de *features* individuais para a tarefa de predição de complexidade sentencial?

3. Qual é a influência do tamanho do córpus de treinamento no desempenho dos métodos para a tarefa de predição da complexidade sentencial?

4. Qual método de seleção de *features* fornece um melhor desempenho para a tarefa de predição de complexidade sentencial?

5. Quais são os erros sistemáticos que persistem usando o melhor método de predição de complexidade sentencial?

6. Qual método de AM atinge o melhor resultado para a tarefa de predição de complexidade sentencial?

Hipótese

As métricas de rastreamento ocular, como representantes da complexidade mensurável durante a leitura das sentenças por humanos, contribuirão com um aumento de cerca de 8% na acurácia do melhor modelo de predição da complexidade sentencial para o PB, assim como contribuíram para o inglês.

1.4 Organização da tese

Esta tese utiliza o modelo de Coleção de Artigos, trazendo os sete artigos principais desenvolvidos durante o período da pesquisa, quatro já publicados e três em processo de revisão por pares em revista e conferência. A apresentação dos artigos não segue a ordem cronológica; eles foram agrupados por tema principal de modo a fornecer uma sequência dos assuntos, facilitando a leitura.

Esse formato de tese é bastante prático, mas apresenta alguns pontos negativos, para os quais o autor desde já se desculpa e pede compreensão. Um deles é a quebra na formatação do texto entre os capítulos, com tamanhos de letra nem sempre ideais para a leitura. Outro ponto é a alternância entre a língua portuguesa e a inglesa, dependendo da localização da conferência ou revista para onde os artigos foram submetidos. O terceiro ponto é a redundância de algumas informações, sobretudo nas introduções, trabalhos relacionados e bibliografia dos artigos. Porém os pontos positivos justificam a decisão pelo formato, sendo os principais: seções com histórias auto-contidas, com início, meio e fim, narrando as etapas de um trabalho maior; e também a apresentação de trabalhos já revisados por pares e com redação mais madura.

Antes de entrar nas seções com os artigos, no Capítulo 2 são apresentados os fundamentos para a tarefa de predição da complexidade, com as principais fontes de *features*, córpus relevantes

e formas de avaliação. Esse capítulo traz um pouco de aprofundamento sobre os temas abordados algumas vezes de maneira superficial nos artigos, por limitação de espaço.

O Capítulo 3 traz o artigo que descreve a primeira grande contribuição deste trabalho: o NILC-Metrix. Desde as primeiras provas de conceito, sempre foram utilizadas as métricas linguísticas e psicolinguísticas já disponibilizadas pelo NILC, porém não existia uma documentação centralizada como referência. Esse artigo resgata a história investida no desenvolvimento dessas métricas e apresenta as novas métricas desenvolvidas, totalizando um pacote com 200 medidas revisadas e centralizadas, com código fonte aberto para que a comunidade utilize e continue seu aprimoramento.

O Capítulo 4 traz três artigos que descrevem o processo de criação da segunda grande contribuição: o córpus RastrOS. O primeiro artigo descreve o método desenvolvido para a seleção dos parágrafos do córpus. O segundo explora os métodos mais recentes para avaliação da similaridade semântica necessária para a implementação das normas de previsibilidade semântica do RastrOS. O terceiro descreve o córpus final, com as contribuições em PLN e computacionais para o projeto, além da descrição dos procedimentos de coleta com os participantes do rastreamento ocular e do teste Cloze[6] para a criação das normas de previsibilidade.

No Capítulo 5 são apresentados três artigos voltados para a tarefa principal abordada na pesquisa: a avaliação da complexidade de sentenças. O primeiro artigo descreve a compilação do PorSimplesSent, primeiro córpus de sentenças alinhadas para a tarefa em PB e os primeiros métodos apresentados como *baseline* com o melhor modelo atingindo a acurácia de 74,2% na tarefa de avaliação da complexidade sentencial. O segundo artigo evolui os métodos e melhora a acurácia no PorSimplesSent para 87,8%, trazendo a avaliação para o domínio rural. O terceiro e mais recente atinge o estado da arte da tarefa para o PB com 97,5% de acurácia, agregando as métricas de rastreamento ocular e os métodos de *Transfer Learning*[7].

Finalmente o Capítulo 6 retoma as questões de pesquisa, resume as contribuições e propõe alguns trabalhos futuros, nascidos das limitações do estudo.

[6] Ver Seção 2.3.3.
[7] Ver Seção 2.4

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

Este capítulo traz os tópicos principais relacionados com a tarefa de predição da complexidade sentencial, para facilitar a leitura dos métodos propostos para a tarefa. Cobre, em cinco seções, as tarefas relacionadas, as métricas utilizadas nos trabalhos da literatura, os córpus disponíveis para avaliação da tarefa, abordagens de AM e métricas de avaliação utilizadas.

2.1 Tarefas relacionadas com a predição da complexidade sentencial

Segue uma breve introdução das principais tarefas da área de PLN que estão diretamente relacionadas com a tarefa principal deste trabalho.

Adaptação textual

A Adaptação Textual é uma área de pesquisa de grande importância dentro da área de PLN, geralmente conectada com práticas educacionais, mas também com aplicações bem diversas como, por exemplo, auxiliar na recuperação de informações biomédicas (JONNALAGADDA; GONZALEZ, 2010). Ela permite alterar o conteúdo de um texto sem mudar seu significado, na maior parte das vezes. Possui duas grandes abordagens: Simplificação e Elaboração Textual (MAYER, 1980).

Simplificação textual

A Simplificação Textual consiste no processo de reduzir a complexidade de um texto, enquanto se preserva o conteúdo informativo e significado, tornando o texto mais fácil de ser compreendido por leitores humanos ou ser processado por programas (SIDDHARTHAN, 2006).

Os primeiros avanços na área de simplificação textual automática surgiram com a ideia de dividir sentenças longas em sentenças menores para melhorar os resultados dos analisadores sintáticos (CHANDRASEKAR; DORAN; SRINIVAS, 1996 apud VAJJALA; MEURERS, 2014a).

Tipos de simplificação textual

Arfé, Mason e Fajardo (2018) definem o objetivo da simplificação textual como a adaptação da complexidade do texto (ou *readability*, em inglês) para as habilidades de um determinado grupo de leitores e, desta forma, *readability measures* (ou medidas de complexidade textual) foram desenvolvidas para alinhar/escolher textos para leitores, pois essas medidas podem predizer o quão difícil um texto será para seus leitores.

Desta relação entre *readability* e simplificação textual, surgem as abordagens profunda e superficial para *readability*, culminando nas abordagens cognitiva, temática (ou topical) e linguística para simplificação textual, explicadas abaixo.

Segundo as autoras acima, a abordagem superficial para complexidade textual se baseia no tamanho das palavras e sua frequência e no tamanho das orações para predizer a complexidade literal dos textos, ou seja, a compreensão do significado estrito de uma única proposição. Enquanto que a abordagem profunda, baseada em *features* como a presença e densidade de marcadores discursivos e correferência no texto, consegue predizer coerência e compreensão no nível inferencial, isto é, a integração entre segmentos de um texto e entre o texto e o conhecimento prévio do leitor.

Abaixo são definidos os três tipos de complexidade — a cognitiva, a linguística (envolvendo os níveis lexical e sintático) e a temática — que levam a três abordagens para simplificação textual, de mesmo nome.

A complexidade cognitiva está relacionada com a capacidade limitada de um leitor de identificar e compreender a estrutura global e local de um texto. A estrutura global é responsável por organizar as informações (ou tópicos) de um texto. As estruturas de textos informativos (jornalísticos, por exemplo) são mais variadas do que textos narrativos, podendo ser uma da seguinte lista: descrição, sequência, comparação e contraste, problema-solução e causa-efeito e, inclusive, aparecer de forma não-exclusiva, dificultando o seu reconhecimento. Essa dificuldade pode impedir um leitor de responder o que é dito no texto e de fazer um resumo dele, por exemplo.

A outra dificuldade se dá no processamento local de um texto, realizado pelo leitor, para conectar sentenças e identificar as suas relações (de contraste, exemplificação, causa, resultado, finalidade, dentre outras). As soluções para essas duas dificuldades são dois conjuntos de simplificações, chamadas de cognitiva no nível global e no nível local. No nível local são usadas para:

1. Aumentar a coesão via conectivos para explicitamente mostrar a relação entre sentenças;

2. Utilizar correferência para conectar as ideias.

Já no nível global temos simplificações para:

1. Facilitar a retenção de novo conhecimento aprendido do texto via organização do conteúdo textual, ajudando o leitor a identificar a estrutura do discurso pelo uso de marcadores discursivos (linguísticos e tipográficos);

2. Organizar fatos e ideias presentes no texto pelo uso de subtítulos/seções que resumem o conteúdo dos parágrafos.

A complexidade temática está associada à falta de conhecimento de mundo necessário para entender alguns temas.

Quanto às simplificações linguísticas, temos a lexical e a sintática. A complexidade lexical está relacionada ao desconhecimento do significado de palavras e expressões. A complexidade sintática está relacionada à capacidade ou não de processar alguns tipos de estrutura de sentenças. Na área de PLN, as simplificações linguísticas foram mais exploradas e muitos métodos criados, para várias línguas. Elas são detalhadas nas próximas seções.

Simplificação lexical

A Simplificação Lexical é uma forma de simplificação por meio da substituição de palavras raras ou complexas por hipônimos, hiperônimos ou sinônimos, equivalentes e mais simples, deixando a leitura com compreensão mais fácil para pessoas com baixo letramento, falantes não nativos de uma dada língua, disléxicos e afásicos, dentre outros (BOITO, 2014). Um exemplo de sentença simplificada lexicalmente pode ser visto na Tabela 3.

A simplificação lexical geralmente é realizada com o apoio de dicionários compilados e recursos como WordNet (MILLER, 1995; FELLBAUM, 1998), de grandes córpus como a Simple Wikipedia (em uma abordagem de *ensembles*), e também de outras abordagens mais recentes baseadas em *word embeddings* (PAETZOLD; SPECIA, 2016) e redes neurais para *ranking* (PAETZOLD; SPECIA, 2017).

Tabela 3 – Exemplo de sentença simplificada lexicalmente

Original	Se acentuada e prolongada, a **hipertermia** pode causar a morte do animal.
Simplificada	Se acentuada e prolongada, a **febre** pode causar a morte do animal.

Fonte: Elaborada pelo autor.

Simplificação sintática

A análise sintática é o estudo da disposição das palavras em uma oração e é dividida em funções sintáticas (sujeito e predicado) e constituintes (sintagmas nominais, verbais, preposicionais, adjetivais e adverbiais) (CANDIDO-JUNIOR, 2013).

A Simplificação Sintática consiste em dividir orações longas (como exemplificado na Tabela 4) ou alterar a estrutura sintática das orações, eliminando fenômenos sintáticos considerados complexos para a inteligibilidade e compreensão de uma classe de leitores. Ainda segundo Candido-Junior (2013), alguns exemplos comuns de fenômenos sintáticos são: reordenação de componentes de uma oração para facilitar a compreensão da informação principal veiculada, mudança de voz passiva para ativa, resolução anafórica de pronomes relativos, reordenação de orações e divisão de orações.

Tabela 4 – Exemplo de sentença simplificada sintaticamente

Original	O uso de forragem conservada, cujas formas mais comuns são: ensilagem e fenação, é uma solução para alimentar o rebanho.
Simplificada	O uso de forragem conservada é uma solução para alimentar o rebanho. As formas mais comuns para conservar forragens são: ensilagem e fenação.

Fonte: Elaborada pelo autor.

Embora haja um compromisso entre simplificação sintática e aumento do texto – é natural que quebrar uma oração longa em várias torne o texto mais longo, pois sujeitos devem ser adicionados –, para vários públicos essa é uma adaptação necessária para permitir o entendimento do texto.

A principal ferramenta de simplificação para o PB foi desenvolvida durante o projeto PorSimples (ALUÍSIO; GASPERIN, 2010), e é chamada Simplifica (CANDIDO-JUNIOR; OLIVEIRA; ALUÍSIO, 2009; SCARTON *et al.*, 2010). Ela apoia autores na redação de textos mais simples, auxiliando tanto na simplificação lexical, que foi baseada em listas de palavras simples, quanto na sintática, realizada via regras baseadas no parser Palavras (BICK, 2000).

Inicialmente a tarefa era solucionada por meio de regras fixas programadas, porém a abordagem mais recente utiliza modelos de redes neurais recorrentes inspiradas na tarefa de Tradução Automática (*Machine Translation*), nos quais o texto original é "traduzido" em sua versão simplificada dentro da própria língua, utilizando o conhecimento adquirido com treinamento em grandes córpus (SCARTON; SPECIA, 2018).

Sumarização automática

A Sumarização Automática pode ser definida como a diminuição da extensão dos textos mantendo os conteúdos principais. Ela tem um papel muito importante na simplificação de textos,

principalmente para os níveis mais baixos de letramento, nos quais o tamanho do texto já é um fator desestimulante para a leitura. Diversos métodos de sumarização, na abordagem extrativa (na qual o sumário é composto de orações retiradas do texto original, sem alterações), foram avaliados no projeto PorSimples (MARGARIDO *et al.*, 2008) e foi escolhido o método Extração de Palavas-Chave por freqüências de Radicais (EPC-R) para ser usado na ferramenta Facilita, desenvolvida no mesmo projeto (WATANABE *et al.*, 2009a; WATANABE *et al.*, 2009b).

Elaboração textual

A Elaboração Textual visa melhorar a compreensão de um texto e/ou ampliar/explorar o vocabulário do leitor, adicionando informações como: sinônimos/antônimos ao lado de palavras ou expressões complexas, definição de conceitos ou ainda tornar explícitas as conexões entre as ideias (MAYER, 1980).

A elaboração lexical, em contraste com a simplificação lexical, não substitui as palavras e sim adiciona uma ou uma lista de palavras para explicar o significado de uma palavra complexa. Também pode inserir uma definição curta conforme exemplificado na Tabela 5. Trata-se de uma abordagem adequada em situações como o aprendizado de crianças e de uma segunda língua, pois explica e enriquece o vocabulário do estudante (URANO, 2000).

Tabela 5 – Exemplo de sentença simplificada por elaboração textual

Original	A ensilagem é o processo de conservação do alimento por fermentação anaeróbia.
Simplificada	A ensilagem é o processo de conservação do alimento por fermentação anaeróbia **(sem a presença de ar)**.

Fonte: Elaborada pelo autor.

2.2 Métricas

As formas de medir automaticamente a complexidade de textos ou sentenças representam por si só uma ampla área de pesquisa de interesse da Linguística Aplicada, Linguística Computacional, Psicolinguística, Educação e Fonoaudiologia. A análise automatizada da complexidade de textos, também conhecida em inglês por *Automatic Readability Assessment* (ARA), tem um viés de aplicação prática, pois ajuda a indicar material de leitura adequado, por exemplo para uma dada série escolar, mas também pode contribuir para um melhor entendimento dos processos de leitura e compreensão em populações com processamento típico e atípico de linguagem.

Graesser, McNamara e Kulikowich (2011) dividem as abordagens de predição e medição da complexidade (ou simplicidade) de textos em:

- **Tradicionais**: que usam uma única métrica ou a combinação linear de poucas métricas de dificuldade;
- **Modernas**: que analisam textos com múltiplas características em vários níveis linguísticos e cognitivos, e foram alavancadas por métodos de AM nas últimas duas décadas.

Um exemplo da primeira abordagem é o Índice Flesch que será visto na Subseção 2.2.1 e outro da segunda abordagem é o Coh-Metrix, apresentado na Seção 2.2.2.

Um dos grandes desafios para a aplicação dos métodos de AM em textos é a criação de córpus grandes e balanceados, anotados com as classes de interesse, por professores ou linguistas. O aprendizado do modelo usa a conversão dos textos em valores, geralmente numéricos, para serem usados nas fases de treinamento e avaliação dos métodos. Isso geralmente é obtido por meio da extração e seleção de métricas dos textos, em diversos níveis da língua, para em seguida utilizá-las como *features* nos métodos de aprendizado.

Há uma crítica frequente a essa abordagem de anotação da complexidade que usa preditores com base em córpus com julgamento de especialistas: o fato de a anotação não ser baseada no desempenho real da leitura de estudantes, por exemplo. Entretanto, se já há grande dificuldade em anotar um grande córpus com avaliação de professores, conseguir um córpus de alunos é mais ainda difícil e demorado (VAJJALA; MEURERS, 2016). O córpus *Touchstone Applied Science Associates* (TASA), na língua inglesa, é o único grande córpus disponível que atende essa crítica, no melhor do nosso conhecimento, por ser formado por 37.520 amostras de textos, com o tamanho de um parágrafo de tamanho médio de 288,6 palavras (desvio padrão de 25,4), cujas dificuldades foram avaliadas via tarefa de leitura de estudantes, sendo anotados também com a métrica DRP (*Degrees of Reading Power*)[1] (GRAESSER; MCNAMARA; KULIKOWICH, 2011). Entretanto, a possibilidade de usar rastreamento ocular para capturar o processo de leitura de estudantes é muito bem-vinda e foi explorada nesta tese.

[1] <http://textcomplexity.questarai.com/getdrp/>

Para facilitar a apresentação, são mostradas nas próximas 4 seções as principais fontes de métricas para a tarefa de predição da complexidade (textual e sentencial): fórmulas clássicas, linguísticas, psicolinguísticas e rastreamento ocular. Dentro de cada seção são descritas as principais métricas citadas na literatura.

2.2.1 Fórmulas clássicas

As primeiras fórmulas para avaliação de inteligibilidade textual surgiram na década de 1920 nos Estados Unidos, e por volta de 1980 já existiam mais de duzentas fórmulas diferentes (DUBAY, 2007).

Mesmo com o advento das abordagens mais modernas para resolver a tarefa, essas fórmulas continuam a ter grande importância para as tarefas de PLN, e são usadas isoladamente ou em conjunto com outras *features*. As principais para o escopo deste trabalho são detalhadas a seguir.

Índice Flesch

A fórmula *Flesch Reading-Ease Score* (FRES), para ser usada com textos em inglês:

$$\mathrm{F} = 206.835 - 1.015\left(\frac{\text{total palavras}}{\text{total sentenças}}\right) - 84.6\left(\frac{\text{total sílabas}}{\text{total palavras}}\right) \tag{2.1}$$

Segue a seguinte escala de interpretação: 90-100: Muito simples, 80-89: Simples, 70-79: Relativamente simples, 60-69: Padrão, 50-59: Relativamente complexo, 30-49: Complexo, 0-29: Muito complexo. É uma das mais antigas e utilizadas fórmulas de inteligibilidade e foi criada por Rudolph Flesch em 1948 (DELL'ORLETTA; MONTEMAGNI; VENTURI, 2011; SJÖHOLM, 2012). Foi adaptada para o PB em 1996 pelo NILC (MARTINS *et al.*, 1996), adicionando 42 pontos a todos os escores da fórmula original em inglês:

$$\mathrm{F} = 248.835 - 1.015\left(\frac{\text{total palavras}}{\text{total sentenças}}\right) - 84.6\left(\frac{\text{total sílabas}}{\text{total palavras}}\right) \tag{2.2}$$

Flesch-Kincaid grade level

A fórmula *Flesch–Kincaid Grade Level* apresenta como resultado um número que corresponde a uma série no sistema educacional americano, facilitando a avaliação do nível de complexidade de livros e textos. Pode ser interpretada como o número de anos de educação necessários para a leitura de um dado texto:

$$\mathrm{FK} = 0.39\left(\frac{\text{total palavras}}{\text{total sentenças}}\right) + 11.8\left(\frac{\text{total sílabas}}{\text{total palavras}}\right) - 15.59 \tag{2.3}$$

Foi desenvolvida por J. Peter Kincaid em 1975 (KINCAID *et al.*, 1975) a partir da anterior criada por Rudolph Flesch. É também uma função linear que utiliza a média de sílabas por palavras e média de palavras por sentença, estimando assim as complexidades lexical e sintática do texto (DELL'ORLETTA; MONTEMAGNI; VENTURI, 2011; SJÖHOLM, 2012).

Dale-Chall

Inspirada pela fórmula Flesch, a fórmula *Dale-Chall* acrescenta validação da dificuldade das palavras contra um dicionário com 3 mil palavras simples, sendo também considerada a média do tamanho das sentenças:

$$\mathrm{DC} = 0.1579 \left(\frac{\text{total palavras difíceis}}{\text{total palavras}} \times 100 \right) + 0.0496 \left(\frac{\text{total palavras}}{\text{total sentenças}} \right) \tag{2.4}$$

Foi criada em 1948 e atualizada posteriormente em 1995 por Edgard Dale e Jeanne Chall (CHALL; DALE, 1995; DELL'ORLETTA; MONTEMAGNI; VENTURI, 2011).

Gunning Fog Index

Gunning Fog Index ou simplesmente FOG Index foi criada em 1952 por Robert Gunning:

$$\mathrm{GF} = 0.4 \left[\left(\frac{\text{total palavras}}{\text{total sentenças}} \right) + 100 \left(\frac{\text{total palavras complexas}}{\text{total palavras}} \right) \right] \tag{2.5}$$

Ao avaliar a dificuldade de inteligibilidade dos jornais por estudantes de graduação, ele escreveu que os textos estavam repletos de incertezas, névoa (*fog* em inglês) e complexidade desnecessária (DUBAY, 2014). Palavras complexas nesse contexto são as que possuem três ou mais sílabas.

Coleman–Liau

Baseada em caracteres em vez de sílabas por palavra, possibilita utilizações mais mecânicas em textos:

$$\mathrm{CLI} = 0.0588\mathrm{L} - 0.296\mathrm{S} - 15.8 \tag{2.6}$$

Na fórmula acima, L é a média da quantidade de letras por 100 palavras e S é a média do número de sentenças por 100 palavras (COLEMAN; LIAU, 1975).

Brunét

O Índice de Brunét é uma variação da TTR (*Type Token Ratio*), mas insensível ao tamanho do texto:

$$W = N^{V^{-0.165}} \tag{2.7}$$

Na fórmula acima N é o número de *tokens* e V é o total de palavras do vocabulário (ou types). Foi criado por Étienne Brunet em 1978 (cf.(CUNHA, 2015; THOMAS *et al.*, 2005)). Os valores típicos da métrica variam entre 10 e 20, sendo que uma fala mais rica produz valores menores.

Honoré

A Estatística de Honoré é outra variação da TTR, também insensível ao tamanho do texto:

$$R = \frac{100 log N}{1 - \frac{V_1}{V}} \tag{2.8}$$

Na fórmula acima N é o número de *tokens* e V_1 é o número de palavras do vocabulário que aparecem uma única vez e V é o número de itens lexicais (ou types). Foi criada por A. Honoré em 1979 (cf. (CUNHA, 2015; THOMAS *et al.*, 2005)), sendo que valores altos da fórmula indicam um vocabulário rico.

2.2.2 Linguísticas

As métricas linguísticas procuram extrair características nos níveis lexical, morfossintático, sintático, semântico e discursivo da língua. Existem ferramentas com esse fim específico, que facilitam bastante o processo:

Coh-Metrix / TERA

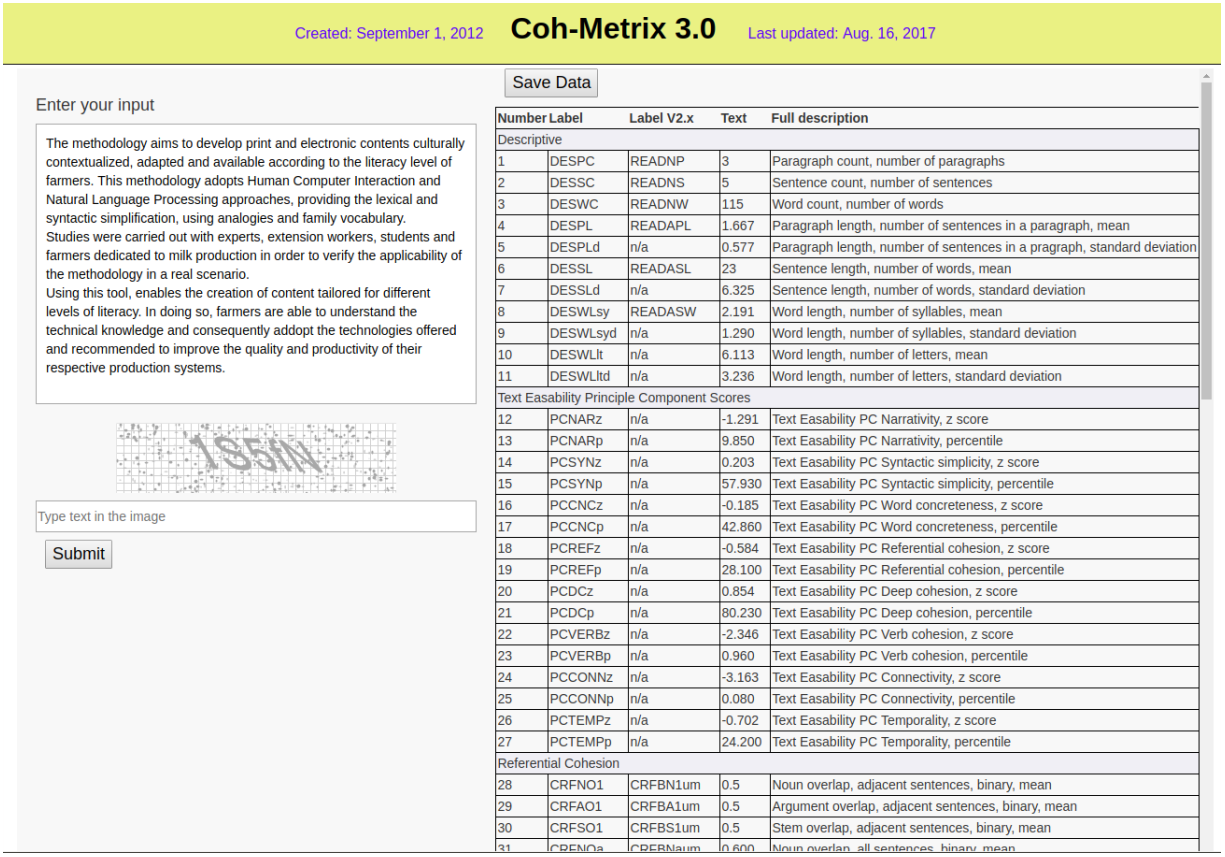
O Coh-Metrix[2] (MCNAMARA *et al.*, 2014) é uma ferramenta desenvolvida para a língua inglesa, que extrai de um texto métricas de coesão e coerência, permitindo avaliar a complexidade da sua leitura.

Os autores definem coesão como a relação entre as características do texto que guiam o leitor para a representação mental do significado, e coerência é a representação mental que o leitor cria durante a leitura (GRAESSER *et al.*, 2004).

[2] <http://cohmetrix.com/>

A versão 3.0 do Coh-Metrix implementa 106 métricas para a língua inglesa, agrupadas nas 11 categorias: *Descriptive, Text Easability Principal Component Scores, Referential Cohesion, LSA, Lexical Diversity, Connectives, Situation Model, Syntactic Complexity, Syntactic Pattern Density, Word Information e Readability*. A tela da ferramenta pode ser vista na Figura 1, que traz no lado direito os valores de diversas métricas do pequeno texto informado no lado esquerdo.

Figura 1 – Tela do Coh-Metrix com um texto de exemplo



Number	Label	Label V2.x	Text	Full description
Descriptive				
1	DESPC	READNP	3	Paragraph count, number of paragraphs
2	DESSC	READNS	5	Sentence count, number of sentences
3	DESWC	READNW	115	Word count, number of words
4	DESPL	READAPL	1.667	Paragraph length, number of sentences in a paragraph, mean
5	DESPLd	n/a	0.577	Paragraph length, number of sentences in a pragraph, standard deviation
6	DESSL	READASL	23	Sentence length, number of words, mean
7	DESSLd	n/a	6.325	Sentence length, number of words, standard deviation
8	DESWLsy	READASW	2.191	Word length, number of syllables, mean
9	DESWLsyd	n/a	1.290	Word length, number of syllables, standard deviation
10	DESWLlt	n/a	6.113	Word length, number of letters, mean
11	DESWLltd	n/a	3.236	Word length, number of letters, standard deviation
Text Easability Principle Component Scores				
12	PCNARz	n/a	-1.291	Text Easability PC Narrativity, z score
13	PCNARp	n/a	9.850	Text Easability PC Narrativity, percentile
14	PCSYNz	n/a	0.203	Text Easability PC Syntactic simplicity, z score
15	PCSYNp	n/a	57.930	Text Easability PC Syntactic simplicity, percentile
16	PCCNCz	n/a	-0.185	Text Easability PC Word concreteness, z score
17	PCCNCp	n/a	42.860	Text Easability PC Word concreteness, percentile
18	PCREFz	n/a	-0.584	Text Easability PC Referential cohesion, z score
19	PCREFp	n/a	28.100	Text Easability PC Referential cohesion, percentile
20	PCDCz	n/a	0.854	Text Easability PC Deep cohesion, z score
21	PCDCp	n/a	80.230	Text Easability PC Deep cohesion, percentile
22	PCVERBz	n/a	-2.346	Text Easability PC Verb cohesion, z score
23	PCVERBp	n/a	0.960	Text Easability PC Verb cohesion, percentile
24	PCCONNz	n/a	-3.163	Text Easability PC Connectivity, z score
25	PCCONNp	n/a	0.080	Text Easability PC Connectivity, percentile
26	PCTEMPz	n/a	-0.702	Text Easability PC Temporality, z score
27	PCTEMPp	n/a	24.200	Text Easability PC Temporality, percentile
Referential Cohesion				
28	CRFNO1	CRFBN1um	0.5	Noun overlap, adjacent sentences, binary, mean
29	CRFAO1	CRFBA1um	0.5	Argument overlap, adjacent sentences, binary, mean
30	CRFSO1	CRFBS1um	0.5	Stem overlap, adjacent sentences, binary, mean
31	CRFNOa	CRFBNaum	0.600	Noun overlap, all sentences, binary, mean

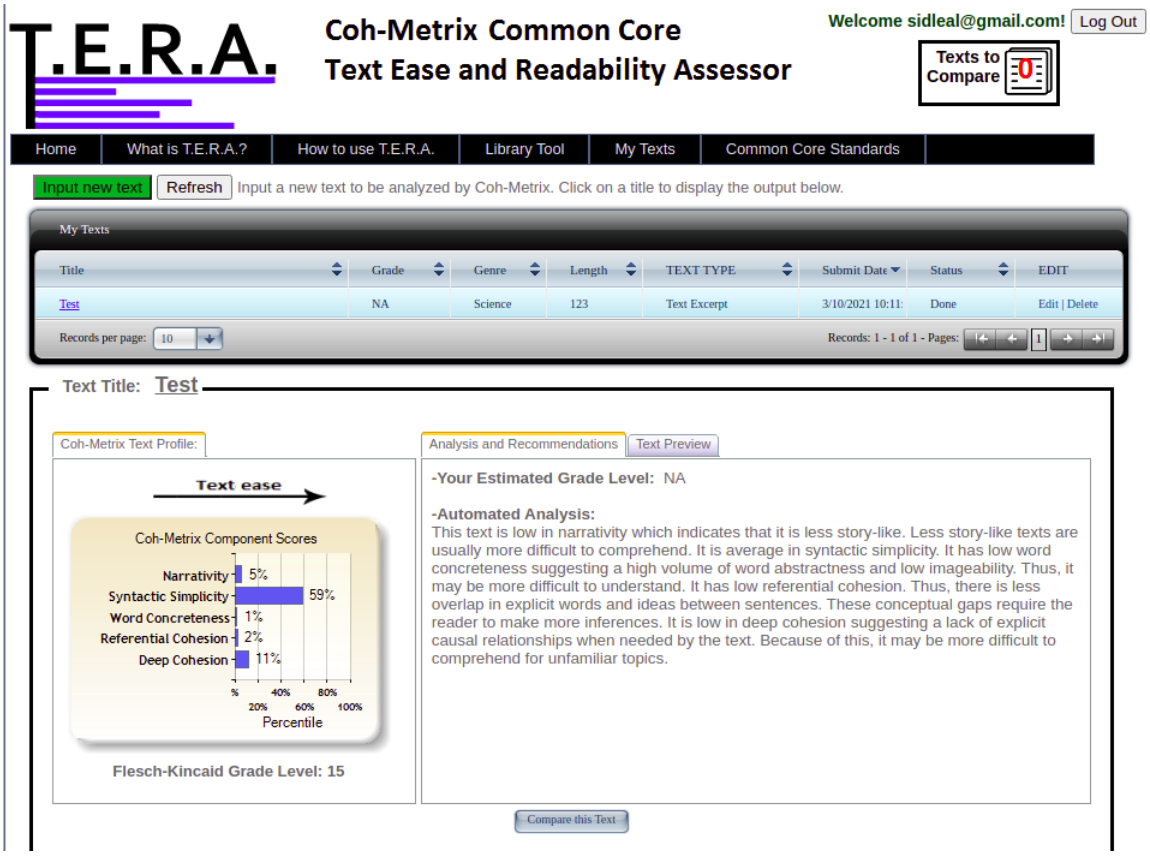
Fonte: Elaborada pelo autor.

Em contraste com as fórmulas clássicas que analisam o texto apenas no nível das palavras e sentenças e geram um único valor para quantificar a complexidade do texto, o Coh-Metrix utiliza uma análise multinível, alinhada com teorias de compreensão textual (GRAESSER; MCNAMARA; KULIKOWICH, 2011):

1. ***Words***: Como o conhecimento do vocabulário de uma língua tem um grande impacto sobre o tempo de leitura e compreensão, Coh-Metrix tem uma grande quantidade de métricas relacionadas a palavras, incluindo: análise de categorias gramaticais ou *Part of Speech* (PoS), frequência, medidas psicolinguísticas como concretude, familiaridade, idade de aquisição, imageabilidade, categorias semânticas obtidas da WordNet de Princeton[3];

[3] https://wordnet.princeton.edu/

Figura 2 – Tela de exemplo do Coh-Metrix-T.E.R.A.



Fonte: Elaborada pelo autor.

2. ***Syntax***: Algumas sentenças do discurso oral são curtas, apresentam poucas orações relativas, poucas palavras nos sintagmas nominais e se apresentam na voz ativa, mas sentenças de textos escritos geralmente aparecem de forma oposta, demandando mais processamento da memória de trabalho. Coh-Metrix computa essas contagens e outras como similaridade de pares de sentenças adjacentes, que facilitam a leitura e compreensão;

3. ***Textbase***: A base textual está relacionada com o significado em vez da análise de palavras e da sintaxe. A co-referência é um mecanismo importante para conectar as proposições, as orações e sentenças na base textual, assim Coh-Metrix traz várias métricas para o cômputo da co-referência como a sobreposição de palavras de conteúdo, de substantivos e de radicais (*content word overlap, noun overlap, e stem overlap*, respectivamente). A diversidade lexical está relacionada com a coesão porque um número elevado de palavras diferentes em um texto significa que as palavras novas precisam ser integradas no contexto do discurso. Coh-Metrix também computa várias métricas relacionadas com o modelo estatístico para cálculo de similaridade chamado *Latent Semantic Analysis* (LSA), pois ele ajuda a medir o conhecimento implícito do leitor em adição às palavras explícitas usadas no texto;

4. ***Situation Model / Mental Model***: Textos narrativos incluem pessoas, objetos, ações,

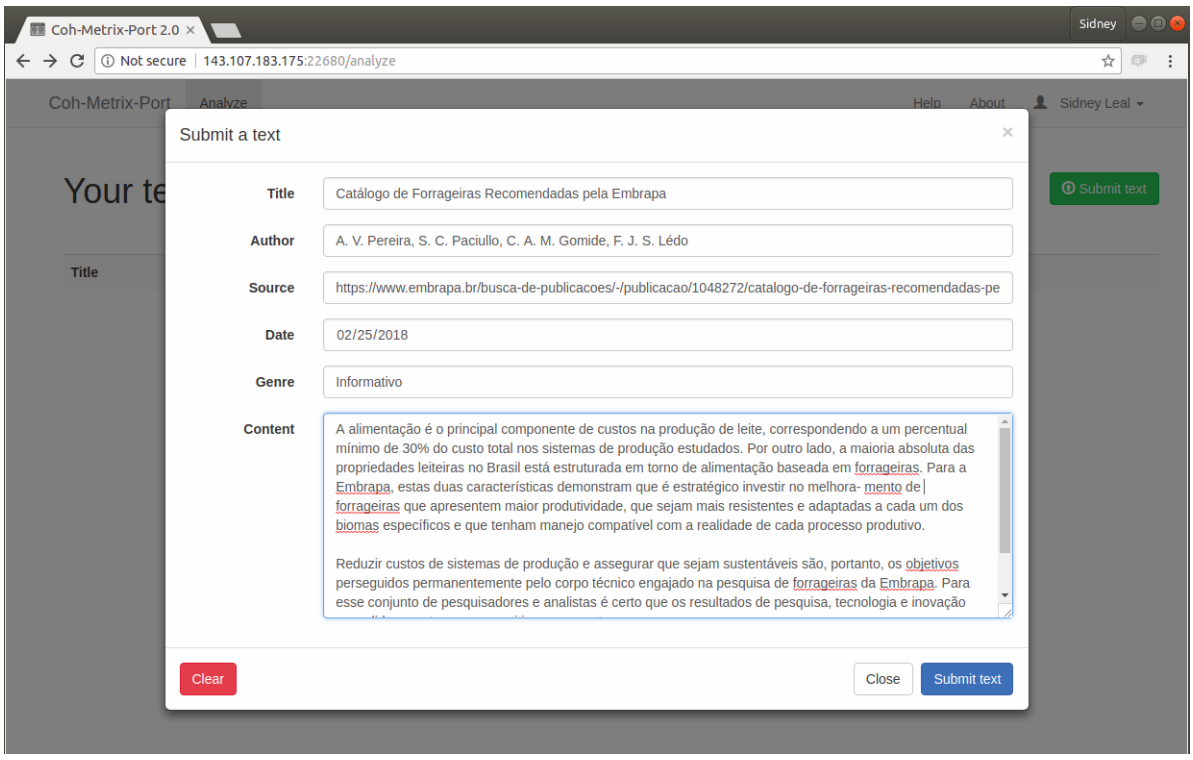
eventos, processos, planos e outros detalhes de uma estória, já em textos informativos o modelo mental é diferente, pois devem ajudar a entender como modelos da física, biologia e outras ciências funcionam. Assim, há métricas para avaliar se há quebras no entendimento desses modelos mentais que emergem de um texto;

5. ***Genre and Rhetorical Structure***: Exemplos de uma tipologia de gêneros são: narrativo, expositivo, persuasivo ou descritivo. Textos narrativos são mais fáceis de se ler, compreender e relembrar do que textos informativos. Coh-Metrix analisa se um texto pode ser classificado como narrativo ou informativo, via uma métrica chamada narratividade.

T.E.R.A.[4] (*Text Ease and Readability Assessor*) é uma ferramenta construída pelos mesmos autores, que usa o Coh-Metrix para avaliar amostras dos textos, reduzindo as métricas em cinco fatores, levantados via *Principal Component Analysis* (PCA) (MCNAMARA *et al.*, 2013; GRAESSER; MCNAMARA; KULIKOWICH, 2011): *Narrativity, Syntactic Simplicity, Word Concreteness, Referential Cohesion (Textbase)* e *Deep Cohesion (Situation Model)*. Na Figura 2 é possível ver um exemplo da análise do texto com as cinco dimensões.

Coh-Metrix-Port

Figura 3 – Tela de avaliação do Coh-Metrix-Port 2



Fonte: Elaborada pelo autor.

4 <http://www.commoncoretera.com/>

O Coh-Metrix-Port[5] (SCARTON; ALUÍSIO, 2010) é uma adaptação para o PB do Coh-Metrix, desenvolvida dentro do projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital), que teve como objetivo promover o acesso a textos da Web a pessoas com baixo letramento.

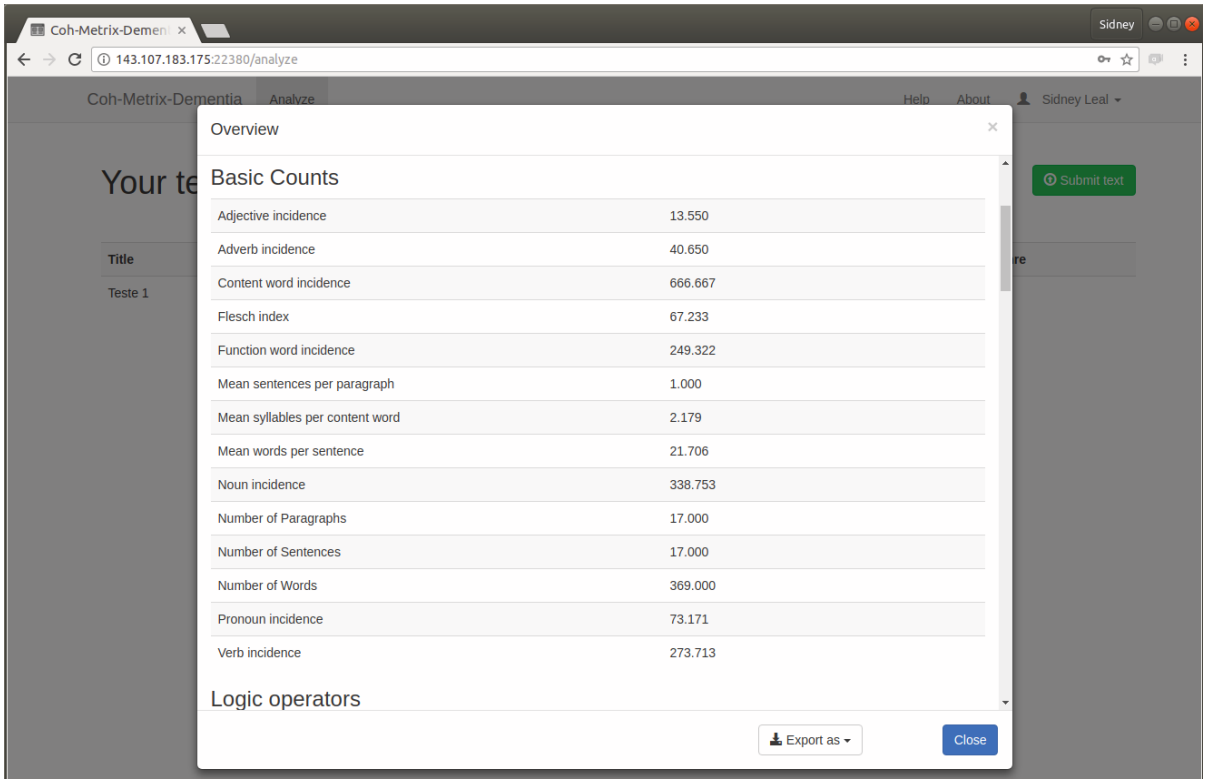
O Coh-Metrix-Port implementa 48 métricas específicas para o PB (SCARTON *et al.*, 2010), divididas nas categorias: contagens básicas, operadores lógicos, frequências, hiperônimos, tokens, constituintes, conectivos, ambiguidade, co-referência e anáforas. A tela de cadastro dos textos, da versão 2.0[6], pode ser vista na Figura 3.

Coh-Metrix-Dementia

O Coh-Metrix-Dementia[7] (CUNHA, 2015) é uma adaptação do Coh-Metrix-Port para análise automática de distúrbios de linguagem nas demências (como Doença de Alzheimer) ou no Comprometimento Cognitivo Leve (CCL).

Ele adiciona 25 novas métricas às 48 do Coh-Metrix-Port, nas categorias: disfluências, análise de semântica latente, diversidade lexical, complexidade sintática e densidade semântica.

Figura 4 – Exemplo da tela de saída do Coh-Metrix-Dementia



Fonte: Elaborada pelo autor.

[5] <http://fw.nilc.icmc.usp.br:22680/>
[6] Refeita por Cunha (2015)
[7] <http://fw.nilc.icmc.usp.br:22380/>

Disponibiliza no total 73 métricas para o PB. Sua tela principal pode ser vista na Figura 4. É importante citar que Treviso (2017) criou um detector de disfluências para facilitar a utilização da ferramenta e suas métricas, que dependem de *parsers* treinados com textos bem escritos, isto é, gramaticalmente corretos.

LIWC

LIWC (*Linguistic Inquiry and Word Count*) é uma ferramenta baseada em dicionários para análise dos vários componentes emocionais, cognitivos e linguísticos em amostras de textos (PENNEBAKER *et al.*, 2015 apud REIS, 2017), com categorias como: estatísticas comuns do texto, dimensão linguística, processos psicológicos, relatividade, assuntos pessoais e miscelânea, totalizando aproximadamente 100 métricas (CUNHA, 2015).

A sua primeira versão foi criada em 1993, a segunda em 2001, a terceira em 2007 e a última em 2015. O dicionário da versão inglesa conta com 6400 palavras, radicais e emoticons (PENNEBAKER *et al.*, 2015).

A tradução e adaptação do dicionário para o PB foi realizada em uma colaboração entre NILC, Checon Pesquisa e Unisinos no período de 2010 a 2012 e está disponível no site do projeto PortLex[8].

AIC

Também criada dentro do contexto do PorSimples (MAZIERO; PARDO; ALUÍSIO, 2008), a ferramenta AIC (Análise Automática de Inteligibilidade de Córpus) traz 39 métricas, com o principal diferencial de utilizar o analisador sintático PALAVRAS (BICK, 2000) para o cálculo delas. Elas estão organizadas em seis classes: estatísticas do texto, voz passiva, características das orações, densidade, personalização e marcadores discursivos (CUNHA, 2015; REIS, 2017). A tela de saída pode ser vista na Figura 5.

2.2.3 Psicolinguísticas

As palavras possuem algumas propriedades subjetivas estudadas pela psicolinguística como: imageabilidade, concretude, familiaridade e idade de aquisição (SANTOS *et al.*, 2017), detalhadas abaixo:

- **Imageabilidade**: Envolve a facilidade e rapidez de evocar uma imagem mental da palavra.
- **Concretude**: É o grau com que uma palavra se refere a objetos, pessoas, lugares ou coisas que podem ser percebidas pelos sentidos, em contraste com os conceitos abstratos.
- **Familiaridade**: É o grau com que pessoas conhecem e usam palavras no dia a dia.

[8] <http://nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc>

Figura 5 – Exemplo da tela de saída da AIC, atualmente não disponível no site do NILC.

Fonte: Maziero, Pardo e Aluísio (2008).

- **Idade de Aquisição**: Estimativa da idade em que uma palavra foi aprendida, calculada via análise feita por adultos.

Essas propriedades têm um grande impacto na complexidade dos textos e sentenças, e trazem melhorias aos resultados de várias tarefas de PLN, como simplificação lexical e tarefas de classificação semântica quando utilizadas em conjunto com as demais métricas (PAETZOLD; SPECIA, 2016).

Santos *et al.* (2017) anotaram automaticamente essas métricas em um banco[9] de 26.874 palavras do PB utilizando um método baseado em regressão e *Multi-View Learning*, com recursos fáceis de se obter em várias línguas.

2.2.4 Rastreamento ocular

As métricas do rastreamento ocular trazem uma abordagem recente e diferente das métricas mostradas anteriormente. Sua contribuição é bastante relevante, uma vez que permitem uma aproximação da percepção mais realista da complexidade pelos leitores.

Os movimentos dos olhos podem ser interpretados como uma janela para o processamento do cérebro, refletindo os tempos cognitivos envolvidos em determinada tarefa. Por exemplo, durante a leitura os movimentos dos olhos são controlados por uma complexa interação entre os

[9] A base está disponível em: <http://143.107.183.175:21380/portlex/index.php/en/component/content/article/2-uncategorised/23-psycholinguistic>

fatores de baixo nível (por exemplo, o quanto o olho consegue ver e interpretar a cada fixação) e de alto nível (por exemplo, o processamento sintático) (BARRETT; AGIC; SØGAARD, 2015).

Rayner (1998) divide a pesquisa sobre os movimentos dos olhos (ou rastreamento ocular) em três grandes eras. A primeira era vai desde as primeiras observações sobre os movimentos dos olhos durante a leitura em 1879 até os anos 1920. Algumas importantes descobertas foram feitas nessa era como, por exemplo, o fato de que não percebemos nenhuma informação durante o reposicionamento do olhar, denominado sacada ou *saccade*, em inglês.

A segunda era coincide com o movimento behaviorista na psicologia experimental, com os trabalhos com focos mais práticos — estudos dos movimentos dos olhos em si ou em aspectos superficiais da tarefa investigada — e menos concentrados na utilização dos movimentos para inferir o processamento cognitivo.

Figura 6 – Rastreamento ocular de plataforma, com óculos simples, e com óculos especial de realidade virtual.

Fonte: Imotions (2017), Fove (2018).

A terceira era começa em meados dos anos 1970, com melhorias nos sistemas de rastreamento que permitiram medidas mais acuradas e simples de obter (Figura 6). Nessa era, juntamente com os avanços das teorias de processamento da linguagem, os movimentos dos olhos começaram a ser utilizados para exame crítico dos processos cognitivos durante a leitura.

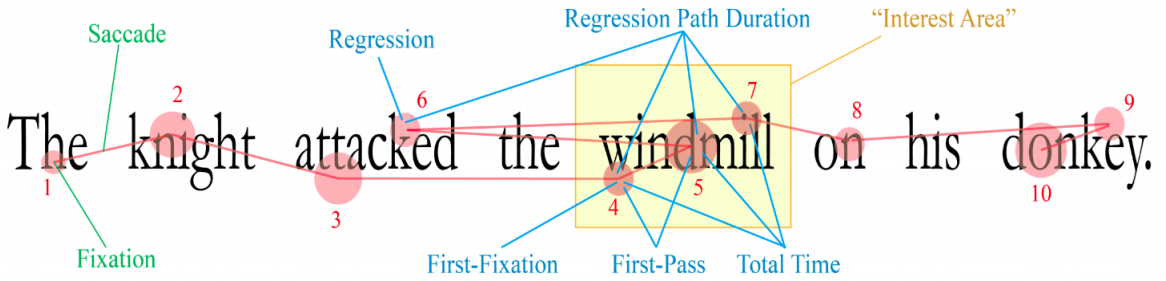
Em PB, o rastreamento ocular já é utilizado há algum tempo na área da psicolinguística. Por exemplo Maia, Lemle e França (2007) utilizaram para investigar o papel do processamento morfológico na identificação de palavras, Leitão, Ribeiro e Maia (2012) utilizaram na investigação do processamento anafórico e Teixeira, Fonseca e Soares (2014) para evidenciar o custo de resolução de pronomes nulos e plenos.

As características básicas dos movimentos dos olhos são:

- **Sacadas (*Saccades* em inglês)**: Os contínuos movimentos oculares, o reposicionamento do olhar (durante uma sacada nenhuma informação é percebida).
- **Fixações (*Fixations* em inglês)**: Os tempos de fixação em um ponto de atenção entre as sacadas.

A partir dessas duas características é possível medir diversas outras informações relevantes para o processo de leitura e interpretação de textos. As principais métricas obtidas pelo movimento dos olhos são exemplificadas na Figura 7, com a simulação do caminho feito pelo olhar em dez fixações numeradas sequencialmente, e detalhadas a seguir:

Figura 7 – Principais métricas de rastreamento ocular



Fonte: Zelenina (2015).

- ***First fixation duration***: Tempo da primeira fixação na palavra.
- ***First pass fixation duration***: Quando uma palavra é longa, pode ser necessário um segundo ponto de fixação dentro da própria palavra. Essa métrica é a soma dos tempos das fixações na primeira passada pela palavra.
- ***Total fixation duration***: Soma de todos os tempos de fixação na palavra.
- ***Average fixation duration***: Tempo médio de fixação, quando se tem mais de um ponto por palavra ou média por sentença.
- ***Regression***: Regressões no texto podem indicar necessidade de rever alguma informação para entendimento do ponto atual, por exemplo, para resolver uma correferência. É uma métrica muito importante para medir complexidade textual e sentencial.
- ***Regression path duration***: Mede a extensão da regressão; quanto maior a regressão, maior o esforço despendido para a leitura, como resultado de um texto mais complexo.
- ***Interest area***: Pontos de interesse, onde o leitor passou mais tempo fixando no texto. Calculado com a soma de todas as fixações.
- ***Skipping rate***: Algumas palavras são naturalmente saltadas durante a leitura, como artigos e preposições. Não saltar essas palavras pode indicar um leitor com menor proficiência na leitura.
- ***Number of fixations***: Quantidade de fixações na palavra; uma palavra simples só deve exigir uma única fixação.

- ***Second pass fixation duration***: Tempo de fixação na segunda vez que o leitor retorna à palavra.
- ***Spillover from previous word***: Nem sempre o processamento de uma palavra é completado antes que o olhar se mova para a próxima. Nesses casos ocorre o efeito de "transbordamento" do tempo para a palavra seguinte.

Existem diversos córpus, principalmente para a língua inglesa, com métricas de rastreamento ocular, descritos na Subseção 2.3.3.

2.3 Córpus

Para a avaliação da complexidade de sentenças, principal tarefa de estudo neste projeto, é importante utilizar córpus com sentenças alinhadas[10] e anotadas em níveis de complexidade. A seguir, são apresentados os principais utilizados na tarefa, para as línguas inglesa, italiana e sueca (Subseção 2.3.1). Também são apresentados o PorSimples — que possui alinhamento sentencial para o PB (Subseção 2.3.2) e, portanto, é uma das bases para este trabalho — e os córpus com métricas de rastreamento ocular disponíveis para outras línguas (Subseção 2.3.3), que motivaram a criação de um córpus nos mesmos moldes para o PB.

2.3.1 Córpus para avaliação da tarefa de predição da complexidade

Esta seção traz os principais córpus utilizados nos trabalhos internacionais revisados para a tarefa de predição da complexidade em nível sentencial: Wikipedia-SimpleWikipedia, OneStopEnglish, Newsela, SimPA, Rep/2Par, PaCCSS-IT e LäSBarT/GP2006.

Wikipedia-SimpleWikipedia - Inglês

Córpus de sentenças alinhadas criado por Zhu, Bernhard e Gurevych (2010) com base na SimpleWikipedia[11] e Wikipedia[12]. Foi utilizado por Vajjala e Meurers (2014a) e se tornou um *benchmark* para a avaliação da tarefa na língua inglesa.

Foi criado com o pareamento das sentenças de 65.133 artigos da SimpleWikipedia e Wikipedia, usando *Term Frequency - Inverse Document Frequency* (TF-IDF). Para a escolha da medida de alinhamento, avaliaram a performance de três medidas de similaridade: TF-IDF, *word overlap* e *word-based maximum edit distance* (MED), contra 120 pares de sentenças anotadas manualmente. A precisão ficou acima de 90% no TF-IDF, conforme Tabela 6.

Tabela 6 – Avaliação das medidas de similaridade para alinhamento das sentenças da Wikipedia e SimpleWikipedia.

Similaridade	Precisão (%)	Recall (%)
TF-IDF	91,3	55,4
Word Overlap	50,5	55,1
MED	13,9	54,7

Fonte – (ZHU; BERNHARD; GUREVYCH, 2010)

Como resultado final, criaram 108.016 sentenças alinhadas, anotadas em duas classes: complexas ou simples, sendo que uma sentença complexa pode estar mapeada para uma ou mais simples, para tratar a divisão de sentenças.

[10] Conforme constatado por Vajjala e Meurers (2016)
[11] <http://simple.wikipedia.org>
[12] <http://en.wikipedia.org>

Esse córpus foi atualizado cinco anos depois de criado por Hwang *et al.* (2015), chegando a 150.000 pares de sentenças alinhadas.

OneStopEnglish - Inglês

Foi compilado por Vajjala e Meurers (2016) a partir do site OneStopEnglish (OSE) e utilizado para avaliação da tarefa de complexidade das sentenças. O córpus foi disponibilizado publicamente em 2018 sob a licença *Creative Commons* para *download*[13] (VAJJALA; LUČIĆ, 2018).

O OneStopEnglish é um site de apoio a professores de inglês, onde são publicadas lições semanais[14] com artigos retirados do jornal *The Guardian* e reescritos por especialistas para três níveis de aprendizado: iniciante, intermediário e avançado.

Possui um total de 228 artigos, compostos por 76 trios (três versões de cada texto, uma para cada nível). As sentenças foram alinhadas com TF-IDF (dentro do mesmo texto, as sentenças que mantiveram a maior parte das palavras após a simplificação foram consideradas um par), resultando em dois conjuntos:

- **OSE3:** Com sentenças nos três níveis (iniciante-intermediário-avançado), totalizando 837 trios.
- **OSE2:** 3.113 pares de sentenças nos níveis iniciante-intermediário, intermediário-avançado e iniciante-avançado.

Newsela - Inglês

O Newsela é um grande córpus (não-público) para o inglês, que foi utilizado por Scarton, Paetzold e Specia (2018b), Scarton e Specia (2018) e Stajner, Ponzetto e Stuckenschmidt (2017) em seus trabalhos de investigação da complexidade textual e sentencial.

A versão utilizada pelos autores acima (2016) é composta de 10.787 artigos jornalísticos em inglês, que incluem 1.911 artigos em sua versão original e também em versões simplificadas por humanos em 4 ou 5 níveis de complexidade. Cada documento é caracterizado por um identificador de versão entre 0 e 5, sendo zero o mais complexo e cinco o mais simples, e também um nível de leitura que vai de 2 a 12, onde 2 é o nível mais simples e 12 o mais complexo.

Scarton, Paetzold e Specia (2018b) fizeram o alinhamento automático dos textos[15], resultando em um córpus com 19.198 pares de documentos alinhados em 300.475 pares de parágrafos e 550.644 pares de sentenças.

[13] <https://zenodo.org/record/1219041>
[14] <http://www.onestopenglish.com/skills/news-lessons/weekly-topical-news-lessons>
[15] A biblioteca para alinhamento e anotação automáticos está disponível em <https://github.com/ghpaetzold/massalign>

SimPA - Inglês

O córpus SimPA foi criado no escopo do projeto SIMPATICO (SCARTON; PAETZOLD; SPECIA, 2018a), para o domínio de textos da administração pública em inglês. Ele contém, atualmente, 1.100 sentenças originais com suas simplificações nos níveis lexical (três anotadores) e sintático (a partir de uma das versões simplificadas).

Na versão atual[16] possui 3.300 sentenças simplificadas lexicalmente e mais 1.100 também simplificadas sintaticamente. Uma das vantagens desse córpus é permitir a análise das simplificações lexicais e sintáticas de forma isolada.

Rep / 2Par - Italiano

Utilizados por Dell'Orletta, Montemagni e Venturi (2011) para avaliar a tarefa de complexidade sentencial, o *La Repubblica*[17] (Rep) é um córpus jornalístico com mais de 380 milhões de tokens disponíveis para o italiano, enquanto o *Due Parole* (2Par) possui textos adaptados por linguistas para pessoas com baixo letramento e comprometimento cognitivo leve. Segundo os autores, o 2Par possui 322 documentos e 73 mil palavras.

PaCCSS–IT - Italiano

O *Parallel Corpus of Complex–Simple Aligned Sentences for ITalian* é o maior córpus de sentenças alinhadas disponível atualmente para o italiano. Foi criado por Brunato *et al.* (2016) e utilizado no trabalho de Bosco, Pilato e Schicchia (2018b).

O córpus foi compilado por um método automatizado que buscou por paráfrases em um grande córpus de textos obtidos por webcrawler (ItWaC com 2 bilhões de palavras) e em seguida selecionou o melhor par e fez *ranking* para identificar o lado simples e o complexo.

Conta com aproximadamente 63.000 pares alinhados e está disponível publicamente para download[18].

LäSBarT / GP2006 - Sueco

Utilizados por Sjöholm (2012) em sua tese, o *LäSBarT*[19] é um córpus considerado de simples leitura enquanto o GP2006[20], do jornal *Göteborgsposten*, foi a fonte de sentenças complexas.

[16] Disponível publicamente em <https://github.com/SIMPATICOProject/simpa>
[17] <http://docs.sslmit.unibo.it/doku.php?id=corpora:repubblica>
[18] <http://www.italianlp.it/resources/paccss-it-parallel-corpus-of-complex-simple-sentences-for-italian/>
[19] <https://spraakbanken.gu.se/resource/lasbart>
[20] <https://spraakbanken.gu.se/resource/gp2006>

Ambos são córpus de textos e não de sentenças, portanto podem ocorrer sentenças simples no segundo e sentenças complexas no primeiro. O *LäSBarT* conta com 104.058 sentenças e o GP2006 possui 1.376.836 sentenças.

2.3.2 Córpus para avaliação da tarefa de predição da complexidade em PB

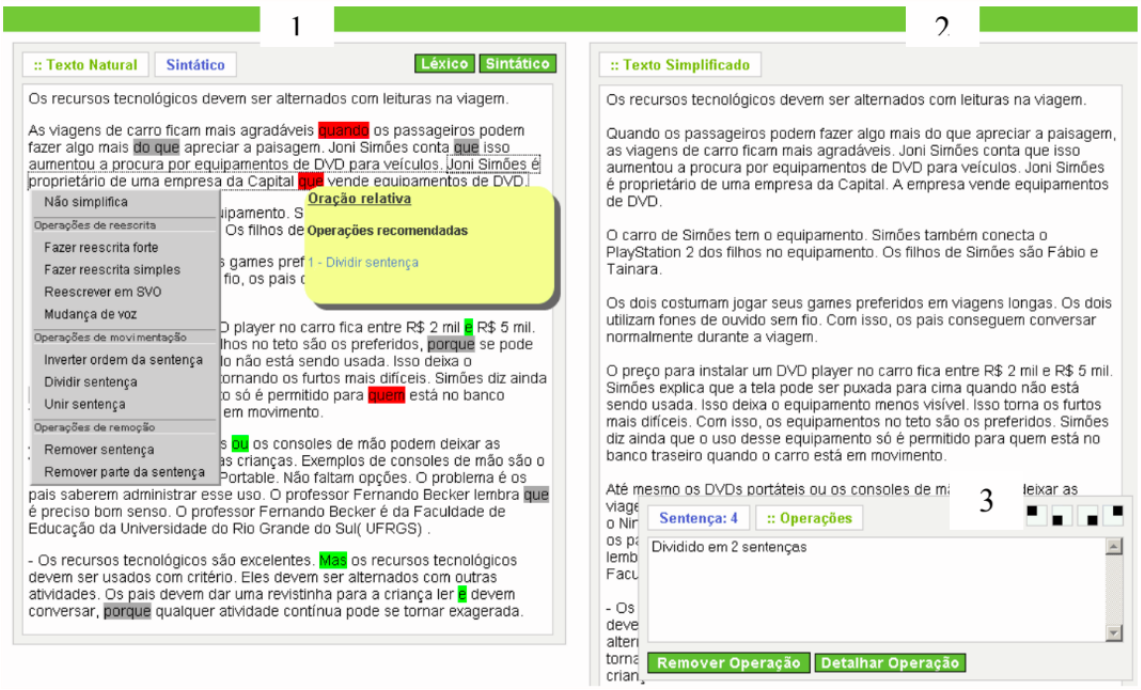
Para o português brasileiro, há poucos recursos prontos para a tarefa de avaliação da complexidade. Até onde foi possível verificar, o primeiro córpus público especificamente para avaliação automática no nível sentencial foi o PorSimplesSent, produzido no contexto deste trabalho.

Nesta seção, são descritos o PorSimples, que deu origem ao PorSimplesSent e também o CorPop, que traz uma amostra representativa do português popular escrito e pode ser utilizado na validação dos modelos.

PorSimples

Córpus paralelo de textos originais e simplificados criado em 2009 no projeto PorSimples[21] (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) do NILC (CASELI *et al.*, 2009).

Figura 8 – Exemplo da tela do editor de anotação do PorSimples



Fonte: Caseli *et al.* (2009).

[21] <http://www.nilc.icmc.usp.br/nilc/index.php/projetos?layout=edit&id=27>

Um editor foi desenvolvido para a tarefa de anotação, cuja tela principal pode ser vista na Figura 8. No lado esquerdo, fica o texto original e à direita, sua versão simplificada.

Os textos jornalísticos foram simplificados em dois níveis por especialistas linguistas:

- **Natural**: textos para os quais o anotador escolheu livremente as operações de simplificação, inclusive podendo escolher não simplificar uma sentença.
- **Forte**: Anotadores seguiram o manual de simplificação também desenvolvido no projeto.

A primeira fase do córpus foi criada a partir de 104 textos do jornal Zero Hora. A documentação é bastante rica, por exemplo as estatísticas principais das simplificações nesta fase são mostradas na Figura 9, onde é possível constatar que a divisão de sentenças e substituição de palavras foram as operações mais utilizadas.

Figura 9 – Estatísticas das simplificações na primeira fase do córpus PorSimples

Syntactic and Lexical Simplification Operations	Number of sentences / (%) / Average sentence length					
	Original to Natural			Natural to Strong		
Non-simplification	418	19.75%	13.1	2,220	71.52%	11.86
Strong rewriting	7	0.33%	19.85	4	0.13%	14.5
Simple rewriting	509	24.05%	21.91	313	10.0%	16.95
Subject-verb-object ordering	31	1.46%	25.06	13	0.42%	14.15
Transformation to active voice	89	4.21%	22.12	65	2.09%	18.95
Inversion of clause ordering	191	9.03%	22.36	74	2.38%	18.89
Splitting sentences	723	34.17%	26.80	380	12.24%	23.58
Joining sentences	5	0.24%	10.83	6	0.19%	18.33
Dropping one sentence	6	0.28%	11	3	0.09%	5.3
Dropping sentence parts	241	11.39%	26.20	49	1.58%	22.20
Lexical Substitution	980	46.31%	23.46	196	6.34%	18.01

Fonte: Caseli *et al.* (2009).

Na segunda fase, foram adicionados 50 textos do Caderno de Ciência do jornal Folha de São Paulo, resultando em 154 trios alinhados num total de 462 textos e mais de 185 mil tokens.

Na Tabela 7 podem ser vistos os números de sentenças de cada nível, e para os tokens confira a Tabela 8. Uma característica importante do córpus PorSimples foi a anotação dos fenômenos linguísticos nas sentenças; uma extração deles pode ser vista para cada nível na Tabela 9. Com essa informação é possível verificar que alguns fenômenos adicionam mais complexidade que outros, por exemplo, as orações apositivas foram as que mais diminuíram em número durante o processo de simplificação, já as subordinadas contraintuitivamente aumentaram nos níveis mais simples.

Esse córpus é muito importante para este trabalho, pois já possui alinhamento no nível sentencial. Esse alinhamento foi utilizado para gerar o córpus PorSimplesSent, detalhado no Capítulo 5.

Tabela 7 – PorSimples - Estatísticas de Sentenças.

	Total	Mínimo/Texto	Máximo/Texto	Média/Texto
Original	2.985	5	46	19
Natural	4.080	5	62	26
Forte	4.974	7	72	32

Fonte: Elaborada pelo autor.

Tabela 8 – PorSimples - Estatísticas de Tokens.

	Total	Mínimo/Sentença	Máximo/Sentença	Média/Sentença
Original	61.026	2	71	21
Natural	61.754	2	60	15
Forte	63.030	2	47	13

Fonte: Elaborada pelo autor.

Tabela 9 – PorSimples - Fenômenos Linguísticos.

	Coordenadas	Subordinadas	Relativas	Passivas	Apositivas
Original	1.443	805	897	319	306
Natural	1.352	899	759	257	105
Forte	1.210	876	527	167	73

Fonte: Elaborada pelo autor.

CorPop

O CorPop (córpus de referência do português popular escrito do Brasil) foi criado em 2018 por Pasqualini (2018) durante seu doutorado. Ele traz uma compilação bem avaliada de textos selecionados com base no nível de letramento médio dos leitores do país, das seguintes fontes:

1. Textos do jornalismo popular do Projeto PorPopular (jornal Diário Gaúcho) consumido maciçamente pelas classes C e D;

2. Textos e autores mais lidos pelos respondentes das últimas edições da pesquisa Retratos da Leitura no Brasil;

3. Coleção “É Só o Começo” (adaptação de clássicos da literatura brasileira para leitores com baixo letramento, realizada por linguistas);

4. Textos do jornal Boca de Rua, produzido por pessoas em situação de rua, com baixa escolaridade e baixo letramento;

5. Textos do Diário da Causa Operária, imprensa operária brasileira produzida também por pessoas dentro da faixa média de letramento do país.

O córpus possui 684 mil tokens, conforme apresentado na Tabela 10 por módulo. Está parcialmente disponível publicamente[22] (ferramentas e listas de palavras).

Tabela 10 – Número total de *types* e *tokens* do córpus CorPop.

Módulo	***Types***	***Tokens***
PorPopular	6.378	30.944
Hora de Santa Catarina	4.118	18.303
Boca de Rua	8.913	71.454
Diário da Causa Operária	7.841	59.785
Retratos da Leitura no Brasil	22.463	430.806
Coleção É Só o Começo	8.161	73.507
Total	**32.138**	**684.799**

Fonte – (PASQUALINI, 2018)

2.3.3 Córpus com métricas de rastreamento ocular

Conforme detalhado na Subseção 2.2.4, as métricas de rastreamento ocular são muito importantes para a tarefa de predição da complexidade sentencial. Uma das formas de utilização dessas métricas nos trabalhos revisados é treinar métodos para predizê-las em córpus já existentes. A seguir, são descritos os córpus internacionais com dados de rastreamento ocular.

Dundee Eye-Tracking

O Córpus Dundee foi criado por Alan Kenned e Joël Pynte em 2003, no Departamento de Psicologia da Universidade de Dundee (KENNEDY; HILL; PYNTE, 2003 apud BARRETT; AGIC; SØGAARD, 2015). As medidas foram registradas durante a leitura de artigos de jornais do *The Independent* para o Inglês e *Le Monde* para o Francês.

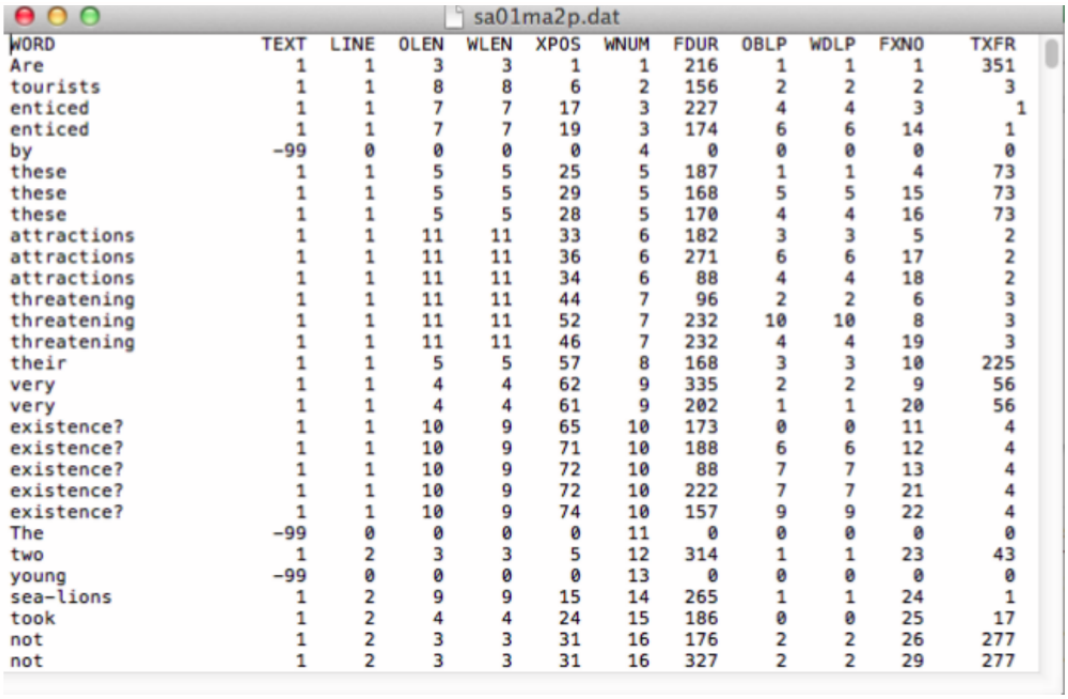
Para o Inglês, contou com dez participantes que falavam a língua de forma nativa, resultando em um córpus com 2.368 sentenças e 56.212[23] tokens após a leitura de vinte artigos, divididos em quarenta telas com cinco linhas, e cada linha com oitenta caracteres (KENNEDY; PYNTE, 2005). Após ler cada texto, o participante respondia a um teste rápido de múltipla escolha para avaliar a compreensão.

O córpus traz as métricas de tempo e ordem de fixação do olhar organizados como mostrado na Figura 10. Nela é possível ver a sentença em inglês: *Are tourists enticed by these attractions threatening their very existence?* (Os turistas são seduzidos por essas atrações que ameaçam sua própria existência?), dividida em uma palavra por linha (repetidas quando fixadas mais de uma vez). A coluna FDUR traz o tempo de fixação em milissegundos.

[22] <http://www.ufrgs.br/textecc/porlexbras/corpop/index.php>

[23] Ou 51.502 segundo Barrett, Agic e Søgaard (2015), ao contar as pontuações e abreviações coladas nas palavras anteriores de acordo com a tokenização do córpus Dundee.

Figura 10 – Exemplo de um arquivo do córpus Dundee


sa01ma2p.dat

WORD	TEXT	LINE	OLEN	WLEN	XPOS	WNUM	FDUR	OBLP	WDLP	FXNO	TXFR
Are	1	1	3	3	1	1	216	1	1	1	351
tourists	1	1	8	8	6	2	156	2	2	2	3
enticed	1	1	7	7	17	3	227	4	4	3	1
enticed	1	1	7	7	19	3	174	6	6	14	1
by	-99	0	0	0	0	4	0	0	0	0	0
these	1	1	5	5	25	5	187	1	1	4	73
these	1	1	5	5	29	5	168	5	5	15	73
these	1	1	5	5	28	5	170	4	4	16	73
attractions	1	1	11	11	33	6	182	3	3	5	2
attractions	1	1	11	11	36	6	271	6	6	17	2
attractions	1	1	11	11	34	6	88	4	4	18	2
threatening	1	1	11	11	44	7	96	2	2	6	3
threatening	1	1	11	11	52	7	232	10	10	8	3
threatening	1	1	11	11	46	7	232	4	4	19	3
their	1	1	5	5	57	8	168	3	3	10	225
very	1	1	4	4	62	9	335	2	2	9	56
very	1	1	4	4	61	9	202	1	1	20	56
existence?	1	1	10	9	65	10	173	0	0	11	4
existence?	1	1	10	9	71	10	188	6	6	12	4
existence?	1	1	10	9	72	10	88	7	7	13	4
existence?	1	1	10	9	72	10	222	7	7	21	4
existence?	1	1	10	9	74	10	157	9	9	22	4
The	-99	0	0	0	0	11	0	0	0	0	0
two	1	2	3	3	5	12	314	1	1	23	43
young	-99	0	0	0	0	13	0	0	0	0	0
sea-lions	1	2	9	9	15	14	265	1	1	24	1
took	1	2	4	4	24	15	186	0	0	25	17
not	1	2	3	3	31	16	176	2	2	26	277
not	1	2	3	3	31	16	327	2	2	29	277

Fonte: Zelenina (2015).

Este córpus foi utilizado por Singh *et al.* (2016) e posteriormente por Gonzalez-Garduño e Søgaard (2017) para treinar os modelos e melhorar a acurácia na predição da complexidade das sentenças.

Potsdam Sentence Corpus (PSC)

Córpus criado por Kliegl *et al.* (2004) e expandido posteriormente pelos autores para a língua alemã (KLIEGL; NUTHMANN; ENGBERT, 2006).

Possui dados da leitura de 144 sentenças por 65 participantes, utilizando o equipamento EyeLink I (*sampling rate* de 250 Hz). Mais detalhes da sua construção podem ser vistos na Tabela 11. As sentenças foram construídas para o experimento, com o objetivo de representar uma grande variedade de estruturas gramaticais em torno de um conjunto de palavras-alvo escolhidas.

Ghent Eye-Tracking Corpus (GECO)

O *Ghent Eye-Tracking Corpus* (GECO) foi construído por Cop *et al.* (2016) a partir da leitura silenciosa do livro *The Mysterious Affair at Styles*, da autora Agatha Christie, em inglês e holandês.

Utilizaram o equipamento EyeLink 1000 para a captura (com *sampling rate* de 1 kHz), em 19 leitores bilíngues (primeira língua holandesa e segunda língua inglesa) e 14 leitores

monolíngues ingleses, todos estudantes universitários da *University of Southampton*. Todos os participantes leram aproximadamente 5.000 sentenças. O córpus está disponível publicamente[24].

O GECO foi utilizado por Gonzalez-Garduño e Søgaard (2018) e resultou em uma acurácia um pouco melhor em comparação ao modelo treinado com o córpus Dundee. Mais detalhes podem ser vistos na Tabela 11.

Provo

Criado em 2017 por Steven Luke e Kiel Christianson, o Provo é um córpus focado em normas de previsibilidade, para estudo do processo preditivo na leitura. Além das métricas de rastreamento ocular, também inclui métricas para previsibilidade morfossintática e semântica para cada palavra (LUKE; CHRISTIANSON, 2018). Está disponível para download no site da *Open Science Framework*[25]. Está dividido em duas partes, com medidas para cada palavra em 55 parágrafos (e 2.689 palavras):

- ***Predictability norms***: Métricas de previsibilidade das palavras, avaliadas por 470 participantes entre 18 e 50 anos (alunos da *Brigham Young University*);
- ***Eye-tracking data***: Métricas de rastreamento ocular de 84 participantes (falantes do inglês de forma nativa) durante a leitura dos parágrafos.

Russian Sentence Corpus (RSC)

O RSC é um córpus bem recente, construído por Laurinavichyute *et al.* (2018) para a língua russa, seguindo o modelo do córpus PSC, também com 144 sentenças.

Possui dados da leitura de 96 russos monolíngues, utilizando o equipamento EyeLink 1000 Plus (com *sampling rate* de 1 Khz). Está disponível publicamente no *Open Science Framework*[26].

Principais córpus de rastreamento ocular comparados

Também estão disponíveis dois outros córpus, para o hindi — o *Postdam-Allahabad Hindi Eyetracking Corpus* (HUSAIN; VASISHTH; SRINIVASAN, 2015) — e para o chinês — o *Beijing Sentence Corpus of Mandarin Chinese* (YAN *et al.*, 2010) — embora não tenham sido descritos na Tabela 11, que tenta mostrar as principais características dos córpus citados na seção anterior.

[24] <http://expsy.ugent.be/downloads/geco>
[25] <https://osf.io/sjefs/>
[26] <https://osf.io/x5q2r/>

Tabela 11 – Córpus de rastreamento ocular com textos e sentenças

Córpus, língua e download	**Aparato, apresentação e gravação dos movimentos dos olhos**	**Forma de apresentação do estímulo**	**Número de Participantes**	**Estatísticas do córpus, gêneros de texto e fontes**
Dundee Corpus (Kennedy e Pynte, 2003) (Kennedy et al., 2013) **Línguas:** Inglês e Francês **Download:** Obtidos com o próprio autor **Dundee Treebank:** Bitbucket[27]	Dr. Bouis Oculometer Eyetracker. Uso de barra de mordida dental de cera e descanso no queixo. **Taxa de Amostragem:** monocular direita de 1000 Hz. Ângulo visual de 0.3º. Resolução efetiva melhor do que 1 caractere. **Apresentação e gravação:** -	**Fonte:** monoespaçada de alta resolução (8x16), branco sobre preto em um monitor monocromático **Texto:** apresentado 5 linhas por vez, espaçamento duplo, linha de 80 caracteres. **Distância da tela:** 50 cm	**Rastreamento:** 10 participantes da língua inglesa e 10 da francesa **Previsibilidade:** 272 participantes, para todas as palavras, com amostragem de 25 respostas por palavra em toda sentença.	**Córpus em inglês:** 56.212 **Tokens**, 9.776 **Types** **Sentenças:** 2.368 **Gênero:** jornalístico, (editoriais de jornais) **Fonte:** *The Independent* e *Le Monde*
Potsdam Sentence Corpus (Kliegl et al., 2004) (Kliegl et al., 2006) **Língua:** alemão **Download:** uni-potsdam[28]	SR EyeLink System com descanso no queixo **Taxa de Amostragem:** binocular de 250 Hz e resolução de posição de olhos de 20 sec-arc. Ângulo visual de 0.35º **Apresentação e gravação:** Dados foram coletados em 2 labs com equipamentos e setup idênticos	**Fonte:** regular courier 12, de um monitor de resolução (832 x 632). **Sentenças:** apresentadas no centro **Distância da tela:** 60 cm	**Rastreamento:** 33 jovens adultos, universitários e 32 adultos mais velhos, de 65-83 anos **Previsibilidade:** 272 alemães nativos; total de 83 protocolos. Divisão do córpus: 116 estudantes do Ensino Médio (17-19 anos), 76 universitários (19-38 anos), 80 adultos idosos (66-80 anos).	**Palavras:** 1.138 (994 depois de excluídas as primeiras) **Tamanho de sentenças:** de 5 a 11, com média de 7,9 palavras. **Sentenças:** 144

[27] <https://bitbucket.org/lowlands/release>
[28] <http://read.psych.uni-potsdam.de/pmr2/>

The Provo Corpus (Luke e Christianson, 2017) **Língua:** Inglês americano **Download:** osf.io[29]	Eyelink 1000 Plus (SR Research, Canada) de montagem na área de trabalho e descanso no queixo. **Taxa de Amostragem:** 1000 Hz, Resolução espacial de 0.01º, Ângulo visual de 1º para 3 caracteres, Leitura binocular e gravação do movimento monocular direito. **Apresentação e gravação:** Experiment Builder (SR Research Ltd.).	**Fonte:** monitor com resolução de 1.600 x 900. **Texto:** Sentenças apresentadas como parte de um texto multilinha, apresentação randômica dos textos **Distância da tela:** 60 cm.	**Rastreamento:** 84 alunos da Brigham Young University **Previsibilidade:** 470 (265 sexo feminino, 203 masculino) alunos da Brigham Young University. Idade variou de 18 a 50 anos, média de 21 anos	**Palavras:** 2.689 **Types:** 1.197 **Tamanho de palavras:** de 1 a 15 letras (média: 4,76) **Sentenças:** 134, com 13,3 palavras em média **Número de parágrafos:** 55 parágrafos curtos (média: 50 palavras e 2,5 sentenças). **Gênero:** jornalístico, divulg. científica, literário **Fontes:** notícias onlines, artigos de divulg. científica, livros de ficção de domínio público

[29] <https://osf.io/sjefs/>

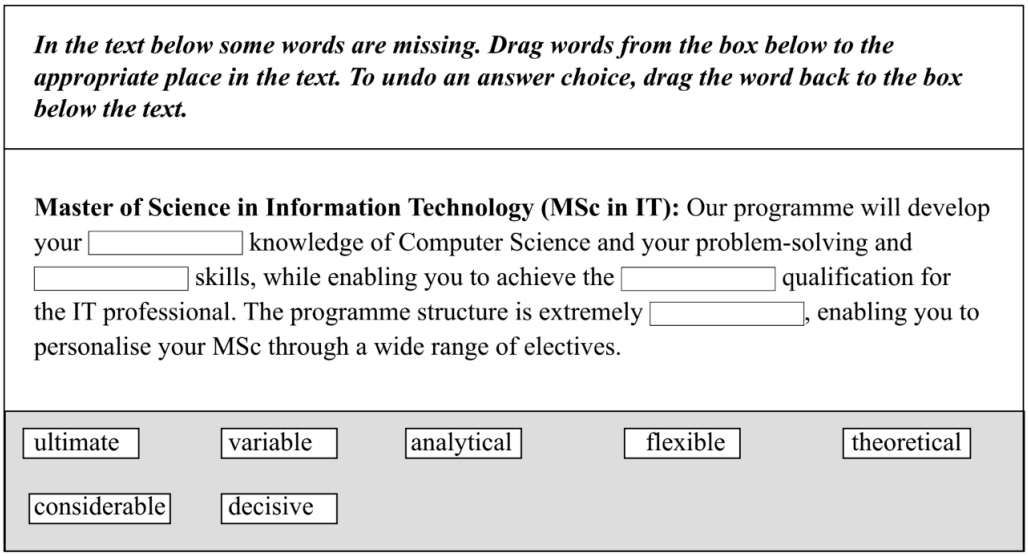
Ghent Eye-Tracking Corpus (GECO) (Cop et al., 2016) **Língua:** Inglês (monolíngue) e Inglês-holandês (bilíngue) **Download:** expsy.ugent.be[30]	Eyelink 1000 (SR Research, Canadá) de montagem na área de trabalho, com descanso no queixo. **Taxa de Amostragem:** 1000 Hz; leitura binocular e gravação do movimento monocular direito. Ângulo visual de 1º para 3 caracteres (ou 30 pixels). **Apresentação e gravação:** Experiment Builder (SR Research Ltd.).	**Fonte:** Courier New 14, em um monitor cinza claro. **Texto:** Linhas com espaçamento triplo; uso de parágrafos para amostragem, com no máximo 145 palavras em cada tela. Sala de leitura fracamente iluminada.	**Rastreamento:** 19 estudantes bilíngues (Inglês-Holandês) da Ghent University, com idade média de 21,2 anos (18–24). 14 estudantes monolíngues (língua inglesa) da University of Southampton, com idade média de 21.8 anos (18–36) Todos do curso de Psicologia. **Previsibilidade:** Não há.	**Palavras:** 59.716 (Holandês), 54.364 (Inglês) **Types:** 5.575 (Holandês), 5.012 (Inglês) **Sentenças:** 5.301 (Monolíngues); 5.190 (Holandês) e 5.300 (Inglês / Bilíngues) **Gênero:** literário **Fonte:** Romance *The Mysterious Case at Styles* de Agatha Christie.
Russian Sentence Corpus (RSC) (Laurinavichyute, et al., 2018) **Língua:** Russo **Download:** osf.io[31]	EyeLink 1000 Plus de montagem na área de trabalho e descanso no queixo. **Taxa de Amostragem:** 1000 Hz, gravação do movimento monocular direito. Ângulo visual de 0.29º **Apresentação e gravação:** Experiment Builder (SR Research Ltd.)	**Fonte:** Courier New 22 **Sentenças:** apresentadas no centro da tela, monitor 24 pol ASUS VG248QE, Resolução de (1.920 x 1.080) **Distância da tela:** 90 cm (55 cm da câmera)	**Rastreamento:** 96 russos, monolíngues (66 mulheres e 30 homens, idade de 24 anos em média (18-80) **Previsibilidade:** 750 pessoas	**Palavras:** 1.362 com exclusão da primeira e última, restando 1.074 **Sentenças:** 144 (5-13 palavras, Média = 9)

Fonte: Elaborada pelo autor.

[30] <http://expsy.ugent.be/downloads/geco>
[31] <https://osf.io/x5q2r>

Teste Cloze e normas de previsibilidade

Figura 11 – Exemplo de teste cloze com aplicação em uma interface computacional



Fonte: McCray e Brunfaut (2018).

Uma adaptação do teste Cloze foi utilizada neste trabalho para a construção do córpus RastrOS, nos mesmos moldes do trabalho de Luke e Christianson (2018), que divulgaram as normas de previsibilidade das palavras em complemento aos dados de rastreamento ocular no córpus Provo. Os autores do Provo afirmam que o método mais comum para estimar a previsibilidade de uma palavra é utilizando este teste.

O Cloze é uma técnica que apresenta ao participante um texto onde uma ou algumas palavras são omitidas e substituídas por um traço que o participante precisa preencher, por meio da escrita/digitação ou escolhendo em uma lista.

O termo Cloze vem do inglês *closure*, com o significado aproximado de "fechamento"que foi utilizado para nomear um procedimento para mensuração de inteligibilidade de texto em 1953 por Wilson Taylor, com base fundamentada na Psicologia da Gestalt e na noção estatística de amostra aleatória (ABREU *et al.*, 2017).

2.4 Abordagens de aprendizagem de máquina

A área de Inteligência Artificial (IA) começou apenas como uma área teórica, porém as abordagens e técnicas de AM começaram a ter mais destaque com o aumento da complexidade de problemas e do volume de dados que demandam tratamento computacional (FACELI *et al.*, 2011).

Geralmente as abordagens são divididas em aprendizagem **supervisionada**, com dados pré-anotados ou classes conhecidas e **não supervisionada**, onde não se conhecem as classes e os algoritmos procuram identificar padrões ou tendências relevantes em conjuntos.

Nos trabalhos dos capítulos seguintes, são utilizadas algumas formas tradicionais de implementação de tarefas de AM como classificação, regressão, clusterização e *ranking*. Classificação e regressão são exemplos de aprendizagem supervisionada e clusterização de não supervisionada. Essas principais abordagens são descritas a seguir.

Classificação

Na classificação, a entrada é um vetor de *features* $x \in \Re^d$ e a saída é um rótulo $y \in \mathscr{Y}$ representando uma classe onde $\mathscr{Y}$ é o conjunto de rótulos de classes. O objetivo do aprendizado é treinar um classificador $f(x)$ que consiga determinar o rótulo y dado um vetor de *features* x (LI, 2014; FACELI *et al.*, 2011).

Para visualizar, um exemplo simples de classificação é um algoritmo que classifica as frutas de um cesto, em maçãs, laranjas ou bananas de acordo com suas cores, formatos e pesos.

Regressão

De forma semelhante, na regressão a entrada é um vetor de *features* $x \in \Re^d$, porém a saída é um número real $y \in \Re$ e o objetivo do aprendizado é treinar uma função $f(x)$ que consiga determinar o número real y dado um vetor de *features* x (LI, 2014; FACELI *et al.*, 2011).

Um exemplo seria estimar o preço de uma fruta de acordo com seu peso, apresentação e época do ano.

Clusterização

Os métodos de clusterização ou agrupamento são parte das técnicas de aprendizagem não supervisionada e permitem analisar um grande número de métricas e dados, gerando sugestões de grupos por afinidade.

Seguindo nos exemplo anteriores, temos um cesto de frutas, mas não sabemos quais tipos de frutas estão no cesto, então é solicitado ao algoritmo que tente agrupá-las por formato, cor e peso.

Segundo Faceli *et al.* (2011) os algoritmos dessas técnicas geralmente são classificados em :

- **Baseados em centróides:** Otimizam o critério de agrupamento de forma iterativa, procurando minimizar o erro quadrático ou variação dentro do *cluster*. O algoritmo mais simples e mais utilizado é o K-Means[32].
- **Hierárquicos:** Geram uma sequência de partições aninhadas a partir de uma matriz de proximidade. Podem ser do tipo **aglomerativo**, que começa com um grupo para cada objeto e vai combinando, ou **divisivo**, que começa com um único grupo e vai dividindo sucessivamente. Um exemplo desse grupo é o AgglomerativeClustering[33].
- **Baseados em densidade**: Assumem que cada *cluster* é uma região de alta densidade de objetos, separada das demais por regiões de baixa densidade, por exemplo o algoritmo DBScan[34].
- **Baseados em grafos:** Os dados são representados em um grafo de proximidade, no qual cada nó representa um objeto e as arestas, a similaridade ou distância.
- **Baseados em redes neurais:** Sistemas paralelos compostos de unidades simples de processamento; por exemplo o algoritmo SOM (*Self-Organizing Map*).
- **Baseados em *Grid*:** Define um *grid* (reticulado) para o espaço de dados. Muito eficiente para grandes conjuntos de objetos.

Ranking

Ranking[35] é a abordagem central para diversas tarefas de Recuperação da Informação e Processamento de Línguas Naturais (LI, 2014), por exemplo: recuperação de documentos, reconhecimento de entidades, perguntas e respostas, busca personalizada, sumarização de documentos, tradução automática, etc. Em vez de esperar uma classe ou um número como saída, o objetivo do *ranking* é uma lista de resultados em uma ordem de importância que seja útil para a solução do problema.

Recentemente, uma nova área chamada *Learning to Rank* surgiu da intersecção das áreas de Aprendizado de Máquina, Recuperação de Informação e Processamento de Línguas Naturais, e consiste basicamente em fazer ranqueamento utilizando técnicas de Aprendizado de Máquina, por exemplo, RankingSVM, RankBoost, BM25, CRank, etc.

[32] https://scikit-learn.org/stable/modules/clustering.html#k-means
[33] https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering
[34] https://scikit-learn.org/stable/modules/clustering.html#dbscan
[35] Tanto *ranking* quanto o neologismo "ranqueamento"são formas aceitas nesse contexto para os trabalhos em PB.

Multi-View Learning

Multi-View Learning busca enriquecer os modelos de aprendizado utilizando uma combinação de vários conjuntos de *features* e/ou vários algoritmos para um determinado problema. É inspirada nas técnicas de *Multi-View Data* como, por exemplo, a utilização de digitais, reconhecimento facial e reconhecimento de íris para identificação de uma pessoa (XU; TAO; XU, 2013).

Desta forma, o modelo final se beneficia dos pontos fortes de cada algoritmo individual e geralmente atinge melhores resultados. Uma das primeiras abordagens consiste em treinar dois ou mais algoritmos em duas ou mais visões distintas dos dados, procurando maximizar a concordância entre eles.

SVM

As *Support Vector Machines* (SVM) são uma técnica de Aprendizado de Máquina que consegue resultados comparáveis, e em algumas aplicações até mesmo superiores, aos obtidos por outros algoritmos como as Redes Neurais Artificiais (LORENA; CARVALHO, 2007).

São embasadas na teoria do aprendizado estatístico, de Vapnik e Chervonenkis (VAPNIK; CHERVONENKIS, 1971; VAPNIK, 1995; VAPNIK, 1998), que estabelece uma série de princípios para a obtenção de classificadores com uma boa generalização.

Existem diversas implementações disponíveis da técnica; um dos principais trabalhos que aborda a tarefa para o inglês (VAJJALA; MEURERS, 2016), utilizou usou o RankSVM da ferramenta SVM$^{\text{rank}}$ (JOACHIMS, 2006).

MLP

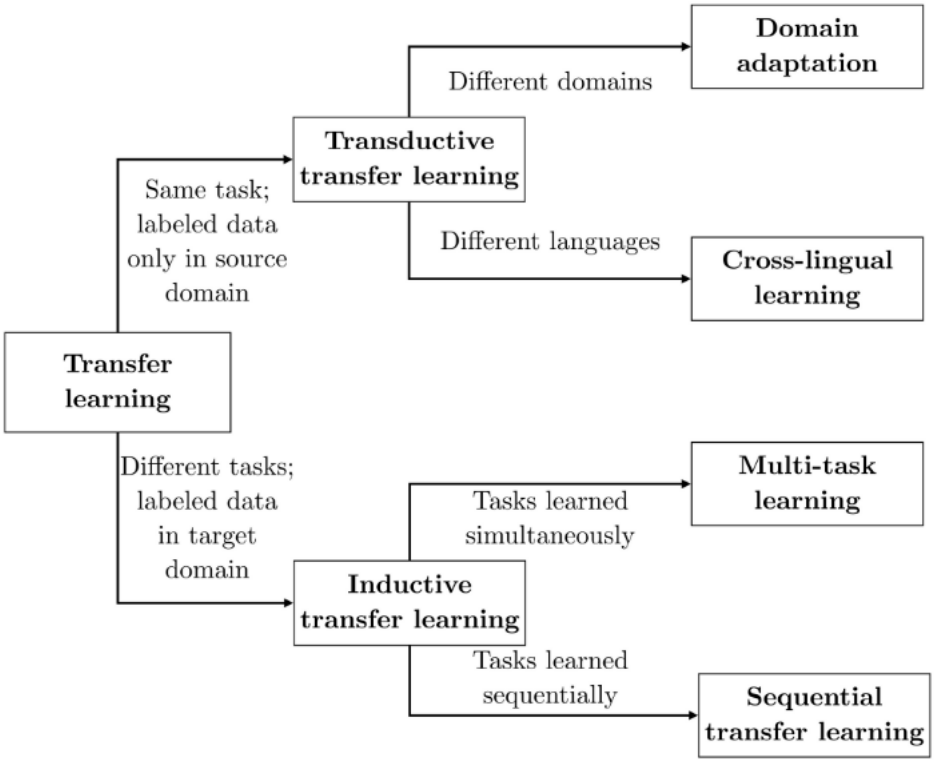
As Redes Neurais Artificiais são técnicas de Aprendizado de Máquina inspiradas no funcionamento dos neurônios naturais (RUMELHART; GROUP, 1986) e consistem numa série de funções encadeadas aplicadas sobre uma entrada de dados para obter uma ou mais saídas estimadas, ou por outro ângulo, organizadas na forma de neurônios artificiais conectados com pesos próprios e funções de ativação.

As *Multi Layer Perceptron* (MLP) são Redes Neurais com várias camadas. Possuem uma camada de entrada com um neurônio artificial para cada *feature*, uma ou mais camadas ocultas intermediárias, e uma camada de saída que retorna o valor final computado. É a técnica utilizada pelo trabalho detentor do estado da arte da tarefa para a língua inglesa (GONZALEZ-GARDUñO; SØGAARD, 2018).

Transfer Learning

Transfer Learning é uma área de AM que estuda a transferência dos pesos treinados em uma determinada distribuição para outra diferente. Tem como foco transferir modelos entre tarefas, domínios ou línguas diferentes, melhorando com isso a generalização do aprendizado (RUDER *et al.*, 2019). A Figura 12 traz uma divisão didática entre as abordagens, que podem ser indutivas ou transdutivas, além de contrastar *Multi Task Learning* e *Sequential Transfer Learning*, sendo a primeira com o treinamento simultâneo e a segunda uma tarefa por vez em sequência.

Figura 12 – Proposta de taxonomia para *Transfer Learning* em Processamento de Língua Natural



Fonte: Ruder (2019).

Multi-Task Learning

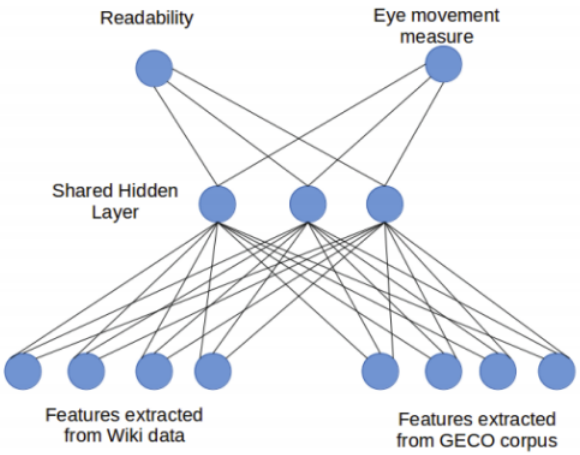
Multi-Task Learning é uma abordagem para melhorar o aprendizado de uma tarefa, utilizando informações contidas no treinamento de outras tarefas relacionadas (CARUANA, 1997). Ao treinar mais de uma tarefa ao mesmo tempo, o modelo é forçado a generalizar mais, obtendo melhores resultados nas amostras nunca vistas.

Em redes neurais isso é conseguido utilizando aprendizado em paralelo e compartilhando os pesos das camadas ocultas. Geralmente são classificadas em dois grupos, de acordo com a estratégia de compartilhamento (RUDER, 2017):

- ***Hard Parameter Sharing***: O compartilhamento forte é o método mais comum, consiste em usar os mesmos pesos (camadas ocultas) para todas as tarefas que serão treinadas, mas mantendo camadas de saída independentes;

- ***Soft Parameter Sharing***: No compartilhamento fraco, cada tarefa possui seu próprio modelo com seus próprios pesos, porém é executada uma etapa de regularização durante o treinamento que encoraja que os pesos das camadas ocultas escolhidas se mantenham similares.

Figura 13 – Exemplo de arquitetura *Multi-Task MLP*



Fonte: Gonzalez-Garduño e Søgaard (2018).

Um exemplo dessa abordagem, utilizando uma rede neural do tipo MLP de 3 camadas, pode ser visto na Figura 13, utilizada nos trabalhos de Gonzalez-Garduño e Søgaard (2017) e Gonzalez-Garduño e Søgaard (2018).

LSTM

As LSTM (*Long Short Term Memory*) (HOCHREITER; SCHMIDHUBER, 1997) são Redes Neurais do tipo recorrente (*Recurrent Neural Network* (RNN)) que possuem conexões recorrentes entre as camadas, o que permite lidar com entradas que não possuem tamanho fixo e também podem capturar relações temporais.

Elas surgiram para lidar com o problema das RNNs, conhecido como desvanecimento do gradiente (*vanishing gradient*) em que o valor do treinamento se torna muito pequeno para as saídas posteriores, perdendo a importância de relações com unidades muito anteriores. Para isso adicionam o conceito de funções *gates* internas nas células LSTM para decidir pela atualização ou esquecimento dos pesos. Essa técnica foi utilizada por um trabalho bem recente para a tarefa na língua italiana (BOSCO; PILATO; SCHICCHIA, 2018b).

LSA

Latent Semantic Analysis (LSA) é uma abordagem de PLN que analisa relações entre conjuntos de documentos e as palavras que eles contém, criando agrupamentos por tópicos de acordo com a distribuição estatística dos termos (LANDAUER *et al.*, 1997).

A análise é efetuada em duas grandes etapas, primeiro se cria uma matriz esparsa utilizando TF-IDF e depois aplica-se a técnica Singular Value Decomposition (SVD) para reduzir o número de dimensões e criar os vetores densos que serão utilizados pelos algoritmos de AM.

Word Embeddings

As *Word Embeddings* são uma aplicação relativamente recente da técnica de vetores densos para representar palavras na área de PLN. Elas são obtidas utilizando Redes Neurais para aprender automaticamente representações vetoriais e relações semânticas em grandes córpus não anotados.

As implementações mais conhecidas são Word2Vec (MIKOLOV *et al.*, 2013), Glove (PENNINGTON; SOCHER; MANNING, 2013) e FastText (BOJANOWSKI *et al.*, 2016).

Segundo Vieira e Santos (2019), para o português temos publicados três conjuntos de embeddings:

- **LX-Center Portugal**[36]: Word2Vec treinado em 1,7 bilhão de tokens;
- **NILC Embeddings**[37]: Word2Vec, FastText, Wang2Vec e Glove treinados em 1,3 bilhão de tokens;
- **PLN-PUCRS**[38]: Word2Vec e FastText treinados em 4,9 bilhão de tokens.

BERT

O *Bidirectional Encoder Representations from Transformers* (BERT) é um modelo de língua baseado em *Transformers* e mecanismos de atenção criado pelos pesquisadores da Google. Ele entrou em evidência em 2018 melhorando o estado da arte de diversas tarefas de PLN (GOOGLEBLOG, 2018; DEVLIN *et al.*, 2018).

A grande vantagem da abordagem com *Transformers* é que o processamento pode ser feito de forma paralela durante o treinamento. Isso representa um grande avanço em relação ao modelo de treinamento sequencial anterior utilizando apenas as RNN's.

O treinamento do BERT é feito em grandes córpus e envolve duas tarefas simultâneas:

[36] <http://lxcenter.di.fc.ul.pt/datasets/en/index.html>
[37] <http://nilc.icmc.usp.br/embeddings>
[38] <https://github.com/jneto04/ner-pt>

- *Masked Language Model*: Predizer lacunas (palavras substituídas por [MASK]) nas sentenças, exemplo: *"The [MASK] brown fox [MASK] over the lazy dog."*;
- *Next Sentence Prediction*: Predizer a próxima sentença inteira, dada a sentença atual. Isso permite que o BERT aprenda contexto que ultrapassa a sentença.

Uma grande diferença em relação aos modelos anteriores é que seu treinamento utiliza a sentença da esquerda para a direita e também da direita para a esquerda (por isso o bidirecional no nome).

O modelo utilizado no trabalho da Seção 4.2 foi treinado em PB e disponibilizado pelos pesquisadores da NeuralMind (NEURALMIND, 2020). Para o treinamento foi utilizado o córpus BrWaC (FILHO *et al.*, 2018), com 2,68 bilhões de palavras.

Estão disponíveis duas versões para o PB[39], uma ***base*** com 110 milhões de parâmetros e uma ***large*** com 330 milhões.

[39] <https://github.com/neuralmind-ai/portuguese-bert>

2.5 Métricas de avaliação

As métricas de avaliação mais utilizadas na tarefa de predição da complexidade sentencial são Acurácia, Precisão e Medida-F; para calculá-las é necessário primeiro criar uma Matriz de Confusão (TAN; STEINBACH; KUMAR, 2006), conforme exemplo na Tabela 12, sendo T_P = Verdadeiro Positivo (*True Positive*), F_N = Falso Negativo (*False Negative*), F_P = Falso Positivo (*False Positive*) e T_N = Verdadeiro Negativo (*True Negative*). Além delas este trabalho utiliza RMSE, Correlação de Pearson, Medida-V e Silhueta, elaboradas mais adiante.

Tabela 12 – Matriz de confusão para duas classes

Classe	Predita A	Predita B
Verdadeira A	T_P	F_N
Verdadeira B	F_P	T_N

Fonte – (TAN; STEINBACH; KUMAR, 2006)

Acurácia

É a porcentagem de amostras positivas e negativas classificadas corretamente sobre a soma de amostras positivas e negativas.

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{2.9}$$

Precisão

É a porcentagem de amostras positivas classificadas corretamente sobre o total de amostras classificadas como positivas.

$$\text{Precision} = \frac{T_P}{T_P + F_P} \tag{2.10}$$

Revocação (Recall)

É a porcentagem de amostras positivas classificadas corretamente sobre o total de amostras positivas.

$$\text{Recall} = \frac{T_P}{T_P + F_N} \tag{2.11}$$

Medida-F

F-Measure, também chamada de *F-Score* é uma medida ponderada de precisão e sensitividade (recall). É uma medida muito importante, principalmente em casos com classes desbalanceadas.

$$\mathrm{F} = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{2.12}$$

MSE e RMSE

Mean Squared Error (MSE) ou Erro Quadrático Médio é uma medida para avaliar o tamanho do erro de um regressor. Em resumo é a média das diferenças entre os valores observados y_i e os valores preditos $f(x_i)$. As diferenças são elevadas ao quadrado para evitar o efeito de cancelamento dos resultados negativos e positivos (FACELI *et al.*, 2011).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{2.13}$$

A *Root Mean Squared Error* (RMSE) ou Raiz do Erro Quadrático Médio pode ser vista como um passo adicional ao cálculo do MSE, que cancela o efeito da potência aplicada e deixa os resultados menores e mais simples de analisar. É obtida simplesmente aplicando a raiz no MSE.

$$RMSE = \sqrt{MSE} \tag{2.14}$$

Coeficiente de Correlação de Pearson

A correlação entre duas variáveis demonstra como elas estão relacionadas. O Coeficiente de Correlação de Pearson (*Pearson Correlation*), também chamado de coeficiente de correlação produto-momento, é uma das principais medidas utilizadas atualmente. Ele mede o grau da correlação entre duas variáveis e a direção (positiva ou negativa) dessa correlação (KENT.EDU, 2021).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.15}$$

Homogeneidade, Completude e Medida-V

Homogeneidade e completude são medidas para avaliação do resultado de agrupamentos em métodos de AM não supervisionados, usando análise da entropia (ROSENBERG; HIRSCHBERG, 2007). Essas medidas são mais úteis quando são conhecidas as classes dos itens, mas

caso não sejam, é possível utilizá-las para medir a concordância entre algoritmos diferentes de clusterização.

Homogeneidade mede se cada grupo contém apenas membros de uma única classe.

Completude mede se todos os membros de uma dada classe estão atribuídos a uma mesma classe.

A Medida-V (*V-Measure*) é a média harmônica entre as duas anteriores, na fórmula a seguir o valor padrão de beta é 1.

$$V = \frac{(1+beta) \times homogeneidade \times completude}{beta \times (homogeneidade + completude)} \tag{2.16}$$

Coeficiente da Silhueta

Outra medida para avaliação de agrupamentos em métodos de AM não supervisionados, o Coeficiente da Silhueta (*Silhouette Coefficient*) é especialmente útil quando não são conhecidas as classes previamente. Valores altos desse coeficiente indica um modelo com grupos melhores definidos (ROUSSEEUW, 1987). Na fórmula abaixo, **a** é a média da distância entre um item e todos os demais pontos na mesma classe e **b** é a média da distância entre um item e todos os demais pontos no grupo mais próximo.

$$S = \frac{b-a}{max(a,b)} \tag{2.17}$$

CAPÍTULO

3

NILC-METRIX: MÉTRICAS LINGUÍSTICAS E PSICOLINGUÍSTICAS

Esta pesquisa se apoiou no grande conjunto de métricas linguísticas e psicolinguísticas desenvolvidas no NILC por mais de uma década, e deixa como contribuição uma ferramenta Web com 200 dessas métricas reunidas, revisadas e documentadas com o apoio dos demais autores do artigo deste capítulo. Essas métricas foram essenciais para os métodos de AM utilizados nos Capítulos 4 e 5.

Título:	***NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese***
Autores:	**Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann e Sandra Maria Aluísio**
Ano:	**2021**
Revista:	**Language Resources and Evaluation - Special Issue: Computational approaches to Portuguese**
Situação:	**Submetido - Sob revisão editorial**

Este artigo apresenta a ferramenta NILC-Metrix, conta a história do seu desenvolvimento, disponibilizando publicamente os códigos fontes e documentação. Apresenta também exemplos de aplicações das métricas para três tarefas diferentes.

Foi submetido para a edição especial da revista LREV “Computational Approaches to Portuguese” e aguarda revisão editorial.

LREV manuscript No.
(will be inserted by the editor)

NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese

Sidney Evaldo Leal[1] · Magali Sanches Duran[1] · Carolina Evaristo Scarton[2] · Nathan Siegle Hartmann[3] · Sandra Maria Aluísio[1,]**

Received: date / Accepted: date

Abstract This paper presents and makes publicly available the NILC-Metrix, a computational system comprising 200 metrics proposed in studies on discourse, psycholinguistics, cognitive and computational linguistics, to assess textual complexity in Brazilian Portuguese (BP). These metrics are relevant for descriptive analysis and the creation of computational models and can be used to extract information from various linguistic levels of written and spoken language. The metrics in NILC-Metrix were developed during the last 13 years, starting in 2008 with Coh-Metrix-Port, a tool developed within the scope of the PorSimples project. Coh-Metrix-Port adapted some metrics to BP from the Coh-Metrix tool that computes metrics related to cohesion and coherence

Sidney Evaldo Leal
sidleal@gmail.com
https://orcid.org/0000-0002-8817-2063

Magali Sanches Duran
magali.duran@gmail.com

Carolina Evaristo Scarton
c.scarton@sheffield.ac.uk
https://orcid.org/0000-0002-0103-4072

Nathan Siegle Hartmann
nathanshartmann@gmail.com

**Corresponding author: Sandra Maria Aluísio
sandra@icmc.usp.br
https://orcid.org/0000-0001-5108-2630

[1] Instituto de Ciências Matemáticas e de Computação - University of São Paulo, São Paulo, Brazil

[2] The University of Sheffield, Sheffield, UK

[3] Itaú Unibanco, São Paulo, Brazil

of texts in English. After the end of PorSimples in 2010, new metrics were added to the initial 48 metrics of Coh-Metrix-Port. Given the large number of metrics, we present them following an organisation similar to the metrics of Coh-Metrix v3.0 to facilitate comparisons made with metrics in Portuguese and English. In this paper, we illustrate the potential of NILC-Metrix by presenting three applications: (i) a descriptive analysis of the differences between children's film subtitles and texts written for Elementary School I[1] and II (Final Years)[2]; (ii) a new predictor of textual complexity for the corpus of original and simplified texts of the PorSimples project; (iii) a complexity prediction model for school grades, using transcripts of children's story narratives told by teenagers. For each application, we evaluate which groups of metrics are more discriminative, showing their contribution for each task.

Keywords Readability · Text Complexity Metrics · Brazilian Portuguese

1 Introduction

A set of metrics called NILC-Metrix was developed both in funded projects, involving multiple researchers, and in master's and doctoral projects at the Interinstitutional Center for Computational Linguistics — NILC[3], from 2008 to 2021. The motivation for developing this large set of metrics, the phases of its development, and also re-implementations of some metrics to make the use of Natural Language Processing (NLP) tools uniform, are summarised below.

The initial motivation for building a set of metrics for automatic evaluation of textual complexity in BP started in the PorSimples project, whose theme was the Simplification of Portuguese Texts for Digital Inclusion and Accessibility (Candido et al, 2009; Aluísio and Gasperin, 2010). The target audience of PorSimples are people with low literacy, who want to obtain information from web texts but have some difficulty as they are literate at rudimentary and basic levels, according to the functional literacy indicator called INAF[4].

In many projects of the reviewed literature, automatic text simplification is implemented as a process that reduces the lexical and/or syntactic complexity of a text while trying to preserve its meaning and information (Carroll et al, 1998; Max, 2006; Shardlow, 2014). However, there are simplification projects, for example, the Terence project, in which the target audience also requires simplifications to improve the understanding of the text both at the local level, helping to establish connections between close sentences and also at the global level of the text, helping in the construction of a mental representation of the text (Arfé et al, 2014). There are still other initiatives, such as the Newsela[5] company, which perform the conceptual simplification, simplifying

[1] Comprises classes from 1st to 5th grade.

[2] Comprises classes from 6th to 9th grade, in an age group that corresponds to the transition between childhood and adolescence.

[3] http://www.nilc.icmc.usp.br/

[4] https://ipm.org.br/inaf

[5] https://newsela.com/

the content, in addition to the form (Xu et al, 2015). Newsela also includes elaborations in the text to make certain concepts more explicit or the use of redundancies to emphasise important parts of the text. In addition, operations reduce and omit information that is not suitable for a given target audience. Based on the aforementioned simplification projects, we realised that textual complexity and textual simplification are strongly associated in the NLP area. We also realised that the type of simplification used in the Terence project aims to improve the coherence of a text, which makes the authors characterise this type of simplification as being at the cognitive level. The simplification done by Newsela is the most complete in terms of different operations, although still without complete automation (but see the advances carried out by Alva-Manchego et al (2017)).

During the project PorSimples, we implemented a system called **Facilita** responsible for adapting web content for low-literacy readers by using lexical elaboration and named entity labeling (Watanabe et al, 2010), and the simplification system was called **Simplifica**. One of Simplifica's particularities was to carry out 2 levels of simplification, called natural and strong, to help people who are literate at basic and rudimentary levels, respectively. To analyse the textual complexity of the resulting text, and thus assess whether the simplification goal had been achieved, a multiclass predictor of textual complexity was built using traditional machine learning methods. This predictor required the extraction of a set of metrics that could assess the complexity of a text and compute proxies to assess the cohesion and coherence of the simplifications supported by Simplifica's automatic rules. In this scenario, the Coh-Metrix-Port (Scarton and Aluísio, 2010; Scarton et al, 2010a,b) project was created.

At NILC, we had already carried out a readability study before PorSimples aiming to adapt the Flesch Index to BP (Martins et al, 1996), based on a corpus created to help identify the weights of the linear formula that evaluates word size and size of sentences in texts of various text genres and sources. The Flesch Index (Flesch, 1948) is based on the theory that the shorter the words and sentences are, the easier a text is to be read. Although it is very practical, as it is a number indicative of the complexity of the text and can be associated with school grades, it does not inform which operations to perform in a given text to reach the sizes of short sentences, for example. In addition, it can lead us to make mistakes, because a short text is not the only characteristic of an easy-to-read text. One of the criticisms of the Flesch Index and other traditional readability formulas (Dale and Chall, 1948; Gunning, 1952; Fry, 1968; Kincaid et al, 1975) is that they are often used to adapt instructional material as prescriptive guides and not as simple predictive tools for textual complexity (Crossley et al, 2008). These mistakes derive from the failure to understand that the traditional readability formulas were not made to explain the reason for the difficulty of a text, as they are not based on theories of text understanding. Instead these formulas were based on the statistical correlation of superficial measures of a text with its level of complexity, previously established by a linguist or specialist in education, for example.

Once the limits of traditional readability formulas at the beginning of the Coh-Metrix-Port project were understood, we chose the Coh-Metrix project as a foundation for the metrics to be developed in PorSimples. Coh-Metrix computes computational cohesion and coherence metrics for written and spoken texts (Graesser et al, 2004, 2011, 2014) based on models of textual understanding and cognitive models of reading (Kintsch and Van Dijk, 1978; Kintsch and Keenan, 1973; Kintsch, 1998) that explain: (i) how a reader interacts with a text, (ii) what types of memories are involved in reading, e.g., how the overload of working memory caused by using too many words before the main verb negatively influences the processing of sentences, (iii) the role of the propositional content of the speech (Kintsch, 1998) which means that if the coherence of a text is improved, so will its comprehension (Crossley et al, 2007), and (iv) how the mechanisms of cohesion, for example, discourse markers and repetition of entities, will help to create a coherent text. In summary, just as the Coh-Metrix tool[6] for the English language does, the textual complexity analysis planned in Coh-Metrix-Port uses a framework of multilevel analysis.

Coh-Metrix-Port provided 48 metrics grouped into 10 classes. However, one of its requirements was the use of open-source NLP tools. Thus, many syntactic metrics were not implemented, given the lack of free parsers with good performance at the time. Then, the AIC tool (Automatic Analysis of the Intelligibility of the Corpus) was created (Maziero et al, 2008) within the scope of PorSimples. AIC has 39 metrics (most of them are syntactic) based on the parser Palavras (Bick, 2000) (see details in Section 2).

After the end of PorSimples, in 2010, new metrics were added to the list of the initial 48 of the Coh-Metrix-Port tool and the 39 of the AIC. This was the case of the 25 new metrics of the Coh-Metrix-Dementia (Cunha et al, 2015; Aluísio et al, 2016), developed in a master's dissertation. During the implementation of Coh-Metrix-Dementia, the first re-implementation of Coh-Metrix-Port was done to standardise interfaces and the use of NLP tools. For example, the use of nlpnet PoS tagger (Fonseca et al, 2015) was set as the default tagger, as Coh-Metrix-Dementia incorporates the Coh-Metrix-Port's 48 metrics. In 2017, during a NILC student's PhD, a large lexical base with 26,874 words in BP was automatically annotated with concreteness, age of acquisition, imageability and subjective frequency (similar to familiarity) (Santos et al, 2017), enabling the implementation of 24 psycholinguistic metrics.

The technology transfer project called *Personalisation of Reading using Automatic Complexity Classification and Textual Adaptation tools* added 72 new metrics, many of them related to lexical and syntactic simplicity, to the already extensive set of metrics built by NILC.

Finally, the RastrOS project[7] brought a new implementation to the 10 metrics based on semantic cohesion, via Latent Semantic Analysis (LSA) (Landauer et al, 1997), as well as for the calculation of lexical frequency metrics, now normalised. For the training of the LSA model with 300 dimensions, a

[6] `http://cohmetrix.com/`

[7] `https://osf.io/9jxg3/?view_only=4f47843d12694f9faf4dd8fb23464ea9`

large corpus of documents from the web, BrWaC (Wagner Filho et al, 2018), was used. This same corpus was used, together with the corpus Brasileiro[8], to calculate the lexical frequency metrics.

NILC-Metrix is, therefore, the result of various research projects developed at NILC. Its metrics were revised (some were rewritten, others discarded, several others had their NLP resources updated) and documented in detail between 2016 and 2017. This documentation is available on the project's website. The metrics can be accessed via Web interface[9] and its code is publicly available for download[10], with an AGPLv3 license. Two of the parsers used by the metrics, Palavras and LX-parser (Silva et al, 2010), need to be installed, for the correct functioning of the metrics that depend on them; Palavras is a proprietary parser; LX-parser has a license that does not allow the parser to be distributed[11].

In this paper, we present NILC-Metrix in detail and illustrate the potential of the tool with three applications of its metrics: (i) an evaluation of texts heard and read by children, showing the differences between the subtitles of films and children's series of the Leg2Kids project[12] and informative texts written for children in Elementary School I and II, compiled during the Coh-Metrix-Port and Adapt2Kids project (Hartmann and Aluísio, 2020); (ii) a new predictor of textual complexity for the corpus of original and simplified texts of the PorSimples project, comparing its results with the predictor developed in Aluisio et al (2010); and (iii) a predictor of textual complexity, using narrative transcripts from the Adole-Sendo project[13].

The remainder of this paper is organised as follows. Section 2 describes two tools developed during PorSimples that provided the basis for NILC-Metrix. Section 3 presents the metrics, grouped into 14 classes, which is very similar to the organisation of the metrics used by Coh-Metrix v3.0, to make the comparative studies easier. Section 4 presents the corpora used in the NILC-Metrix applications and also the results of the three experiments with the metrics. Section 5 carries out a review analysing studies that used sets of metrics available in NILC-Metrix, in several research areas — Natural Language Processing, Neuropsychological Language Tests, Education, Language and Eye-tracking studies. Finally, Section 6 presents some concluding remarks and suggests future work.

2 Background: Coh-Metrix-Port and AIC tool Metrics

In this section, we present details of the two tools developed in the PorSimples project: Coh-Metrix-Port and AIC. The Coh-Metrix-Port provided 48 metrics

[8] http://corpusbrasileiro.pucsp.br/

[9] http://fw.nilc.icmc.usp.br:23380/nilcmetrix

[10] https://github.com/nilc-nlp/nilcmetrix

[11] http://lxcenter.di.fc.ul.pt/tools/en/conteudo/LX-Parser_License.pdf

[12] http://www.nilc.icmc.usp.br/leg2kids/

[13] adole-sendo.info/

grouped into 10 classes, shown below with the NLP tools and resources used in their implementation:

1. Basic Counts contains 14 metrics related to basic statistics (average number of words per sentence and per paragraph, average number of letters per word, number of words and sentences in the text, and average number of syllables per content word), Flesch Index, and PoS related counts, using a model trained with the MXPOST tagger and the Nilc tagset[14];
2. Logic operators contains 5 metrics related to the counting of logical operators AND, OR, IF, Negation;
3. Content word frequencies contains 2 metrics that use the largest lexicon that existed at the beginning of PorSimples, the *Banco do Português*[15], with 700 million words. These two metrics have been maintained in the current version of NILC-Metrix, but new frequency metrics, using larger corpora, have also been included;
4. Hypernyms and Ambiguity bring a metric that calculates the average number of hypernyms per verbs in sentences using the BP Wordnet.Br v.1.0[16] and 4 metrics that calculate the impact of the number of senses (calculated based on the Electronic Thesaurus of Portuguese TeP 2.0[17]) for content words (verbs, nouns, adjectives and adverbs);
5. Tokens groups 3 metrics of lexical richness and level of formality: the well-known Type-Token Ratio and two more related to personal pronouns in phrases and text, implemented using a partial parser to identify noun phrases;
6. Constituents deal with 3 metrics related to the workload in working memory, computing modifiers within noun phrases, the number of noun phrases and the number of words before main verbs;
7. Connectives brings 9 metrics related to discursive markers that help to explain the temporal, causal, additive and logical relationships in the text, implemented based on the work of Pardo and Nunes (2006);
8. Coreferences and Anaphoras bring 7 metrics that address referential cohesion, implemented using the MXPOST tagger, a stemmer and the Unitex-PB dictionary[18].

AIC has 39 metrics, implemented mainly with information extraction from the Palavras parser (Bick, 2000) and grouped into 5 classes, which deal with Basic Counts, Syntactic Information on Clauses, Density of Morphosyntactic Categories, Personalisation, and Discourse Markers:

1. Basic Counts contains 6 metrics related to basic statistics on: number of characters, number of words and number of sentences in the text; average number of characters per word, average number of words per sentence,

[14] http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html

[15] https://www.pucsp.br/pesquisa-seleta-2011/projetos/047.php

[16] http://www.nilc.icmc.usp.br/wordnetbr/

[17] http://www.nilc.icmc.usp.br/tep2/

[18] http://www.nilc.icmc.usp.br/nilc/projects/unitex-BP/web/index.html

and number of simple words, based on the Biderman (1998) children's dictionary;
2. Syntactic Information brings 13 metrics about clause information in sentences, mainly extracted from the parser Palavras (Bick, 2000), such as: number of sentences in the passive voice, mode and average number of clauses per sentence, number of clauses, number of sentences (separated by the number of its clauses), number of clauses that start with coordinating conjunctions, number of clauses that start with subordinating conjunctions, number of coordinating conjunctions, number of subordinating conjunctions, number of verbs in the gerund, participle, infinitive and all 3 together;
3. Density of Syntactic and Morphosyntactic Categories, extracted using the parser Palavras (Bick, 2000), contains 8 metrics: number of adverbs, number of adjectives, number of prepositional objects and their average by clause and sentence, number of relative clauses, number of appositive clauses, number of adverbial adjuncts;
4. Personalisation contains 10 metrics related to the number of personal and possessive pronouns and their division by person and number;
5. Discourse Markers contains two metrics related to discursive markers, based on the work of Pardo and Nunes (2006): number of discursive markers and number of ambiguous discursive markers in the text. The latter are those that indicate more than one discourse relation. For example, in English "since" can function as either a temporal or causal connective.

3 NILC-Metrix Presentation

NILC-Metrix gathers 200 metrics developed over more than a decade for Brazilian Portuguese. The main objective of these metrics is to provide proxies to assess cohesion, coherence and textual complexity. Among other uses, NILC-Metrix may help researchers to investigate: (i) how text characteristics correlate with reading comprehension; (ii) which are the most challenging characteristics of a given text, that is, which characteristics make a text or corpus more complex; (iii) which texts have the most adequate characteristics to develop target learners' skills; and (iv) which parts of a text are disproportionately complex and should be simplified to meet a given audience. We hope that making the metrics available will stimulate new applications to validate them. For the sake of presentation, the metrics are grouped into 14 categories, following their similarity and theoretical grounds. They are: Descriptive Index, Text Easability Metrics, Referential Cohesion, LSA-Semantic Cohesion, Lexical Diversity, Connectives, Temporal Lexicon, Syntactic Complexity, Syntactic Pattern Density, Semantic Word Information, Morphosyntactic Word Information, Word Frequency, Psycholinguistic Measures and Readability Formulas.

3.1 Descriptive Index

Under this category we grouped the metrics that describe basic text statistics: number of words in the text; number of paragraphs in the text; number of sentences in the text; mean number of sentences per paragraph; mean number of syllables per content word; mean number of words per sentence; maximum number of words per sentence; minimum number of words per sentence; standard deviation of number of words per sentence; proportion of subtitles in relation to the number of sentences in the text. The length of words, sentences and paragraphs correlates with the effort required to read a text. The standard deviation of words per sentence, as well as the maximum and minimum number of words per sentence, indicate how homogeneous a text is under this parameter. A large standard deviation is suggestive of large variations in terms of the number of words per sentence. If a text has many subtitles, this may affect the standard deviation. These metrics do not require sophisticated resources to be processed: it is sufficient to have a tokeniser and sentence segmentation that recognise tokens, sentences and paragraph boundaries.

3.2 Text Easability Metrics

This category brings together the metrics that measure how easy a text is. There are four measures that calculate the proportion of short, medium, long and very long sentences in relation to all sentences in the text (the four add up to 100%). The classification of sentences according to their length is based on the following parameters: up to 11 words = short; between 11 and 12 = medium; between 12 and 15 = long; above 15 = very long. Two other metrics of text easability accounts for the proportion of easy and difficult conjunctions to total words. The classification of conjunctions according to their easability is based on an informed lexicon. Another metric of text easability is the proportion of first-person personal pronouns in relation to all personal pronouns in the texts. First-person personal pronouns indicate proximity to the reader. Finally, the dictionary of Simple Words by Biderman (1998) and a list of 909 concrete words from Janczura et al (2007) provided the lexicon used to calculate the proportion of simple content words to all content words in the text. Content words (nouns, verbs, adjectives and adverbs) constitute the variable vocabulary a reader has to know to understand the text (they oppose to function words, such as determiners, conjunctions, prepositions, numbers and pronouns, which do not point to extra linguistic referents). The greater the proportion, the simpler the text.

3.3 Referential Cohesion

There are nine metrics in this category and they capture the presence of elements necessary to construct coreference chains. These metrics calculate the

overlap of content words in adjacent sentences and among all sentences of the text. Stem overlap is also calculated (such as in abolish-abolition). The longer the text, the greater the need of coreference chains to help the reader to make connections between parts of the text, rendering the text easier to understand.

3.4 LSA-Semantic Cohesion

The metrics that calculate semantic cohesion are grounded in Latent Semantic Analysis (LSA)[19] (Landauer et al, 1997), which considers the overlap of semantically related words. Co-occurrence is the basis to capture semantic relations. LSA uses Singular Value Decomposition (SVD) to reduce the complex matrix of words co-occurrences in a document to approximately 100-500 functional dimensions. Therefore, by representing the similarity of words in a vector space and computing the cosine of the angle between vectors of pairs of words, one can represent greater similarity with high cosines. The LSA model for NILC-Metrix was trained on BrWaC[20], with 300 dimensions. BrWac is the largest Brazilian corpus publicly available today (53 million documents, 2.68 billion words, and 5.79 million unique forms).

NILC-Metrix has eleven metrics of semantic cohesion. Six of them calculate the mean and the standard deviation of semantic overlap between: adjacent sentences, adjacent paragraphs and all sentence pairs in the text. The language model is also used to calculate the mean and the standard deviation of givenness (previous given information) and span (Hu et al, 2003) (an alternative and better method to capture given information) in the current sentence. Finally, the cross-entropy calculates the mean difference of the probability distribution of sentence pairs in the language model.

3.5 Psycholinguistic Measures

NILC-Metrix brings six indices for each of the following psycholinguistic measures: age of acquisition, concreteness, familiarity and imageability, totalling 24 metrics. These measures are related to text easability: the lower the words' age of acquisition, the easier the text, and the higher the words' concreteness, familiarity and imageability, the easier the text. The lexical resource used by these metrics (Santos et al, 2017) contains 26,874 words (content words), therefore if a word of the text is not included in the resource, these metrics are affected.

3.6 Lexical Diversity

Lexical diversity is a measure obtained through the type-token ratio (TTR), that is, the number of types (all words, disregarding repetitions) divided by the

[19] `http://lsa.colorado.edu/`

[20] `https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC`

number of tokens (all words, considering repetitions). Lexical diversity is inversely proportional to cohesion: the lower the lexical diversity, the higher the cohesion. As explained by McNamara et al (2014) *TTR is correlated with text length because as the number of word tokens increases, there is a lower likelihood of those words being unique.* NILC-Metrix includes TTR for: all words, content words, function words, nouns, verbs, adjectives, pronouns, indefinite pronouns, relative pronouns, prepositions and punctuation. Again, the detailed metrics are intended to investigate where the difficulty of the text lies.

3.7 Connectives

Connectives are words that help the reader to establish cohesive links between parts of the text. NILC-Metrix provides metrics for the proportion of all connectives in the text, as well as for the proportion of four different types of connectives: additive, causal, logical and temporal. Temporal connectives, however, are within the temporal lexicon category. For each type, there is a distinct metric specifying the positive and negative ones. Besides that, the most frequent connectives, "e" (and), "ou" (or) and "se" (if) are focused on specific metrics.

3.8 Temporal Lexicon

The eleven indices gathered in this item detail the relative occurrences of each verb tense and mood in relation to the total verb tenses and moods in the text. Temporal connectives, positives and negatives, are also included in this category. The temporal lexicon is the first step towards enabling the construction of temporal cohesion metrics.

3.9 Syntactic Complexity

NILC-Metrix contains a series of metrics using syntactic analyses, both from dependencies and constituent parsers. Some of them focus on syntax characteristics associated to the demand on working memory, as the number of words before the main verb. Using data from dependence trees, there are three metrics: distance in the dependency tree and two syntactic complexity indexes: Yngve (Yngve, 1960) and Frazier (Frazier, 1985). Yngve's index is based on the premise that the syntactic trees tend to branch to the right, and that deviations from this pattern correspond to greater complexity in the language.

Frazier proposed a bottom-up approach, starting from the word and moving up the syntactic tree until it finds a node that is not the leftmost child of its parent. Each node in the tree receives a score of 1, and the nodes that are children of nodes of sentence type, 1.5. The score of each word is given by the sum of the scores of the nodes belonging to its branch.

In addition, the category of syntactic complexity brings various proportion measures involving clauses, enabling an in-depth investigation on where the complexity of a text lies: clauses with postponed subject; clauses in non-canonical order (canonical order is SVO: subject-verb-object), clauses in passive voice, infinite verb clauses, subordinate clauses, relative clauses, adverbial clauses, etc.

3.10 Syntactic Pattern Density

In this category, there are four metrics correlated with text processing difficulty: gerund clauses, mean number of words per noun phrase, maximum and minimum number of words per noun phrase.

3.11 Morphosyntactic Word Information

In this category, one can find the traditional measures of content and functional word densities, in the text and per sentence, as well as a series of break-downs of these densities: adjectives, adverbs, verbs (inflected and non-inflected), nouns, prepositions, pronouns (detailed by type and inflection). Altogether there are 42 metrics that, although they do not individually give a measure of complexity, may be useful to investigate in detail where the difficulty of a text lies.

3.12 Semantic Word Information

This category has eleven metrics. Two of them use Brazilian Portuguese LIWC 2007 Dictionary[21] to calculate the proportion of words with negative/positive polarity in relation to all words in the text. Five measures of ambiguity (of content words, and in detail by nouns, adjectives, verbs and adverbs) are calculated according to their respective number of senses in TeP (Portuguese Electronic Thesaurus). The average amount of hypernyms per verb in sentences uses information extracted from Wordnet.Br. Finally, there are three metrics relating to the proportion of abstract nouns and proper nouns in sentences and in the text.

3.13 Word Frequency

This category presents ten frequency measures. The two oldest present frequencies (not normalised) of all content words and of the rarer words in the text. They were extracted from Corpus do Português, which was the largest corpus at that time, with 700 thousand words. More recently, four frequency

[21] `http://143.107.183.175:21380/portlex/index.php/en/liwc`

measures were extracted from Corpus Brasileiro (Sardinha, 2004), which has around one billion tokens and four from BrWaC, which has around 2.68 billion tokens (Wagner Filho et al, 2018). The four measures are the same for the two corpora: average frequency of content words and rare content words; average frequency of all words and all rare words. The resulting eight measures were first normalised using fpm (frequency per million) and then normalised using the zipf logarithm scale. The difference between the two corpora is that Corpus Brasileiro assigned the PoS tags to the words out of context and for BrWaC, we assigned the PoS tags in context.

3.14 Readability Formulas

This category gathers five classic formulas used to assess text readability:

The Brunet readability index (Thomas et al, 2005) is a kind of type/token ratio that is less sensitive to the text length. It raises the number of types to the constant -0.165 and then raise the number of tokens to the result.

The Dale Chall adapted formula (Dale and Chall, 1948) combines the percentage of unfamiliar words with the average number of words per sentence. Unfamiliar words are those not included in the Dictionary of Simple Words (Biderman, 1998). The calculus is: (0.1579 * percentage of unfamiliar words) + (0.0496 * average amount of words per sentence) + 3.6365.

The Flesch readability index (Kincaid et al, 1975) looks for a correlation between average word and sentence lengths. The formula after adaption is: 248.835 - [1.015 x (average words per sentence)] - [84.6 x (average syllables per word)].

Gunning's Fog index[22] adds the average sentence length to the percentage of difficult words and multiplies this by 0.4. Difficult words are those with more than two syllables. The result is directly related to the 12 American grade levels.

Honore's Statistics (Thomas et al, 2005) is a type/token ratio that takes into account, besides the number of types and tokens, the number of hapax legomena, that is, types that have only one token in the text.

4 NILC-Metrix Applications

In this section, we present three applications of NILC-Metrix metrics. Section 4.2 provides a comparison of texts heard and read by children, showing the differences between the legends of children's films and series from the Leg2Kids project and informational genre texts written for children in Elementary School I and II, compiled during the Coh-Metrix-Port and Adapt2Kids projects. Section 4.3 presents a new predictor of textual complexity for the corpus of original and simplified texts of the PorSimples project, comparing the

[22] `https://core.ac.uk/reader/77238827`

results of the trained model with the 200 metrics of Nilc-Metrix with the predictor developed in Aluisio et al (2010), retrained with 38 metrics developed in the Coh-Metrix-Port project. Section 4.4 presents a predictor of textual complexity using transcripts of narratives from the Adole-sendo project to predict school grades. Section 4.1 describes the corpora used in the three experiments.

4.1 Corpora used in the experiments

4.1.1 Transcribed Legends of the Leg2Kids and Nonfiction Texts for Early School Years of the Adapt2kid projects

The Leg2Kids corpus comprises 36,413 subtitles of films and a series of the genres Family and Animation in Brazilian Portuguese, made available by Open Subtitles[23] in 2019. The corpus was preprocessed to remove the existing time stamps in each subtitle (these markers define the time interval in which a subtitle will be displayed on the screen). Markings from the subtitle editors, such as web page addresses, acknowledgments, sponsorship, among others, were also removed. The corpus was then sentenced and tokenised by the NLTK[24] tool. Leg2Kids contains a total of 153,791,083 *tokens* and 452,312 *types*, and a *type-token ratio* (TTR) of 0.29%. This TTR value implies greater lexical richness than SUBTLEX-PT-BR (Tang, 2012) (0.22% TTR), a similar subtitle corpus in BP.

In order to build the Adapt2kids corpus for research on textual simplification for children (Hartmann and Aluísio, 2020), we took advantage of some corpus already compiled in the PorSimples project, such as *Ciência Hoje das Crianças* (CHC)[25], *Folhinha*[26], *Para Seu Filho Ler*[27]. To enlarge this corpus created during PorSimples, we selected the following sources: SARESP tests[28] and textbooks for specific grades. SARESP tests are generally administered once a year; the test contains several textual genres – that is, there are few informative texts. We obtained only 72 texts, distributed in five grades, from SARESP tests. Regarding textbooks, we selected 178 informative texts from textbooks about the Portuguese language written in Portuguese. Because of the small amount of texts which had information about grade level, new sources were included in the corpus: NILC corpus[29] and the magazine *Mundo Estranho*[30], which contains 7,645 texts. The source distribution of Adapt2Kids corpus is shown in Table 1.

[23] https://www.opensubtitles.org

[24] https://www.nltk.org/

[25] http://chc.org.br/

[26] http://www.folha.uol.com.br/folhinha

[27] https://zh.clicrbs.com.br/rs

[28] https://sites.google.com/site/provassaresp

[29] http://nilc.icmc.usp.br/nilc/images/download/corpusNilc.zip

[30] http://mundoestranho.abril.com.br

Textbooks	NILC corpus	SARESP tests	*Ciência Hoje das Crianças*	*Folhinha* Issue of Folha de São Paulo	*Para seu Filho Ler* Issue of Zero Hora	*Mundo Estranho*
492	262	72	2.589	308	166	3.756

Table 1: Distribution of Adapt2Kids texts by source.

From these 2 large corpora, we selected 2 samples with the same number of texts (see Table 2) by: (i) selecting Adapt2Kids texts whose number of tokens is greater than 100, totalling 7,136 texts; (ii) selecting 7,136 texts of Leg2Kids longer than 600 tokens. Leg2Kids has a *type-token ratio* (TTR) of 0.29%, but the sample selected of this corpus has a TTR of 0.012%. The sample selected of Adapt2Kid has a TTR of 0.04% implying greater lexical richness than Leg2Kids' sample but less lexical richness than Escolex (Soares et al, 2014) (1.5 % TTR), which comprises 171 textbooks in European Portuguese for children attending the 1st to 6th grades (6- to 11-year old children) in the Portuguese education system.

Corpus	**Texts**	**Sent**	**ASL**	**Types**	**Tokens**	**TTR**
Leg2Kids	7,136	2,170,971	6.18	148,004	11,972,556	0.012
Adapt2Kids	7,136	133,685	17.37	85,063	2,148,929	0.04

Table 2: Description of samples of Leg2Kids and Adapt2kids corpora

4.1.2 Original and Simplified Texts of the PorSimples Project

The PorSimples project has 154 original texts, considered complex for the target public, which were manually simplified on 2 levels, called natural simplification and strong simplification (see Table 3). The result of the process is a parallel corpus with 462 texts. These two types of simplifications were proposed to attend the needs of people with different levels of literacy.

Level	**Texts**	**Sent**	**ASL**	**Types**	**Tokens**	**TTR**
Original	154	2960	19.99	11,106	57,237	0.19
Natural	154	4078	15.76	9,792	59,420	0.17
Strong	154	4918	12.76	9,647	60,760	0.16
Total	462	11,956	16.17	12,053	177,417	0.06

Table 3: Description of PorSimples corpus

In PorSimples, the human annotator was free to choose which operations to use when performing a natural simplification, among the ones available, and when to use them. The annotator could decide not to simplify a sentence, for example. Strong simplification, on the other hand, was driven by explicit rules from a manual of syntactic simplification also developed in the project,

which states when and how to apply the simplification operations. The simplifications were supported by an Annotation Editor (Caseli et al, 2009). The Annotation Editor has two modes to assist the human annotator: a Lexical and a Syntactical mode. In the Lexical mode, the editor proposes changes in words and discourse markers by simpler and/or more frequent ones, using two linguistic resources: (1) a list of simple words extracted from Biderman (1998) and a list of concrete words from Janczura et al (2007) and (2) a list of discourse markers extracted from the work developed by Pardo and Nunes (2006). The Syntactical mode has 10 syntactic operations based on syntactic information provided by the parser Palavras (Bick, 2000). The syntactic operations, which are accessible via a pop-up menu, are the following: (1) non simplification; (2) simple or (3) strong rewriting; (4) putting the sentence in its canonical order (subject-verb-object); (5) putting the sentence in the active voice; (6) inverting the clause ordering; (7) splitting or (8) joining sentences; (9) dropping the sentence or (10) dropping parts of the sentence.

4.1.3 Transcribed Narratives of the Adole-sendo Project

Adole-sendo is a project being developed at the Federal University of São Paulo (UNIFESP) that aims to assess biopsychosocial factors that affect the development of teenage (from 9 to 15 years old) behavior according to biological maturation measures. Here, we use only chronological age and related grades to train a complexity predictor of the narratives the teenagers produced setting a baseline for the Adole-sendo project. Currently, there are data collected from 271 participants, according to the distribution shown in Table 4.

Stages of Education	**Grade**	**Texts**	**Sent**	**ASL**	**Types**	**Tokens**	**TTR**
Elementary	4th	9	188	16.89	572	2,844	0.20
School I	5th	34	749	16.38	1,089	11,026	0.10
Elementary	6th	70	1,234	20.15	1,502	22,137	0.07
School II	7th	43	973	20.13	1,368	16,090	0.09
	8th	15	323	23.13	791	5,724	0.14
	9th	59	718	25.15	1,323	15,204	0.09
High School	1st	41	603	26.58	1,271	12,615	0.10
Total		271	4,788	21.80	3,129	85,640	0.04

Table 4: Description of Adole-sendo corpus

The data for this project comprises transcribed narratives obtained from the task of telling children's stories from memory (referred to as retelling herein) for each adolescent. There are two stories used by participants in the collection: Jack and the Beanstalk and Little Red Riding Hood. The participant chooses one of the two.

The process of creating the corpus of the Adole-Sendo project included three steps: (i) transcription of the retelling audios and annotation of six linguistic phenomena at the word level; (ii) linguistic annotation of five types of

disfluencies; (iii) automatic generation of narratives without the disfluencies and the phenomena annotated in the transcripts. These stages are summarised below.

Transcription of Children's Story Retelling Audios. The transcripts were obtained from retelling audios recorded during interviews with the participants at their own schools in the city of São Paulo, Brazil, (through convenience sampling), following the standard Portuguese spelling and sentence segmentation according to the rules of the language, based on prosodic and syntactic rules for written texts. The transcripts also included syntactic errors and filled pauses. Pronunciation variations were not transcribed. Linguistic phenomena at the word level, such as intentional word repetition, three types of pauses, filled pauses and interrupted words were annotated during transcription.

Linguistic Annotation of Disfluencies. Once the transcription stage was completed, 237 texts were imported into the web platform Inception[31] for annotating five phenomena of disfluencies, described in an annotation manual: (i) Discourse markers; (ii) Comments not related to the story were annotated; (iii) Repetitions of unintended words; (iv) Self-corrections; and (v) Filled pauses.

After a pilot annotation, three pairs of annotators each received a third of the transcriptions. The annotation comprises two tasks: (i) finding and correctly segmenting the phenomenon and (ii) correctly labeling the delimited phenomenon. We used these two agreement measures as Cohen's Kappa only analyses chunks that the pair has segmented equally and excludes incomplete or missing annotations. The annotation obtained had high Kappas but the pairs had problems in finding the phenomena and segmenting them correctly. For this reason, we analysed the Alpha measure which considered the character-by-character annotation of the narrative, penalising cases in which the annotator exceeded one character (or more characters) or forgot to annotate a given phenomenon. In the annotation made, Alphas were negative for all pairs in the performed annotation, indicating that the annotators had difficulties in following the annotation manual or that they adopted different rules, consistently. The Alpha measure showed that the manual should be revised and agreed upon between the pairs of annotators to proceed with the curation stage to produce the gold standard annotation. The corpus was curated by the pair's most experienced annotator and the most difficult cases were discussed among all annotators.

Automatic Generation of Narratives without Disfluencies. The linguistic annotations were computed to be used in the analysis of narratives, and a modified narrative — without the six annotations taken during the transcription and without the five annotations of disfluencies — was generated automatically. Afterwards, 34 new transcripts were annotated by an experienced annotator, totalling 271 narratives in the corpus.

[31] `https://inception-project.github.io/`

4.2 Comparison between spoken and written texts targeting children

Written and spoken language are distinctive in nature. Whilst written texts are usually self-contained and well-structure, spoken language can use extra knowledge (information from pragmatics) in order to be unambiguous. The question we aim to answer in this section is whether or not we can quantify the differences between the two modalities by using metrics from NILC-Metrix. As in the original Coh-Metrix, NILC-Metrix's metrics can capture different aspects of textual complexity that we will use to compare written (Adapt2Kids) and spoken (Leg2Kids) language. Our experiments are similar to the work of Louwerse et al (2004) that uses Coh-Metrix to compare written and spoken language in English.

In order to perform the analysis, we use the Welch's t-test (Welch, 1947): an extension of the t-test for distributions with unequal variances. We consider a statistically significant difference between the means of the distributions when the $p-$value is smaller than 0.001. Each metric is analysed in isolation and we discard metrics that cannot be applied to our texts (for example, paragraph-related metrics). Our analysis is divided into the 14 categories of NILC-Metrix.

4.2.1 Descriptive Index

Although these are the most basic metrics in NILC-Metrix, they already provide some insight into the main differences between the two corpora. For instance, as expected, Leg2Kids subtitles have, on average, a significantly higher number of words (1,638.26) and sentences (292.06) than texts in Adapt2Kids (300.28 and 18.75, respectively). On the other hand, the word per sentence ratio is smaller in Leg2Kids (6.38) than in Adapt2Kids (17.17). In addition, the standard deviation of the sentence length is significantly higher in Adapt2Kids (9.57) than in Leg2Kids (5.11). This analysis highlights some of the main characteristics of Leg2Kids: subtitles consist of longer texts in terms of sentences, although they have less words per sentence (which is expected for dialogues, mainly subtitles with screen display constraints). Interestingly, there are no significant differences in terms of the maximum sentence length, i.e., on average, the longest sentences are similar in both corpora (around 37 words). Table 5 shows the metrics with significantly higher values in Adapt2Kids than Leg2Kids (first row) and vice-versa (second row).

Adapt2Kids	syllables per content word, words per sentence, min sentence length, sentence length standard deviation
Leg2Kids	number of sentences, number of words

Table 5: Descriptive Index: first line means higher values in Adapt2Kids and second line means higher values in Leg2Kids.

4.2.2 Text Easability Metrics

All metrics from this category show significant differences. In particular, Leg2Kids (0.69) has a higher ratio of personal pronouns than Adapt2Kids (0.23), which is expected in the dialogue modality. Long sentences are more frequent in Adapt2Kids (ratio=0.51) than Leg2Kids (ratio=0.04), which is also a characteristic of subtitles. Finally, Leg2Kids has a higher ratio of simple words (0.76) than Adapt2Kids (0.74). Table 6 summarises the metrics with significantly higher values in Adapt2Kids than Leg2Kids (first row) and vice-versa (second row).

Adapt2Kids	ratio of easy(hard) conjunctions, ratio of very long(long/medium) sentences
Leg2Kids	ratio of personal pronouns, ratio of short sentences, ratio of simple words

Table 6: Text Easability Metrics: first line means higher values in Adapt2Kids and second line means higher values in Leg2Kids.

4.2.3 Referential Cohesion

Three metrics in this category do not show statistically significant differences: proportion of adjacent references, argument overlap and mean of co-referent pronouns. For both corpora, the values of the metrics are low, suggesting they are both not complex. All other metrics show higher values in Adapt2Kids than Leg2Kids, indicating that written texts have more ambiguous pronouns, although they also have lexical repetition, which is a characteristic of simple texts. On the other hand, in dialogue, pronouns are usually easily solved, as most of them address the interlocutor/speaker. In addition, dialogue is dependent on extra-textual context and elements from pragmatics that may impact their intelligibility. Therefore, we cannot clearly conclude that written texts are simpler than subtitles. Instead, our analysis shows that written texts use more artifacts of referential cohesion than dialogues in subtitles.

4.2.4 LSA-Semantic Cohesion

In general, Adapt2Kids shows higher values of LSA-Semantic than Leg2Kids, which may suggest that the written texts present high semantic similarity among their sentences. This is not surprising, given that in Leg2Kids scenes they are not explicitly identified and, therefore, changes in topics may occur with higher frequency than in written texts. One exception is the LSA metric measured using the sentence span, where Leg2Kids shows, on average, a higher value (0.93) than Adapt2Kids (0.86). Differently from simply calculating the cosine similarity between a sentence vector and the average vector of its predecessors, in the span case the previous sentences are used to form a

vector sub-space. The current sentence vector is decomposed in two components: one in the previous sentence sub-space and another perpendicular to this sub-space. The similarity score is then calculated between these two components and it is expected to measure similarity beyond the explicit content presented in previous sentences. This is an interesting result, suggesting that LSA semantic cohesion in subtitles needs to be measured using context beyond explicit clues. Table 7 summarises the metrics with significantly higher values in Adapt2Kids than Leg2Kids (first row) and vice-versa (second row).

Adapt2Kids	LSA adjacent (all/givenness) mean, LSA adjacent (all/givenness) standard deviation
Leg2Kids	LSA span mean, LSA span standard deviation, cross entropy

Table 7: LSA-Semantic Cohesion: first line means higher values in Adapt2Kids and second line means higher values in Leg2Kids.

4.2.5 Psycholinguistic Measures

Concreteness Adapt2Kids has, on average, a higher concreteness score than Leg-2Kids (4.30 and 4.08, respectively). This happens mainly because Adapt2Kids has a high ratio of concreteness of words with scores between 4 and 5.5, whilst Leg2Kids has a high ratio of concreteness of words with scores between 2.5 and 4. Therefore, written texts in Adapt2Kids use significantly more concrete words than spoken language in Leg2Kids.

Familiarity In terms of familiarity, Leg2Kids has a higher average score than Adapt2Kids (5.12 and 4.84, respectively). Leg2Kids shows a significantly higher ratio of words with familiarity scores between 5.5 and 7, whilst the highest ratio for Adapt2Kids for words with familiarity scores is between 4 and 5.5. Contrary to the concreteness results, Leg2Kids subtitles have significantly more familiar words than Adapt2Kids.

Age of Acquisition Adapt2Kids has a higher mean value of age of acquisition score (4.54) than Leg2Kids (3.72), suggesting that the subtitles are more accessible for younger ages. This happens because Leg2Kids has a high number of words with age of acquisition scores below 4, while most words in Adapt2Kids have scores higher than 4.

Imageability Leg2Kids and Adapt2Kids are not significantly different in terms of imageability. Although Leg2Kids shows a significantly higher value of words with scores between 4 and 5.5 (0.69) than Adapt2Kids (0.65); the absolute difference is rather small to draw any conclusions. Similarly, on the range of scores between 2.5 and 4, Adapt2Kids shows a significantly higher score than Leg2Kids, but the absolute difference is also very small (0.25 versus 0.23).

4.2.6 Lexical Diversity

Most metrics in the lexical diversity category show significantly higher values in the Adapt2Kids texts. This indicates a higher complexity for written texts than subtitles. For instance, the type-token ratio scores for Adapt2Kids were 0.75, whilst Leg2Kids scored 0.74. In terms of content word diversity, Adapt2Kids also showed a higher score (0.84) than Leg2Kids (0.79). The exceptions where content density (that measures the proportion of content words in relation to functional words) and maximum proportion of content words (that shows the proportion of content words in the most complex sentence of a document). For these two metrics, Leg2Kids shows significantly higher values (1.74 for content density and 0.84 for maximum proportion content words) than Adapt2Kids (1.48 and 0.73 for content density and maximum proportion of content words, respectively).

4.2.7 Connectives and Temporal Lexicon

Except for the ratio of positive causal connectives and ratio of negative logical connectives, all other metrics showed statistically significant differences between both corpora. However, most values in this category are considerably small (all ratio values are below 0.1), suggesting that connectives are scarce in both written and spoken language. Adapt2Kids shows the highest ratio of all connectives (0.09 vs 0.08), whilst Leg2Kids has the highest ratio of negations (0.03 vs 0.01).

In the Temporal Lexicon category, the only metric that does not show statistically significant differences is the ratio of positive temporal connectives. Similar to connectives, most values are below 0.1 and Adapt2Kids has the highest values for the majority of metrics, indicating that written texts make more use of this type of connectives. Adapt2Kids also has the highest proportion of verbs in the present tense, suggesting that written texts have more frequent verb inflexions (0.57 vs. 0.22). On the other hand, Adapt2Kids has the highest proportion of auxiliary verbs followed by a verb in the past participle tense (0.14 vs. 0.01), which is a sign of higher complexity, but may also indicate that written texts are more formal than subtitles. Finally, Adapt2Kids also shows a higher proportion of different verb tenses (4.38) than Leg2Kids (3.61), which may also be capturing the characteristic of narrativity in subtitles, which implies in using the past tense frequently (Graesser et al, 2014).

4.2.8 Syntactic Complexity and Syntactic Pattern Density

In general, Adapt2Kids shows higher syntactic complexity than Leg2Kids with a significant difference. The only exception is the metric measuring the distance in a parse tree, which did not show any statistically significant differences in the results. Table 8 shows a selection of metrics in this category and their mean values for both Adapt2Kids and Leg2Kids (higher values mean higher syntactic complexity). These results highlight that subtitles (and dialogue in general) use

simplified syntax. However, it is worth mentioning that, in Leg2Kids, sentences were automatically devised, as the subtitles were divided into the frames they appear in the screen. Therefore, more investigation is needed to draw further conclusions.

	Adapt2Kids	Leg2Kids
words before main verb	1.51	0.80
adverbs before main verb	0.26	0.09
clauses per sentence	2.35	0.46
coordinate conjunctions per clauses	0.04	0.23
frazier	7.06	5.99
proportion of non-SVO clauses	0.33	0.11
proportion of relative clauses	0.13	0.02
proportion of subordinate clauses	0.44	0.11
yngve	2.48	1.60

Table 8: Results for selected syntactic complexity metrics.

Similarly to the results in the Syntactic Complexity category, Adapt2Kids also shows the highest values for Syntactic Pattern Density metrics. For instance, the mean size of noun phrases is significantly higher in Adapt2Kids (4.91) than in Leg2Kids (2.11).

4.2.9 Morphosyntactic Word Information, Semantic Word Information and Word Frequency

All metrics in the Morphosyntactic Word Information category show statistically significant differences. Leg2Kids has the highest proportion of content words (0.62 vs. 0.59), while Adapt2Kids shows the highest proportion of functional words (0.41 vs. 0.38). Adapt2Kids has the highest noun (0.33 vs. 0.25) and adverb (0.77 vs. 0.37) ratios, whilst the ratio of pronouns (0.15 vs. 0.08) and verbs (0.24 vs. 0.16) are highest in Leg2Kids. Adapt2Kids also has the highest values for the proportion of infinitive verbs (0.18 vs. 0.07), inflected verbs (0.61 vs. 0.27) and non-inflected verbs (0.34 vs. 0.10). The ratio of prepositions per clause and per sentence is considerably higher in Adapt2Kids (1.35 and 2.73, respectively) than in Leg2Kids (0.17 and 0.21 respectively). The proportion of relative pronouns is also higher in Adapt2Kids (0.27) than in Leg2Kids (0.03). Finally, whilst the proportion of third person pronouns is the highest in Adapt2Kids (0.57 vs. 0.30), Leg2Kids shows the highest values for the proportions of second (0.32 vs. 0.2) and first person (0.37 vs. 0.05) pronouns.

In the Semantic Word Information category, the only metric that does not show statistically significant differences is the proportion of negative words. Leg2Kids shows the highest values for metrics measuring the ambiguity of adjectives (5.01 vs. 3.60), nouns (2.49 vs. 2.29), verbs (10.95 vs. 9.75) and content words (6.17 vs. 4.47). The mean value of verb hypernyms and the proportion of positive words are higher in Adapt2Kids (0.56 and 0.39, respectively) than in Leg2Kids (0.38 and 0.34, respectively).

Finally, in the Word Frequency category, all metrics show statistically significant differences. The log of the mean frequency values for content words extracted from Corpus Brasileiro and BrWac are slightly higher in Leg2Kids (4.53 and 4.43, respectively) than in Adapt2Kids (4.51 and 4.28, respectively). When considering all words for the same metrics, Adapt2Kids shows slightly higher values than Leg2Kids.

4.2.10 Readability Formulas

Table 9 shows the average scores for each metric in this category for the different corpora (the differences are statistically significant). All metrics suggest that Leg2Kids is simpler than Adapt2Kids. However, it is worth emphasising that these readability metrics may not be capturing simplicity in our case. When analysing the Descriptive Indexes, we show that Leg2Kids has smaller sentences and smaller words than Adapt2Kids (words per sentence and syllables per content words metrics). Since readability metrics rely heavily on these two factors, it cannot be concluded that Leg2Kids is simpler than Adapt2Kids without any further analysis.

	Adapt2Kids	Leg2Kids
Brunet (↑)	11.03	12.87
Adapted Dale-Chall (↓)	9.85	8.99
Flesch Reading Ease (↑)	51.72	76.35
Gunning Fog (↓)	7.00	2.65
Honoré statistics (↓)	1,040.01	933.04

Table 9: Results for readability metrics (arrows indicate the simplicity direction).

4.3 Complexity prediction of original and simplified texts using PorSimples corpus

The PorSimples corpus of simplified texts was used to train a textual complexity model for the Simplifica (Scarton et al, 2010b) tool, which helped in the manual simplification process, supported by simplification rules. The model helps a professional to know when to stop the simplification process. In PorSimples, we had the mapping: natural - literate at a basic level; and strong - literate at a rudimentary level (Aluisio et al, 2010). The objective of the following experiment is to exemplify the use of NILC-Metrix metrics to classify these complexity levels.

In Aluisio et al (2010), the 42 Coh-Metrix-Port metrics are presented that are used for training a classifier for three levels of textual complexity. Here, we used 38 of these 42 metrics as four of them were discontinued due to a project decision in parser changing. The four discontinued metrics were: *Incidence of NPs*, *Number of NP modifiers*, *Number of high level constituents* and *Pronoun-NP ratio*.

Here, we try to answer two questions via machine learning experiments: (i) whether new features, described in Section 3, developed after the Coh-Metrix-Port project, add value to the task textual complexity prediction using the parallel corpus of PorSimples; and (ii) which categories of features best describe the characteristics that distinguish texts of the PorSimples project (original texts, naturally simplified and strongly simplified).

The method used was the Multinomial Logistic Regression, which has as its premise the ordinal relationship between classes (levels of simplification) (Heilman et al, 2008). This was the same method used in the original article of the Coh-Metrix-Port (Aluisio et al, 2010) project. In order to better refine the analysis, we used the F1 metric by class and we also presented the F1 Macro, which provided us with a greater degree of detail regarding the difficulty of the task of classifying textual complexity. All experiments followed the stratified 10-fold cross-validation methodology when splitting the data between the training and testing sets. The stratified strategy ensures that all training and test folds contain all text levels, increasing the experiment's robustness. The division into 10 folds for training and testing is a good proxy for the leave-one-out methodology, ensuring good generalisation of the results achieved and greater confidence in a non-overfit or underfit result. We are aware of the small number of texts available for this experiment and the bias of such data volume analysis. Thus, it is essential to be careful about data usage.

Category	Strong	Natural	Original	F1 Macro
All	0.655	**0.568**	**0.888**	**0.704**
Coh-Metrix-Port	0.719	0.514	0.806	0.679
Readability Formulas	**0.720**	0.402	0.782	0.635
Syntactic Complexity	0.675	0.409	0.813	0.632
Text Easability Metrics	0.661	0.413	0.763	0.612
Morphosyntactic Word Information	0.679	0.408	0.739	0.609
Descriptive Index	0.701	0.284	0.734	0.573
LSA-Semantic Cohesion	0.637	0.349	0.721	0.569
Lexical Diversity	0.592	0.384	0.714	0.563
Referential Cohesion	0.663	0.323	0.689	0.558
Semantic Word Information	0.422	0.331	0.671	0.475
Connectives	0.506	0.286	0.623	0.472
Syntactic Pattern Density	0.577	0.269	0.551	0.466
Word Frequency	0.477	0.318	0.582	0.459
Temporal lexicon	0.552	0.232	0.530	0.438
Psycholinguistic Measures	0.394	0.250	0.593	0.412

Table 10: Performance on PorSimples dataset. Results presented by category of features.

Table 10 presents the results of the automatic text classification experiment by the feature's category. This division gives us better visibility regarding the categories that most contribute to automatic classification, that is, those that best describe the characteristics that distinguish the original texts and their two levels of simplification. When comparing the use of all the features concerning the 38 of the Coh-Metrix-Port, we noticed it again in the macro F1 and also in the Natural and Original Classes, despite a slight worsening

concerning the classification of the Strong class. Regarding feature categories, we noticed that the combination of all features presented the best F1 Macro for the task and also the best F1 micro for the Natural and Original classes. Regarding F1 for the *Strong* class, we noticed that the individual use of the *Readability Formulas* category presented a better result than its aggregated usage with other features. This result is interesting, as it presents us with a scenario in which the other groups of features confuse the classifier concerning the classification of this class. This confusion can occur due to the improvement in the distinction of the other classes (*Natural* and *Original*), causing a trade-off in relation to the *Strong* class. In both evaluations, we noticed that the aggregate use of all features produces a slight worsening in the classification of the *Strong* class, although it produces better results in general, which is positive in the end.

Category	#
Syntactic Complexity	13
Word Frequency	6
Descriptive Index	5
Readability Formulas	5
LSA-Semantic Cohesion	5
Lexical Diversity	4
Text Easability Metrics	3
Psycholinguistic Measures	3
Connectives	2
Referential Cohesion	2
Morphosyntactic Word Information	2
Semantic Word Information	1
Syntactic Pattern Density	1

Table 11: Features by category resulting from a Boruta procedure.

We carried out a feature selection step to better understand which features are relevant in explaining the phenomenon of classification of PorSimples texts. We know that not all features are necessarily useful: some may not differentiate between simple and complex texts and others may be correlated with each other, that is, redundant. Therefore, we run the Boruta (Kursa et al, 2010; Kursa and Rudnicki, 2010) method for feature selection. Boruta checks which features are more informative to explain the event of interest than a random variable produced from the shuffling of the feature itself. If a feature explains an event, it is correlated with the fact that a text is simple or complex, but if we scramble that feature, it loses its correlation with the event and no longer explains it. Boruta eliminated 147 of the 200 features, resulting in a subset of 53 features. Table 11 shows the count of resulting features by category of features.

The justification for choosing Boruta among other selection methods was because the algorithm was designed to classify what the original article calls the "all relevant problem": finding a subset of features that are relevant to a given classification task. This is different from the "minimum-optimal problem", which is the problem of finding the minimum subset of features that

perform in a model. Although the machine learning models in production should ultimately aim at selecting optimal minimum features, Boruta's thesis is that, for exploration purposes, minimal optimisation goes too far. Moreover, the method is robust to the correlation of features. In scenarios with a large number of features, dealing with their correlation can be a very costly task. Thus, using Boruta can also speed up the stage of preparing features, justifying our choice.

Category	Strong	Natural	Original	F1 Macro
All	0.708	0.508	**0.860**	**0.692**
Coh-Metrix-Port	0.719	**0.514**	0.806	0.679
Readability Formulas	**0.720**	0.402	0.782	0.635
Syntactic Complexity	0.687	0.414	0.796	0.632
Text Easability Metrics	0.644	0.389	0.752	0.595
Descriptive Index	0.691	0.302	0.716	0.570
Morphosyntactic Word Information	0.614	0.330	0.708	0.551
Lexical Diversity	0.586	0.359	0.699	0.548
LSA-Semantic Cohesion	0.590	0.295	0.672	0.519
Word Frequency	0.468	0.321	0.600	0.463
Syntactic Pattern Density	0.557	0.271	0.554	0.461
Connectives	0.500	0.219	0.555	0.425
Referential Cohesion	0.307	0.340	0.540	0.396
Psycholinguistic Measures	0.410	0.227	0.531	0.389
Semantic Word Information	0.266	0.243	0.531	0.346
Temporal lexicon	0.148	0.098	0.254	0.167

Table 12: Performance on PorSimples dataset using only feature selected by Boruta. Results presented by category of features.

We replicated the PorSimples text classification experiment using only the features selected by Boruta. Table 12 presents the results obtained. Once more, we noticed that all feature usage (now the 53 selected ones) performed better in the classification of textual complexity concerning the 38 features replicated from Coh-Metrix-Port. We noticed a minimal difference in performance in the *Strong* and *Natural* classes but a significant gain in the *Original* class, demonstrating value when using the new features. When comparing the use of the 53 selected features concerning the 200 features developed, we noticed a slight drop in the F1 Macro obtained, which can be justified by the small size of the dataset and weak correlations between the features, as well as between a feature and the target of the task. This kind of phenomenon tends to be irrelevant as the increase in the dataset causes effects such as these to be considered statistically insignificant. When we analyse the performance of the categories of features, the data show us that the difference in performance in the prediction of the *Strong* class decreased between the use of all selected features and the use of the selected features of the *Readability Formulas* category. While this difference tends to a rounding error, the combined performance of all the features selected in the prediction of the *Natural* and *Original* classes, as well as in the F1 Macro, stands out regarding the individual use of the feature categories. We realised, therefore, that the development of new linguistic features adds value in predicting the textual complexity of PorSimples texts.

4.4 Complexity prediction of transcribed speech narratives of Adole-Sendo project

This experiment was performed to validate and exemplify the use of NILC-Metrix metrics applied to transcribed speech texts, using the 271 narratives of the Adole-Sendo corpus (see Section 4.1.3), grouped by grades and stages of education. As we are focusing on grades 6 to 9 of Elementary School II, our dataset comprises two new sets of narratives from the grades of Elementary School I, grouped, and of narratives from Secondary School. This division in six classes also helped to balance the samples: 4th and 5th grades were grouped in ESI (Elementary School I) totalling 43 texts; 6th grade has 70 texts, 7th grade has 43 texts, 8th grade has 15 texts, 9th grade has 59 texts and SC (Secondary School) has 41 texts. As can be seen in the Figure 1, the task is not trivial, as there is no clear separation between classes in two dimensions, for example.

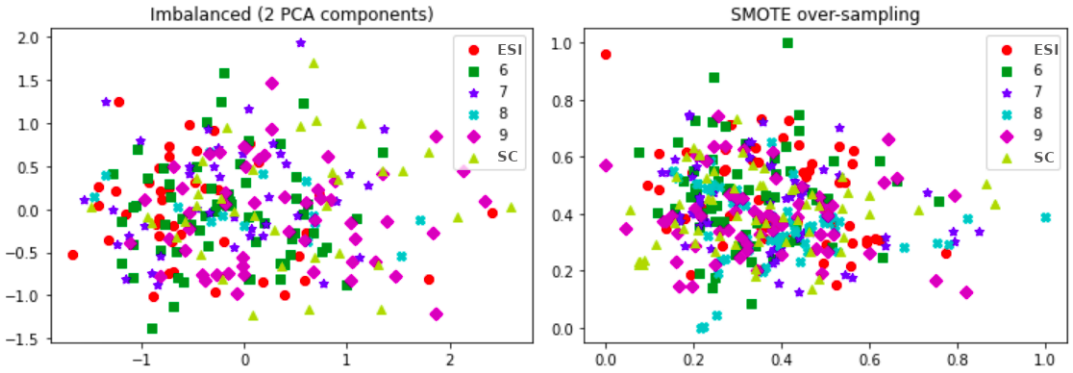


Fig. 1: Adole-Sendo classes distribution plotted using PCA, before and after data-augmentation with SMOTE

We proceed the experiment with the normalisation of the 200 features using the MinMaxScaler which leaves all values between 0 and 1. Then, the ANOVA technique was used to select features (Brownlee, 2019), reducing the number of relevant columns to 194 correlated with the classes; the top 20 more relevant features can be seen in Table 13. 10% of each class of the dataset was also separated for validation (26 samples). For the remaining 245 samples, the classes were balanced using the SMOTE Over-Sampling (Chawla et al, 2002) data-augmentation method. The result of this process can be seen in Figure 1 where 63 samples were assigned per class.

Five classification methods from the Scikit-Learn[32] library were chosen, using standard hyperparameters: a) Linear SVM with C = 0.025 ; b) SVB RBF with C = 1; c) Random Forest with max_depth = 5; d) Neural Network MLP with 100 neurons in the hidden layer and e) Gaussian Naive Bayes. The best F1-Score method was the Neural Net with 0.62, but very close to SVM

[32] https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

Name	Group	Weight
cross_entropy	LSA-Semantic Cohesion	9.30
prepositions_per_sentence	Morphosyntactic Word Information	7.58
first_person_pronouns	Morphosyntactic Word Information	6.02
long_sentence_ratio	Text Easability Metrics	5.76
content_density	Lexical Diversity	5.75
verbs_max	Morphosyntactic Word Information	5.75
prepositions_per_clause	Morphosyntactic Word Information	5.65
content_words	Morphosyntactic Word Information	5.56
adverbs_standard_deviation	Morphosyntactic Word Information	5.51
function_words	Morphosyntactic Word Information	5.47
ratio_function_to_content_words	Morphosyntactic Word Information	5.29
sentences_with_one_clause	Syntactic Complexity	5.19
adj_arg_ovl	Referential Cohesion	4.82
dalechall_adapted	Readability Formulas	4.79
content_word_max	Lexical Diversity	4.65
idade_aquisicao_mean	Psycholinguistic Measures	4.61
arg_ovl	Referential Cohesion	4.58
non-inflected_verbs	Morphosyntactic Word Information	4.50
pronouns_min	Morphosyntactic Word Information	4.45

Table 13: Top 20 features ordered by weight after selection with ANOVA technique on Adole-Sendo classification task

(Table 14). The CV F-Score was calculated using 10-Fold Cross Validation and the Val F-Score was calculated from the prediction values in the validation dataset. Confusion matrices of test and validation data can be seen in Figure 2.

Classifier	CV F-Score	Val. F-Score
Linear SVM	0.28	0.13
RBF SVM	0.61	**0.88**
Random Forest	0.39	0.68
Neural Net	**0.62**	**0.88**
Naive Bayes	0.38	0.48

Table 14: ML methods evaluated in the Adole-Sendo classification task, CV is 10-Fold Cross Validation and Val. is the result in reserved validation samples

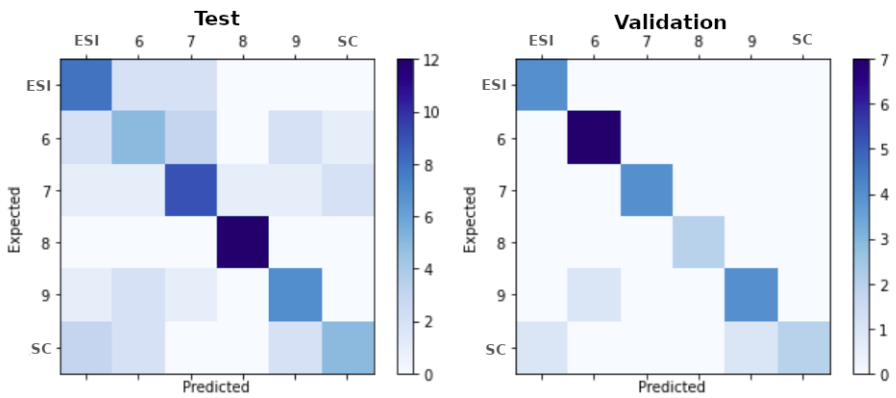


Fig. 2: Confusion Matrix for test and validation samples from Adole-Sendo corpus

Group	CV F1-Score	Std
Lexical Diversity	0.23	0.06
Text Easability Metrics	0.21	0.04
Morphosyntactic Word Information	0.21	0.05
Psycholinguistic Measures	0.19	0.03
Semantic Word Information	0.18	0.05
Descriptive Index	0.15	0.03
LSA-Semantic Cohesion	0.15	0.04
Temporal lexicon	0.15	0.06
Readability Formulas	0.15	0.04
Syntactic Complexity	0.14	0.04
Connectives	0.14	0.03
Word Frequency	0.14	0.03
Referential Cohesion	0.12	0.02
Syntactic Pattern Density	0.11	0.02

Table 15: Evaluation of each group isolated features on Adole-Sendo classification task. CV F1-Score is the average of F1 with 10-Fold Cross Validation and Std is the standard deviation.

Finally, the weight of each group of metrics was evaluated in the classification, using MLP Neural Net (the best method of the previous step). The set of metrics that performed best in isolation was **Lexical Diversity**, with 0.23 F1-Score, followed by **Text Easability Metrics** and **Morphosyntactic Word Information**. The complete list can be seen in the Table 15.

5 Uses of NILC-Metrix Metrics

In this section, we review 5 published studies in several research areas — Natural Language Processing, Neuropsychological Language Tests, Education, Language and Eye-tracking studies — to illustrate the wide-ranging use of sets of metrics available in NILC-Metrix.

Santos et al (2020) used 165 metrics of NILC-Metrix to evaluate their contribution to detect fake news for the BP language. The focus of the study was on 17 metrics of this large set, from 4 categories (Classic Readability Formulas, Referential Cohesion, Text Easability Metrics and Psycholinguistics), named as readability features by the authors. The authors selected the following classic readability formulas: Flesch Index, Brunet Index, Honore Statistic, Dale Chall Formula, and Gunning Fog Index. From the set of 9 metrics of Referential Cohesion of NILC-Metrix, 7 of them were used: 4 metrics from the Psycholinguistic Measures and one from the set of Text Easability Metrics. In their study the authors used an open access and balanced corpus called Fake.Br corpus[33], with aligned texts totalling 3,600 false and 3,600 true news. SVM with the standard parameters of Scikit-learn[34] was used, along with traditional evaluation measures of precision, recall, F-measure and general accuracy in a 5-fold cross-validation strategy. The results of their study showed that readability features were relevant for detecting fake news in BP, achieving, alone, up to 92% classification accuracy.

[33] `https://github.com/roneysco/Fake.br-Corpus`

[34] `https://scikit-learn.org/stable/index.html`

Aluísio et al (2016) evaluated classification and regression methods to identify linguistic features for dementia diagnosis, focusing on Alzheimer Disease (AD) and Mild Cognitive Impairment (MCI), to distinguish them from Control Patients (CT). In their paper, a narrative language test was used based on sequenced pictures (Cinderella story) and features extracted from the resulting transcriptions, using the Coh-Metrix-Dementia tool. It is important to note that the NILC-Metrix includes 18 metrics from Coh-Metrix-Dementia, 11 metrics from the LSA-Semantic Cohesion class, 4 from the Syntactic Complexity class, 2 Readability Formulas and one from the class Lexical Diversity. For the classification results, they obtained 0.82 F1-score in the experiment with three classes (AD, MCI and CT), and 0.90 for two classes (CT *versus* (MCI+AD)), both using the CFS-selected features; for regression, they obtained 0.24 MAE for three classes, and 0.12 for two classes, both using all features available in the Coh-Metrix-Dementia tool.

Gazzola et al (2019) investigated the impact of textual genre in assessing text complexity in BP educational resources. Their final goal was to develop methods to assess the stage of education for the Open Educational Resources (OER) available on the platform MEC-RED (from the Brazilian Ministry of Education)[35]. For this purpose, a corpus with textbooks for Elementary School I, Elementary School II, Secondary School and Higher Education was compiled. A set of 79 metrics from NILC-Metrix was selected, based on the study by Graesser and McNamara (2011). Using those 79 metrics, they found correspondence which 53 metrics of Coh-Metrix, and grouped them into: *Metrics Related to Words*, *Related to Sentences* and *Related to Connections between Sentences*. After selecting the features, 5 Machine Learning methods were tested: SVM, MLP, Logistic Regression and Random Forest from scikit learn[36]. SVM performed better with 0.804 F-Measure, therefore it was used in an extrinsic evaluation with two sets of OER, reaching 0.518 F-Measure in the set with text genres similar from the training set (textbook corpus) and 0.389 F-Measure for the animation/simulation and practical experiment resources, which are very common in the MEC-RED platform.

Finatto et al (2011) evaluated the differences in text complexity of popular Brazilian newspapers (aimed at a public with a lower education) with traditional ones (aimed at more educated readers), using cohesion, syntax and vocabulary metrics, including ellipsis. In their contrastive analysis, the authors used 48 metrics from Coh-Metrix-Port and included 5 new ones related to the co-reference of ellipses, based on a corpus annotation. The annotation involved identifying ellipses of three types: nominal, verbal and sentential. The study selected a balanced corpus of texts seeking the widest possible range of themes and editorials. They used 80 texts from the traditional Zero Hora newspaper from 2006 and 2007 and 80 texts from the popular Diário Gaucho from 2008[37]. The authors found out that the most discriminative features between

[35] `https://plataformaintegrada.mec.gov.br/home`

[36] `https://scikit-learn.org/stable/`

[37] `https://gauchazh.clicrbs.com.br/`

both newspapers were a set of 14 features grouped into 5 classes: Referential Cohesion, Word Frequency, Syntactic Complexity, Descriptive Index, Morphosyntactic Word Information, extracted using Coh-Metrix-Port, but ellipsis did not have a distinctive role.

Leal et al (2019) used NILC-Metrix metrics to propose a less subjective model for choosing texts and paragraphs for a project in the area of Psycholinguistics called RastrOS. In their study, the objective was to select 50 paragraphs with a wide range of language phenomena for RastrOS, a corpus with predictability norms and eye tracking data during silent reading of short paragraphs. First, 58 metrics with great relevance to the task were manually selected (grouped into structural complexity, types of sentences, co-reference and morphosyntactics). Next, these metrics were extracted from all the paragraphs to help with grouping together texts with similar types of features by K-Means and Agglomerative Clustering methods. To assess the quality of the groups, the Elbow method, V-Measure and Silhouette techniques were used. After grouping, the paragraphs went through a human selection to find a few examples from each large text group.

6 Concluding Remarks and Future Work

The objective of this paper was to introduce and make the NILC-Metrix, a computational system comprising 200 metrics for BP, publicly available. We presented the motivation for developing this large set of metrics and also illustrated the wide-ranging uses of NILC-Metrix published in studies of several research areas. We also presented three experiments based on corpora, using NILC-Metrix: an analysis of the differences between children's film subtitles and texts written for children, a new predictor of textual complexity for the PorSimples corpus, and a complexity prediction model for school grades, using transcripts of children's story narratives. For each case of study, we showed the robustness of NILC-Metrix, highlighting the importance of having a large number of metrics, that cover multiple linguistic aspects, available for textual analysis.

Regarding future studies, we foresee two lines of research. The first one is related to implementing existing and new metrics and the NLP resources used for implementation. For example, instead of using three parsers (LX-Parser, Malt and Palavras) when implementing syntactic metrics, in the near future we will be able to use robust parsing models for Portuguese, available in the POeTiSA project[38]. As for new metrics, we also have a long list of suggestions. Idea Density is a metric that computes the number of propositions of a text, divided by its number of words; it was implemented in Coh-Metrix-Dementia (Cunha et al, 2015), using a set of rules over dependency parsing[39]. Once a

[38] `https://sites.google.com/icmc.usp.br/poetisa`

[39] The metric uses a tool called IDD32 (Idea Density from Dependency Trees), which can extract propositions from well-written English and Portuguese texts, which is a drawback for its general use.

robust parsing model is made available, the robustness of this metric can be evaluated and implemented in the NILC-Metrix. Flor et al (2013) defined a metric called lexical tightness that measures global cohesion of content words in a text. According to the authors, this metric represents the degree to which a text tends to use words that are highly inter-associated in the language. This metric is a candidate to be evaluated and compared with the semantic cohesion metrics based on LSA, already implemented in the NILC-Metrix. Duran et al (2007) evaluated temporal indices available in the Coh-Metrix in order to investigate temporal coherence. Six of the indices are available in the Coh-Metrix v3.0 and are related to the grammatical function (PoS, connectives that are already implemented in the NILC-Metrix, and temporal adverbial phrases); three other temporal indices were also proposed in their work. Temporal cohesion is also a candidate to be investigated and implemented for BP. Other sets of metrics related to causal and intentional cohesion available in Coh-Metrix should be studied and evaluated for BP. The second line of research is related to the validation of sets of metrics for several NLP tasks. We hope this paper can encourage researchers to work in this line of validating sets of metrics as we have made available both the access and code of the 200 metrics developed.

7 Acknowledgments

This work is part of the RastrOS Project supported by the São Paulo Research Foundation (FAPESP—Regular Grant #2019/09807-0). The authors would like to thank all the members of the PorSimples project that provided the basis for building Coh-Metrix-Port and AIC metrics. We would also like to thank all the students who contributed (after PorSimples finished) to enlarging the set of metrics, revising it, applying it in various NLP tasks and, finally to making NILC-Metrix publicly available.

8 Declarations

Funding: This research was supported by The São Paulo Research Foundation (FAPESP) (*Fundação de Amparo à Pesquisa do Estado de São Paulo*, in Portuguese), Regular Grant #2019/09807-0. **Conflicts of interest/Competing interests:** The authors have no conflicts of interest to declare. **Availability of data and material (data transparency):** Four datasets used in the applications of NILC-Metrix are available, in tsv format, in the file DATA at `https://github.com/nilc-nlp/nilcmetrix`. **Code availability (software application or custom code):** Source Code of NILC-Metrix is available at `https://github.com/nilc-nlp/nilcmetrix` under AGPLv3 license.

Authors' contributions: Sidney Leal: Conceptualisation, Investigation, Methodology, Resources, Software Development, Validation, Writing –

original paper; **Magali Duran:** Conceptualisation, Data curation, Investigation, Resources, Writing – original paper; **Carolina Scarton:** Conceptualisation, Data curation, Investigation, Methodology, Resources, Software Development, Validation, Writing – original paper; **Nathan Hartmann:** Conceptualisation, Data curation, Investigation, Methodology, Resources, Software Development, Validation, Writing – original paper; **Sandra Aluisio:** Conceptualisation, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original paper.

References

Aluísio S, Gasperin C (2010) Fostering digital inclusion and accessibility: The PorSimples project for simplification of Portuguese texts. In: Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, Association for Computational Linguistics, Los Angeles, California, pp 46–53, URL https://www.aclweb.org/anthology/W10-1607

Aluisio S, Specia L, Gasperin C, Scarton C (2010) Readability assessment for text simplification. In: Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Los Angeles, California, pp 1–9, URL https://www.aclweb.org/anthology/W10-1001

Aluísio SM, Cunha A, Scarton C (2016) Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In: Silva JR, Ribeiro R, Quaresma P, Adami A, Branco A (eds) Computational Processing of the Portuguese Language - 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings, Springer, Lecture Notes in Computer Science, vol 9727, pp 109–114, DOI 10.1007/978-3-319-41552-9_10, URL https://doi.org/10.1007/978-3-319-41552-9_10

Alva-Manchego F, Bingel J, Paetzold G, Scarton C, Specia L (2017) Learning how to simplify from explicit labeling of complex-simplified text pairs. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Asian Federation of Natural Language Processing, Taipei, Taiwan, pp 295–305, URL https://www.aclweb.org/anthology/I17-1030

Arfé B, Oakhill J, Pianta E (2014) The text simplification in terence. In: Mascio TD, Gennari R, Vitorini P, Vicari R, de la Prieta F (eds) Methodologies and Intelligent Systems for Technology Enhanced Learning, Springer International Publishing, Cham, pp 165–172

Bick E (2000) The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. University of Arhus, Århus

Biderman MTC (1998) Dicionário Didático de Português. Editora Ática, São Paulo

Brownlee J (2019) How to choose a feature selection method for machine learning. URL https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/, [Online; accessed 2021.03.01]

Candido A, Maziero E, Specia L, Gasperin C, Pardo T, Aluisio S (2009) Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In: Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Boulder, Colorado, pp 34–42, URL https://www.aclweb.org/anthology/W09-2105

Carroll J, Minnen G, Canning Y, Devlin S, Tait J (1998) Practical simplification of english newspaper text to assist aphasic readers. In: In Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology, pp 7–10

Caseli H, de Freitas Pereira T, Specia L, Pardo TAS, Gasperin C, Aluísio SM (2009) Building a brazilian portuguese parallel corpus of original and simplified texts. In: Advances in Computational Linguistics, Research in Computer Science (CICLing-2009), vol 41, pp 59–70

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. Journal Of Artificial Intelligence Research 16:321–357, URL http://arxiv.org/abs/1106.1813

Crossley SA, Dufty DF, McCarthy PM, McNamara DS (2007) Toward a new readability: A mixed model approach. In: Proceedings of the Cognitive Science Society, vol 29, pp 197–202, URL https://escholarship.org/uc/item/39r3d755

Crossley SA, Greenfield J, McNamara DS (2008) Assessing text readability using cognitively based indices. Tesol Quarterly 42(3):475–493

Cunha ALVd, Sousa LBd, Mansur LL, Aluisio SM (2015) Automatic proposition extraction from dependency trees: helping early prediction of alzheimer's disease from narratives. In: International Symposium on Computer-Based Medical Systems - CBMS, IEEE, DOI 10.1109/CBMS.2015.19

Dale E, Chall JS (1948) A formula for predicting readability: Instructions. Educational research bulletin pp 37–54

Duran ND, McCarthy PM, Graesser AC, McNamara DS (2007) Using temporal cohesion to predict temporal coherence in narrative and expository texts. Behavior Research Methods, Instruments, & Computers 39:212—-223, DOI 10.3758/BF03193150

Finatto MJB, Scarton CE, Rocha A, Aluísio S (2011) Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero (characteristics of popular news: the evaluation of intelligibility and support to the genre description) [in Portuguese]. In: Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, URL https://www.aclweb.org/anthology/W11-4506

Flesch R (1948) A new readability yardstick. Journal of applied psychology 32(3):221

Flor M, Beigman Klebanov B, Sheehan KM (2013) Lexical tightness and text complexity. In: Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility, Association for Computational Linguistics, Atlanta, Georgia, pp 29–38, URL https://www.aclweb.org/anthology/W13-1504

Fonseca ER, Rosa JLG, Aluisio SM (2015) Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. Journal of the Brazilian Computer Society 21(2), DOI 10.1186/s13173-014-0020-x

Frazier L (1985) Syntactic complexity. In: Dowty DR, Karttunen L, Zwicky AM (eds) Language Parsing: Psychological, Computational, and Theoretical Perspectives, Cambridge University Press, pp 129–189

Fry E (1968) A readability formula that saves time. Journal of reading 11(7):513–578

Gazzola M, Leal S, Aluísio S (2019) Predição da complexidade textual de recursos educacionais abertos em português. In: 12th Brazilian Symposium in Information and Human Language Technology (STIL 2019), Brazilian Computer Society (SBC), pp 1–10

Graesser AC, McNamara DS (2011) Computational analyses of multilevel discourse comprehension. Topics in cognitive science 3(2):371–398

Graesser AC, McNamara DS, Louwerse MM, Cai Z (2004) Coh-metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers 36:193—-202, DOI 10.3758/BF03195564

Graesser AC, McNamara DS, Kulikowich JM (2011) Coh-metrix: Providing multilevel analyses of text characteristics. Educational Researcher 40(5):223–234, DOI 10.3102/0013189X11413260, URL https://doi.org/10.3102/0013189X11413260, https://doi.org/10.3102/0013189X11413260

Graesser AC, McNamara DS, Cai Z, Conley M, Li H, Pennebaker J (2014) Coh-metrix measures text characteristics at multiple levels of language and discourse. The Elementary School Journal 115(2):210–229

Gunning R (1952) {The Technique of Clear Writing}. McGraw-Hill, New York

Hartmann NS, Aluísio SM (2020) Adaptação lexical automática em textos informativos do português brasileiro para o ensino fundamental. Linguamática 12(2):3–27, DOI 10.21814/lm.12.2.323, URL https://linguamatica.com/index.php/linguamatica/article/view/323

Heilman M, Collins-Thompson K, Eskenazi M (2008) An analysis of statistical models and features for reading difficulty prediction. In: Proceedings of the third workshop on innovative use of NLP for building educational applications, pp 71–79

Hu X, Cai Z, Louwerse M, Olney A, Penumatsa P, Graesser A (2003) A revised algorithm for latent semantic analysis, Morgan Kaufman Publishers, pp 1489–1491. 18th International Joint Conference of Artificial Intelligence , IJCAI'03 ; Conference date: 09-08-2003 Through 15-08-2003

Janczura GAA, Castilho GMAd, Rocha NO, van Erven TdJC, Huang TP (2007) Normas de concretude para 909 palavras da lÃ-ngua portuguesa. Psicologia: Teoria e Pesquisa 23:195 – 204, URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-37722007000200010&nrm=iso

Kincaid JP, Fishburne Jr RP, Rogers RL, Chissom BS (1975) Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., DTIC Document

Kintsch W (1998) Comprehension: A paradigm for cognition. Cambridge university press

Kintsch W, Keenan J (1973) Reading rate and retention as a function of the number of propositions in the base structure of sentences. Cognitive psychology 5(3):257–274

Kintsch W, Van Dijk TA (1978) Toward a model of text comprehension and production. Psychological review 85(5):363

Kursa MB, Rudnicki WR (2010) Feature selection with the boruta package. J Stat Softw 36(11):1–13

Kursa MB, Jankowski A, Rudnicki WR (2010) Boruta–a system for feature selection. Fundamenta Informaticae 101(4):271–285

Landauer TK, Laham D, Rehder B, Schreiner ME (1997) How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In: Shafto MG, Langley P (eds) Proceedings of the 19th annual meeting of the Cognitive Science Society, pp 412–417

Leal SE, Aluísio SM, dos Santos Rodrigues E, Vieira JMM, Teixeira EN (2019) Métodos de clusterização para a criação de corpus para rastreamento ocular durante a leitura de parágrafos em português. In: JDP 2019 - Jornada de Descrição do Português, Salvador, Bahia, Brasil, p 270–278

Louwerse MM, McCarthy PM, McNamara DS, Graesser AC (2004) Variation in Language and Cohesion across Written and Spoken Registers. In: Proceedings of the twenty-sixth annual conference of the Cognitive Science Society, Mahwah, NJ, pp 843–848

Martins T, Ghiraldelo C, Nunes M, Jr O (1996) Readability formulas applied to textbooks in brazilian portuguese. Série Computação 28, ICMSC-USP, martins, T.B.F.; Ghiraldelo, C.M.; Nunes, M.G.V.; Oliveira Jr., O.N. Readability Formulas Applied to Textbooks in Brazilian Portuguese. Notas do ICMSC-USP, Série Computação, nro. 28, 1996, 11p

Max A (2006) Writing for language-impaired readers. In: In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2006. Lecture Notes in Computer Science, vol 3878, Springer, Berlin, Heidelberg, pp 7567–570, DOI 10.1007/11671299_59

Maziero EG, Pardo TAS, Aluísio SM (2008) Ferramenta de análise automática de inteligibilidade de córpus (aic). Tech. rep., Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional (NILC-TR-08-08), 14 p., Julho 2008, University of São Paulo, ICMC/USP, São Carlos-SP.

McNamara DS, Graesser AC, McCarthy PM, Cai Z (2014) Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press, DOI 10.1017/CBO9780511894664

Pardo TAS, Nunes MGV (2006) Review and evaluation of dizer - an automatic discourse analyzer for brazilian portuguese. In: Vieira R, Quaresma P, das Graças Volpe Nunes M, Mamede NJ, Oliveira C, Dias MC (eds) Computational Processing of the Portuguese Language, 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006, Proceedings, Springer, Lecture Notes in Computer Science, vol 3960, pp 180–189, DOI 10.1007/11751984_19, URL https://doi.org/10.1007/11751984_19

Santos LBd, Duran MS, Hartmann NS, Candido Junior A, Paetzold GH, Aluísio SM (2017) A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In: International Conference on Text, Speech, and Dialogue - TSD 2017, Proceedings, Springer, Lecture Notes in Artificial Intelligence, vol 10415, pp 281–28, DOI 10.1007/978-3-319-64206-2_32

Santos R, Pedro G, Leal S, Vale O, Pardo T, Bontcheva K, Scarton C (2020) Measuring the impact of readability features in fake news detection. In: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, pp 1404–1413, URL https://www.aclweb.org/anthology/2020.lrec-1.176

Sardinha APB (2004) Corpus brasileiro. URL http://corpusbrasileiro.pucsp.br/cb/Inicial.html, [Online; accessed 2021.03.21]

Scarton C, Aluísio S (2010) Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. Linguamática 2(1):45–61

Scarton C, Gasperin C, Aluísio S (2010a) Revisiting the readability assessment of texts in portuguese. In: Advances in Artificial Intelligence – IBERAMIA - Volume 6433 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp 306–315

Scarton C, Oliveira-Junior O, Candido-Junior A, Gasperin C, Aluísio SM (2010b) Simplifica: a tool for authoring simplified texts in brazilian portuguese guided by readability assessments. In: Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies, Los Angeles, CA, pp 41–44

Shardlow M (2014) A survey of automated text simplification. International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014 4(1), DOI 10.14569/SpecialIssue.2014.040109, URL http://dx.doi.org/10.14569/SpecialIssue.2014.040109

Silva JR, Branco A, Castro S, Reis R (2010) Out-of-the-box robust parsing of portuguese. In: Pardo TAS, Branco A, Klautau A, Vieira R, de Lima VLS (eds) Computational Processing of the Portuguese Language, 9th International Conference, PROPOR 2010, Porto Alegre, RS, Brazil, April 27-30, 2010. Proceedings, Springer, Lecture Notes in Computer Science, vol 6001, pp 75–85, DOI 10.1007/978-3-642-12320-7_10, URL https:

//doi.org/10.1007/978-3-642-12320-7_10

Soares A, Medeiros JC, Simões A, Machado J, Costa A, Álvaro Iriarte, Almeida J, Pinheiro A, Comesaña M (2014) Escolex: A grade-level lexical database from european portuguese elementary to middle school textbooks. Behavior Research Methods 46:240–253

Tang K (2012) A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research. UCL Working Papers in Linguistics 24:208–214

Thomas C, Keselj V, Cercone N, Rockwood K, Asp E (2005) Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In: IEEE International Conference Mechatronics and Automation, 2005, vol 3, pp 1569–1574 Vol. 3, DOI 10.1109/ICMA.2005.1626789

Wagner Filho JA, Wilkens R, Idiart M, Villavicencio A (2018) The brWaC corpus: A new open resource for Brazilian Portuguese. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, URL https://www.aclweb.org/anthology/L18-1686

Watanabe WM, Candido A, Amâncio MA, de Oliveira M, Pardo TAS, Fortes RPM, Aluísio SM (2010) Adapting web content for low-literacy readers by using lexical elaboration and named entities labeling. In: Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A), Association for Computing Machinery, New York, NY, USA, W4A '10, DOI 10.1145/1805986.1805998, URL https://doi.org/10.1145/1805986.1805998

Welch BL (1947) The generalization of "student's" problem when several different population variances are involved. Biometrika 34(1-2):28–35

Xu W, Callison-Burch C, Napoles C (2015) Problems in current text simplification research: New data can help. Transactions of the Association for Computational Linguistics 3:283–297, DOI 10.1162/tacl_a_00139, URL https://www.aclweb.org/anthology/Q15-1021

Yngve VH (1960) A model and hypothesis for language structure. Proceedings of the American Philosophical Association 104(5):444–466

CAPÍTULO

4

RASTROS: MÉTRICAS DE RASTREAMENTO OCULAR E MÉTODOS RELACIONADOS

Este capítulo relata a criação do RastrOS, um córpus com dados de rastreamento ocular durante a leitura silenciosa de parágrafos e as normas de previsibilidade (coletadas por meio do teste Cloze aplicado em mais de 400 estudantes da graduação em seis universidades), com base em três artigos, descritos abaixo.

A Seção 4.1 traz o método criado para a escolha dos parágrafos do córpus, como uma alternativa mais objetiva para a seleção visando a diversidade de fenômenos da língua. A Seção 4.2 traz os métodos avaliados para medição da similaridade semântica das palavras do córpus RastrOS, usando modelos de língua recentes. Traz também um córpus criado a partir do teste Cloze para a tarefa de *Sentence Completion*, usado para ranquear os melhores métodos acima e assim criar um método híbrido de similaridade semântica. Finalmente a Seção 4.3 detalha a infraestrutura computacional desenvolvida para a criação do córpus e também disponibiliza a primeira versão dos dados, que foram utilizados para atingir o estado da arte para a tarefa de avaliação da complexidade sentencial, descrito no Capítulo 5.

4.1 Métodos de clusterização para a criação de córpus

Título:	**Métodos de Clusterização para a Criação de Corpus para Rastreamento Ocular durante a Leitura de Parágrafos em Português**
Autores:	**Sidney Evaldo Leal, Erica dos Santos Rodrigues, João Marcos Monguba Vieira, Elisângela Nogueira Teixeira e Sandra Maria Aluísio**
Ano:	**2019**
Conferência:	**VI Jornada de Descrição do Português - Salvador - BA**
Situação:	**Publicado**

A primeira etapa para a criação do córpus RastrOS foi a seleção dos parágrafos que fariam parte dele. Esse artigo reporta o método criado para a escolha desses parágrafos, com o objetivo de fazer uma seleção de forma menos subjetiva e procurando evitar parágrafos muito parecidos.

O método avaliado foi a clusterização ou agrupamento, usando algoritmos de aprendizagem não supervisionada e quatro grupos de métricas linguísticas extraídas com o NILC-Metrix.

Os resultados foram bastante satisfatórios e o trabalho foi apresentado no *workshop* JDP 2019, que aconteceu em conjunto com o STIL e BRACIS em Salvador - BA. Este artigo foi premiado como ***Best Paper*** do *workshop*.

Métodos de Clusterização para a Criação de Corpus para Rastreamento Ocular durante a Leitura de Parágrafos em Português

Sidney Evaldo Leal[1], **Sandra Maria Aluísio**[1],
Erica dos Santos Rodrigues[2],
João Marcos Munguba Vieira[3], **Elisângela Nogueira Teixeira**[3]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
[2] Departamento de Letras - Pontifícia Universidade Católica do Rio de Janeiro (PUC)
[3] Departamento de Letras Vernáculas - Universidade Federal do Ceará (UFC)

[1]sidleal@gmail.com, [1]sandra@icmc.usp.br, [2]ericasr@puc-rio.br
[3]joaomvieira@gmail.com, [3]elisteixeira@letras.ufc.br

Abstract. *This paper presents a method for automating the process of choosing a short passages subset of a large corpus to be used in psycholinguistic research that investigates reading using eye-tracking. To show the method effectiveness, a corpus with 100 short passages of 3 textual genres was used to choose a smaller corpus with 50 passages, using clustering methods and 58 metrics of several linguistic levels. The groups resulting from clustering were evaluated by similarity criteria and the method proved to be useful in supporting the selection of material to be used in psycholinguistic studies.*

Resumo. *Este trabalho apresenta um método para automatização do processo de escolha de um subconjunto de parágrafos de grandes corpora a ser utilizado em pesquisas psicolinguísticas que investigam a leitura usando rastreamento ocular. Para mostrar a efetividade do método, foi utilizado um corpus com 100 parágrafos de 3 gêneros textuais para a escolha de um corpus menor com 50 parágrafos, via métodos de clusterização, usando 58 métricas linguísticas. Os grupos resultantes da clusterização foram avaliados com base em critérios de similaridade e o método mostrou-se útil para apoiar a seleção de material para estudos psicolinguísticos.*

1. Introdução

Atualmente, corpora de rastreamento ocular são frequentemente utilizados no estudo de custos de processamento de estruturas linguísticas para, por exemplo, (i) avaliar modelos e métricas de dificuldade sintática [González-Garduño and Søgaard 2017], (ii) para melhorar ou avaliar modelos computacionais de simplificação via compressão sentencial [Klerke et al. 2016] e (iii) avaliar a qualidade da tradução automática com métricas objetivas [Klerke et al. 2015]. No entanto, existem poucos destes recursos, para um pequeno número de idiomas, por exemplo, inglês [Luke and Christianson 2017, Cop et al. 2016], inglês e francês [Kennedy et al. 2003], alemão [Kliegl et al. 2004] e russo [Laurinavichyute et al. 2018].

Para o português do Brasil, o rastreamento ocular já é utilizado há algum tempo nas pesquisas da área de Psicolinguística. Por exemplo, [Maia et al. 2007]

utilizaram para investigar o papel do processamento morfológico na identificação de palavras; [Leitão et al. 2012] utilizaram na investigação do processamento anafórico; [da Silva e Forster 2013] investigou o processamento incremental de orações relativas restritivas de objeto; e [Teixeira et al. 2014] para evidenciar o custo de resolução de pronomes nulos e plenos. Entretanto, não há nenhum grande corpus do português, publicamente disponível, com dados de rastreamento ocular de jovens adultos e com normas de previsibilidade para a tarefa de leitura silenciosa. Essa é uma grande lacuna que restringe as possibilidades de pesquisa nas áreas de Psicologia Cognitiva, Psicolinguística e Processamento de Línguas Naturais (PLN).

Pesquisas na área de Psicolinguística, especificamente de processamento da sentença, podem se beneficiar de corpora de textos autênticos linguisticamente anotados, que permitem fazer uma correlação dos tempos de leitura com fenômenos linguísticos, por exemplo os elencados abaixo:

1. complexidade estrutural do período (períodos simples vs. compostos);
2. transitividade verbal;
3. animacidade do sujeito e do objeto;
4. tipos de sentenças (ativas/passivas/relativas);
5. mecanismos de construção de relações de correferência, entre outros.

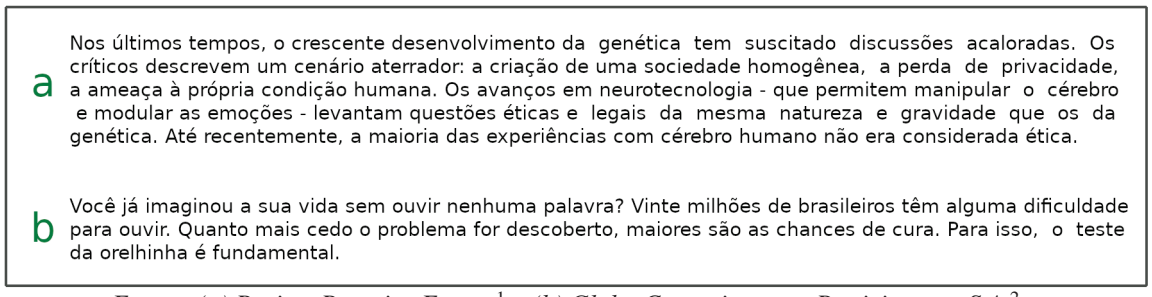
Nos experimentos psicolinguísticos, estímulos são construídos para examinar o efeito de fatores/variáveis independentes no comportamento do participante e, assim, poder investigar hipóteses de trabalho. Uma crítica muitas vezes feita a esses trabalhos diz respeito ao nível de naturalidade dos estímulos experimentais, com consequências em termos do grau de validade ecológica das pesquisas. Assim, projetos atuais utilizam textos autênticos, envolvendo diferentes gêneros textuais (jornalísticos, científicos, literários, etc.), para permitir uma avaliação da influência conjugada de um conjunto de fatores linguístico-textuais que podem afetar o processamento linguístico durante a leitura, em condições menos artificiais de realização da tarefa. Esses corpora são compilados para trazerem uma rica diversidade de fenômenos linguísticos, como, por exemplo, os cinco tipos de descrição das estruturas sintáticas elencados acima, para que se possa correlacionar a diversidade destes fenômenos com tempos de leitura, o comportamento durante a leitura e a avaliação de modelos complexos de controle dos movimentos dos olhos durante a leitura (por exemplo, o E-Z reader - modelo de processamento lexical serial [Reichle et al. 2006] - e o Swift - modelo de processamento lexical paralelo [Engbert et al. 2002]), implementados por simulações computacionais.

Entretanto, uma dificuldade para a compilação desses corpora é a anotação manual destes fenômenos, que idealmente deveria usar mais de um anotador para a avaliação do nível de concordância entre eles [Carletta 1996]. De posse deste corpus anotado com os fenômenos, se pode escolher aquele subconjunto com os atributos variados dos fenômenos linguísticos para a adequação do estudo. Por exemplo, os parágrafos da Figura 1 são do gênero de divulgação científica e jornalístico, respectivamente, e apresentam o mesmo número de sentenças, mas eles diferem em vários níveis linguísticos, por exemplo, na complexidade de seu léxico, na complexidade sintática e tamanho das sentenças, no nível de formalidade.

Dada a disponibilização pública de várias métricas automáticas para avaliação da coesão e coerência de textos escritos ou falados para a língua portuguesa

([Scarton and Aluísio 2010]; [Aluísio et al. 2016]), várias métricas, além do número de sentenças, poderiam ser analisadas para que a escolha dos parágrafos seja adequada para uma dada pesquisa em psicolinguística.

Figura 1. Parágrafos de gêneros diferentes, com mesmo número de sentenças



Fontes: (a) Revista Pesquisa Fapesp[1] *e (b) Globo Comunicação e Participações S.A.*[2]

Esta pesquisa apresenta um método para automatização do processo de escolha de um subconjunto, tomado de um grande corpus de parágrafos para pesquisas que utilizam rastreamento ocular durante a leitura destes parágrafos. Ela é parte integrante do projeto RastrOS[3].

Para mostrar a efetividade do método, utilizamos, como exemplo, um corpus com 100 parágrafos de três gêneros (jornalístico, divulgação científica e literário) (Seção 3), para a escolha de um subcorpus que traga 50 parágrafos, sendo 35 dos gêneros jornalístico e literário e 15 de divulgação científica, via métodos de clusterização, detalhados na Seção 2. O método proposto faz uso de um grande conjunto de métricas de vários níveis linguísticos (Seção 4), disponíveis publicamente na Plataforma Simpligo (https://simpligo.sidle.al/nilcmetrixdoc). Particularmente, foram escolhidas 58 métricas, agrupadas em quatro conjuntos; três destes conjuntos – tipos de sentenças (7 métricas), complexidade da estrutura sintática (22 métricas) e análise de correferência (8 métricas) foram escolhidos para modelar diretamente os três estudos de comparação dos tempos de leitura abaixo: (i) complexidade estrutural do período (períodos simples vs. compostos); (ii) tipos de sentenças (ativas/passivas/relativas); (iii) mecanismos de construção de relações de correferência, entre outros. E o conjunto denominado morfossintaxe (21 métricas) foi escolhido para modelar indiretamente os estudos sobre transitividade verbal e animacidade do sujeito e do objeto. Finalmente, a Seção 5 mostra o conjunto de agrupamentos resultante, juntamente com métodos para avaliar sua qualidade.

2. Aprendizado de Máquina e Métodos de Clusterização

Inicialmente, a área de Inteligência Artificial (IA) era considerada uma área teórica, mas nas últimas décadas com o crescimento do volume de dados e complexidade de problemas que necessitam de tratamento computacional, as técnicas de Aprendizagem de Máquina (AM) começaram a se destacar [Faceli et al. 2011]. Elas são boas ferramentas na criação

[1] `https://revistapesquisa.fapesp.br/2002/07/01/manipuladores-de-cerebros/`

[2] `https://g1.globo.com/bemestar/noticia/mais-de-20-milhoes-de-brasileiros\-tem-alguma-dificuldade-para-escutar.ghtml`

[3] Um grande corpus com medidas de RASTReamento Ocular e normas de previsibilidade durante a leitura de estudantes do ensino Superior no Brasil - `http://www.nilc.icmc.usp.br/nilc/index.php/rastros`

de hipóteses (ou funções) a partir da experiência passada, para predizer respostas ou descrever dados dos problemas que se deseja tratar. Hoje são utilizadas em tarefas tão diversas quanto reconhecimento de fala, detecção de fraudes financeiras, condução autônoma de automóveis, diagnóstico de doenças, dentre outras.

Dentro da AM, existem algoritmos que procuram identificar padrões ou tendências relevantes em conjuntos de dados sem necessidade de um elemento externo servindo de guia do aprendizado. Essas técnicas são chamadas de aprendizagem não supervisionada. Destas, as de clusterização (ou agrupamento) são de especial interesse deste trabalho, pois permitem analisar um grande número de métricas e dados, gerando sugestões de grupos por afinidade.

Os algoritmos dessas técnicas geralmente são classificados em [Faceli et al. 2011]:

- **Baseados em centróides:** Otimizam o critério de agrupamento de forma iterativa, procurando minimizar o erro quadrático ou variação dentro do *cluster*.
- **Hierárquicos:** Geram uma sequência de partições aninhadas a partir de uma matriz de proximidade. Podem ser do tipo **aglomerativo**, que começa com um grupo para cada objeto e vai combinando, ou **divisivo**, que começa com um único grupo e vai dividindo sucessivamente.
- **Baseados em densidade**: Assumem que cada *cluster* é uma região de alta densidade de objetos, separada das demais por regiões de baixa densidade.
- **Baseados em grafos:** Os dados são representados em um grafo de proximidade, no qual cada nó representa um objeto e as arestas, a similaridade ou distância.
- **Baseados em redes neurais:** Sistemas paralelos compostos de unidades simples de processamento; por exemplo o algoritmo SOM (*Self-Organizing Map*).
- **Baseados em *Grid*:** Define um *grid* (reticulado) para o espaço de dados. Muito eficiente para grandes conjuntos de objetos.

O algoritmo mais simples e mais utilizado é o K-Means[4] que utiliza técnica baseada em centróides. Nesta pesquisa, além do K-means, também foram avaliados dois outros algoritmos – o AgglomerativeClustering[5], do tipo hierárquico e o DBScan[6], baseado em densidade. Este trabalho utilizou a implementação deles na biblioteca scikit-learn em python. O DBScan não teve bons resultados no nosso cenário devido ao tamanho e distribuição do conjunto de dados.

3. Conjunto de dados separados por gêneros de texto

Foram selecionados manualmente 100 parágrafos de três gêneros e várias fontes, procurando incluir uma boa amostra para abranger o máximo dos fenômenos do português brasileiro escrito. Os parágrafos do gênero jornalístico foram obtidos de portais de notícias bem conhecidos como G1, Metro, BBC, Reuters, Terra, Estadão, Folha de São Paulo, Jornal da USP, dentre outros. Os parágrafos do gênero literário vieram de romances em domínio público. Os parágrafos de divulgação científica vieram das fontes: Revista Pesquisa Fapesp, Galileu, Aventuras na História, Época, Exame, Isto é, caderno ciência e

[4]https://scikit-learn.org/stable/modules/clustering.html#k-means
[5]https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering
[6]https://scikit-learn.org/stable/modules/clustering.html#dbscan

tecnologia do Jornal do Brasil, Mente e Cérebro, National Geographic Brasil, Piauí, Scientific American Brasil, dentre outras.

A distribuição dos parágrafos pode ser vista na Tabela 1. O objetivo deste trabalho foi selecionar, dentre os 100 parágrafos, um subconjunto, com 50 parágrafos, que mantivesse a maior variância possível dos fenômenos da língua, relacionados com os cinco estudos de comparação dos tempos de leitura, descritos na Seção 1.

Tabela 1. Distribuição dos parágrafos por gênero

Gênero	Quantidade disponível	Alvo da seleção
Jornalístico	43	35
Literário	9	
Divulgação científica	48	15
Total	100	50

4. Métricas Selecionadas

Para representar cada parágrafo do conjunto de dados, foram escolhidas 58 métricas calculadas com o apoio da ferramenta NILC-Metrix[7]. Essas métricas foram agrupadas em 4 conjuntos, resultando em 22 sobre complexidade estrutural/sintática (e.g. períodos simples vs compostos), 7 com tipos de orações (e.g.ativas/passivas, relativas), 8 com mecanismos de construção de relações de correferência e 21 relacionadas com a morfossintaxe (e.g. categorias gramaticais e flexão de substantivos e verbos); a lista completa pode ser vista na Tabela 2.

5. Método para Escolha de Subconjuntos via Clusterização e Avaliação

Após selecionar as métricas, os três conjuntos de parágrafos foram processados e foram executados diversos experimentos, buscando a melhor divisão de grupos, dentro de cada conjunto. Os melhores resultados foram obtidos utilizando a técnica chamada "Método do Cotovelo"[8] (do inglês *Elbow Method*) para encontrar o número ideal de agrupamentos. Esta técnica simula diversas divisões em número crescente de grupos e calcula as variâncias internas de cada grupo, buscando o ponto de equilíbrio [Dangeti 2017]. O gráfico com o cálculo do "cotovelo" para o gênero jornalístico pode ser visto na Figura 2, com o título "Cotovelo Kmeans", no exemplo ele indica 7 grupos ótimos, que foram plotados no gráfico com título "Grupos", com números de 0 a 6.

5.1. Redução de Dimensionalidade via Análise de Componentes Principais

Outra técnica utilizada para melhorar os resultados dos experimentos foi a Análise de Componentes Principais ou PCA (do inglês *Principal Component Analysis*). PCA é um procedimento matemático que cria novas métricas (ou variáveis) que são uma combinação linear das métricas originais e é utilizado para reduzir a dimensionalidade dos dados, sendo aplicado como uma etapa de pré-processamento antes de métodos de clusterização, como os apresentados na Seção 2. Ele permite visualizar os dados no espaço (cf. na Figura 2, os gráficos com títulos "Gênero Jornalístico" e "Textos - por índice") e também melhorar a generalização dos algoritmos [Dangeti 2017].

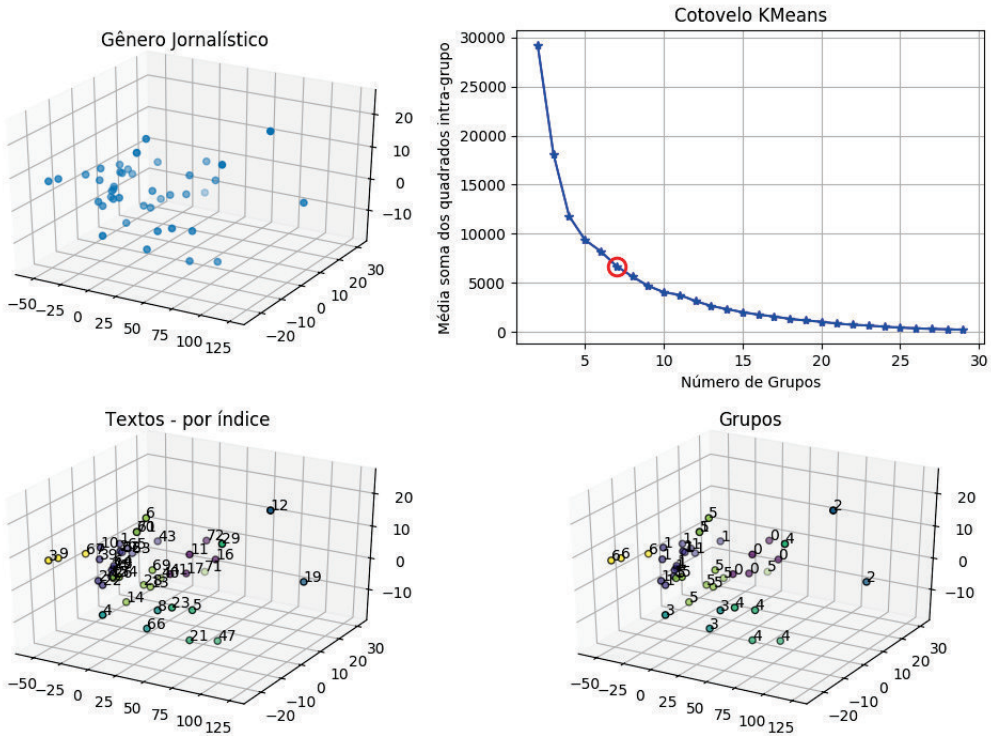
[7]https://simpligo.sidle.al/nilcmetrix

[8]A denominação vem do fato que se o gráfico relembra um braço, então o "cotovelo" (ponto de inflexão da curva) é uma boa indicação de que o modelo subjacente se encaixa melhor naquele ponto.

Tabela 2. Lista de todas as métricas utilizadas.

Nome	Descrição
	Complexidade Estrutural
words_per_sentence	Média de palavras por sentença
sentences	Quantidade de sentenças no parágrafo
words	Quantidade de palavras no parágrafo
sentence_length_max	Quantidade máxima de palavras por sentença
sentence_length_min	Quantidade mínima de palavras por sentença
sentence_length_std	Desvio padrão da quantidade de palavras por sentença
yngve	Complexidade sintática de Yngve (árvores sintáticas fora do padrão de ramificação à direita)
frazier	Complexidade sintática de Frazier (baseada na profundidade das árvores sintáticas)
dep_distance	Distância na árvore de dependência
words_before_main_verb	Quantidade média de palavras antes dos verbos principais das orações principais das sentenças
clauses_per_sentence	Quantidade média de orações por sentença
sentences_with_zero_clause	Proporção de sentenças sem verbos em relação a todas as sentenças do parágrafo
sentences_with_one_clause	Proporção de sentenças com uma oração em relação a todas as sentenças do parágrafo
sentences_with_two_clauses	Proporção de sentenças com duas orações em relação a todas as sentenças do parágrafo
sentences_with_three_clauses	Proporção de sentenças com três orações em relação a todas as sentenças do parágrafo
sentences_with_four_clauses	Proporção de sentenças com quatro orações em relação a todas as sentenças do parágrafo
sentences_with_five_clauses	Proporção de sentenças com cinco orações em relação a todas as sentenças do parágrafo
sentences_with_six_clauses	Proporção de sentenças com seis orações em relação a todas as sentenças do parágrafo
sentences_with_7+_clauses	Proporção de sentenças com sete ou mais orações em relação a todas as sentenças do parágrafo
punctuation_diversity	Proporção de *types* de pontuações em relação à quantidade de *tokens* de pontuações no parágrafo
punctuation_ratio	Proporção de sinais de pontuação em relação à quantidade de palavras do parágrafo
non_svo_ratio	Proporção de orações que não estão no formato SVO (sujeito-verbo-objeto) em relação a todas as orações
	Tipos de orações
passive_ratio	Proporção de orações na voz passiva analítica em relação à quantidade de orações do parágrafo
relative_clauses	Proporção de orações relativas em relação à quantidade de orações do parágrafo
relative_pronouns_div_ratio	Proporção de *types* de pronomes relativos em relação à quantidade de *tokens* de pronomes relativos
subordinate_clauses	Proporção de orações subordinadas pela quantidade de orações do parágrafo
infinite_subordinate_clauses	Proporção de orações subordinadas reduzidas pela quantidade de orações do texto
coordinate_conj_per_clauses	Proporção de conjunções coordenativas em relação a todas as orações do texto
apposition_per_clause	Quantidade média de apostos por oração do texto
	Correferência
adjacent_refs	Média das proporções de candidatos a referentes na sentença anterior em relação aos pronomes pessoais do caso reto nas sentenças
anaphoric_refs	Média das proporções de candidatos a referentes nas cinco sentenças anteriores em relação aos pronomes anafóricos das sentenças
arg_ovl	Quantidade média de referentes que se repetem nos pares de sentenças do texto
adj_arg_ovl	Quantidade média de referentes que se repetem nos pares de sentenças adjacentes
stem_ovl	Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças
adj_stem_ovl	Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças adjacentes
adj_cw_ovl	Quantidade média de palavras de conteúdo que se repetem nos pares de sentenças adjacentes
coreference_pronoum_ratio	Média de candidatos a referente, na sentença anterior, por pronome anafórico do caso reto
	Morfossintáticas
verbs	Proporção de verbos em relação à quantidade de palavras do parágrafo
verbs_max	Proporção máxima de verbos por palavras em relação à quantidade de palavras das sentenças
verbs_min	Proporção mínima de verbos por palavras em relação à quantidade de palavras das sentenças
verbs_standard_deviation	Desvio padrão das proporções entre verbos e a quantidade de palavras das sentenças
verbal_time_moods_diversity	Quantidade de diferentes tempos-modos verbais que ocorrem no texto
adverbs	Proporção de advérbios em relação à quantidade de palavras do texto
adverbs_max	Proporção máxima de advérbios em relação à quantidade de palavras das sentenças
adverbs_min	Proporção mínima de advérbios em relação à quantidade de palavras das sentenças
adverbs_standard_deviation	Desvio padrão das proporções entre advérbios e a quantidade de palavras das sentenças
noun_ratio	Proporção de substantivos em relação à quantidade de palavras do parágrafo
nouns_max	Proporção máxima de substantivos em relação à quantidade de palavras das sentenças
nouns_min	Proporção mínima de substantivos em relação à quantidade de palavras das sentenças
nouns_standard_deviation	Desvio padrão das proporções entre substantivos e a quantidade de palavras das sentenças
pronoun_ratio	Proporção de pronomes em relação à quantidade de palavras do parágrafo
pronouns_max	Proporção máxima de pronomes em relação à quantidade de palavras das sentenças
pronouns_min	Proporção mínima de pronomes em relação à quantidade de palavras das sentenças
pronouns_standard_deviation	Desvio padrão das proporções entre pronomes e a quantidade de palavras das sentenças
adjective_ratio	Proporção de adjetivos em relação à quantidade de palavras do parágrafo
adjectives_standard_deviation	Desvio padrão das proporções entre adjetivos e a quantidade de palavras das sentenças
preposition_diversity	Proporção de *types* de preposições em relação à quantidade de *tokens* de preposições
syllables_per_content_word	Quantidade média de sílabas por palavra no parágrafo

Figura 2. Visualização dos parágrafos e grupos do gênero jornalístico.



5.2. Avaliação do Método

Neste trabalho, os agrupamentos foram gerados utilizando o algoritmo K-Means e AgglomerativeClustering, em seguida foram calculadas as medidas de silhueta (*Silhouette*) para os grupos e *V-Measure* [Rosenberg and Hirschberg 2007] para medir a concordância entre os dois algoritmos. Os resultados podem ser vistos na Tabela 3. A silhueta mede o quão similar é um objeto em seu grupo, em comparação com os demais grupos, e varia de -1 a +1. No nosso cenário, o valor médio 0,38 pode ser considerado bom, tendo em vista que os parágrafos já possuem certa similaridade pela seleção prévia (parágrafos curtos). Já a *V-Measure* obtida reforça que os algoritmos concordam com a divisão dos objetos nos grupos em mais de 90%. A Homogeneidade (*Homogeneity*) avalia se cada grupo contém somente membros de uma única classe, a Completude (*Completeness*) avalia se todos os membros de uma classe estão no mesmo grupo, sendo a *V-Measure* a média harmônica entre elas duas.

Tabela 3. Resultados

Gênero	Número de Grupos	Itens por grupo Med (Min-Max)	K-Means Silhouette	Agglomerative Silhouette	Homogeneity	Completeness	V-Measure
Jornalístico	7	7 (2-14)	0,38	0,38	0,93	0,92	0,92
Literário	4	2 (1-4)	0,39	0,39	1,00	1,00	1,00
Divulgação científica	7	5 (2-15)	0,38	0,35	0,82	0,79	0,81
Média	6	11 (1,6-4,6)	0,38	0,37	0,92	0,90	0,91

A Tabela 1 mostra os alvos de seleção para montar o corpus de 50 parágrafos a partir do corpus inicial de 100 parágrafos (35 parágrafos dos gêneros jornalísticos e literários

e 15 do gênero de divulgação). O experimento realizado selecionou 7, 4 e 7 grupos (cf. Tabela 3). Assim, o trabalho final para montar o corpus de pesquisa pode ser realizado pela escolha manual, apoiada por algum critério importante para a pesquisa, como o tamanho dos parágrafos. No exemplo deste artigo, de 11 grupos serão selecionados 35 parágrafos e de 7 grupos de divulgação, os 15 parágrafos finais.

6. Considerações finais

A possibilidade de selecionar textos com características linguísticas específicas pode ser muito relevante em estudos de natureza experimental na área de psicolinguística. A contribuição desta pesquisa com um método de clusterização que atende esse propósito se mostrou bastante eficiente, pois o conjunto de métricas automáticas ajudou a agrupar parágrafos com características semelhantes, realizando uma anotação dirigida ao agrupamento. Com a lista dos parágrafos agrupados, a tarefa de selecionar manualmente a amostra final tornou-se bem mais simples e informada. É possível selecionar um número de itens de cada grupo, de forma aleatória, ou com alguma forma de ranqueamento (maiores ou menores parágrafos, por exemplo).

Acreditamos que a utilização dos recursos de PLN e Aprendizagem de Máquina não supervisionada podem ajudar bastante em tarefas trabalhosas como a anotação manual dos fenômenos da língua em um corpus. Como continuação deste trabalho, os autores pretendem disponibilizar uma ferramenta web com o método apresentado, para automatizar a análise e permitir que outros pesquisadores consigam replicar o experimento sem esforço de codificação.

7. Agradecimentos

À Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP, processo número 2019/09807-0, pelo apoio financeiro.

Referências

Aluísio, S. M., Cunha, A., Toledo, C., and Scarton, C. (2016). Computational tool for automated language production analysis aimed at dementia diagnosis. In *International Conference on Computational Processing of the Portuguese Language, Demonstration Session*.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, 22(2):249–254.

Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2016). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49:602–615.

da Silva e Forster, R. A. M. (2013). *Aspectos do Processamento de Orações Relativas: Antecipação de Referentes e Integração de Informação Contextual*. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio).

Dangeti, P. (2017). *Statistics for Machine Learning*. Packt Publishing, E-Book.

Engbert, R., Longtin, A., and Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, 42(5):621 – 636.

Faceli, K., Lorena, A. C., Gama, J., and de Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina*. LTC - Livros Técnicos e Científicos, Rio de Janeiro.

González-Garduño, A. V. and Søgaard, A. (2017). Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, Copenhagen, Denmark. Association for Computational Linguistics.

Kennedy, A., Hill, R., and Pynte, J. (2003). The dundee corpus. *Proceedings of the 12th European conference on eye movement*.

Klerke, S., Castilho, S., Barrett, M., and Søgaard, A. (2015). Reading metrics for estimating task efficiency with mt output. In *Conference on Empirical Methods in Natural Language Processing*, pages 6–13. Association for Computational Linguistics.

Klerke, S., Goldberg, Y., and Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Kliegl, R., Grabner, E., Rolfs, M., and Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*, pages 262–284.

Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K., and Klieg, R. (2018). Russian sentence corpus: Benchmark measures of eye movements in reading in russian. *Behavior Research Methods*, pages 1–18.

Leitão, M. M., Ribeiro, A. J. C., and Maia, M. (2012). Penalidade do nome repetido e rastreamento ocular em português brasileiro. *Revista LinguíStica*, v8 n2.

Luke, S. G. and Christianson, K. (2017). The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*.

Maia, M., Lemle, M., and França, A. I. (2007). Efeito stroop e rastreamento ocular no processamento de palavras. *Ciências e Cognição 2007*, 12:02–17.

Reichle, E. D., Pollatsek, A., and Rayner, K. (2006). E-z reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cogn. Syst. Res.*, 7(1):4–22.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL)*, pages 410–420.

Scarton, C. E. and Aluísio, S. M. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61.

Teixeira, E. N., Fonseca, M. C. M., and Soares, M. E. (2014). Resolução do pronome nulo em português brasileiro: Evidência de movimentação ocular. *VEREDAS: Sintaxe das Línguas Brasileiras*, 18.

4.2 Construção de métodos de previsibilidade semântica

Título:	***Evaluating Semantic Similarity Methods to build Semantic Predictability Norms of Reading Data***
Autores:	**Sidney Evaldo Leal, Edresson Casanova, Gustavo Henrique Paetzold e Sandra Maria Aluísio**
Ano:	**2021**
Conferência:	**The twenty-fourth International Conference on Text, Speech and Dialogue (TSD 2021)**
Situação:	**Aceito para publicação.**

Duas das colunas da planilha disponibilizada pelo córpus Provo (principal inspiração do RastrOS) são a similaridade semântica entre a palavra-alvo e o contexto anterior e entre a palavra-alvo e a palavra-resposta do teste Cloze, utilizando o método Latent Semantic Analysis (LSA) (Ver Seção 2.4).

Em vez de simplesmente replicar a solução com LSA, aproveitamos o cenário do projeto RastrOS para avaliar também modelos mais recentes para similaridade semântica, como as Word Embeddings estáticas disponibilizadas para o português e o modelo BERT, como exemplo de Word Embeddings contextualizadas (Ver Seção 2.4). Este artigo apresenta essa investigação, e também o benchmark criado para *Sentence Completion* que é uma tarefa similar à tarefa específica de similaridade semântica para normas de previsibilidade. Para o córpus RastrOS, foi criada também uma terceira medida de similaridade semântica entre a palavra-resposta e o contexto prévio. O artigo foi submetido para a conferência TSD 2021[1] e aceito para publicação em 24.05.2021, após revisão de pares.

[1] <https://www.kiv.zcu.cz/tsd2021/index.php>

Evaluating Semantic Similarity Methods to Build Semantic Predictability Norms of Reading Data

Sidney Leal[1], Edresson Casanova[1], Gustavo Paetzold[2], and Sandra Aluísio[1]

[1] Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (USP)
{sidleal,edresson}@usp.br / {sandra}@icmc.usp.br

[2] Universidade Tecnológica Federal do Paraná (UTFPR) - Campus Toledo
ghpaetzold@outlook.com

Abstract. Predictability corpora built via Cloze task generally accompany eye-tracking data for the study of processing costs of linguistic structures in tasks of reading for comprehension. Two semantic measures are commonly calculated to evaluate expectations about forthcoming words: (i) the semantic fit of the target word with the previous context of a sentence, and (ii) semantic similarity scores that represent the semantic similarity between the target word and Cloze task responses for it. For Brazilian Portuguese (BP), there was no large eye-tracking corpora with predictability norms. The goal of this paper is to present a method to calculate the two semantic measures used in the first BP corpus of eye movements during silent reading of short paragraphs by undergraduate students. The method was informed by a large evaluation of both static and contextualized word embeddings, trained on large corpora of texts. Here, we make publicly available: (i) a BP corpus for a sentence-completion task to evaluate semantic similarity, (ii) a new methodology to build this corpus based on the scores of Cloze data taken from our project, and (iii) a hybrid method to compute the two semantic measures in order to build predictability corpora in BP.

Keywords: Semantic Predictability, Cloze Test, Language Models

1 Introduction

Predictability is defined as how much a word can be predicted based on its previous context, i.e., as the probability of knowing a word before reading it. According to Bianchi et al. (2020) [1] the task of predictability of a given word is fundamental to analyse the behaviour that humans process information during reading. Predictability corpora are commonly built via Cloze task and generally accompany eye-tracking data for the study of processing costs of linguistic structures in tasks of reading for comprehension [4]. However, only few resources exist, for a small number of languages, for example, English [16], English and French [12], and German [13]. For Brazilian Portuguese, there was no large eye-tracking corpus with predictability norms such as those cited above. In order to fulfill this gap, a large corpus with eye movements during silent reading by undergraduate students was built. The RastrOS corpus is composed of short paragraphs of authentic texts in Brazilian Portuguese (BP) taken from different textual genres. Thus, it allows an assessment of the combined influence of a set of linguistic-textual factors that can affect linguistic processing during reading, in less artificial conditions for carrying out the task.

In the Provo project [15], using Cloze probabilities as predictors of reading via eye movement patterns, Luke and Christianson (2016) found that for the English language, highly predictable words were quite rare (5% of content words and 20% of function words), and most words had a more-expected candidate consistently available from context even when word identity was not. A highly predictable word is the one which has a high Cloze probability. In Provo, words with .67 or higher probability of being completed by a specific single word were considered highly predictable. In addition, they found that predictability of partial semantic and syntactic information influenced reading times above and beyond Cloze scores, suggesting that predictions generated from context are most often graded rather than full lexical prediction.

The RastrOS corpus was inspired by the Provo project methodology and, therefore, it has three types of Cloze scores: full-orthographic form, PoS and inflectional properties, and semantic predictability scores for all 2494 words in the 50 text paragraphs. In this study, we deal with the semantic predictability, which is divided in two measures: (i) the semantic predictability between the target word of a sentence and the participant's responses, to inform predictability studies that evaluate the role of knowledge of the paradigmatic structure of language and (ii) the semantic fit of the target word with the previous context of a sentence, used in eye-tracking analyses on the knowledge of the language's syntagmatic structure. Although the Provo project has chosen to use Latent Semantic Analysis (LSA) [14] to provide both semantic scores, for BP we have decided to evaluate word embeddings models from two families of methods: (i) those that work with a co-occurrence word matrix, such as LSA, and (ii) predictive methods such as Word2Vec [17] and FastText [2]. We also evaluate one contextualized word representation model — BERT [6], recently trained for BP [23]. Moreover, in the Provo project, the implementation of the semantic predictability between the target word and the participant's responses is made by using the cosine of the angle between the vectors of responses and targets which is a very used distance measure to quantify semantic similarity between words in distributional semantics models. However, it is important to evaluate different distributional semantics models than those that combine single word vectors into a sentence context vector using the sum or average of the words' individual embeddings (see [8]) to implement the other semantic measure — semantic fit of the target word with the previous context of a sentence. This decision is important and was pursued to inform Psycholinguistic studies in BP and to allow a great diversity of uses for the RastrOS corpus. To the best of our knowledge, there are no publicly available sentence completion task corpora for BP. Therefore, we propose to evaluate semantic similarity using a new dataset for sentence completion task [30, 31], which was created in this study. As the dataset of the RastrOS project[3] is composed of sentences from three different genres, we decided to use the same dataset to evaluate several similarity methods.

The remainder of this paper is organized as follows. Section 2 reviews approaches to sentence completion task. Section 3 describes an initial version of the RastrOS corpus used in the evaluations of semantic similarity methods for building the semantic predictability scores of our final corpus. Section 4 presents the process of building our Sentence Completion dataset using Cloze data. Section 5 presents the evaluation of static

[3] http://www.nilc.icmc.usp.br/nilc/index.php/rastros

and contextualized word embeddings using the new dataset created, human benchmark performance, and also the hybrid method to calculate semantic scores proposed for the RastrOS project.

2 Review on Sentence Completion Task and Datasets

The Sentence Completion task was defined in 2011 [30] and consists in, given a sentence with a gap, guessing what word or phrase would best fit the gap.

Early approaches to Sentence Completion employ classic language modelling techniques, such as n-gram language models [29], backoff and class-based models [31, 3]. There are also examples of strategies that rely on hand-crafted metrics [27] and word similarity metrics that measure the similarity between each possible answer and the words in the sentence to be completed [31]. More recently, different variants of Recurrent Neural Networks (RNN) models have been used as an alternative to hand-crafted metrics and features. [24] addresses the task using Recurrent Memory Networks, of which the structure is more easily interpretable than regular RNNs. Mirowski and Vlachos (2015) [18] also finds that incorporating dependency relations throughout the training process also improves on the performance of regular RNNs for the task. Current state-of-the-art approaches rely on BERT models. Devlin et al (2019) achieved impressive results in a large Sentence Completion dataset with 113,000 instances [6]. Using another BERT variant that operates at word-level, Park and Park (2020) achieves even more impressive results in two separate Sentence Completion datasets [20].

Because they are relatively easy to obtain, Sentence Completion datasets are abundant for the English language. Some of them are the Microsoft Research Sentence Completion (MRSC) dataset [30], composed of questions from Project Gutenberg novels, and the SAT dataset [27], composed of questions from practice SAT exams. One can also find datasets with questions taken from children's books [11], online newspaper articles [10], summaries [5], mundane everyday stories [19], and college/high-school entrance exams [28]. Nevertheless, datasets for languages other than English are much scarce. The TOPIK dataset [20] for Korean is the only example we could find.

3 The initial Version of the RastrOS Corpus

The RastrOS project is multicentric and has two phases of data collection with participants. First, cloze scores were collected via an online survey for each word, except the first one, in our 50-paragraph corpus. Each participant read 5 paragraphs from the pool of 50 ones. Second, these same paragraphs were presented to a different set of participants, who read them while their eye movements were being tracked. In the RastrOS project, a high-accuracy eye-tracker was used — the EyeLink 1000 Desktop Mount.

Cloze. For this study we are using project's initial data, without applying exclusion criteria, to inform the implementation of semantic measures of the final project. In total, data from 314 people (172 females) were included. Participants' ages ranged from 17 to 73 years (Average: 22; SD: 7.4); 309 participants filled in the age field. Participants were university undergraduate students and approximately 5.71% were graduated or

initiating postgraduate education. Participants completed an online survey administered through a web-based software developed in the RastrOS project. Participants first answered a few demographic questions (gender, age, education level, language history), then proceeded to complete the main body of the survey. For the first question of a paragraph, only the first word in the text was visible. For each question of the survey, participants were instructed to fill each gap that, in their assessment, allows them to follow the linguistic material previously read. As soon as they typed a word, the text with the word expected in the gap was displayed and a new gap appeared. For each new sentence in the paragraph the initial word was displayed. This same procedure was repeated until the paragraph was completed. Each participant was assigned to complete five random paragraphs, giving responses for an average of 245 different words. For each word in each text, an average of 26 participants provided a response (range: 20 - 33). Responses were edited for spelling. When a response contained multiple words, the first word was coded.

Data. Fifty short passages were taken from a variety of sources, including online news articles, popular science magazines, and literary texts. These 50 different text paragraphs are composed of 120 sentences and their 257 segmented clauses (see Table 1). We also annotated clauses formed by non-finite verbs or small non-finite verb clauses together with finite clauses of our corpus, in order to indicate the number of non-finite verb clauses in the corpus. Moreover, we segmented appositives and parenthetical expressions, adding them to the total of different clause types. It is important to note that several small clauses of several types were kept together with others in the process of segmentation, although their types were annotated. The paragraphs were an average of 49 words long (range: 36–70) and contained 2 sentences on average (range: 1–5). Sentences were on average 20 words long (range: 3–60). Across all texts, there were 2494 words total (2831 tokens including punctuation), including 1237 unique word forms. Table 1 summarizes some statistics of our corpus.

Table 1. Statistics of the RastrOS Corpus.

Number of Paragraphs	50
Number of Sentences	120
Number of Clauses	257
Independent clauses (simple clauses)	27
Main clauses (in complex or compound-complex sentences)	72
Coordinate clauses	55
Subordinate Nominal and Adverbial Clauses (dependent clauses)	73
Nonrestrictive and restrictive relative Causes (dependent clauses)	40
Non-finite Clauses	26
Appositives and parenthetical expressions	17
Number of Words	2494
Number of Types	1237

The words were tagged for part of speech (PoS) using the nlpnet tagger [7]. In total, the 50 passages contained 1438 content words (186 adjectives, 119 adverbs, 756 nouns, and 377 verbs) and 992 function words (143 conjunctions, 234 determiners, 445 prepositions, 170 pronouns), and 67 other words and symbols. In addition, inflectional

information was also coded for the words within each class where appropriate, using the Unitex-PB Dictionary [25]. Nouns and adjectives were coded for gender and number and verbs were coded for person. Target words ranged from 1 to 18 letters (or hyphen) long (M: 4). The frequency of each target word was obtained from the brWaC corpus [26], a large web corpus for Brazilian Portuguese. Word frequencies ranged from 0 to 120 (M: 17) words per million. Transitional probabilities, which have been implicated as a possible source of contextual information in reading, were computed from the brWaC corpus by dividing the frequency of the collocation of the target word and the previous word (e.g., the frequency of "I agree") by the frequency of the previous word alone (e.g., the frequency of "I" in all contexts). This provides an estimate of the predictability of the upcoming word, given that the reader has just encountered the previous word.

4 Building the Sentence Completion Dataset via Cloze Data

In this work, the list of alternatives for the target words of a given sentence was obtained from the responses of the students who participated in the Cloze test of the RastrOS project. The Cloze test guarantees a left-context to fill each new gap in a given sentence, using the student's language model. In paragraphs with more than one sentence, Cloze's final sentences gain an even greater context than is provided by the preceding sentences, but it is still a local context; only the last gap in the last sentence of a paragraph provides the overall context for the student to fill in the missing word. The number of sentences in the corpus used by the Cloze test of the RastrOS project is smaller than the one of MRSC dataset which has 1040 sentences, and although small it is a challenging dataset of 113 sentences. Seven sentences were discarded as their contexts were too restrictive for presenting several good distractors. The automatic procedure presented in Section 4.1 generates a report to help the human judgement about the four distractors, which is described in Section 4.2.

4.1 Automatically Generating Alternates

For each sentence in the dataset, the answers of the cloze test were used as candidate words, following the rules below.

1. Only content target words were selected, from the middle of the sentence to its end, as this interval provides a better decision context for the participants of Cloze. For example, Sentence 1 in Table 2 has 15 words, therefore starting from the word "foi", the content words were presented with their frequency in the brWaC. In addition, two metrics that help to analyze the scenario of participant's choice are also presented: (i) Orthographic Match (Match), which is the total human hits divided by total human responses; and (ii) Certainty (Cert), the amount of modal response divided by total human responses (humans may have been wrong, but they agreed with the error);
2. The responses of the participants were grouped and ordered from the most frequent to the least frequent;

3. Application of a filter to remove responses with grammatical classes (PoS tag) other than the one of the target word;
4. In the second filter, all synonyms of the target word were removed, using a BP thesaurus (sinonimos.com.br) as a resource;
5. Then, the candidate words and their answers were ranked by frequency (from the rarest to the most frequent);
6. The generated report shows the sentence paragraphs to facilitate the initial human assessment;
7. The report also shows synonyms from the chosen thesaurus sinonimos.com.br of each alternative. The synonyms are ranked by frequency in brWaC to facilitate the choice of the less frequent ones.

Table 2. Excerpt of the report generated by the rules in Section 4.1. It shows a sentence of the dataset and 5 target words for human evaluation, ordered by frequency in brWaC. Next to each target word there is its PoS, and the values of the Freq, Match and Cert metrics, followed by the list of responses from the Cloze test, already filtered by PoS and synonyms. The target "ciência" is presented in bold as it was chosen by the rules presented in Section 4.2.

Sentence 1: *A invenção do zero pelos humanos foi crucial para a matemática e a ciência modernas.* (The invention of zero by humans was crucial to modern mathematics and science.)				
crucial (ADJ)	F:28041	M:0	C:0.4	[responsável]
modernas (ADJ)	F:38839	M:0	C:0.09	[moderna, exata, comum]
matemática (N)	F:66257	M:0.25	C:0.25	[evolução, humanidade, compreensão, ciência, história]
ciência (N)	F:234713	M:0.03	C:0.28	[física, invenção, geometria, abstração, sociedade, calculadora]
foi (N)	F:8232437	M:0.68	C:0.68	[é, trouxe, contribuiu]

4.2 Human Judgment

The initial human choice is related to the triplet: **target word (PoS) // Metrics of the target word (Freq, Match, Cert) // list of alternatives**, which must meet the following 8 rules, which are based on the rules used in the MRSC dataset.

1. Give preference to the rarest target word, with a long list of alternatives.
2. Use the metrics Match and Cert to generate a more challenging dataset.
3. The list of alternatives must contain options that are grammatically correct.
4. Choose alternatives that require some analysis to arrive at the answer.
5. Alternatives that require understanding properties of entities that are mentioned in the sentence are desirable.
6. Dictionary use is encouraged, if necessary.
7. The correct answer should always be a significantly better fit for that sentence than each of the four distractors; it should be possible to write down an explanation as to why the correct answer is the correct answer, that would persuade most reasonable people.
8. The partial list of candidates for a given sentence may include synonyms (taken from sinonimos.com.br) of the candidates if it is necessary to complete the list when the number of alternates is less than four.

Following the Rule 1 for the sentence in Table 2, the words “matemática/Maths” and “ciência/Science” are strong candidates, as they have 8 and 9 alternatives, respectively. With Rule 2, we can see that “matemática” is a relatively predictable word (Match 0.25) and many students agreed on the prediction (Cert 0.25), so for a challenging dataset it would be better not to use it. “ciência” is a better choice, since few students got it right (0.03) and many chose it wrong, but they agreed on the choice that the answer was “física/Physics” (0.28). Checking Rule 3: In the list of alternatives of the word “ciência”, all alternatives generate a grammatically correct sentence. Rule 4 helps to exclude the alternatives “física” and “computação/Computer Science”, as they are large areas of “ciência”; it also helps to exclude“abstração/abstraction” and “calculadora/calculator”, as they are words closely related to the target word “ciência”, although not synonyms. Words related to the target word via semantic similarity that are correct in the global context of the sentence must be eliminated. Applying Rule 5, we understand that the alternatives “geometria/Geometry” and “álgebra/Algebra” require analysis of the part-whole relation, since they are subjects of “matemática”, so they are chosen; “sociedade/society” remains a viable alternative. Applying Rule 7, the alternative “sociedade” is eliminated, as it fits the sentence very well. It is also noted that “geometria” and “álgebra” are not correct answers, as they are disciplines of “matemática”, a word that appears previously in the sentence. Following the seven rules above, the alternatives chosen for “ciência” were: [geometria, língua, álgebra]. Applying Rule 8, another synonym of “língua/language” was chosen. Thus, for the sentence in Table 2, the final list is: [geometria, álgebra, língua, fala]/[geometry, algebra, language, speech] and the final question is presented in Figure 1.

A invenção do zero pelos humanos foi crucial
para a matemática e a _____ modernas.
(A) geometria
(B) língua
(C) fala
(D) ciência
(E) álgebra

Fig. 1. An example question included in the sentence completion dataset via Cloze data.

5 Evaluation Results and the Hybrid Method Proposed

Corpora and Models Evaluated. As the sentence completion task involves training a method on a large corpus of plain text to then try to predict the missing words in the test set, in this work we are using pretrained static word embeddings available in two large word embeddings repository for BP, described below.

1. NILC Embeddings repository[4] contains models trained on a large corpus composed of 17 Portuguese subcorpora, including literary works in public domain and a collection of classic fiction books, totalizing 1.3 billion tokens. The models were generated using Word2Vec, FastText, Wang2Vec (all three available in both Skip-gram and CBOW versions) and Glove. The models are available in 50, 100, 300, 600 and 1000 dimensions [9].

[4] http://www.nilc.icmc.usp.br/embeddings

2. PUCRS BBP Embeddings repository[5] uses a corpus of 4.9 billion tokens, composed of three publicly available resources: brWaC [26], a large multi-domain web corpus for Brazilian Portuguese, with 2.7 billion tokens, BlogSet-BR [21], a Brazilian Portuguese corpus containing 2.1 billions words extracted from 7.4 millions posts over 808 thousand different Brazilian blogs, and a dump of Wikipedia articles in Brazilian Portuguese from 2019-03-01. There are four models available: 300-d Skip-gram and CBOW trained with Word2Vec and 300-d Skip-gram and CBOW trained with FastText [22].

$\text{BERT}_{\text{Base}}$ and $\text{BERT}_{\text{Large}}$ and LSA models were trained on part of BBP Embeddings corpus — the brWaC corpus, cited above. LSA model was trained in this work with 300 dimensions. For static word representation vectors, the Word2Vec [17] and FastText [2] models were used, both with 300 dimensions and trained with the CBOW architecture, that predicts a word given a context, on both BBP and NILC corpora.

Following Zweig et al (2012), we used Total Word Similarity method to evaluate the static word embeddings models presented in Table 3. First the embeddings of the candidate answers (four distractors) and the correct answer for the sentence were calculated, then the embeddings of the rest of the words in the sentence were obtained. For each candidate answer, the total similarity between it and the words of the rest of the sentence was calculated using the cosine distance (using the sum of word vectors). The candidate answer with the shortest distance was chosen.

Table 3. Evaluation results (accuracy) on the Sentence Completion dataset of the RastrOS project.

Method	Corpus	Hits	Accuracy
LSA	brWaC	26/113	23.01%
Word2Vec	BBP	36/113	31.86%
Word2Vec	NILC	32/113	28.32%
FastText	BBP	39/113	34.51%
FastText	NILC	26/113	23,01%
$\text{BERT}_{\text{Base}}$	brWaC	65/113	57.52%
$\text{BERT}_{\text{Large}}$	brWaC	66/113	**58.41%**

To use BERT trained in Portuguese by Souza et al (2020), we chose the model in the Masked Language Model task, where the objective is to predict the masked word. We proposed the following method. First we pass the sentence over to BERT and mask the word that we must complete. Then, we activate the model to obtain the predicted list of tokens from the BERT's trained model vocabulary, ordered by probability. Finally, we choose the highest probability among the five alternatives of the sentence completion dataset of the RastrOS project.

The BERT Large model achieved the best results with **58.41%** accuracy. Although the NILC embeddings corpus contains a wide variety of textual genres, including the literary genre, the models trained on it performed below the same models trained on the BBP corpus. Finally, although the BERT models were trained on a smaller corpus than the largest corpus we used (BBP), they stood out for the task, in the dataset created.

[5] `https://www.inf.pucrs.br/linatural/wordpress/pucrs-bbp-embeddings/`

Human Performance. To provide human benchmark performance, we asked six Brazilian native speaking graduate researchers to answer the questions on a small subset with 20 questions. The average accuracy was 76% (range: 60% - 85%) and Kappa Fleiss value of 0.59. Zweig and Burges (2011) cite a human performance of 91% (an unaffiliated human answer) on the MRSC dataset in a random subset of 100 questions and Zweig et al (2012) cite a human performance for high-scholar's (six of them) of 87% accuracy and for graduate students (five students) of 95%, on a dev-test with 95 questions.

Hybrid Method to Calculate Semantic Scores. In the RastrOS project, we calculated the semantic similarity scores — the semantic fit of the target word with the previous context of a sentence, and semantic similarity scores that represent the semantic similarity between the target word and Cloze task responses — differently than Provo project. We calculated both semantic measures taking the previous context in consideration in order to take advantage of BERT results in our sentence completion dataset. We used BERT models trained in Portuguese by Souza et al (2020) in the task of Masked Language Model, where the objective is to predict the masked word. We proposed the following method to calculate the similarity between two words (target and response predicted) given a context:

1. We pass a sentence to BERT and mask the target word. For example, for the Sentence 1 of our dataset: *A invenção do zero pelos humanos foi crucial para a matemática e a ciência modernas.*, in the task semantic fit we use the context, the target word and the highest probability response predicted by BERT (Task 1). To calculate the semantic similarity between the target word and Cloze task responses we use the context, the target word and a student' response, each time (Task 2) (see Table 4).
2. Then, we activate the model obtaining the probability **p** of the prediction for each vocabulary token of the BERT model.
3. Using these probabilities, for each task shown above we calculate the distance between two possible candidates using the following equation: $dist(p1, p2) = \|p1 - p2\|$, considering $p1$ and $p2$ the probabilities of predicted model for candidate 1 and 2, respectively.
4. After calculating these values for each of the instances of our corpus, we normalize the values using the following equation: $s(p1, p2, max\_dist) = 1 - (dist(p1, p2)/max\_dist)$, considering max_dist as the largest of the distances obtained for the given task. Thus, obtaining a value between 0 and 1 that shows how similar two words are given a previous context; we consider 1 the most similar.

However, BERT has a limited vocabulary since low frequency words (rare words) of the training corpus are grouped in the token *UNK* during the training phase. Therefore, the token *UNK* brings the inflated probability of a group of words. The results of this fact are that our proposed method does not provide good results for about 29% of words in the dataset of sentence completion when using the model $\text{BERT}_{\text{Large}}$, for example. To solve this limitation, we proposed a hybrid method. For those words that are not present in the dictionary of the trained BERT model, the similarity is calculated using

Table 4. Example using BERT model to predict the values for the two tasks addressed in this study. P1 is the first word, P2 is the second word, O is the Model output before normalization and N after normalization.

Context: **[A invenção do zero pelos humanos foi crucial para a]**	
Task 1: Semantic Fit	P1: **ciência** (Target Word) P2: **física** (BERT Prediction) O: 0.53 N: **0.97**
Task 2: Semantic Similarity	P1: **ciência** (Target Word) P2: **geometria** (Student response) O: 3.37 N: **0.83**

the cosine distance of our best static embedding model, evaluated in the dataset of the Sentence Completion task (see Table 3): the FastText trained on the BBP corpus.

6 Conclusions

In this paper, we presented the Cloze Task in the RastrOS project, which has Cloze scores for the full-orthographic form, PoS and inflectional properties and semantic scores for all 2494 words in the 50 text paragraphs. Here, we contributed with: (i) the sentence completion task for BP making available a new **test set** with 113 questions[6]; (ii) a new methodology to build the test set for the sentence completion task, using the scores of Cloze data; and (iii) a hybrid method to calculate the semantic scores based on an evaluation of static and contextualized word embeddings. Although there are several approaches to create a sentence completion dataset (see Section 2) we took advantage of Cloze results, using the most difficult answers for students. While our test dataset is small compared with others for English, it is very challenging, given that average human accuracy was 76% in comparison to MRSC's 91%. The procedure created to produce this new test set can be applied in other eye-tracking projects in BP which provide Cloze scores for all the words of a sentence. As far we know, there is no other study evaluating the proposed approach here to build semantic predictability norms, using contextualized word representations like BERT, ELMo, GPT-2. Our work is the most similar to Bianchi et al. (2020) that evaluate different word embeddings models (LSA, Word2Vec, FastText) and N-gram-based language models to estimate how humans predict words in a Cloze test in order to understand eye movements in long Spanish texts. In contrast to Bianchi et al. (2020), we gave a step further regarding the choice of pretrained models as we evaluated a contextualized word representation (BERT) which was important to implement the semantic fit in our work. We understand that BERT is a better fit as this measure requires the calculation of the distance between the embedding of the target word and its entire previous context. We hope that our study opens a bridge between Psycholinguistics and Natural Language Processing to inform new projects in BP.

At the end of the RastrOS project, in July 2021, both sets of data (predictability norms and eye-tracking data) will be publicly available in the OSF platform as part of the RastrOS Corpus.

[6] Dataset and evaluation sources are available at: `https://github.com/sidleal/TSD2021`

References

1. Bianchi, B., Monzón, G.B., Ferrer, L., Slezak, D.F., Shalom, D.E., Kamienkowski, J.E.: Human and computer estimations of predictability of words in written language. Scientific Reports **10**(4396), 1–11 (2020)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
3. Correia, R., Baptista, J., Eskenazi, M., Mamede, N.: Automatic generation of cloze question stems. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (eds.) Proceedings of the Computational Processing of the Portuguese Language - 10th International Conference, PROPOR. Lecture Notes in Computer Science, vol. 7243, pp. 168–178. Springer (2012)
4. Demberg, V., Keller, F.: Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition **109**(2), 192–210 (2008)
5. Deutsch, D., Roth, D.: Summary cloze: A new task for content selection in topic-focused summarization. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3711–3720 (2019)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423
7. Fonseca, E.F., Garcia Rosa, J.L., Aluísio, Maria, S.: Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. J Braz Comput Soc, Open Access **21**(2), 1340 (2015)
8. Frank, S.: Word embedding distance does not predict word reading time. In: Proceedings of the 39th Annual Conference of the Cognitive Science Society (CogSci). pp. 385–390. Austin, TX : Cognitive Science Society (2017)
9. Hartmann, N.S., Fonseca, E.R., Shulby, C.D., Treviso, M.V., Rodrigues, J.S., Aluísio, S.M.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. pp. 122–131. SBC, Porto Alegre, RS, Brasil (2017)
10. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Advances in neural information processing systems. pp. 1693–1701 (2015)
11. Hill, F., Bordes, A., Chopra, S., Weston, J.: The goldilocks principle: Reading children's books with explicit memory representations. arXiv preprint arXiv:1511.02301 (2015)
12. Kennedy, A., Pynte, J., Murray, W.S., Paul, S.A.: Frequency and predictability effects in the dundee corpus: an eye movement analysis. Quarterly journal of experimental psychology, 66(3) pp. 601–18 (2013). https://doi.org/10.1080/17470218.2012.676054
13. Kliegl, R., Grabner, E., Rolfs, M., Engbert, R.: Length, frequency, and predictability effects of words on eye movements in reading. European Journal of Cognitive Psychology, 16 **16**(1/2), 262–284 (2004)
14. Landauer, T.K., Laham, D., Rehder, B., Schreiner, M.E.: How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In: Shafto, M.G., Langley, P. (eds.) Proceedings of the 19th annual meeting of the Cognitive Science Society. pp. 412–417 (1997)
15. Luke, S.G., Christianson, K.: Limits on lexical prediction during reading. Cognitive Psychology **88**, 22 – 60 (2016). https://doi.org/https://doi.org/10.1016/j.cogpsych.2016.06.002

16. Luke, S.G., Christianson, K.: The provo corpus: A large eye-tracking corpus with predictability norms. Behavior Research Methods **50**, 826–833 (2018)
17. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings (2013)
18. Mirowski, P., Vlachos, A.: Dependency recurrent neural language models for sentence completion. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 511–517 (2015)
19. Mostafazadeh, N., Roth, M., Louis, A., Chambers, N., Allen, J.: Lsdsem 2017 shared task: The story cloze test. In: Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics. pp. 46–51 (2017)
20. Park, H., Park, J.: Assessment of word-level neural language models for sentence completion. Applied Sciences **10**(4), 1340 (2020)
21. Santos, H., Woloszyn, V., Vieira, R.: BlogSet-BR: A Brazilian Portuguese blog corpus. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC). pp. 661–664 (2018)
22. Santos, J., Consoli, B., dos Santos, C., Terra, J., Collonini, S., Vieira, R.: Assessing the impact of contextual embeddings for portuguese named entity recognition. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS). pp. 437–442 (2019)
23. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Lecture Notes in Computer Science, vol 12319. Springer, Cham. pp. 403–417 (2020)
24. Tran, K., Bisazza, A., Monz, C.: Recurrent memory networks for language modeling. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 321–331. Association for Computational Linguistics, San Diego, California (Jun 2016)
25. Vale, O.A., Baptista, J.: Novo dicionário de formas flexionadas do unitex-pb avaliação da flexão verbal. In: Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. pp. 171–180. Brazilian Computer Society, Porto Alegre, RS, Brasil (2015)
26. Wagner Filho, J.A., Wilkens, R., Idiart, M., Villavicencio, A.: The brWaC corpus: A new open resource for Brazilian Portuguese. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC). pp. 4339–4344 (2018)
27. Woods, A.: Exploiting linguistic features for sentence completion. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 438–442 (2016)
28. Xie, Q., Lai, G., Dai, Z., Hovy, E.: Large-scale cloze test dataset created by teachers. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2344–2356. Association for Computational Linguistics (2018)
29. Yuret, D.: Ku: Word sense disambiguation by substitution. In: Proceedings of the 4th International Workshop on Semantic Evaluations. pp. 207–213. Association for Computational Linguistics (2007)
30. Zweig, G., Burges, C.J.C.: The microsoft research sentence completion challenge. Tech. rep., Microsoft Research, Technical Report MSR-TR-2011-129 (2011)
31. Zweig, G., Platt, J.C., Meek, C., Burges, C.J., Yessenalina, A., Liu, Q.: Computational approaches to sentence completion. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 601–610. Association for Computational Linguistics, Jeju Island, Korea (Jul 2012)

4.3 O córpus RastrOS

Título:	***RastrOS Project: Natural Language Processing contributions to the development of an eye-tracking corpus with predictability norms for Brazilian Portuguese***
Autores:	**Sidney Evaldo Leal, Katerina Lukasova, Maria Teresa Carthery-Goulart e Sandra Maria Aluísio**
Ano:	**2021**
Revista:	**Language Resources and Evaluation**
Situação:	**Submetido - Sob revisão editorial**

Este artigo detalha o processo de criação do córpus RastrOS, da coleta dos dados de rastreamento ocular(VIEIRA, 2020) e das normas de previsibilidade. Também disponibiliza publicamente os resultados atuais, e destaca as contribuições de PLN e ferramentas computacionais desenvolvidas para o projeto, sendo as principais:

- O método de clusterização para seleção dos parágrafos, já abordado na Seção 4.1;
- O método híbrido para avaliação da previsibilidade semântica, detalhado na Seção 4.2;
- A ferramenta web para aplicação do teste Cloze.

Os *datasets* desta primeira versão do RastrOS estão no repositório do *Open Science Framework* e o código das aplicações e *scripts* estão no Github. O artigo foi submetido para a revista LREV no início de 2021 e encontra-se sob revisão.

LREV manuscript No.
(will be inserted by the editor)

RastrOS Project: Natural Language Processing contributions to the development of an eye-tracking corpus with predictability norms for Brazilian Portuguese

Sidney Evaldo Leal[1] · Katerina Lukasova[2] · Maria Teresa Carthery-Goulart[2] · Sandra Maria Aluísio[1,]**

Received: date / Accepted: date

Abstract This article presents RastrOS, a new corpus of eye-tracking while university students practice silent reading of paragraphs of texts in Brazilian Portuguese (BP). The article shows the potential of the corpus for Natural Language Processing (NLP) using it to evaluate the task of sentence complexity prediction in BP and makes available NLP resources and methods developed to create the corpus. Specifically, we present: (i) the method used to select the corpus paragraphs from large corpora, using linguistic metrics and clustering algorithms; (ii) a platform for collecting the Cloze test, which is also responsible for creating the project datasets, and (iii) the hybrid semantic similarity method, based on word embeddings models and contextualized word representations, used to generate semantic predictability norms. RastrOS can be downloaded from the Open Science Framework repository with the computational infrastructure mentioned above. Datasets with predictabil-

Sidney Evaldo Leal
sidleal@gmail.com
https://orcid.org/0000-0002-8817-2063

Katerina Lukasova
katerina.lukasova@ufabc.edu.br
https://orcid.org/0000-0002-1137-7298

Maria Teresa Carthery-Goulart
teresa.carthery@ufabc.edu.br
https://orcid.org/0000-0002-2751-4541

**Corresponding author: Sandra Aluísio
sandra@icmc.usp.br
https://orcid.org/0000-0001-5108-2630

[1] Instituto de Ciências Matemáticas e de Computação - University of São Paulo, São Paulo, Brazil

[2] Center of Mathematics, Computing and Cognition, Federal University of ABC, São Paulo, Brazil

ity norms and eye-tracking data are available in the OSF repository, link at http://www.nilc.icmc.usp.br/nilc/index.php/rastros

Keywords Natural Language Processing, Eye-tracking Corpus, Predictability Norms, Brazilian Portuguese, Sentence Complexity Prediction

1 Introduction

In the area of Psycholinguistics, specifically in sentence processing, some studies defend models of syntactic-semantic processing, seeking evidence in the processing costs for syntactic structures in reading comprehension experiments using eye-tracking corpora. These studies have been carried out based on controlled experiments, which take, in general, the sentence as the main unit of analysis. A criticism often made about these works concerns the ecological validity of experimental stimuli. The two oldest eye-tracking corpora — Dundee (Kennedy et al, 2003) and Potsdam Sentence Corpus (PSC) (Kliegl et al, 2004, 2006) — present different proposals regarding the text to be read, and thus provided the basis for the dichotomy corpus of sentences *versus* corpus of texts. The PSC has yet another peculiarity, because in order to investigate the combined effects of word length, frequency and contextual predictability[1], sentences with a variety of grammatical structures were constructed instead of bringing sentences from naturally-occurring texts. There are several arguments for and against the use of authentic texts for the purpose of creating a corpus with eye-tracking measures (see Laurinavichyute et al (2019)). For example, using texts has the advantage of ecological validity, as reading becomes more natural, enabling us to record the regressive movements for previous sentences and to observe the time of integration of information between sentences. Those who advocate the use of a sentence corpus recall that the capture is clean without recording eye movement between lines and there is more precision in the reading time data. There is also the argument about limiting the use of a single genre, as in the case of Dundee and GECO (Cop et al, 2017) that impacts on the variability of text characteristics, which is quite reasonable. However, this restriction can be combated using short paragraphs from various genres and sources, as Provo (Luke and Christianson, 2018, 2016) does.

Eye-tracking corpora have also been used in NLP tasks to, for example, (i) evaluate models and metrics of sentence complexity (Gonzalez-Garduño and Søgaard, 2017; Singh et al, 2016), (ii) improve or evaluate computational models of simplification via sentence compression (Klerke et al, 2016) and (iii) evaluate the quality of machine translation with objective metrics (Klerke et al, 2015). However, only few resources exist, for a small number of languages, for example, English (Luke and Christianson, 2018; Cop et al, 2017), Russian (Laurinavichyute et al, 2019), Hindi (Husain et al, 2014), Chinese (Yan et al, 2010), German (Kliegl et al, 2004, 2006) and English and French (Kennedy et al, 2013, 2003).

[1] Predictability is a measure of how successfully a word can be guessed on the basis of the previous context.

For BP, there is no large eye-tracking corpus with predictability norms such as those cited above. In order to fulfill this gap, we built a corpus of eye movements in silent reading of short paragraphs in BP. We also collected Cloze scores for every word, except the first in a paragraph, across the sentences of the above short paragraphs. Our corpus is called RastrOS and deals with paragraphs of authentic texts taken from different textual genres. Thus, it allows an assessment of the combined influence of a set of linguistic-textual factors that can affect linguistic processing during reading, in less artificial conditions for carrying out the task.

One of the goals of RastrOS was to study lexical predictability in a morphologically rich language, such as Portuguese, and to understand the role of partial predictability, i.e. if there is a more-expected candidate available from context even when word identity is not. Therefore, we made available the predictability of the Part-of-speech (PoS), inflectional attributes and semantic similarity information for each word in the RastrOS corpus to replicate the investigation on the graded nature of prediction carried out by Luke and Christianson (2016).

Hoping that RastrOS can be used in a myriad of NLP tasks in Portuguese, we also made available two metrics used in the sentence processing literature to quantify the complexity of a sentence: lexical surprisal and entropy reduction. While surprisal measures the relative unexpectedness of a word in context, entropy reduction is based on the concept of entropy, which is a measure proposed to quantify the degree of uncertainty about what is being communicated as a sentence unfolds (Lowder et al, 2018).

RastrOS was created by a multicenter project lasting 2 years, which started in August 2019, with the support of a Brazilian research support agency[2] and has been used in two studies: (i) lexical and partial prediction in BP and its effects on reading (Vieira, 2020), and (ii) evaluation of automatic methods of predicting sentence complexity in BP, using a large set of linguistic, psycholinguistic and metrics from eye-tracking data (Leal et al, 2020). For the first study, the eye-tracking data collection will continue when classes at the universities participating in the project return to in-person instruction. For the second, the current eye-tracking dataset has already brought benefits to performing the task of predicting sentence complexity, generating a new method in the state-of-the-art.

In this paper, we present the current status of RastrOS in order to make the corpus publicly available. The paper also presents NLP resources and methods developed for collecting data and generating RastrOs datasets and using the corpus for the task of evaluating sentence complexity in BP.

The remainder of this paper is organised as follows. Section 2 presents eye-tracking corpora related to the RastrOS corpus. Section 3 presents the content

[2] The six research centers are the Federal University of Ceará (UFC), the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), the Institute of Mathematical and Computer Sciences, University of São Paulo (ICMC/USP), the Federal Technological University of Paraná (UTFPR), Toledo campus, the State University of Rio de Janeiro (UERJ) and the Federal University of ABC (UFABC).

of the RastrOS corpus. Section 4 presents the computational infrastructure used to develop RastrOS we have also made available. To demonstrate the range of potential applications of the RastrOS corpus, Section 5 illustrates the creation of an automatic method for predicting sentence complexity in PB. Finally, Section 6 concludes the paper and points out some future works.

2 Related Work

The constitution of the RastrOS corpus is partially based on the structure and methods of the Provo Corpus (Luke and Christianson, 2018, 2016). The Provo Corpus contains eye-tracking data from 84 participants and predictability norms for the words of 55 paragraphs taken from online news articles, popular science magazines, and public-domain works of fiction, in English, created with 470 Cloze test participants.

Provo Corpus differs from its predecessors by combining some characteristics. The stimuli for the Cloze and eye-tracking procedures are connected sentences, instead of loose sentences, such as those presented, for example, in the PSC and in the Russian Sentence Corpus (RSC) (Laurinavichyute et al, 2019). Predictability norms, in turn, are provided for all words of the sentence (except the first of each paragraph), whether they are content words or functional words. Provo Corpus thus distinguishes itself from works in which the norms were presented only for the final words of the sentences (eg. Bloom and Fischler (1980); Schwanenflugel and Rey (1986)) or only for content words. In addition, Provo Corpus contains predictability scores not only for full-orthographic forms, but also for morphosyntactic classes (PoS), inflectional forms and semantic similarity relationships.

The RastrOS corpus incorporates these characteristics from the Provo Corpus. Therefore, RastrOS also has three types of Cloze scores: (i) full-orthographic form, (ii) PoS and inflectional properties and (iii) semantic predictability scores, for all 2494 words, in the 50 short text paragraphs. However, RastrOS differs from Provo and from other corpora both because of aspects of its constitution and due to some methodological contributions. To begin with, it is the first eye-tracking corpus with predictability norms for BP. Regarding the selected stimuli, in the same way as Provo Corpus or Dundee Corpus, paragraphs of sentences connected as stimuli were presented in RastrOs. However, RastrOs started from a motivated selection of texts. A computational method was developed to select the 50 paragraphs that make up the present corpus based on specific linguistic properties (see Section 4). From the corpora mentioned above, GECO (Cop et al, 2017) also selected novels motivated by linguistic features, but used only three measures[3], while RastrOS used 58 linguistic metrics.

[3] To evaluate the textual difficulty, the Flesch Reading Ease, and SMOG grade were used; to evaluate the naturalness of the language of the novels, the Kullback–Leibler divergence measure was used based on the Subtlex database (Keuleers et al, 2010).

Another important difference between RastrOS and Provo Corpus is the method to calculate semantic similarity scores. Although the Provo Corpus chose to use Latent Semantic Analysis (LSA) (Landauer et al, 1997), in RastrOS we developed a hybrid semantic similarity method, based on other families of methods (see Section 4) and also calculated a new semantic similarity score — the semantic fit of the Cloze task response with the previous context of a sentence besides the two already provided by Provo.

Finally, in RastrOS we also made available two information complexity metrics — lexical surprisal and entropy reduction — using only lexical predictability scores, the same way these metrics were implemented in Lowder et al (2018). Table 1 presents a comparison among the eye-tracking corpora cited here and RastrOS.

Corpus	**Language**	**Stimulus**	**Corpus Stats**	**Participants**
Dundee Corpus Kennedy et al (2003, 2013)	English	Connected Sentences	**Words:** 56,212 **Sentences:** 2,368 **Texts:** 20 newspaper editorials	**Eye-tracking:** 20 **Predictability:** 272
Potsdam Sentence Corpus (PSC) Kliegl et al (2004, 2006)	German	Isolated Sentences	**Words:** 1,138 **Sentences:** 144 **Texts:** N/A	**Eye-tracking:** 65 (Kliegl et al, 2004) 222 (Kliegl et al, 2006) **Predictability:** 272
Ghent Eye-Tracking Corpus (GECO) Cop et al (2017)	English and Dutch	Connected Sentences	**Words:** 59,716 (Dutch) 54,364 (English) **Types:** 5,575 (Dutch) 5,012 (English) **Sentences:** 5,301 (Monolingual) 5,190 (Dutch) 5,300 (English) (Bilingual) **Texts:** 1 entire novel	**Eye-tracking:** 33 **Predictability:** N/A
Provo Corpus Luke and Christianson (2016, 2018)	English	Connected Sentences	**Words:** 2,689 **Types:** 1,197 **Sentences:** 134 **Texts:** 55 paragraphs from online news articles, popular science magazines, and public-domain works of fiction	**Eye-tracking:** 84 **Predictability:** 470

Russian Sentence Corpus (RSC) Laurinavichyute et al (2019)	Russian	Isolated Sentences	**Words:** 1,362 **Sentences:** 144 **Texts:** N/A	**Eye-tracking:** 96 **Predictability:** 750
RastrOS	Brazilian Portuguese	Connected Sentences	**Words:** 2,494 **Types:** 1,237 **Sentences:** 120 **Texts:** 50 paragraphs from news articles, literary texts and popular science articles	**Eye-tracking:** 37 **Predictability:** 393

Table 1: Eye-tracking corpora and RastrOS numbers

3 Content of the RastrOS corpus

The eye-tracking data collection started in November 2019 at UFC. We were going to start at the other 4 centers from March 2020 onwards, but were interrupted due to the pandemic. Data collection using the Cloze test for predictability norms began in January 2020 and ended in November 2020. This collection was also impacted by the pandemic, albeit on a smaller scale, as it was applied online, via a website, allowing students to complete the test at home, when classes migrated online.

The studied population are undergraduate students from Literature /Linguistics and Computing courses. They are all speakers of BP. The two collections (Cloze and eye-tracking) were performed by different participants; without intersections. The next sections describe the two data collections; the eye-tracking data collection is described in Vieira (2020), in detail.

3.1 Predictability norms

3.1.1 Participants

Four hundred and seventeen students (200 men, 217 women), from the six universities in Brazil, cited above, answered an online Cloze task on the Simpligo-Cloze Platform, developed for the RastrOS project, and described in detail in Section 4. Participants were recruited by invitation from lecturers and members of the project team. All participants read and signed an online Informed Consent Form prior to taking the tests. All tests were answered on computers. Participants did not receive any kind of compensation. All participants were at least doing an undergraduate degree and were native speakers of BP. Before

starting the data collection, the current project was approved by the Research Ethics Committee at each of the six participant and co-participant universities involved[4].

3.1.2 Criteria for data exclusion

For the Cloze test, two exclusion criteria were used: age and commitment to the task. The age exclusion criterion led to the exclusion of 13 participants, using the 2.5 * standard deviation criterion, removing participants over 43 years of age. This age criterion was used because the invitation to undergraduate students was also made in research laboratories that have undergraduate, master´s and doctoral students. The invitation was thus accepted by 28 students who had already graduated, naturally older than the average age of undergraduate students. From these 28 students, after exclusion by age, 23 remained.

The exclusion criterion concerning attention to the task eliminated paragraphs and not participants, in principle. Forty two paragraphs were excluded from the dataset whose student responses contained more than 10% of random responses[5]. However, the application of this criterion resulted in the exclusion of 12 more participants, leaving 393 participants, (191 men, 202 women, Mean Age: 22.6 (17-43)).

The number of paragraphs answered per participant ranged from 1 to 5 (M: 4.41, SD: 1.28). Thus, researchers who are going to use the dataset can choose to use only the answers of the students who completed the 5 paragraphs or use all the answers, as the number of paragraphs answered per participant was indicated in the corpus[6].

3.1.3 Materials

For RastrOS, the corpus comprises 50 paragraphs that sum in total 120 sentences, and 2494 words total (2831 tokens including punctuation), out of which 1237 were unique. Words per paragraph range from: 36-70 (average of 49). Word length range: 1 - 18 (average of 4.96). The average size of function words is 2.5, and of content words is 6.7. In accordance with the hyphen rules in BP, we decided that hyphenated words would be one word, hence the 18 letters long words. The average number of sentences per paragraph is 2.4 (range: 1-5), and the average word per sentence is 20.8 (range: 3-60).

[4] Human Research Ethics Committee - CEP of the School of Arts, Sciences and Humanities (EACH) at the University of São Paulo (USP); UFABC Research Ethics Committee; Research Ethics Committee of the Federal University of Ceará (CEP/UFC/PROPESQ); Research Ethics Committee involving Human Beings (CEP) at the Federal Technological University of Paraná (UTFPR); Research Ethics Committee at the State University of Rio de Janeiro (CEP-UERJ), Research Ethics Committee on Human Beings at the Veiga de Almeida University (CEP-UVA)

[5] Typing a random sequence such as “asdf”, expletives, and English words

[6] This information was included in the dataset Rastros_Corpus_Cloze_FULL.tsv, in the variable Qty_Paragraphs_Partic

The 50 paragraphs of the corpus were taken from various sources in journalistic, literary and popular science genres, at a rate of 40% for newspaper articles, 20% for literary texts and 40% for popular science communication. The paragraphs were selected from a corpus larger than 100 paragraphs to account for the greatest diversity of linguistic factors relevant for processing cost assessment, reflected in the reading process: structural complexity of the period (simple vs. compound periods); verbal transitivity; sentence types (active/passive/relative); coreference relations, among others. The computational method developed to support the choice of the subset of paragraphs of a large corpus is described in detail in Section 4.

3.1.4 Procedure

All participants completed the Cloze task online using the Simpligo-Cloze Platform. A separate link was provided for each of the six universities. First, participants read and signed an Informed Consent Form. Then they filled in a demographic questionnaire containing the following questions: name, ID, age, sex, undergraduate course, current semester, languages other than BP, e-mail and phone for contact. Next, participants went through one practice paragraph. The same practice paragraph was used for every participant. All participants were instructed to fill in the gap with a word they thought would fit with the previous content of the paragraph.

We assigned each participant to 5 random paragraphs out of the 50. The criteria for the paragraph selection was sorting the one with the lowest answer count in each genre, making sure all paragraphs would be answered before repeating any. Therefore, at least one of each genre was selected randomly, then the 2 paragraphs with the least number of answers were added, making it a total of 5. We collected an average of 33 answers for each word (range: 25-43) and 34 answers per paragraph (range: 25-43).

3.1.5 Content of predictability norms file

The responses of the participants and the target words were compared to analyse the correspondence in three ways: (i) orthographically (traditional Cloze score), (ii) using the morphosyntactic class (PoS), and (iii) comparing the inflection. For correspondence in graphic form, all words were converted to lower case and correspondence was considered if the target words and responses were graphically identical. To assess the correspondence between PoS, the two classes should be identical, as well as for inflection.

The students' responses were edited to manually correct typing errors. For multiple word responses, only the first was chosen. To annotate the 50 paragraphs of the RastrOS corpus and the responses of the participants regarding the morphosyntactic class, content *versus* functional word, and inflection information (or morphological attributes), two approaches were evaluated:

1. using the morphosyntactic tagger nlpnet[7] (Fonseca and Rosa, 2013; Fonseca et al, 2015), which was inspired by the SENNA software (Collobert et al, 2011), in conjunction with the UNITEX-PB dictionary[8]; and
2. using of syntactic parser Palavras (Bick, 2000).

Although nlpnet is one of the best morphosyntactic taggers for BP, it does not provide information about the inflection of words. The Palavras parser, on the other hand, has information on morphosyntactic tags and word inflection, in addition to syntactic tags[9], however, it makes the text tagging process more computationally costly.

Figure 1 shows the distribution of words in PoS classes in RastrOS. For the content words, we have: 21,763 nouns, 12,545 verbs, 6,802 adjectives, and 4,888 adverbs.

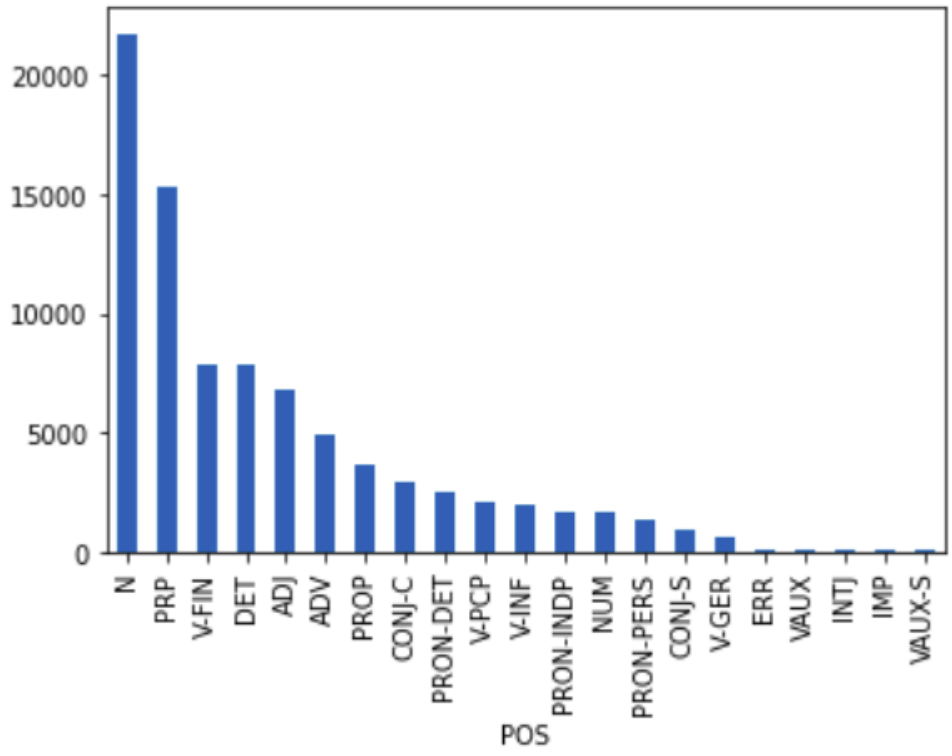


Fig. 1: N stands for noun, PRP for preposition, V-FIN for finite verb, DET for article, ADJ for adjective, ADV for adverb, PROP for proper noun, CONJ-C for coordinating conjunction, V-PCP for past participle, V-INF for infinitive, PRON-INDP for independent pronoun (or substantive pronoun), NUM for number, PRON-PERS for personal pronoun, CONJ-S for subordinating conjunction, V-GER for gerund, ERR for error, VAUX for auxiliary verb, INTJ for interjection, IMP for command/imperative, VAUX-S for auxiliary verb.

[7] http://nilc.icmc.usp.br/nlpnet/

[8] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html

[9] https://visl.sdu.dk/

The decision in the RastrOS project was to use the Palavras parser, as it is a single system that provides the 3 types of information used in RastrOS, and it is not necessary to map the nlpnet tags to the UNITEX-PB dictionary tags, which use sets in different granularities. Thus, we were able to simplify the use of annotation tools/resources in the project.

In order to provide an estimate of the similarity of the responses and targets we decided to evaluate word embedding models and one contextualised word representation model recently trained for BP. Details of the proposed method based on this evaluation are presented in Section 4. RastrOS provides, besides semantic similarity between the target word and Cloze task responses, two other measures: (i) the semantic fit of the target word with the previous context of a sentence and (ii) the semantic fit of the Cloze task response with the previous context of a sentence. All 3 were calculated by the method developed in this project.

Table 2 describes the variables of the predictability norms of RastrOS, which have the same names as those used in Provo, to facilitate comparisons. We also make available, in the OSF repository, a file with all the answers from all Cloze participants, to facilitate additional predictability studies, in addition to the two main files of the corpus, described in Table 2 and Table 3, in Appendix A.

Variable	**Description**
Word_Unique_ID	The ID number for each word in the dataset
Text_ID	The text number of RastrOS corpus (paragraph 1–50)
Text	The paragraphs from which the target word is taken
Word_Number	The position of the word in the text
Sentence_Number	The number of the sentence (1-120) in which the current word is located
Word_In_Sentence_Number	The position of the current word within the current sentence
Word	The target word, with punctuation, capitalization and contractions removed
Response	The response produced by the participant in the Cloze task
Response_Count	Number of participants who produced a given response
Total_Response_Count	The total number of responses provided on the Cloze task for this word token
Response_Proportion	How often a given response was provided, as a proportion of all responses. Response_Proportion = Response_Count/Total_Response_Count

Table 2: Predictability norm variables, and their explanations.

3.2 Eye-tracking data

3.2.1 Participants and data exclusion

Forty-six undergraduate students (20 men, 26 women, Mean Age: 22, range: 18-40, laterality: 43 right, 3 left) from the UFC, Brazil, participated in the Eye-tracking reading task (Vieira, 2020). Nine participants had to be removed for different reasons, 37 remained. Two were removed for not completing the task, 1 for skim reading, and 6 were removed for having unusual fixations and saccades, probably due to calibration errors. None of the participants took part in the Cloze task. Participants were recruited by e-mail, phone, or face-to-face invitation. All participants were at least doing an undergraduate degree and were native speakers of BP. All had normal or corrected-to-normal vision. An Informed Consent Form was signed by every participant. None received any kind of compensation for participating.

3.2.2 Apparatus and Procedure

The corpus used for the eye-tracking reading task was the same as described in Section 3.1.3. All participants read all 50 paragraphs. Eye movements were recorded on an Eye Link 1000 Hz (SR Research), desktop version with a chin rest. The experiment was programmed on the Experiment Builder (SR Research). Paragraphs were presented using the monospaced Courier New font, size 18-point with double space between lines. Text was in black and the background was in light gray. The distance between the participant's eye and camera was 65 cm. The room was lit for the participant's well-being. A nine-point grid calibration was executed before practice trials, and after roughly 10 minutes intervals. Before each trial, a drift correction was made before the paragraph was revealed, and we ran a full recalibration if fixations deviated more than 0.5 degrees from the focal point. Before starting the actual test, participants read 2 practice paragraphs, and then they read all 50 paragraphs, one by one in a random order plus 2 practice paragraphs at the beginning. After finishing a paragraph, the participant had to press a button on a joystick to continue. Yes-no comprehension questions appeared in 20 paragraphs, to ensure the participants had to look at their answer for 2 seconds and use the confirmation button on a joystick to continue. The participants were asked to move as little as possible, and they were instructed to read silently. The total run time was approximately 25 minutes.

3.2.3 Content of eye-tracking data file and data exclusion

Data was processed using the Data Viewer (SR Research). First, all fixations shorter than 80 ms were merged with fixations that were longer than 80 ms and within the distance threshold of 0.5 degrees. Afterwards, the process was repeated, except the fixation duration threshold was 40 ms and the distance threshold was 1.25 degrees. Then, all fixations under 80 ms and over 800 ms

were removed, as well as fixations outside interest areas (words). Each trial was thoroughly examined for tracking loss and errors such as the participants accidentally skipping a trial, which meant a removal of 3.3% of trials. Lastly, for reading times, outliers were also removed (2.5 standard deviations from the mean, roughly 3% of the data).

In Table 3, presented in Appendix A, the columns that appear in the file RastrOS_Corpus_Eytracking_Data.tsv of Corpus RastrOS are listed and described.

First, the participant and word identification variables are listed (11 variables). We also indicate the textual genre of each paragraph and sentence, for assessments related to predictability in different text genres. Then there were variables associated with traditional predictability measures (Cloze scores). Following these are the variables associated with morphosyntactic predictability (the predictability of PoS and inflection). Variables associated with semantic predictability appear below. RastrOS proposed a new semantic score, which is a measure of the semantic fit of the response with the previous context of a sentence, without a correspondent in Provo.

Following the three semantic predictability measures, 4 word frequency measures appear using 2 large corpora in BP: the Corpus Brasileiro[10] and the Corpus BrWac[11]. We used both the normalised frequency (or frequency per million), which is the original frequency of the words in a given corpus multiplied by 1 million, divided by the size of the corpus, and the frequency on the Zipf scale that is calculated as log 10 (normalised frequency) + 3.

The Corpus Brasileiro is a collection of approximately one billion words of written and spoken Portuguese, characterised by the diversity of text genres, for example, the academic, encyclopedic, journalistic, literary, technical genre, among others, such as those of politics, representing spoken language. The BrWaC corpus (Wagner Filho et al, 2018) was made available in January 2017; it has 3.53 million web documents, 2.68 billion words and 5.79 million unique forms (TTR 0.0021). We made the frequency lists of words used in our project available for downloads in the OSF repository, for future research use in the field of Psycholinguistics in Brazil.

We also made available two metrics from the sentence processing literature — lexical surprisal and entropy reduction. The surprisal metric, which is defined as the negative log probability of a word w, given its preceding context (see Eq. 1), was calculated using the probability of human correctness of the Cloze test response. This probability is available in the column Ortographic_Match and shows the number of correct answers divided by the total answers for each word. To avoid errors in calculating the log, we adopt the same approach used by Lowder et al (2018) — substituting probabilities 0 for half the lowest probability of our corpus (our lowest value is 0.023): every value 0 has been replaced by 0.0115. For each word, the log of the value of the

[10] http://corpusbrasileiro.pucsp.br/cb/Inicial.html

[11] https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC

Ortographic_Match column was calculated and multiplied by -1 so that the numbers were positive.

$$surprisal(w_i) = -log\, P(w_i|w_1...w_{i-1}) \quad (1)$$

The Entropy Reduction metric was calculated according to the procedure described in Lowder et al (2018): for each word, the distribution of all answers (right and wrong) was obtained and the Shannon Entropy formula (see Eq. 2) was applied to calculate the entropy H of the probability distribution over X, which is represented as a function of the probabilities of the various possible outcomes.

$$H(X) = -\sum_{x \in X} P(x) log_2 P(x) \quad (2)$$

To obtain the reduction value, we subtract the entropy of the previous word from the entropy of the current word. The result is negative when there is a reduction in entropy and positive if there is an increase. Unlike Lowder et al (2018), we chose not to normalise the positive results to 0, since this can be easily done in future studies using this metric.

The next three measures are related to the time of typing the Cloze answers, indicating the initial time of typing, after presenting the gap, the duration of typing and the time the typing ended. Finally, there are 36 eye-tracking variables, in which three[12] of them were used in the sentence complexity prediction method, described in Section 5. These eye-tracking variables are the output of the SR Research Data Viewer (SR Research).

4 NLP resources and methods used to develop RastrOS

When creating the infrastructure of NLP resources and methods of the RastrOS project, we evaluated several options of taggers, parsers, semantic similarity methods and also options of word frequency lists to make the best choices to create the predictability norms. This section describes three resources used to create the computational infrastructure of the project: (i) the method of selecting paragraphs from a large corpus to perform the Cloze test and data collection via eye-tracking (Section 4.1); (ii) the platform for collecting the Cloze test data (Section 4.2); and (iii) the method of calculating the semantic predictability norms (Section 4.3).

4.1 Using linguistic metrics and clustering methods to select paragraphs

Research on the costs of human sentence processing during reading, in the area of Psycholinguistics, can benefit from corpora of authentic texts linguistically annotated, which allow a correlation between reading times and

[12] First Pass Reading Time (IA_FIRST_RUN_DWELL_TIME), Total Regression Duration (IA_REGRESSION_PATH_DURATION) and Total Fixation Duration (IA_DWELL_TIME).

linguistic phenomena. Examples of phenomena of interest are the structural complexity of the period (simple *versus* compound periods); verbal transitivity; the animacity of the subject and the object; the types of sentences (active/passive/relative); coreference relations, among others. Having a large corpus annotated with linguistic phenomena, one can choose the subset with the appropriate attributes for a given study. For example, two paragraphs with the same number of sentences and words can differ on several linguistic levels, for example, in the complexity of their lexicon, in the syntactic complexity, in the level of formality, in the mechanisms of cohesion and coherence used. Therefore, there is a need for more linguistic metrics to inform the choice of a given paragraph for a certain study. However, one difficulty in compiling these corpora is the manual annotation of these phenomena, which ideally should use more than one annotator to be able to assess the level of agreement between them (Carletta, 1996). There is, however, an option for this scenario, which was adopted in the RastrOS project, and is described below.

Given the public availability of NILC-Metrix[13] with 200 automatic metrics for assessing the cohesion and coherence of texts written or spoken in Portuguese, a method based on these metrics was created in the RastrOS project (Leal et al, 2019a).

NILC-Metrix was developed by the Interinstitutional Center for Computational Linguistics (NILC), from 2008 to 2020 (Scarton and Aluísio, 2010; Aluísio et al, 2016; Scarton et al, 2010; Santos et al, 2020). It was based on the Coh-Metrix (Graesser et al, 2011) project whose version 3.0 makes publicly available 108 metrics for the English language, grouped into 11 sets, among which cohesion and coherence metrics stand out. The method (see Figure 2) proved to be useful for selecting the 50 paragraphs of the RastrOS corpus.

The computational method to support the choice of a subset of large corpora paragraphs was implemented in python and used the clustering method implementations of the scikit-learn library[14].

It requires texts from the large corpus as input, already processed by NILC-Metrix, with the ID's of the texts in each row and the metrics in columns. Having this entry, the script calculates the ideal number of groups and outputs the list of similar clustered paragraphs, in addition to the group quality assessment measures: V-Measure (Homogeneity and Completeness) and Silhouette. Having the output, a number of items can be selected from each group, at random, or even using ranking on the text size.

The simplest clustering algorithm and most used in the literature is K-Means[15] which uses the centroid-based technique. To create the paragraph selection method of the RastrOS project, K-means was used, and two other algorithms were also evaluated: AgglomerativeClustering [16], of the hierarchical

[13] http://fw.nilc.icmc.usp.br:23380/nilcmetrix

[14] `https://scikit-learn.org/stable`

[15] `https://scikit-learn.org/stable/modules/clustering.html#k-means`

[16] `https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering`

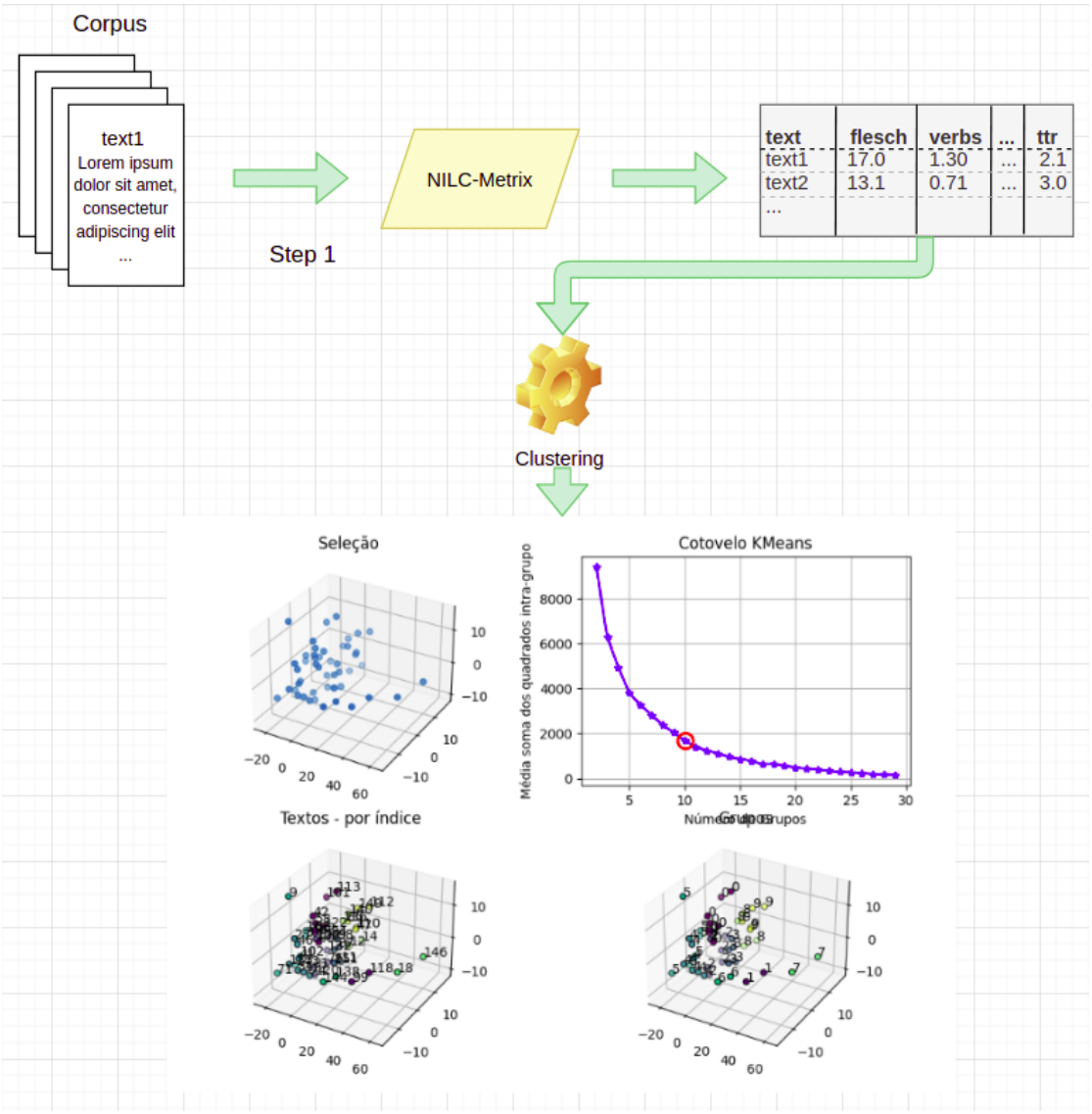


Fig. 2: Pipeline for using the clustering method. Step 1 represents the extraction of metrics from the texts in the corpus, using the NILC-Metrix tool, and generates a file with all the features. Step 2 uses these features to generate the ideal number of groups, using the elbow technique, and presents the distribution of texts and the proposed groups, as well as measures for assessing confidence in these groups.

type and DBScan[17], based on density. DBScan did not perform well in our scenario due to the size and distribution of the data set. AgglomerativeClustering was used to validate the choices of the main method — the K-means.

Fifty-eight NILC-Metrix metrics were chosen, grouped into four sets. Three of these sets: sentence types (7 metrics), syntactic structure complexity (22 metrics) and coreference analysis (8 metrics) were chosen to directly model, respectively:

[17] `https://scikit-learn.org/stable/modules/clustering.html#dbscan`

1. the structural complexity of the period (simple *versus* compound periods);
2. the types of sentences (active/passive/relative); and
3. coreference relations.

The set called morphosyntax (21 metrics) was chosen to indirectly model verbal transitivity and animacity of the subject and the object, as these two linguistic features are not implemented as metrics in NILC-METRIX. More details on each of the 58 metrics can be found at Leal et al (2019a).

To show the effectiveness of the clustering method for selecting a subset of paragraphs from a large corpus, in Leal et al (2019a) the following setup was used:

- a testing corpus with 100 paragraphs of 3 textual genres (journalistic, popular science and literary) for choosing a corpus with 50 paragraphs;
- the K-means and AgglomerativeClustering clustering methods mentioned above; and
- the 58 linguistic metrics also mentioned above, from the set of 200 metrics from NILC-Metrix.

The possibility of selecting texts with specific linguistic features can be very relevant in experimental studies in the field of Psycholinguistics. The RastrOS project thus contributed with the clustering method by automating the process of choosing a subset of large corpora paragraphs, informed by linguistic metrics. The set of automatic metrics helped to group paragraphs with similar features, making an automatic annotation directed to the grouping. The computational method to support the choice of a subset of large corpora paragraphs is available in the OSF repository.

4.2 The Simpligo-Cloze platform to collect data from Cloze tests

An evaluation of free and paid web applications for applying Cloze tests resulted in options that did not meet the needs of the RastrOS project; all the applications found required the registration of each gap as a separate test, requiring a great effort to register the 2494 words in the dataset, which should be predicted. Therefore, a platform was created that allows the registration of all paragraphs at once. The platform automatically tokenises the paragraphs and also has an algorithm for making draws, so that each participant receives a minimum number of paragraphs by text-genre and always the least answered.

The platform was created as a Web application (see Figure 3), thus reaching a larger audience of participants. The response procedure is: i) first, the Free and Informed Consent Form (ICF) is presented, and the name and document for the issuance of the personalised PDF with the agreement of the term are requested; ii) a sociodemographic questionnaire is then presented for statistical purposes; iii) after filling in the data, a training paragraph with the filling instructions is presented; iv) after completing the training, each of the 5 paragraphs is presented for the response, always providing the first word and

the participant responding from the second; and v) after finishing, a thank you message is displayed.

After finishing the collection, the Simpligo-Cloze platform exports the participants' sensitive data, the answers and the typing times during the tests in CSV format. A series of scripts were also developed to process these exported data, from cleaning outliers, processing predictability values (full-orthographic form, PoS, inflection, semantic similarity generated by the method described in Section 4.3, frequency and typing times) to merging the Cloze test data with the eye-tracking data output from the experiment on the Eye-link (see Figure 4).

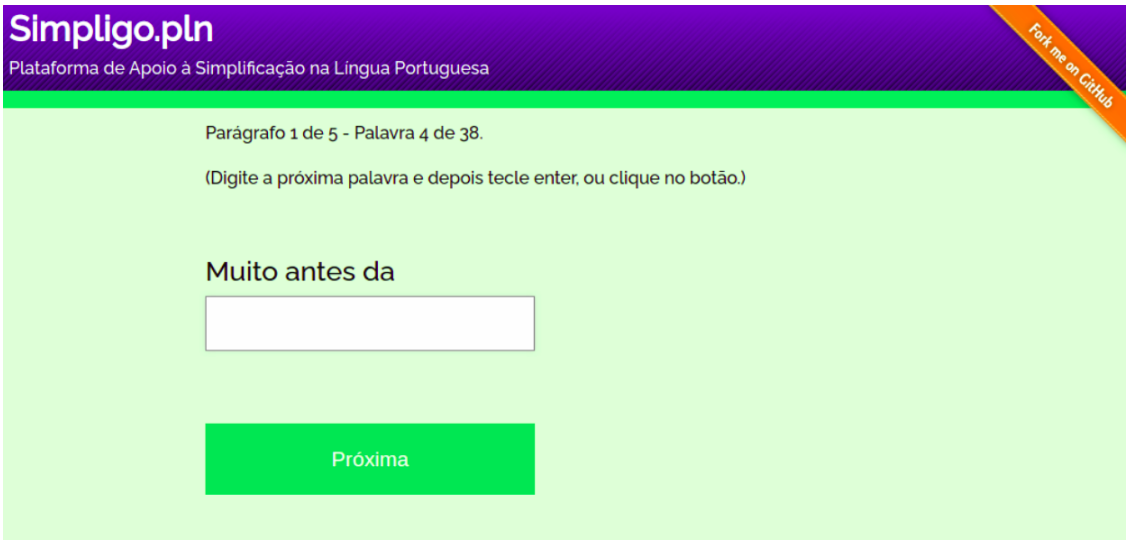


Fig. 3: Screenshot of Simpligo-Cloze platform screen while collecting responses for the Cloze test; in the example above two words have already been answered, since the first word of the paragraph is always provided. The participant must then try to predict the fourth word of the paragraph.

4.3 The hybrid method to create semantic predictability norms for BP

In the Provo corpus study, they found out that although it is very difficult that the lexical prediction in reading is high, there is room for the prediction of the PoS of the word being guessed or even the prediction of a similar word to complete a given gap. Therefore, there was a need to provide a semantic similarity method to evaluate the estimates to (i) the semantic fit of the target word with the previous context of a sentence, and (ii) the semantic similarity score between the target word and Cloze task responses, used by Provo. Although the Provo project chose to use Latent Semantic Analysis (LSA) (Landauer

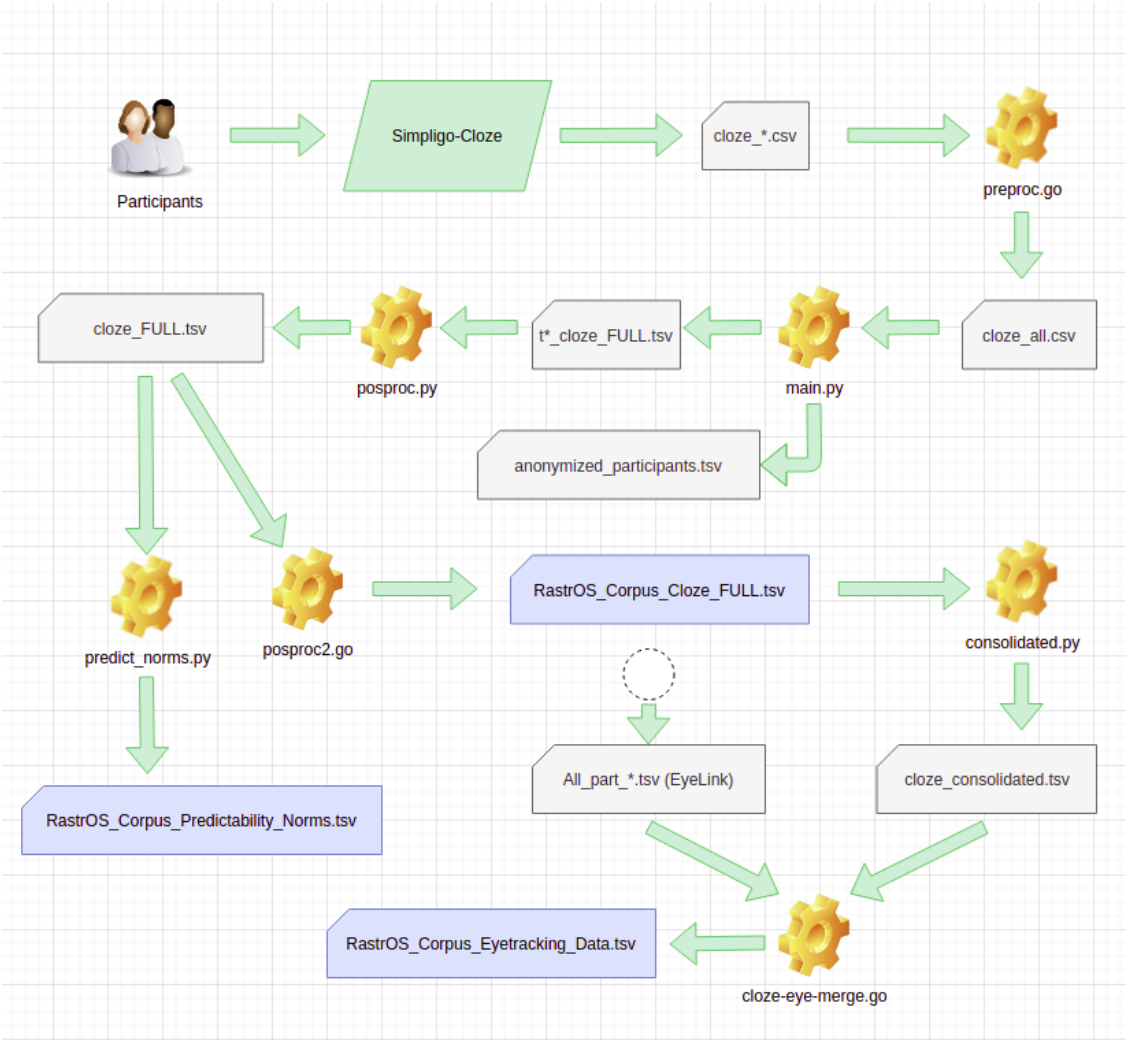


Fig. 4: Pipeline for processing data collected in the RastrOS project. The Simpligo-Cloze platform integrates all methods for creating predictability and eye-tracking datasets. Both the source code of the Simpligo-Cloze platform and the developed scripts are publicly available[18].

et al, 1997) to provide these estimates, in RastrOS we decided to modify the design of the semantic predictability scores.

First, we evaluated word embedding models from two families of methods: (i) those that work with a co-occurrence word matrix, such as LSA, and (ii) predictive methods such as Word2Vec (Mikolov et al, 2013) and FastText (Bojanowski et al, 2016). We also evaluated one contextualised word representation model — BERT (Devlin et al, 2019), recently trained for BP (Souza et al, 2020, 2019) to choose the best ones to estimate the predictability of semantic information in light of the success of the new pre-trained deep language models in the NLP area. Second, we calculated three semantic similarity scores: (i) the semantic fit of the target word with the previous context of a sentence, (ii) the semantic fit of the Cloze task response with the previous con-

[18] `https://github.com/sidleal/simpligo-cloze`

text of a sentence, and (iii) the semantic similarity between the target word and Cloze task responses, differently than Provo project that used (i) and (iii). Moreover, we calculated all three semantic measures taking the previous context into consideration in order to take advantage of the BERT results.

In order to evaluate the models above for the semantic similarity task, we created a new dataset for the sentence completion task (Zweig and Burges, 2011; Zweig et al, 2012), based on the dataset of 50 paragraphs used in RastrOS. We used project's initial Cloze data, without applying exclusion criteria — totalling data of 315 participants, who completed the online Cloze survey of our project.

The Sentence Completion task consists of, given a sentence with a gap, guessing what word or phrase would best fit the gap and, therefore, is very adequate to evaluate a semantic similarity method. We used 5 answers for each sentence with a gap, in which 4 of them were distractors for the correct answer. A set of 14 sentences of our sentence completion dataset is provided in Appendix B. The proposed method is detailed in a paper under review and is summarised below.

The hybrid semantic similarity method is composed of 2 models: (i) the BERT large model trained in the task of Masked Language Model, where the objective is to predict the masked word, and (ii) the FastText model with 300 dimensions, trained with the CBOW architecture. This approach was proposed to solve BERT's limitation to deal with words that are not present in its dictionary. In this case, the similarity is calculated using the cosine distance of our best static embedding model (FastText model).

We proposed the following four steps to calculate the similarity between two words (target and response predicted) given a context.

1. We send a sentence to BERT and mask the target word. For example, for the following sentence of our dataset: *Pesquisadores americanos passaram os últimos tempos estudando um assunto bastante peculiar: baratas. (American researchers have recently studied a very peculiar subject: cockroaches)*, in the semantic fit task we use the context, the target word, and the highest probability response predicted by BERT (Task 1). To calculate the semantic similarity between the target word, and Cloze task responses, we use the context, the target word and a student´s response, each time (Task 2).

Context: [**Pesquisadores americanos passaram os últimos tempos estudando um assunto bastante**]	
Task 1: Semantic Fit Target	P1: **peculiar** (Target Word) P2: **interessante** (BERT Prediction) O: 2.96 N: **0.13**
Task 1: Semantic Fit Response	P1: **importante** (Student Response) P2: **interessante** (BERT Prediction) O: 2.61 N: **0.11**
Task 2: Semantic Similarity	P1: **peculiar** (Target Word) P2: **importante** (Student response) O: 0.34 N: **0.02**

2. Then, we activate the model obtaining the probability p of the prediction for each vocabulary token of the BERT model.
3. Using these probabilities, for each task shown above we calculate the distance between two possible candidates using the following equation: $dist(p1, p2) = \|p1 - p2\|$, considering $p1$ and $p2$ the probabilities of predicted model for candidate 1 and 2, respectively.
4. After calculating these values for each of the instances of our corpus, we normalise the values, using the following equation: $s(p1, p2, max\_dist) = 1 - (dist(p1, p2)/max\_dist)$, considering max_dist as the largest of the distances obtained for the given task. Thus, obtaining a value between 0 and 1 that shows how similar two words are given a previous context; we consider 1 the most similar.

However, BERT has a limited vocabulary since low frequency words (rare words) of the training corpus are grouped in the token *UNK* during the training phase. Therefore, the token *UNK* brings the inflated probability of a group of words. The results of this fact are that our proposed method does not provide good results for about 29% of words in the dataset of sentence completion when using the $\text{BERT}_{\text{Large}}$ model. To solve this limitation, for those words that are not present in the dictionary of the trained BERT model, the similarity is calculated using the cosine distance of our best static embedding model, evaluated in the dataset of the Sentence Completion task: the FastText.

To use the hybrid semantic similarity method, the **get_similarity** or **get_similarity_match** python scripts should be called, which are available in the OSF repository. The first receives the word that the user wants to measure the similarity as a parameter and the preceding text passage that will serve as a context. The second method receives two words and the preceding passage and returns the similarity between them also considering the context. Examples of use with the expected outputs are available in the scripts.

5 Using RastrOS Corpus in an NLP task

This section provides a summary of an automatic method for predicting sentence complexity in BP based on the RastrOS corpus, reported in detail in (Leal et al, 2020).

One of the objectives of automatic assessment of sentence complexity is to indicate which sentences are more complex in a text for a given target audience (for example, children, users with cognitive disabilities, non-native speakers of a language and readers with low level literacy) to support the simplification of content for a target audience (Scarton et al, 2018). This task is evaluated using datasets of aligned sentence pairs, including the complex and simple version of the same sentence. Although the same approach to predict text readability can be used to assess the complexity of sentences, Dell'Orletta et al (2014) demonstrated that a greater number of metrics are needed for readability prediction at the sentence level.

A study conducted by Gonzalez-Garduño and Søgaard (2018) achieved a state-of-the-art performance in readability prediction for English sentences, using multi-task learning and eye-tracking measures. Leal et al (2020) presented a thorough evaluation of sentence readability prediction for BP, using an initial version of the RastrOS eye-tracking corpus, with 30 participants.

The metrics of the RastrOS corpus used were First-Pass Duration, Total Regression Duration and Total Fixation Duration. The best model proposed in Leal et al (2020) reaches the new state-of-the-art for BP with 97.5% accuracy. The previous state-of-the-art was 87.8% with a model that only uses linguistic metrics (Leal et al, 2018). Thus, the improvement in the performance of the new method, obtained using metrics from the RastrOS project, was 10 points, showing an application of the resources generated in the RastrOS project.

The state-of-the-art method — Sequential Transfer Learning — works by transferring learning about complexity resulting from eye-tracking measures to classifying sentences. This can be imagined as the method learning which features contribute to human difficulty during reading, not through texts annotated with complexity classes, but through real data of readers' difficulty. Once trained in this step, the method generalises the difficulty for new sentences that do not have eye-tracking data, estimating the values of the three metrics used for them. Adding these new estimated measures to the linguistic and psycholinguistic metrics already obtained for the sentences, it was possible to achieve 97.5% accuracy in judging which side is complex and which side is simple, given a pair of aligned sentences from the PorSimplesSent dataset (Leal et al, 2018).

To be able to classify the complexity of a single sentence never seen, this model was later used to create a ranking of all sentences in the dataset, from the simplest to the most complex, with a normalised index between 1 and 100. This ranking enabled us to train a regressor with the same features, which estimates the sentence complexity between 1 and 100, in which 1 is the simplest and 100 the most complex (Leal et al, 2019b).

6 Conclusions and Future Work

In this paper, we described a new eye-tracking corpus with predictability norms for BP and the complexity metrics surprisal and entropy reduction implemented with our word-by-word predictability data. We presented its potential with one of the current uses of the corpus — the evaluation of an NLP task, the prediction of sentence complexity in BP. However, there are other psycholinguistic studies planned for the corpus, for example, to examine the processing costs of various types of linguistic structures inserted in different textual genres, to replicate the study of Lowder et al (2018) for BP, and to analyse the data of eye movement, predictability of words and their syntactic positions, within the scope of the paragraph, thus evaluating the role of anticipatory processes during reading. This latter study, in particular, started during a master thesis at Vieira (2020) and will be finished after the new collection of eye-tracking

data is completed. In this article, we made available the infrastructure of NLP resources and methods used to develop RastrOS hoping that it may be useful for other research groups. We also made available the three datasets that make up the RastrOS corpus: two are related to predictability data and the third one to eye-tracking data. Moreover, we provided the dataset with the answers of the Cloze participants that were revised for spelling, indicating the correction made.

7 Data Availability

The datasets generated during the current project are available in the Open Science Framework repository:
Link at `http://www.nilc.icmc.usp.br/nilc/index.php/rastros`.

The file **RastrOS_Corpus_Predictability_Norms.tsv** is a tab-separated values file that contains traditional Cloze scores (lexical predictability), in the format described in Table 2. This file can be used to explore how different factors influence the Cloze task responses.

The file **RastrOS_Corpus_Eyetracking_Data.tsv** is also a tab-separated values file, which contains the eye-tracking data. This file also contains summary predictability values described in Table 2.

The file **Rastros_Corpus_Response_Annotation.tsv** is also a tab-separated value file, with three columns: Response/Correction/Is_RANDOM, that is, all the responses that were corrected for spelling.

The **Rastros_Corpus_Cloze_FULL.tsv** file is also a tab-separated values file, comprising all the responses of all the participants, to make it easier to analyse and study predictability questions.

8 Acknowledgments

This research project received financial support from The São Paulo Research Foundation (FAPESP) (*Fundação de Amparo à Pesquisa do Estado de São Paulo*, in Portuguese), Grant Number 2019/09807-0. The authors would like to thank all the members of the RastrOS project for making the collaboration possible between Psycholinguistics and Natural Language Processing, thus generating a new dataset and new possibilities of studies.

9 Declarations

Funding: This research was supported by The São Paulo Research Foundation (FAPESP) (*Fundação de Amparo à Pesquisa do Estado de São Paulo*, in Portuguese), Grant Number 2019/09807-0.

Conflicts of interest/Competing interests: The authors have no conflicts of interest to declare.

Availability of data and material (data transparency): Datasets are available in the Open Science Framework repository.

Code availability (software application or custom code): Both the source code of the Simpligo-Cloze platform and the developed scripts are publicly available at `https://github.com/sidleal/simpligo-cloze`. The computational method to support the choice of a subset of large corpora paragraphs is available in the OSF repository.

Authors' contributions: Sidney Leal: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software Development, Validation, Writing – original draft;

Katerina Lukasova and Maria Teresa Carthery-Goulart: Data curation, Investigation, Resources, Writing – original draft; and

Sandra Aluisio: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft.

References

Aluísio S, Cunha A, Scarton C (2016) Evaluating progression of Alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In: Silva J, Ribeiro R, Quaresma P, Adami A, Branco A (eds) Computational Processing of the Portuguese Language, Springer International Publishing, Cham, pp 109–114

Bick E (2000) The parsing system Palavras: Automatic grammatical analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press

Bloom PA, Fischler I (1980) Completion norms for 329 sentence context. Memory and Cognition, 8 pp 631–642, DOI doi.org/10.3758/BF03213783

Bojanowski P, Grave E, Joulin A, Mikolov T (2016) Enriching word vectors with subword information. `1607.04606`

Carletta J (1996) Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics 22(2):249–254, URL `https://www.aclweb.org/anthology/J96-2004`

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. J Mach Learn Res 999888:2493–2537, URL `http://dl.acm.org/citation.cfm?id=2078183.2078186`

Cop U, Dirix N, Drieghe D, Duyck W (2017) Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. Behavior Research Methods 49:602–615, DOI 10.3758/s13428-016-0734-0

Dell'Orletta F, Wieling M, Cimino A, Venturi G, Montemagni S (2014) Assessing the readability of sentences: Which corpora and features? Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications pp 163–173

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186, DOI 10.18653/v1/N19-1423, URL https://www.aclweb.org/anthology/N19-1423

Fonseca EF, Garcia Rosa JL, Aluísio, Maria S (2015) Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. J Braz Comput Soc, Open Access 21(2):1340

Fonseca ER, Rosa JLG (2013) A two-step convolutional neural network approach for semantic role labeling. In: IJCNN, IEEE, pp 1–7, URL http://dblp.uni-trier.de/db/conf/ijcnn/ijcnn2013.html#FonsecaR13

Gonzalez-Garduño AV, Søgaard A (2017) Using gaze to predict text readability. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, p 438–443

Gonzalez-Garduño AV, Søgaard A (2018) Learning to predict readability using eye-movement data from natives and learners. In: Proceedings of the The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), pp 5118–5124

Graesser AC, McNamara DS, Kulikowich JM (2011) Coh-metrix: Providing multilevel analyses of text characteristics. Educational researcher 40(5):223–234

Husain S, Vasishth S, Srinivasan N (2014) Integration and prediction difficulty in hindi sentence comprehension: Evidence from an eye-tracking corpus. Journal of Eye Movement Research 8(2), DOI 10.16910/jemr.8.2.3, URL https://bop.unibe.ch/JEMR/article/view/2400

Kennedy A, Hill R, Pynte J (2003) The dundee corpus. Proceedings of the 12th European conference on eye movement

Kennedy A, Pynte J, Murray WS, Paul SA (2013) Frequency and predictability effects in the dundee corpus: an eye movement analysis. Quarterly journal of experimental psychology, 66(3) pp 601–18, DOI 10.1080/17470218.2012.676054

Keuleers E, Brysbaert M, New B (2010) Subtlex-nl: A new measure for dutch word frequency based on film subtitle. Behavior Research Methods 42 p 643–650, DOI doi.org/10.3758/BRM.42.3.643

Klerke S, Castilho S, Barrett M, Søgaard A (2015) Reading metrics for estimating task efficiency with MT output. In: Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning, Association for Computational Linguistics, Lisbon, Portugal, pp 6–13, DOI 10.18653/v1/W15-2402, URL https://www.aclweb.org/anthology/W15-2402

Klerke S, Goldberg Y, Søgaard A (2016) Improving sentence compression by learning to predict gaze. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, pp 1528–1533, DOI 10.18653/v1/N16-1179,

URL https://www.aclweb.org/anthology/N16-1179

Kliegl R, Grabner E, Rolfs M, Engbert R (2004) Length, frequency, and predictability effects of words on eye movements in reading. European Journal of Cognitive Psychology, 16 p 262–284

Kliegl R, Nuthmann A, Engbert R (2006) Tracking the mind during reading: The influence of past, present, and future words on fixation durations. Journal of Experimental Psychology: General, v 135 pp 12–35

Landauer TK, Laham D, Rehder B, Schreiner ME (1997) How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In: Shafto MG, Langley P (eds) Proceedings of the 19th annual meeting of the Cognitive Science Society, pp 412–417

Laurinavichyute AK, Sekerina IA, Alexeeva S, Bagdasaryan K, Klieg R (2019) Russian sentence corpus: Benchmark measures of eye movements in reading in russian. Behavior Research Methods 51:1161–1178, DOI https://doi.org/10.3758/s13428-018-1051-6

Leal SE, Duran MS, Aluísio SM (2018) A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, pp 401–413

Leal SE, Aluísio SM, Rodrigues EdS, Vieira JMM, Teixeira EN (2019a) Métodos de clusterização para a criação de corpus para rastreamento ocular durante a leitura de parágrafos em português. In: Symposium in Information and Human Language Technology - STIL, SBC

Leal SE, Magalhães VMAd, Duran MS, Aluísio SM (2019b) Avaliação automática da complexidade de sentenças do português brasileiro para o domínio rural. In: Symposium in Information and Human Language Technology - STIL, SBC, pp 94–103

Leal SE, Munguba Vieira JM, dos Santos Rodrigues E, Nogueira Teixeira E, Aluísio S (2020) Using eye-tracking data to predict the readability of Brazilian Portuguese sentences in single-task, multi-task and sequential transfer learning approaches. In: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), pp 5821–5831, DOI 10.18653/v1/2020.coling-main.512, URL https://www.aclweb.org/anthology/2020.coling-main.512

Lowder MW, Choi W, Ferreira F, Henderson JM (2018) Lexical predictability during natural reading: Effects of surprisal and entropy reduction. Cognitive science 42 Suppl 4:1166—1183, DOI 10.1111/cogs.12597, URL https://europepmc.org/articles/PMC5988918

Luke SG, Christianson K (2016) Limits on lexical prediction during reading. Cognitive Psychology 88:22 – 60, DOI https://doi.org/10.1016/j.cogpsych.2016.06.002, URL http://www.sciencedirect.com/science/article/pii/S0010028516301384

Luke SG, Christianson K (2018) The Provo Corpus: A large eye-tracking corpus with predictability norms. Behavior Research Methods

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Bengio Y, LeCun Y (eds) 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, URL http://arxiv.org/abs/1301.3781

Santos R, Pedro G, Leal S, Vale O, Pardo T, Bontcheva K, Scarton C (2020) Measuring the impact of readability features in fake news detection. In: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, pp 1404–1413, URL https://www.aclweb.org/anthology/2020.lrec-1.176

Scarton C, Gasperin C, Aluísio SM (2010) Revisiting the readability assessment of texts in portuguese. In: Morales ÁFK, Simari GR (eds) Advances in Artificial Intelligence - IBERAMIA 2010, 12th Ibero-American Conference on AI, Bahía Blanca, Argentina, November 1-5, 2010. Proceedings, Springer, Lecture Notes in Computer Science, vol 6433, pp 306–315, DOI 10.1007/978-3-642-16952-6_31, URL https://doi.org/10.1007/978-3-642-16952-6_31

Scarton C, Paetzold GH, Specia L (2018) Text simplification from professionally produced corpora. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) pp 3504–3510

Scarton CE, Aluísio SM (2010) Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. Linguamática 2(1):45–61

Schwanenflugel P, Rey M (1986) Evidence for a common representational system in the bilingual lexicon. Journal of Memory and Language, 25(5) pp 605–618, DOI doi.org/10.1016/0749-596X(86)90014-8

Singh AD, Mehta P, Husain S, Rajkumar R (2016) Quantifying sentence complexity based on eye-tracking measures. In: Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity, pp 202–212

Souza F, Nogueira R, Lotufo R (2019) Portuguese named entity recognition using bert-crf. arXiv preprint arXiv:190910649 URL http://arxiv.org/abs/1909.10649

Souza F, Nogueira R, Lotufo R (2020) BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)

Vieira JMM (2020) The brazilian portuguese eye tracking corpus with a predictability study focusing on lexical and partial prediction. Master's thesis, Federal University of Ceará (UFC), Universidade Federal do Ceará, Biblioteca Universitária, URL http://www.repositorio.ufc.br/handle/riufc/55798

Wagner Filho JA, Wilkens R, Idiart M, Villavicencio A (2018) The brWaC corpus: A new open resource for Brazilian Portuguese. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, URL https://www.aclweb.org/anthology/L18-1686

Yan M, Kliegl R, Richter EM, Nuthmann A, Shu H (2010) Flexible saccade-target selection in chinese reading. The Quarterly Journal of Experimental Psychology, 63(4) p 705–725

Zweig G, Burges CJC (2011) The microsoft research sentence completion challenge. Tech. rep., Microsoft Research, Technical Report MSR-TR-2011-129

Zweig G, Platt JC, Meek C, Burges CJ, Yessenalina A, Liu Q (2012) Computational approaches to sentence completion. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Jeju Island, Korea, pp 601–610, URL https://www.aclweb.org/anthology/P12-1063

Appendix A. Eye-tracking dataset

Variable	Description
RECORDING_SESSION_LABEL	Session ID (Participant). The ID differentiates participants starting the undergraduate course, finishing the course and taking intermediate semesters.
Word_Unique_ID	A ID number for each word (each token) in the dataset, composed of the information about Text_ID and Word_Number (for example, UID_13_69)
Text_ID	The text number of RastrOS corpus (paragraph 1–50)
Genre	The text genre. RastrOS has three genres: journalistic (JN), literary (LT) and popular science (DC).
Word_Number	The position of the word in the text. It varies from 1 to the length of the paragraph.
Sentence_Number	The ordinal number of the sentence in which the current word is located in the paragraph. This number varies from 1 to 5 as the length of the paragraphs in RastrOS is short.
Word_In_Sentence_Number	The ordinal position of the current word within the current sentence. It varies from 1 to the length of the sentence.
Word_Place_In_Sent	Word position in quartiles of a sentence: 0-25% = 1, 25% -50% = 2, 50% to 75% = 3 and 75% -100% = 4.
Word	The word as it appeared on the screen
Word_Cleaned	The word, with punctuation and capitalisation removed
Word_Length	The length of the current word, in letters
Total_Response_Count	The total number of responses provided on the Cloze task for this word token
Unique_Count	The total number of unique responses provided on the Cloze task for this word token
OrthographicMatch	Cloze probability: The proportion of responses that were an orthographic match with the target word
IsModalResponse	Whether the target word was the most commonly produced response (1) or not (0)
ModalResponse	The modal response. If IsModalResponse is 1, this is the same as Word (see above). If IsModalResponse is 0, this is whichever response was provided most frequently.

ModalResponseCount	A count of how many times the modal response was provided in the Cloze procedure
Certainty	The Cloze probability of the modal response. Certainty = ModalResponseCount/ResponseCount
POS	The part of speech tag of the target word (See https://visl.sdu.dk/visl/pt/info/symbolset-manual.html for more information on the meaning of the specific tags.)
Word_Content_Or_Function	Whether the word is a content word or a function word, based on POS
Word_POS	A more general grouping of parts of speech, based on POS, which includes the following categories (in Portuguese): Adjetivo, Advérbio, Artigo, Conjunção, Interjeição, Nome, Numeral, Preposição, Pronome, Verbo. In English they are: Adjective, Adverb, Article, Conjunction, Interjection, Noun, Number, Preposition, Pronoun, Verb, respectively.
POSMatch	The proportion of responses with the same POS as the target, using POS column.
Word_Inflection	RastrOS evaluates inflection of the following Word_PoS: noun, verb, adjective, pronoun and article, using Palavras tags (https://visl.sdu.dk/visl/pt/info/symbolset-manual.html).For nouns there is gender and number; for finite verbs, person, tense and mode; for infinitive verbs, tense and mode;for past participle verbs, gender and number; for adjectives, gender and number; for personal pronouns, gender, number, case and person; for adjective and substantive pronouns, gender and number; for articles, gender and number.
InflectionMatch	The proportion of responses that carried the same inflection as the target. RastrOS evaluates inflection of the following Word_PoS: noun, verb, adjective, pronoun and article.
Semantic_Word_Context_Score	A measure of the semantic association between the target word and the entire preceding passage context. This score is a measure of the semantic fit of the target word with the previous context of a sentence. It was obtained with the hybrid method created in this project, which uses one word embedding model and the contextualised word representation (BERT) which is described in detail in Section 4).

Semantic_Response_Match_Score	The mean match score between the target and all provided responses. This measure is an estimate of the semantic predictability of a given target word, i.e. it evaluates if the participants can grasp the general meaning of the upcoming word. It was obtained using the hybrid method created in this project, which uses one word embedding model and the contextualised word representation (BERT) which is described in detail in Section 4).
Semantic_Response_Context_Score	A measure of the semantic association between the response and the entire preceding passage context. This score is a measure of the semantic fit of the response with the previous context of a sentence. This metric was proposed in RastrOS, with no correspondent in Provo. It was obtained using the hybrid method created in this project, which uses one word embedding model and the contextualised word representation (BERT) which is described in detail in Section 4).
Freq_brWaC_fpm	Normalised frequency (or frequency per million) of Corpus BrWac words. Corpus BrWac was made available in January 2017 and consists of 3.53 million web documents, 2.68 billion tokens and 5.79 million types (TTR 0.0021).
Freq_Brasileiro_fpm	Normalised frequency (or frequency per million) of the words of the Corpus Brasileiro. The Corpus Brasileiro (`http://corpusbrasileiro.pucsp.br/cb/Inicial.html` and `https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS`) is a collection of approximately one billion words from Brazilian Portuguese, the result of a project coordinated by Tony Berber Sardinha, (GELC, LAEL, Cepril, PUCSP), with funding from Fapesp and CNPq.
Freq_brWaC_log	Frequency on the Zipf scale, which is log 10 (normalised frequency) + 3 of the words using Corpus BrWac. Corpus BrWac was made available in January 2017 and consists of 3.53 million web documents, 2.68 billion tokens and 5.79 million types (TTR 0.0021).

Freq_Brasileiro_log	Frequency on the Zipf scale, which is log 10 (normalised frequency) + 3 of the words using the Corpus Brasileiro. The Corpus Brasileiro (`http://corpusbrasileiro.pucsp.br/cb/Inicial.html` and `https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS`) is a collection of approximately one billion words from Brazilian Portuguese, the result of a project coordinated by Tony Berber Sardinha, (GELC, LAEL, Cepril, PUCSP), with funding from Fapesp and CNPq.
Time_to_Start	Time (in seconds) between the presentation of the gap and when the participant started typing.
Typing_Time	Time between the start of typing and the submission of the response.
Total_time	Sum of Time_to_Start and Typing_Time.
IA_ID	Identification number for each interest area in the text. Note that because of typos and text parsing errors, this number may not correspond to the Word_Number.
IA_LABEL	The string of letters (w/ punctuation) contained within the interest area
TRIAL_INDEX	The order that the text was presented within the experiment for a given participant
IA_LEFT	The left boundary of the interest area, in pixels from the left of the screen
IA_RIGHT	The right boundary of the interest area, in pixels from the left of the screen
IA_TOP	The top boundary of the interest area, in pixels from the top of the screen
IA_BOTTOM	The bottom boundary of the interest area, in pixels from the top of the screen
IA_AREA	The total screen area of the interest area, in pixels
IA_FIRST_FIXATION_DURATION	First Fixation Duration: The duration of the first fixation on the interest area, in milliseconds.
IA_FIRST_FIXATION_INDEX	Ordinal sequence of the first fixation that was within the current interest area
IA_FIRST_FIXATION_VISITED_IA_COUNT	The number of interest areas visited prior to first fixation on the current interest area
IA_FIRST_FIXATION_X	The X position of the first fixation event that was within the current interest area, in pixels

IA_FIRST_FIXATION_Y	The Y position of the first fixation event that was within the current interest area, in pixels
IA_FIRST_FIX_PROGRESSIVE	Checks whether later interest areas have been visited before the first fixation enters the current interest area. 1 if NO higher IA ID in earlier fixations before the first fixation in the current interest area; 0 otherwise. This measure is useful in reading to check whether the first run of fixations in this interest area is in fact first-pass fixations.
IA_FIRST_FIXATION_RUN_INDEX	This counts how many runs of fixations have occurred when a first fixation is made to an interest area. The current run is also included in the tally.
IA_FIRST_FIXATION_TIME	Start time of the first fixation to enter the current interest area
IA_FIRST_RUN_DWELL_TIME	Gaze duration: Dwell time (i.e., summation of the duration across all fixations) of the first run within the current interest area
IA_FIRST_RUN_FIXATION_COUNT	Number of all fixations in a trial falling in the first run of the current interest area
IA_FIRST_RUN_START_TIME	Start time of the first run of fixations in the current interest area
IA_FIRST_RUN_END_TIME	End time of the first run of fixations in the current interest area
IA_FIRST_RUN_FIXATION_%	Percentage of all fixations in a trial falling in the first run of the current interest area
IA_DWELL_TIME	Total Reading Time: Dwell time (i.e., summation of the duration across all fixations) on the current interest area
IA_FIXATION_COUNT	Total fixations falling in the interest area
IA_RUN_COUNT	Number of times the Interest Area was entered and left (runs)
IA_SKIP	An interest area is considered skipped (i.e.,IA_SKIP = 1) if no fixation occurred in first-pass reading.
IA_REGRESSION_IN	Whether the current interest area received at least one regression from later interest areas (e.g., later parts of the sentence). 1 if the interest area was entered from a higher IA_ID (from the right in English); 0 if not.
IA_REGRESSION_IN_COUNT	Number of times interest area was entered from a higher IA_ID (from the right in English)

IA_REGRESSION_OUT	Whether regression(s) was made from the current interest area to earlier interest areas (e.g., previous parts of the sentence) prior to leaving that interest area in a forward direction. 1 if a saccade exits the current interest area to a lower IA_ID (to the left in English) before a later interest area was fixated; 0 if not.
IA_REGRESSION_OUT_COUNT	Number of times an interest area was exited to a lower IA_ID (to the left in English) before a higher IA_ID was fixated in the trial
IA_REGRESSION_OUT_FULL	Whether regression(s) was made from the current interest area to earlier interest areas (e.g., previous parts of the sentence). 1 if a saccade exits the current interest area to a lower IA_ID (to the left in English); 0 if not. Note that IA_REGRESSION_OUT only considers first-pass regressions whereas IA_REGRESSION_OUT_FULL considers all regressions, regardless of whether later interest areas have been visited or not.
IA_REGRESSION_OUT_FULL_COUNT	Number of times interest area was exited to a lower IA_ID (to the left in English)
IA_REGRESSION_PATH_DURATION	Go-Past Time: The summed fixation duration from when the current interest area is first fixated until the eyes enter an interest area with a higher IA_ID
IA_FIRST_SACCADE_AMPLITUDE	Amplitude (in degree of visual angle) of the first saccade entering into the current interest area
	NOTE: Saccade data have not been cleaned, and so include return sweeps (large eye movements from the end of one line to the beginning of the next). Excluding saccades ¿15 deg removes these return sweeps without impacting other reading-related saccades.
IA_FIRST_SACCADE_ANGLE	Angle between the horizontal plane and the direction of the first saccade entering into the current interest area
IA_FIRST_SACCADE_START_TIME	Start time of the saccade that first landed within the current interest area
IA_FIRST_SACCADE_END_TIME	End time of the saccade that first landed within the current interest area

Table 3: Eye-Tracking Dataset Variables.

Appendix B. Excerpt of the Sentence Completion Dataset for Brazilian Portuguese, created in the RastrOS Project

There is a blank in the place of the target word in each sentence; target word and the 4 distractors are presented in the list after each sentence where the target word is in boldface.

1. A invenção do zero pelos humanos foi crucial para a matemática e a _____ modernas.
   (álgebra, fala, língua, geometria, **ciência**)
   The invention of zero by humans was crucial to modern mathematics and _____.
   *(algebra, speech, language, geometry, **science**)*

2. Papagaios e macacos entendem o conceito de zero, e agora as abelhas também se _____ ao clube.
   (exibiram, manifestaram, expuseram, **juntaram**, mostraram)
   Parrots and monkeys understand the concept of zero, and now bees are also _____ at the club.
   *(exhibited, manifested, exposed, **joined**, showed)*

3. Entre os tipos de exposição à radiação que afetam a população mundial, a maior parcela corresponde a exposições médicas, isto é, exames que _____ radiação ionizante para diagnóstico e tratamento.
   (**empregam**, difundem, comunicam, anunciam, divulgam)
   Among the types of radiation exposure that affect the world population, the largest share corresponds to medical exposures, that is, tests that _____ ionising radiation for diagnosis and treatment.
   *(**employ**, disseminate, communicate, announce, disseminate)*

4. Dentre as exposições médicas, os diagnósticos feitos com raios X são a fonte mais significativa para a exposição da _____ mundial.
   (**população**, imagem, recordação, metáfora, reputação)
   Among medical exposures, diagnoses made with X-rays are the most significant source for the exposure of the world _____.
   *(**population**, image, memory, metaphor, reputation)*

5. Pesquisadores americanos passaram os últimos tempos estudando um _____ bastante peculiar: baratas.
   (método, pássaro, remédio, **assunto**, vírus)
   American researchers have recently studied a very peculiar _____: cockroaches.
   *(method, bird, medicine, **subject**, virus)*

6. Especificamente, a capacidade impressionante desses insetos de se espremerem por qualquer espaço e aguentarem _____ de até 900 vezes seu próprio peso sem sofrer grandes danos.
   (magnitudes, ambientes, **pressões**, temperaturas, situações)
   Specifically, the impressive ability of these insects to squeeze themselves into any space and withstand up to 900 times their own weight _____ without suffering major damage.
   *(magnitudes, environments, **pressures**, temperatures, situations)*

7. O prazer é a sombra da felicidade, diz um provérbio hindu, para se referir a esse efeito efêmero da exposição a _____ sensoriais, estéticos ou intelectuais.
   (**estímulos**, problemas, distúrbios, complicações, enigmas)
   Pleasure is only the shadow of happiness, says a Hindu proverb, to refer to this ephemeral effect of exposure to sensory, aesthetic or intellectual _____.
   *(**stimuli**, problems, disorders, complications, puzzles)*

8. Embora intrinsecamente satisfatória, a sensação não se sustenta e, muito rapidamente, tende a se tornar _____ ou mesmo desagradável.
   (envelhecida, **neutra**, atrasada, primitiva, obsoleta)
   Although intrinsically satisfying, the sensation is not sustained and, very quickly, tends

to become _____ or even unpleasant.
*(aged, **neutral**, delayed, primitive, obsolete)*

9. Ainda que saibamos disso, a maioria de nós _____ atrás dessa vivência, insistindo em repeti-la a todo custo.
(estaciona, empaca, **corre**, estica, resiste)
Although we know that, most of us _____ behind this experience, insisting on repeating it at all costs.
*(park, pack, **run**, stretch, resist)*

10. O próprio conceito de verdade, sua flexibilidade, torna-se verdade provisória, o que muito se aproxima estruturalmente dos produtos da ciência e da arte na busca do _____ da vida no Planeta.
(**significado**, pensamento, ensaio, experimento, teste)
The very concept of truth, its flexibility, becomes provisional truth, which is very similar structurally to the products of science and art in the search for the _____ of life on the Planet.
*(**meaning**, thought, assay, experiment, test)*

11. Assim, ao objetivar sentimentos, a arte permite ao espectador uma melhor compreensão de si próprio, dos padrões e da _____ dos sentimentos.
(palavra, cadeia, **natureza**, verdade, genuinidade)
Thus, by objectifying feelings, art allows the viewer to better understand himself or herself, from patterns and _____ of feelings.
*(word, chain, **nature**, truth, genuineness)*

12. O que se conhece a respeito do cérebro e de seu funcionamento é retirado de pesquisas com pessoas que têm acesso à _____ e foram alfabetizadas desde crianças.
(saúde, comodidade, notícia, ultrassonografia, **leitura**)
What is known about the brain and its functioning is taken from research with people who have access to _____ and have been literate since they were children.
*(health, convenience, news, ultrasound, **reading**)*

13. As funções do cérebro e as regiões dele onde ocorrem mais _____ neurais refletem a influência da formação cultural e educacional dos seres humanos.
(lesões, degenerações, danificações, deteriorações, **conexões**)
The neural functions of the brain and the regions where they occur most _____ reflect the influence of the cultural and educational formation of human beings.
*(injuries, degenerations, damage, deteriorations, **connections**)*

14. A evolução ocorre na medida em que o sucesso reprodutivo desigual dos indivíduos adapta a _____ ao ambiente.
(convivência, situação, **população**, seleção, comunhão)
Evolution occurs to the extent that the unequal reproductive success of individuals adapts the _____ to the environment.
*(coexistence, situation, **population**, selection, communion)*

CAPÍTULO

5

AVALIAÇÃO DA COMPLEXIDADE SENTENCIAL: DATASETS E MÉTODOS

Os capítulos anteriores serviram de base e recursos para a tarefa principal desta pesquisa, apresentada neste capítulo por meio de três artigos já publicados e apresentados aqui em ordem cronológica.

A Seção 5.1 traz a criação do PorSimplesSent um córpus de pares de sentenças alinhadas para permitir a avaliação dos métodos na tarefa em PB. A Seção 5.2 evolui os métodos de classificação usando as métricas linguísticas e psicolinguísticas, apresenta uma nova abordagem para avaliação da complexidade em uma aplicação real e avalia a robustez do modelo em um domínio diferente do de treinamento. A Seção 5.3 traz o melhor método desenvolvido durante esta pesquisa, que atingiu o estado da arte atual da tarefa para o PB com 97,5% de acurácia no PorSimplesSent, acrescentando ao treinamento os dados de rastreamento ocular do córpus RastrOS.

5.1 O córpus PorSimplesSent

Título:	***A Nontrivial Sentence Corpus for the Task of Sentence Readability Assessment in Portuguese***
Autores:	**Sidney Evaldo Leal, Magali Sanches Duran e Sandra Maria Aluísio**
Ano:	**2018**
Conferência:	**COLING - 27th International Conference on Computational Linguistics - New Mexico - USA**
Situação:	**Publicado**

Durante a revisão da literatura, foi constatado que era necessário um *dataset* específico para treinamento e avaliação da tarefa de complexidade sentencial. Os trabalhos na língua inglesa apontaram que somente usando sentenças de textos simples e sentenças de textos complexos não era suficiente, pois acontecem sentenças simples em textos complexos e vice-versa. Para garantir a avaliação no nível das sentenças era necessário que o *dataset* possuísse pares de sentenças alinhadas, com a indicação de qual das sentenças do par é simples e qual é complexa (VAJJALA; MEURERS, 2016).

Como já estava sendo cogitado o uso do córpus de textos PorSimples (ver Seção 2.3.2) para a tarefa e ele já trazia o alinhamento entre as sentenças simplificadas, o trabalho foi extrair esses alinhamentos e propor uma solução para os casos de divisão de sentenças, para que o *dataset* resultante não ficasse muito trivial para a classificação automática.

O córpus resultante foi chamado de PorSimplesSent, o artigo que o descreve recebeu boas avaliações e foi publicado no COLING em 2018.

A Nontrivial Sentence Corpus for the Task of Sentence Readability Assessment in Portuguese

Sidney Evaldo Leal sidleal@gmail.com
Magali Sanches Duran magali.duran@uol.com.br
Sandra Maria Aluísio sandra@icmc.usp.br
Institute of Mathematical and Computer Sciences - University of São Paulo
Av. do Trabalhador Saocarlense, 400, São Carlos - SP - Brazil

Abstract

Effective textual communication depends on readers being proficient enough to comprehend texts, and texts being clear enough to be understood by the intended audience, in a reading task. When the meaning of textual information and instructions is not well conveyed, many losses and damages may occur. Among the solutions to alleviate this problem is the automatic evaluation of sentence readability, task which has been receiving a lot of attention due to its large applicability. However, a shortage of resources, such as corpora for training and evaluation, hinders the full development of this task. In this paper, we generate a nontrivial sentence corpus in Portuguese. We evaluate three scenarios for building it, taking advantage of a parallel corpus of simplification, in which each sentence triplet is aligned and has simplification operations annotated, being ideal for justifying possible mistakes of future methods. The best scenario of our corpus PorSimplesSent is composed of 4,888 pairs, which is bigger than a similar corpus for English; all the three versions of it are publicly available. We created four baselines for PorSimplesSent and made available a pairwise ranking method, using 17 linguistic and psycholinguistic features, which correctly identifies the ranking of sentence pairs with an accuracy of 74.2%.

Title and Abstract in Portuguese

Um Corpus Não Trivial de Sentenças para a Tarefa de Avaliação de Complexidade Sentencial em Português

Uma comunicação textual eficaz depende de os leitores serem proficientes o suficiente para compreenderem o texto e de o texto ser claro o suficiente para ser compreendido pelo público-alvo, em uma tarefa de leitura. Quando o significado das informações e instruções textuais não é bem transmitido, muitas perdas e danos podem ocorrer. Entre as soluções para aliviar este problema está a avaliação automática da complexidade sentencial, tarefa que vem recebendo muita atenção devido a sua grande aplicabilidade. No entanto, a escassez de recursos, como corpora para treinamento e avaliação, dificulta o pleno desenvolvimento dessa tarefa. Neste artigo, geramos um corpus de sentenças não triviais em Português. Avaliamos três cenários para construí-lo, aproveitando um corpus paralelo de simplificação textual, no qual cada trio de sentenças está alinhado e possui operações de simplificação anotadas, sendo ideal para justificar possíveis erros de métodos futuros. O nosso melhor cenário do corpus PorSimplesSent é composto por 4.888 pares, que é maior que um corpus similar para o inglês; todas as três versões do corpus PorSimplesSent estão disponibilizadas publicamente. Criamos quatro métricas *baselines* para o PorSimplesSent e um método de ranqueamento por pares, utilizando 17 métricas linguísticas e psicolinguísticas, que identificam corretamente o ranqueamento dos pares de sentenças com uma acurácia de 74.2%.

Proceedings of the 27th International Conference on Computational Linguistics, pages 401–413
Santa Fe, New Mexico, USA, August 20-26, 2018.

1 Introduction

Readability is an issue of great social and economic impact. Effective textual communication depends on readers being proficient enough to comprehend texts, and texts being clear enough to be understood by the intended audience. When the meaning of textual information and instructions is not well conveyed, many losses and damages may occur (Dubay, 2007). In Brazil, for example, only 8% of adult population has reading proficiency (IPM, 2016). The situation is worse in the agriculture and livestock sectors, where only 1% of the surveyed are proficient readers. For this reason, most of rural producers do not have access to new technologies, undermining the development of agribusiness, which accounts for 22% of gross internal product and 30% of Brazilian jobs[1]. Research investments in these sectors, therefore, do not cause as much impact as they potentially might. Identifying which sentences of a text are more complex may help writers of newsletters, manuals and instructions, for example, to adequate their texts to their audiences.

Among the solutions to alleviate this problem is the simplification or adaptation of complex texts, a task that has been partially or fully automatized by Natural Language Processing (NLP) applications. For Brazilian Portuguese, various applications, methods and resources aiming to support simplification in several levels of readability were developed in the Project PorSimples (Aluísio and Gasperin, 2010). Among these resources there is a parallel and aligned corpus with two levels of simplification and annotated simplification operations (Caseli et al., 2009). PorSimples corpus has been used to train readability classifiers for texts (Scarton et al., 2010). Table 1 shows examples of an original sentence of PorSimples corpus (O), its natural simplification (N) and its strong simplification (S). The natural simplification had a substitution of "Uma parcela" by "Alguns" and the strong simplification, shorter than the natural, had a clause removed.

(O) Uma parcela critica o uniforme, porque acredita que ele ameaçaria a individualidade de cada um. (One parcel criticizes the uniform, because it believes that it would threaten the individuality of each one.)
(N) Alguns criticam o uniforme, porque acreditam que ele ameaça a individualidade de cada um. (Some criticize the uniform because they believe that it threatens the individuality of each one.)
(S) Alguns acreditam que o uniforme ameaça a individualidade de cada um. (Some believe that the uniform threatens the individuality of each one.)

Table 1: Examples of simplification in PorSimples.

However, we know that even complex texts have simple sentences, what makes it difficult to identify precisely where complexity lies. In an automatic simplification task, as well, it is difficult to decide which sentence is complex and requires simplification. To address these difficulties, a new task has received attention recently: the prediction of sentences readability, also known by sentence-based readability or sentential complexity task. The first studies on this subject emerged in the beginning of the last decade (Dell'Orletta et al., 2011; Sjöholm, 2012; Del'Orletta et al., 2014).

This task may support simplification systems at least in three applications: (i) to evaluate whether the simplification of a sentence (manual or automatic) is truly simpler than the original sentence or not; (ii) to inform the level of complexity of an original sentence; (iii) to rank the results of several simplification methods, according to their level of complexity. Besides supporting text simplification applications, computer-aided language learning (CALL) systems can benefit from sentence-level readability methods to predict which sentences of a text the students will struggle to read. Furthermore, Open Educational Resources repositories Wiley et

[1] http://www.ibge.gov.br/home/estatistica/economia/agropecuaria/censoagro/

al. (2014) may also take profit of such methods in order to return not merely relevant educational resources, but documents appropriate to the reading level of the user.

Due to its several applications, sentential complexity has been a focus of interest in the NLP studies in recent years, such as Vajjala and Meurers (2014), Vajjala and Meurers (2016), Ambati et al. (2016), Singh et al. (2016), Howcroft and Demberg (2017), Gonzalez-Garduño and Søgaard (2017).

The lack of a sentence-based corpus annotated with regards to readability is a major obstacle to research in this area for Portuguese. Even the English language suffers some drawbacks in what concerns the evaluation of sentential complexity. One of them is the use of benchmarks built from adapted corpora which are automatically aligned, such as Wikipedia and Simple Wikipedia (Zhu et al., 2010). This corpus has some problems to be used as benchmark for text simplification which also prevents its use for the sentential complexity task, for example, automatic sentence alignment errors, inadequate simplifications generating sentences which are not simple, and poor generalization for other genre than encyclopedia (Xu et al., 2015). Other benchmarks for sentential complexity, such as OneStopEnglish corpus (Vajjala and Meurers, 2016), have several positive points — the use of news articles which generalize better for other genres, not having sentence length as high predictive feature, as well as being available by requisition — but also can suffer from errors generated by automatic alignment. Newsela parallel corpus (cf. (Xu et al., 2015)), composed of news articles rewritten by professional editors to be read for children at multiple grade levels, is very beneficial for studying text simplification and could serve as benchmark for sentential complexity if the resulting sentence corpus could be publicly available. Moreover, Scarton et al. (2018) made available the SimPA, an English sentence level corpus for the Public Administration domain with 1,100 original sentences simplified in the lexical (3,300 pairs) and syntactic levels (another 1,100 pairs), annotated by 176 volunteers.

In this paper, we aim at obtaining nontrivial sentence pairs in Portuguese in order to create a gold standard corpus, publicly available. By nontrivial we mean that the pairs are not significantly different in length to avoid the easy judgment that the shorter sentences are the simpler ones. Although it is natural to expect that the simplified sentences are smaller, we found that it is not always true. An example of this is when, in order to simplify a content, one inserts an explanation, examples, or a list of synonyms.

We evaluated three scenarios for building our gold standard corpus from PorSimples corpus, with special care for the split operation, because splitting can generate several short sentences from an original one. The first scenario is a corpus formed of pairs of original and simplified sentences in which, if the split operation is used, we repeat the original sentence to form pairs with each of the simplified sentences. In the second scenario we include pairs with all but the simplified sentences from the split operation. The last scenario is a corpus in which all simplification operations are allowed, but for splitting we only bring the longest simplified sentences to compose the pair original-simplified.

The remainder of this paper is organized as follows. Section 2 reviews the literature on sentence-based readability assessment and its evaluation corpora. Section 3 presents the parallel and aligned corpus of the PorSimples project and explains how we built three evaluation scenarios to create the PorSimplesSent, our corpus for sentence-based readability assessment in Portuguese. In Section 4 we discuss our baselines, our method and features extracted to evaluate the three evaluation scenarios. Conclusions and future work are presented in Section 5.

2 Sentence-based Readability Assessment and its Evaluation Corpora

Initially, sentence-based readability task was considered in isolation by several authors, each one studying a set of features and evaluating in specific corpora. Dell'Orletta et al. (2011) were the first to consider the task of complexity for the sentential level, comparing its difficulty in relation to the textual level, for Italian. They used the SVM method of the LIBSVM library to train a model with 7,000 sentences, half selected in the newspaper *La Repubblica* and half of

the newspaper *Due Parole*, the latter considered simple reading. Interestingly, features at the syntactic level had little influence on the classification of documents, but were very important for the sentential level. Training with 6,000 and testing against 1,000 sentences, they reached 78.2% accuracy at the sentential level. Sjöholm (2012) addressed the task for the Swedish, also using two sets of sentences. For evaluation, 3,500 sentences were taken from the Swedish corpus LäSBarT, considered simple, and 3,500 from the GP2006 (Göteborgsposten journal), considered complex, divided into seven parts, each part used for testing with the model trained in the other six. The best method was Sequential Minimal Optimization (SMO), which reached 83% accuracy. It is important to mention that using the same set of features to evaluate documents (simple and complex) instead of sentences, in the same corpus, they obtained 97% accuracy. Dell'Orletta et al. (2014) returned to the task, addressing the issue of textual genres. They used the same sets of features from the previous article (Dell'Orletta et al., 2011), but now adding three new corpora of different genres to the original journalistic genre: literary, didactic and scientific.

Vajjala and Meurers (2014) made the first evaluation using Wikipedia-Simple Wikipedia corpus, automatically aligned by Zhu et al. (2010). This corpus became the most-used resource for sentential complexity evaluation in the English language. It was created with the matching of the sentences of 65,133 articles of Simple Wikipedia and Wikipedia, using the measure TF-IDF with cosine similarity. For the choice of the alignment measure, they evaluated the performance of three similarity measures: TF-IDF, word overlap and Minimum Edit Distance (MED), against 120 pairs of manually annotated sentences. The accuracy of TF-IDF was above 90%. As a final result, they created 108,016 aligned sentences, annotated in two classes: complex or simple, and a complex sentence may be mapped to one or more simple sentences to handle sentence splitting. This corpus was updated by Hwang et al. (2015), reaching 150,000 pairs of aligned sentences.

Table 2 shows the state-of-the-art (SotA) results we were able to compile, which use Wikipedia-Simple Wikipedia corpus. In the table, the name of each study is listed with the method/baseline used and the accuracy results.

Study	**Method**	**Accuracy (%)**
Flesch-Kincaid	Baseline	72.30
Vajjala and Meurers (2014)	SMOReg	66.00
Vajjala and Meurers (2016)	RankSVM	74.58
Ambati et al. (2016)	SMO	78.87
Singh et al. (2016)	Logistic Regression	75.21
Howcroft and Demberg (2017)	Rank as Classification (RasC)	73.22
Gonzalez-Garduño and Søgaard (2017)	MultiTask MLP	**86.45**

Table 2: SotA results using Wikipedia-SimpleWikipedia corpus.

Vajjala and Meurers (2014) trained a SMO regression model for document complexity, which reached about 90% accuracy. They then applied the model at the sentence level, and even testing in datasets of several sizes, they only achieved 66% accuracy, creating a new *baseline* for the task. They concluded the reason for this low accuracy lies in the incorrect assumption that all Wikipedia sentences are more complex than Simple Wikipedia. Even so, this dataset has been used by several studies of sentence readability. As far as we could see, Gonzalez-Garduño and Søgaard (2017) presents the state-of-the-art for the task, using eye-tracking features together with linguistic and psycholinguistic ones.

Vajjala and Meurers (2016) returned to the task, proposing a new method for evaluating paired sentences based on *ranking*. They contributed with a new corpus of English sentences aligned in three levels, called OneStopEnglish (OSE), used for training and testing. The OSE corpus is a corpus of aligned sentences created from articles rewritten by teaching experts for English language learners at three reading levels (elementary, intermediate, advanced). They used 76 triplets of articles published between 2012 and 2014, resulting in a total of 837 written

sentences with three levels (OSE3). For the alignment, TF-IDF and cosine similarity were used, with values above of 0.7. In addition to OSE3, a second corpus (OSE2) was compiled, which resulted in 3,113 sentence pairs: elementary-intermediate, intermediate-advanced, and elementary-advanced. This corpus was divided in two parts: 65% of pairs for training and the rest for testing.

In addition to significantly improving the accuracy of the task (over 80%), they assessed the impact of linguistic (lexical, syntactic, morphosyntactic) and psycholinguistic features, confirming the importance of eight features in OSE2: AoA (Age of acquisition), CTTR (corrected Type-token ratio), number of subtrees, average length of clause, average word imagery rating, average word familiarity rating, average Colorado meaningfulness rating of words, average concreteness rating. It is important to note that sentence length was not predictive in OSE2 corpus, as in this dataset rewriting and paraphrasing were the most used simplification operations.

As may be seen in Section 3, for our corpus, traditional psycholinguistic features such as AoA, imagery, concreteness, familiarity, have not been used to rank the three types of sentence pairs of PorSimplesSent. We have, indeed, analyzed their contribution to distinguish the three sentence levels, using the resource created by Santos et al. (2017). However, the results were not discriminative. We hypothesize two reasons for this. One of them is related to characteristics of the resource, which has been created automatically based on existing psycholinguistic norms and may contain some bias. The other reason is related to characteristics of the corpus. The corpus PorSimples contain a lot of explanation relating to difficult words (this is a simplification strategy to deal with lexical complexity). However, once explained, the difficult words are repeated along the text. In PorSimplesSent, when there is a split operation, the explanations remain isolated, benefiting only the sentence they appear, whereas the other sentences containing the repetitions of difficult words remain lexically complex. In fact, the psycholinguistic features did not perform well in our corpus and, therefore, they were not chosen as best features for our method.

Table 3 shows SotA results we were able to compile, which use OSE2 corpus, automatically aligned by Vajjala and Meurers (2016). In the table, the name of each study is listed with the method used and the accuracy results, separated by OSE2 subcorpus. OSE(A-E) stands for pairs at the levels Advanced and Elementary; OSE(A-I) for pairs at Advanced and Intermediate levels; OSE(I-E) for Intermediate and Elementary, and OSE(All) for all three pairs. Howcroft and Demberg (2017) joined the subcorpus OSE(A-I) and OSE(I-E), calling it OSE_{near}.

Study	**Method**	OSE(A-E)	OSE(A-I)	OSE(I-E)	OSE(All)
			OSE_{near}		
Flesch-Kincaid	Baseline				69.6
Vajjala and Meurers (2016)	RankSVM				**81.5**
Howcroft and Demberg (2017)	RasC	**85.3**	74.6		77.9
Gonzalez-Garduño and Søgaard (2017)	Multitask MLP	68.5	61.9		

Table 3: SotA accuracy results using OSE2 corpus.

Vajjala and Meurers (2016) explored whether the types of simplification operations are different between Advanced sentences simplified to Intermediate, and Intermediate sentences simplified to Elementary, using OSE3 corpus. That is why we don't have explicit evaluation between these pairs nor between Advanced and Elementary sentence pairs in Table 3.

3 PorSimplesSent Corpus

3.1 PorSimples Corpus

In order to create the PorSimplesSent, our corpus for sentence-based readability assessment in Portuguese, and to train and evaluate methods to predict sentential complexity for this language, we took advantage of PorSimples corpus (Caseli et al., 2009; Aluísio and Gasperin, 2010).

PorSimples corpus consists of 2,915 original sentences simplified into two levels of complexity:

Natural and Strong. All the sentences are from informational texts, being 30% of scientific issues from newspaper Folha de São Paulo[2] and 70% of other issues from newspaper Zero Hora[3].

PorSimples corpus contains complete annotation of each operation made during the simplification process. This was facilitated by the Simplification Annotation Editor, developed in PorSimples project (Caseli et al., 2009). The editor allows the human simplifier to register decisions of lexical and syntactic simplifications, which include substituting words, merging and splitting sentences, deleting part of the sentence, rewriting sentences with other words, and changing constituents order. The editor has a list of operations that may be chosen by the human simplifier. Simplification process in PorSimples was instructed by simplification guidelines, advising how to turn sentences simpler (Specia et al., 2008). Examples show how to tackle with complex structures, like apposition, subordinate clauses, clauses initiated by non-finite verbs, passive voice, inversion of constituents order and embedded clauses.

In a totally annotated process, the alignment between the simplified sentences and their respective simplifications is systematically ensured. This ensured alignment, added to the fact that the corpus contains a large variety of simplification strategies, makes PorSimples a unique corpus, entirely appropriate to evaluate readability predictors.

3.2 Methodology

We created 4,968 pairs and 1,141 triplets of sentences, combining the three levels of PorSimples corpus: Original, Natural and Strong. Pairs and triplets have two or three different sentences aligned, being the Original the more complex in Original-Natural and Original-Strong pairs, and Natural the more complex in Natural-Strong pairs.

In theory, there should be 8,745 pairs (an original-natural, an original-strong and a natural-strong pairing for each of the 2,915 sentences) and 2,915 triplets (original-natural-strong). However, it occurred 3,777 pairs and 1,774 triplets containing at least two identical sentences, because some of the sentences were simplified only in one level or were not simplified at all (they were considered originally simple). Such pairs and triplets were removed from the corpus, which remained with 4,968 pairs and 1,141 triplets.

Table 4 shows what happened with the original sentences of the texts during the simplification process that gave origin to PorSimples corpus. Part of the sentences has not been simplified, possibly because the sentences were considered already simple. The other part is composed of the simplified sentences, which followed one of three possible paths: simplification in both levels (Natural and Strong) or in only one of them (Natural or Strong).

Application of Simplification Operations in PorSimples Sentences	Number of Sentences
NOT simplified in any level	372
Simplified in two levels	1,105
Simplified only in Natural Level	1,268
Simplified only in Strong Level	170
TOTAL	2,915

Table 4: Distribution of original sentences according to the level of simplification.

Additionally, in the PorSimples corpus, 3,873 sentences were simplified into two or more sentences, generating 5,938 sentences, distributed as shown in Table 5. The split leads to an increase of 53% in the overall quantity of simplified sentences.

Each of the resulting sentences is obviously simpler than the split sentence, however, differently from the other pairs, the sentences deriving from split are part and not an integral simplified version of the respective simplified sentence. To evaluate the effect of splitting on the accuracy of

[2]https://www.folha.uol.com.br
[3]https://gauchazh.clicrbs.com.br

Input/Output Levels	Input (A+B)	Non-split sentences (A)	Split sentences (B)	Sentences resulting from split (C)	Output (A+C)	Percentage Increase
Original/ Natural	2,372	1,543	829	1,992	3,535	49%
Natural/ Strong	1,501	782	719	1,621	2,403	60%
TOTAL	3,873	2,325	1,548	3,613	5,938	53%

Table 5: Distribution of sentences increase due to split.

the complexity assessment task, we created three versions of PorSimplesSent. The three versions are very similar, as they pair all the sentences with their respective simplified sentences. They differ in what concerns split sentences.

As we can see in Table 6, the first version, PorSimplesSent1, has 10,616 pairs, including a pair for each sentence resulting from split. The second version, PorSimplesSent2, has 4,968 pairs and, for split sentences, selects only the simplification with greatest score after applying a linear combination of total number of words and word overlapping count, as exemplified in the following. The third version, PorSimplesSent3, disregard all the split sentences and has 2,600 pairs.

Types of Pairs	PorSimplesSent1	PorSimplesSent2	PorSimplesSent3
Original-Natural	3,535	2,372	1,543
Natural-Strong	4,976	1,501	782
Original-Strong	2,105	1,095	275
TOTAL	10,616	4,968	2,600

Table 6: Distribution of pairs by level in the three versions of PorSimplesSent.

For example, given an Original sentence (O) simplified into two sentences in Natural level (N1 and N2):

- (O): O dormitório, de aproximadamente cinco metros por cinco metros, completa-se com um guarda-roupas de duas portas, uma mesa, um frigobar e um aparelho de ar-condicionado. (The dormitory, approximately five meters by five meters, is complete with a two-door wardrobe, a table, a minibar and an air-conditioner.)
- (N1): O dormitório tem mais ou menos cinco metros por cinco metros. (length: 11 words; overlapping: 7 words; score: 11+7=18) (The dormitory is about five meters by five meters.)
- (N2): O dormitório se completa com um guarda-roupas de duas portas, uma mesa, um frigobar e um aparelho de ar-condicionado. (length: 19 words; overlapping: 19 words; score: 19+19=38) (The dormitory is complete with a two-door wardrobe, a table, a minibar and an air-conditioning unit.)

For PorSimplesSent1, we generated 2 pairs: O-N1 and O-N2. For PorSimplesSent2, we generated 1 pair: O-N2. The original was paired with the sentence N2, which presented a score of 38, against a score of 18 of the sentence N1. For PorSimplesSent3 we did not generate any pair with these sentences.

4 Corpus Validation

4.1 Method

To validate the corpus and to contribute with an initial baseline for the task in Portuguese, we evaluated a simple, but successful approach, inspired by Vajjala and Meurers (2016) —

the pair-wise ranking. For sentential complexity, each sentence should receive a score from an ordinal list of complexity, which could be 1 to n, being n the most difficult. Once the ranking method receives a pair of sentences (with feature vectors) it will predict which one is simpler than the other. The problem of sentential complexity is reduced to the comparison of sentences pairs taken from a pool of sentences where the objective is to rank them according to their complexity, trying to minimize inversion of ranks. As these authors, we also chose the RankSVM algorithm implemented in SVM$^{\text{Rank}}$ (Joachims, 2006)[4], which presented the best results among the algorithms tested for the task in English. We gave the rank value 2 to the complex side and value 1 to the simplified side of each sentence pair.

4.2 Features

For this experiment, we evaluated previously the sets of Original, Natural and Strong simplified sentences of PorSimples Corpus, using two publicly available NLP tools for Portuguese to extract textual metrics, which can be used to aid the automated analysis of text readability: Coh-Metrix-Port 2.0[5] (Scarton et al., 2010; Aluísio and Gasperin, 2010) and Coh-Metrix-Dementia[6] (Aluísio et al., 2016), both based on Coh-Metrix (Graesser et al., 2004). Also, we were inspired by another tool named AIC[7], built in PorSimples project which defined several syntactic metrics to be used in evaluation of text readability. Then we chose the 17 features that presented a clear tendency (increase or decrease, depending on the feature) in the three levels compared (see Table 7 and 8) in order to train a predictor.

Table 7 shows mean values of syntactic metrics for Original (O), Natural (N) and Strong (S) sentence levels in PorSimples corpus. In the table, S stands for Number of Sentences, CpS for Clauses per Sentence, ApC for Apposition per Clause, DD for Dependency Distance, MaxNP and MeanNP for Max and Mean Noun Phrase, SC for Subordinate Clauses, MVPpS for Mean Verb Phrase per Sentence, NIV for Non Inflected Verbs, PSR for Postponed Subject Ratio and ISC for Infinite Subordinate Clauses.

Table 8 shows mean values of lexical and psycholinguistic metrics for Original (O), Natural (N) and Strong (S) sentence levels in PorSimples corpus. In the table, WpS stands for Words per sentence, SpCW for Syllables per Content Words and WbMV for Words before Main Verbs.

L	S	CpS	ApC	DD	MaxNP	MeanNP	SC	MVPpS	NIV	PSR	ISC
O	2372	2.62	0.07	48.24	9.87	5.84	0.38	2.24	0.31	0.085	0.179
N	3535	1.95	0.02	28.39	7.35	4.79	0.26	1.71	0.22	0.051	0.124
S	2402	1.74	0.01	22.16	6.48	4.39	0.24	1.55	0.21	0.052	0.117

Table 7: Distribution of corpus sentences according to the level (L) of simplification - Syntactic Metrics.

L	WpS	SpCW	WbMV	Yngve	Frazier	Honoré	Brunet
O	21.01	2.86	6.16	2.89	7.38	1214.16	40.29
N	14.77	2.74	4.09	2.43	6.64	727.87	51.44
S	12.79	2.76	3.73	2.32	6.48	563.98	52.14

Table 8: Distribution of corpus sentences according to the level (L) of simplification - Lexical, Psycholinguistic and the Classic Syntactic Metrics of Yngve and Frazier.

The features are from three different groups: 1-4 are lexical; 5-16 measures syntactic complexity, and the last one is a psycholinguistic measure of working memory overload:

[4] `https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`
[5] http://143.107.183.175:22680
[6] http://143.107.183.175:22380
[7] http://conteudo.icmc.usp.br/pessoas/taspardo/NILCTR0808.pdf

1. **Syllables per content word**: Average number of syllables per content word;
2. **Words per sentence**: Number of words in the sentence;
3. **Brunet**: Classic formula, its a type token ratio form less sensitive to text size (Thomas et al., 2005);
4. **Honoré**: Classic formula similar to Brunet but vocabulary-based (Thomas et al., 2005);
5. **Mean verb phrase per sentence**: Measures the quantity of verb phrases per sentence (implemented via tagger, counts verbs in a sentence);
6. **Yngve**: Measures how much a syntactic tree escapes from the pattern that tend to have branches to the right (Yngve, 1960);
7. **Frazier**: A bottom-up approach to calculate syntactic complexity of a sentence (Frazier, 1985);
8. **Dependency distance**: Calculates dependency distances in the syntactic tree; as dependency distances grows, the text complexity grows together;
9. **Apposition per clause**: Number of appositions in the sentence divided per number of clauses;
10. **Clauses per sentence**: Number of clauses in a sentence (implemented via parser Palavras (Bick, 2000); counts main verbs, excluding auxiliary verbs);
11. **Max noun phrase**: Maximum length of noun phrase in a sentence, calculated in words;
12. **Mean noun phrase**: Mean of noun phrase length in a sentence, calculated in words;
13. **Postponed subject ratio**: Occurrence of Verb-Subject order instead of canonical Subject-Verb order, calculated in relation to the total number of clauses;
14. **Subordinate clauses**: Proportion of subordinate clauses to the total number of clauses;
15. **Infinite subordinate clauses**: Proportion of subordinate clauses made by verbs in infinitive, gerund and past participle form;
16. **Non-inflected verbs**: Number of verbs that have not been inflected, that is, which are in infinite form: infinitive, gerund and past participle;
17. **Words before main verb**: Number of words before the main verbal phrase.

4.3 Evaluation

The 10-fold cross validation accuracy results are displayed in Table 9. As baselines for our tests, we chose four unique features and evaluated them individually on SVM^{Rank}: a) Words before main verb, b) Clauses per sentence, c) Syllables per content word and d) Tokens per sentence. The last line shows the results of our method with 17 features, detailed in Section 4.2.

Features	PorSimplesSent1	PorSimplesSent2	PorSimplesSent3
Words before main verb	45.13%	36.29%	23.06%
Clauses per sentence	59.02%	41.28%	11.32%
Syllables per content word	54.80%	50.90%	46.33%
Tokens per sentence	80.74%	69.35%	40.76%
All 17 features	**83.39%**	**74.20%**	**53.67%**

Table 9: Baselines and first experiment results (accuracy), using SVMRank.

In **PorSimplesSent1**, as expected, using just the number of tokens per sentence it is possible to achieve more than 80% of accuracy. This is because this dataset includes all sentences that are result of split operations, so the majority of simplified sentences are small parts from the original ones. The **PorSimplesSent3**, which has only full sentences, disregarding those that suffered split, is the most difficult to rank. Besides having the smallest number of pairs, PorSimplesSent3 has some simplified sentences that are bigger than the original ones. The **PorSimplesSent2**, on its turn, is a middle term between the previous two: it has split sentences, but only the longest sentence derived from the split is paired with the original sentence. Therefore, we have chosen the dataset PorSimplesSent2 to be our gold standard for sentential complexity task in Portuguese.

Our model with 17 features presents improvement over the strongest baseline (Tokens per Sentence): 2.65 in PorSimplesSent1, achieving 83.39% accuracy; 4.85 in PorSimplesSent2, achieving 74.20% accuracy; and 12.91 in PorSimplesSent3, achieving 53.67% accuracy.

4.4 Error Analysis

We performed a manual analysis, trying to understand the errors made by our model, in order to improve it with new features. Building on the syntactic and lexical operations used to annotated the PorSimples corpus, but now with focus on operations at the sentence level, we proposed a set of 14 labels to annotate the errors. Table 10 shows the errors found after this analysis.

Label Description	Qty	%
1 Replacement by word of the same grammatical class, including multiword discourse markers	169	28.89
2 Replacement by word of different grammatical class, without specifying the classes involved	19	3.25
3 Replacement by paraphrase (one word by several words)	111	18.97
4 Removal of clause	6	1.03
5 Removal of syntactic constituent (subject, adverbial adjunct, etc.)	8	1.37
6 Removal of words	31	5.30
7 Removal of parentheses	10	1.71
8 Insertion of words	33	5.64
9 Change in the order of constituents (such as putting the subject first and the adverb last)	44	7.52
10 Change to active voice	21	3.59
11 Change to synthetic (shortest) passive voice form (by means of passivizating particle “se”)	3	0.51
12 Change from direct to indirect speech	2	0.34
13 Rephrasing	48	8.21
14 ERROR (equal sentences or alignment error, which will be excluded from the corpus)	80	13.68

Table 10: List of Errors used to annotate 418 sentence pairs of PorSimplesSent3.

We annotated 209 of the 418 sentence pairs of PorSimplesSent3 for which our model missed the prediction. The annotation performed by two annotators was double blind and multi-label. A discussion on the pairs presenting annotation disagreement helped to clarify doubts on the annotation process and to assign commonly agreed labels. After that, the remaining sentence pairs were divided into two parts and each part was assigned to only one annotator.

The analysis of these numbers lead us to cogitate which features and metrics might be significant to improve the performance of our ranking model, initially trained with 17 linguistic and psycholinguistic features. Both most frequent labels, 1 and 3, relate to lexical substitution. Example 1 below shows a pair of sentences annotated only with the label 1.

Example 1

- (O): Quem é contra diz que os cães sujam a praia e colocam em risco a saúde dos veranistas. (Those who are against say that the dogs dirty the beach and put at risk the health of the vacationers.)

- (N): Quem é contra diz que os cães sujam a praia e colocam em risco a saúde das pessoas. (Those who are against say that the dogs dirty the beach and put at risk the health of the people.)

The only difference between the two sentences is the pair of words "veranistas" versus "pessoas", in a hyponym relationship. Example 2 brings a pair annotated with label 3. It shows 2 substitutions by paraphases, here understood as a word replaced by several ones, similar in meaning: "possibilitar" by "tornará possível" and "hepática" by "do fígado".

Example 2

- (O): A descoberta possibilitará que pessoas com dano no fígado usem as próprias células-tronco para produzir células hepáticas. (The discovery will enable people with liver damage to use their own stem cells to produce hepatic cells.)
- (N): A descoberta tornará possível que pessoas com dano no fígado usem as próprias células-tronco para produzir células do fígado. (The discovery will make it possible for people with liver damage to use their own stem cells to produce liver cells.)

As many sentence pairs differ by only one word, readability measures to compare words are essential to decide which is the easiest sentence. Word frequency and psycholinguistic properties of words (as age of acquisition, familiarity, concreteness, imageability) may be useful for this purpose. Additionally, there are several resources that may be used to design new metrics to deal with similar words and paraphases. For Portuguese, there are different similar projects of wordnets, among which stand out the OpenWordNet-PT (de Paiva et al., 2012), as the most complete with manual revision, and the CONTO.PT (Gonçalo Oliveira, 2016), built semi-automatically in order to comprise a greater number of words, and which describes itself as a diffuse wordnet. There is also the PPDB (Paraphrase Database), a resource that contains paraphrases in several languages, including Portuguese, automatically extracted from bilingual corpora (Ganitkevitch and Callison-Burch, 2014). Paraphrase in the context of PPDB refers to expressions or equivalent words. As it was generated automatically, the PPDB also contains some false positives. The resource is available in six different sizes: the difference is that larger sets extracted paraphrase rules with less confidence.

For features other than the lexical ones, a very promising research avenue is to test simplified sentences with human readers to confirm whether they are simpler than their original counterparts or not (using eye-trackers). This is relevant because many simplification operations we use are inspired in the literature regarding English language simplification and we need more evidence related to Portuguese language. The error analysis, therefore, provided important insights for future work aiming to increase the accuracy of our model in the dataset made available with this paper. Besides that, 80 pairs were dropped from our dataset because they contain nearly identical sentences or completely different sentences (improperly paired due to alignment error). Therefore, all the three totals in Table 6 were reduced by 80, resulting in 10,536, 4,888, and 2,520 sentences, respectively.

5 Conclusions

In this paper, we presented a new resource to evaluate the task of sentence readability for Portuguese language - the corpus PorSimplesSent. This dataset is larger, in terms of sentence pairs, than a similar corpus for the English language (cg. (Vajjala and Meurers, 2016)), and it is the first resource of this kind for Portuguese language, therefore we believe we can have a blossom of future research for this task. Moreover, we made available four baselines for the corpus and an approach based on pairwise ranking to compare two versions of a sentence. Our model uses 17 lexical, syntactic and psycholinguistic features and identifies the readability level of sentence pairs with an accuracy of 74.2%; an improvement of 2.65 on the strongest baseline. We believe there is plenty of room for improvement of our model and we hope this task receive a lot of attention from researchers devoted to Portuguese language NLP as well. The corpus is made publicly available at `http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources`. As for future work, we will enlarge the number of features to build an improved model to evaluate the task and organize a shared task using it in an NLP conference.

References

Sandra M. Aluísio, Andre Cunha, and Carolina Scarton. 2016. Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In *12th International Conference on Computational Processing of the Portuguese Language (PROPOR 2016)*, volume 9727 of *Lecture Notes in Computer Science*, pages 109–114. Springer Cham.

Sandra M. Aluísio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the Porsimples project for simplication of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas - Association for Computational Linguistics*, pages 46–53, Stroudsburg, PA.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental CCG parser. In *HLT-NAACL*, pages 1051–1057, Stroudsburg, PA. The Association for Computational Linguistics.

Eckhard Bick. 2000. The parsing system Palavras: Automatic grammatical analysis of Portuguese in a Constraint Grammar Framework. *Aarhus University Press.*

Helena M. Caseli, Tiago F. Pereira, Lúcia Specia, Thiago A. S. Pardo, Caroline Gasperin, and Sandra M. Aluísio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *Advances in Computational Linguistics, Research in Computer Science (CICLing-2009)*, volume 41, pages 59–70.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, December. The COLING 2012 Organizing Committee. Published also as Techreport http://hdl.handle.net/10438/10274.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of Italian texts with a view to text simplification. In *SLPAT '11 Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Stroudsburg, PA. Association for Computational Linguistics.

Felice Del'Orletta, Simonetta Montemagni, and Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *International Journal of Applied Linguistics (ITL). Special Issue on Readability and Text Simplification.*

Leandro Borges dos Santos, Magali Sanches Duran, Nathan Siegle Hartmann, Gustavo Henrique Paetzold Arnaldo Candido, and Sandra Maria Aluisio. 2017. A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In *International Conference on Text, Speech, and Dialogue (TSD 2017)*, volume 10415 of *Lecture Notes in Computer Science*, pages 281–289. Springer, Cham.

William H. Dubay. 2007. *Smart Language: Readers, Readability, and the Grading of Text.* Impact Information, Costa Mesa, CA.

Lyn Frazier. 1985. Syntactic complexity. *D.R. Dowty, L. Karttunen and A.M. Zwicky (eds.), Natural Language Parsing, Cambridge University Press.*

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Ana Valeria Gonzalez-Garduño and Anders Søgaard. 2017. Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, Stroudsburg, PA. The Association for Computational Linguistics.

Hugo Gonçalo Oliveira. 2016. Conto.pt: Groundwork for the automatic creation of a fuzzy portuguese wordnet. In *12th International Conference on Computational Processing of the Portuguese Language (PROPOR 2016)*, volume 9727 of *Lecture Notes in Computer Science*, pages 283–295. Springer Cham.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36:193–202.

David M. Howcroft and Vera Demberg. 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 958–968, Stroudsburg, PA. The Association for Computational Linguistics.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, Stroudsburg, PA. The Association for Computational Linguistics.

IPM. 2016. Inaf brasil 2015: Indicador de alfabetismo funcional - alfabetismo no mundo do trabalho. *Instituto Paulo Montenegro*.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, volume 3, pages 217–226. ACM Press.

Carolina Scarton, Caroline Gasperin, and Sandra Aluísio. 2010. Revisiting the readability assessment of texts in Portuguese. In Simari G.R. Kuri-Morales A., editor, *12th Ibero-American Conference on AI, Advances in Artificial Intelligence – IBERAMIA 2010*, volume 6433 of *Lecture Notes in Computer Science*, pages 306–315, Berlin, Heidelberg. Springer.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. Simpa: A sentence-level simplification corpus for the public administration domain. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajakrishnan Rajkumar. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee.

Johan Sjöholm. 2012. *Probability as readability: A new machine learning approach to readability assessment for written Swedish*. LiU Electronic Press.

Lucia Specia, Sandra M. Aluísio, and Thiago A. S. Pardo. 2008. Manual de simplificação sintática para o português. NILC Technical Report 08-06, ICMC-USP, jun. Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional (NILC-TR-08-06), 27 p.

Calvin Thomas, Vlado Keselj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. 2005. Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In *Proceedings of IEEE ICMA 2005*, volume 3, pages 1569–1574. IEEE.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 288–297.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *CoRR*, abs/1603.06009.

David Wiley, T.J. Bliss, and Mary McEwen. 2014. Open educational resources: A review of the literature. In Spector J., Merrill M., Elen J., and Bishop M., editors, *Handbook of Research on Educational Communications and Technology: Fourth Edition*, pages 781–789, New York, NY. Springer.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Victor H Yngve. 1960. A model and hypothesis for language structure. *Proceedings of the American Philosophical Association*, 104(5):444–466.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING), August 2010. Beijing, China*, pages 1353–1361. The COLING 2010 Organizing Committee.

5.2 Primeira avaliação da tarefa com *Pairwise Ranking*

Título:	**Avaliação Automática da Complexidade de Sentenças do Português Brasileiro para o Domínio Rural**
Autores:	**Sidney E. Leal, Vanessa M. A. Magalhães, Magali S. Duran e Sandra M. Aluísio**
Ano:	**2019**
Conferência:	**STIL - Symposium in Information and Human Language Technology - Salvador - BA**
Situação:	**Publicado**

No artigo anterior, foram definidos alguns *baselines* para a tarefa, usando contagens simples e SVM. Este artigo pode ser visto como a continuação da investigação da tarefa iniciada naquele, usando redes neurais artificiais e mais *features*.

Com mais métricas linguísticas e psicolinguísticas obtidas com o NILC-Metrix e seleção de *features*, o melhor método conseguiu atingir 87,80% de acurácia no PorSimplesSent.

Fora o método principal, o artigo traz um método para utilização posterior à classificação dos pares de sentença, usando *ranking* para o treinamento de um regressor que consegue julgar uma sentença isolada nunca vista.

O método foi testado no domínio rural com o apoio da pesquisadora Vanessa Magalhães da Embrapa e demonstrou boa robustez.

O artigo foi apresentado no STIL em 2019, que aconteceu em conjunto com o BRACIS em Salvador - BA, e foi premiado como **terceiro *Best Paper*** da conferência.

Avaliação Automática da Complexidade de Sentenças do Português Brasileiro para o Domínio Rural

Sidney E. Leal[1], Vanessa M. A. Magalhães[2], Magali S. Duran[1], Sandra M. Aluísio[1]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13560-970 – São Carlos – SP
[2]Núcleo de Gestão da Informação e Conhecimento, Embrapa Gado de Leite
Juiz de Fora - MG
[1]sidleal@gmail.com, magali.duran@uol.com.br, sandra@icmc.usp.br
[2]vanessa.magalhaes@embrapa.br

Abstract. *Low literacy is a common problem in the Brazilian dairy sector that may undermine productivity. Hence the importance of simplifying newsletters, technical texts and instructions addressed to this public. The task of automatic evaluation of sentential complexity is new to Portuguese and allows us, for example, to identify which sentences in a text should be simplified. This paper presents a 3-step method for this task, using classical machine learning with MLP neural networks for ranking and regression. The model was trained in a public corpus of sentences collected from journalistic texts and its generalization to other scenarios was evaluated for the rural domain. We obtained accuracy of 87.80% in the ranking, root-mean-square error (RMSE) of 0.06 in the regressor and F-measure of 88.4% in the robustness test.*

Resumo. *A maioria dos produtores de leite possuem baixo letramento, o que prejudica seu acesso às tecnologias com vistas à melhoria das condições de trabalho, bem como ao aumento na produção e na renda. Esse é o motivo da importância de simplificar informativos e textos técnicos dirigidos a esse público. A tarefa de avaliação automática da complexidade sentencial é nova para o português e permite identificar, por exemplo, quais sentenças em um texto devem ser alvo de simplificação. Este artigo apresenta um método de 3 passos para essa tarefa, utilizando redes neurais do tipo MLP para ranqueamento e regressão. O modelo foi treinado em um córpus público de sentenças do gênero jornalístico e sua generalização para outros cenários foi avaliada para o domínio rural. Foram obtidas uma acurácia de 87,80% no ranqueamento, raiz do erro quadrático médio (ou RMSE, em inglês) de 0,06 no regressor e F-measure de 88,4% no teste de robustez.*

1. Introdução

Segundo o INAF Brasil (Indicador de Alfabetismo Funcional), apenas um em cada dez brasileiros adultos é considerado letrado de forma proficiente [IPM 2018]. Esse indicador é bastante alarmante e explicita um dos grandes desafios brasileiros: o acesso à evolução econômica e tecnológica pela população. Percebe-se que a situação é ainda mais crítica quando isolamos certos setores da economia, como o da agropecuária, em que apenas 1% dos entrevistados foram considerados proficientes. Isso significa que a quase totalidade dos produtores rurais pode não ser capaz de usufruir das novas tecnologias desenvolvidas

e publicadas por entidades de pesquisa. A falta de acesso ao conhecimento prejudica bastante esse setor, um dos mais importantes do Brasil, que é responsável por 23% do Produto Interno Bruto (PIB)[1] e 40% da renda da população economicamente ativa[2].

Uma alternativa viável na atualidade é simplificar essas publicações, como fez a Embrapa Gado de Leite, no projeto APP@Rural [Magalhães et al. 2017], que simplificou os comunicados técnicos e partes do Manual de Bovinocultura de Leite, tornando-os mais acessíveis aos produtores, estudantes e extensionistas. A Embrapa utiliza os métodos de classificação textual e de simplificação lexical e sintática do projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) [Aluisio and Gasperin 2010], com adaptação da simplificação lexical para atender aos termos técnicos do domínio rural. O PorSimples teve como objetivo promover o acesso a textos em português brasileiro (PB) por analfabetos funcionais e crianças ou adultos em fase de alfabetização e criou os modelos automáticos com base em textos jornalísticos.

O trabalho presente propõe uma evolução nesses métodos ao criar um método para indicar automaticamente as sentenças alvos de simplificação, permitindo a sua classificação nos quatro níveis indicados pelo relatório de 2018 do INAF: Proficiente, Intermediário, Elementar e Rudimentar.

Na Seção 2, são apresentados a tarefa e os principais trabalhos da área de complexidade sentencial. Na Seção 3, é apresentado o método de avaliação da complexidade proposto neste trabalho, juntamente com o córpus de sentenças alinhadas PorSimplesSENT, um resumo das *features* e o modelo de aprendizado de máquina escolhido. Finalmente, a Seção 4 mostra o teste de robustez feito para o modelo treinado com textos jornalísticos e avaliado em textos produzidos pela Embrapa Gado de Leite, chamados aqui de textos do domínio rural.

2. Avaliação da Complexidade Sentencial

A inteligibilidade[3] de um texto, do inglês *text readability* é, segundo [Dubay 2007, pg.6], *a facilidade de leitura de um texto criada pela escolha de conteúdo, estilo, estruturação e organização que atende ao conhecimento prévio, habilidade de leitura, interesse e motivação da audiência*. As primeiras fórmulas de inteligibilidade foram criadas há quase um século, na década de 1920, nos Estados Unidos e consideravam que a complexidade poderia ser inferida por métricas de palavras e sentenças, baseadas na frequência e tamanho (quantidade de letras) das palavras e na média da quantidade de palavras por sentença. Desde então, a Inteligibilidade Textual tornou-se uma grande área de pesquisa multidisciplinar, com uma vasta bibliografia, e ganhou novas abordagens neste século com o uso de métodos de PLN e AM.

Tradicionalmente, a tarefa tem sido aplicada no nível textual, atribuindo uma nota (ou nível de *ranking*, de proficiência) para um documento inteiro. Porém, em um documento classificado como simples, podem ocorrer sentenças complexas, assim como existem sentenças simples em um documento complexo. Uma sentença é uma unidade

[1] http://www.agricultura.gov.br/noticias/agropecuaria-puxa-o-pib-de-2017

[2] http://www.mda.gov.br/sitemda/noticias/agricultura-familiar-do-brasil-\%C3%A9-8%C2%AA-maior-produtora-de-alimentos-do-mundo

[3] Neste trabalho, usamos os termos complexidade e inteligibilidade (seja no nível sentencial ou textual) como sinônimos.

importante que traz, na maioria das vezes, informação suficiente para inferência e análise da sua complexidade. Embora seja possível usar a mesma abordagem de avaliação da complexidade dos textos para o nível das sentenças, [Dell'Orletta et al. 2014] demonstraram que é necessário um número maior de *features* para a segunda tarefa.

A Tabela 1 mostra: a) uma simplificação por meio da substituição lexical, em que um termo técnico é substituído por outro mais frequente; b) uma sentença simplificada no nível sintático por meio da sua divisão em duas sentenças e c) um caso de elaboração textual, incluindo uma breve explicação de um termo técnico.

Tabela 1. Exemplos de sentenças simplificadas

a) Lexicalmente	Original	Se acentuada e prolongada, a **hipertermia** pode causar a morte do animal.
	Simplificada	Se acentuada e prolongada, a **febre** pode causar a morte do animal.
b) Sintaticamente	Original	O uso de forragem conservada, cujas formas mais comuns são: ensilagem e fenação, é uma solução para alimentar o rebanho.
	Simplificada	O uso de forragem conservada é uma solução para alimentar o rebanho. As formas mais comuns para conservar forragens são: ensilagem e fenação.
c) Elaboração textual	Original	A ensilagem é o processo de conservação do alimento por fermentação anaeróbia.
	Simplificada	A ensilagem é o processo de conservação do alimento por fermentação anaeróbia **(sem a presença de ar)**.

A avaliação do nível de inteligibilidade de sentenças é uma tarefa de pesquisa recente e visa analisar e avaliar individualmente as sentenças de um texto, permitindo uma informação mais acurada dos pontos complexos de um texto. Essa abordagem é relevante, pois, como afirmam [Dell'Orletta et al. 2014] as abordagens de classificação de inteligibilidade que levam em consideração os textos inteiros não trazem grandes vantagens para a posterior aplicação de métodos automáticos de simplificação. Além disso, considerar como complexas todas as sentenças de um texto classificado como complexo, pode prejudicar o treinamento dos métodos, principalmente quando essas sentenças são utilizadas para avaliar a tarefa de predição da complexidade sentencial. Isso foi demonstrado por [Vajjala and Meurers 2014] durante a investigação dos motivos da baixa acurácia que obtiveram ao utilizar o córpus Wikipedia-SimpleWikipedia (sem alinhamento sentencial).

O primeiro trabalho a considerar a tarefa de complexidade especificamente para o nível sentencial foi [Dell'Orletta et al. 2011], comparando a sua dificuldade em relação ao nível textual. Porém, a definição da forma de avaliação da tarefa só foi consolidada por [Vajjala and Meurers 2016] e permitiu que os trabalhos posteriores aperfeiçoassem os resultados comparativamente. [Ambati et al. 2016] conseguiram melhorar significativamente os resultados utilizando um parser do tipo *Combinatory Categorial Grammar* (CCG), e [Gonzalez-Garduño and Søgaard 2018] chegaram no estado da arte para o inglês, utilizando métricas de rastreamento ocular aliadas às linguísticas e psicolinguísticas. [Howcroft and Demberg 2017] e [Singh et al. 2016] também publicaram trabalhos explorando a tarefa com novas métricas; o primeiro trabalho exclusivamente

com métricas psicolinguísticas e o segundo com métricas de rastreamento ocular. Uma comparação dos resultados obtidos para a tarefa de avaliação da complexidade sentencial na língua inglesa é mostrada na Tabela 2.

Tabela 2. Avaliação da tarefa na Wikipedia-SimpleWikipedia

Trabalho	Método	Acurácia
Flesch-Kincaid	Método Clássico	72,30
[Vajjala and Meurers 2016]	RankSVM	74,58
[Ambati et al. 2016]	SMO/Incr CCG	78,87
[Singh et al. 2016]	Regressão Log.	75,21
[Howcroft and Demberg 2017]	RankAsClass.	73,22
[Gonzalez-Garduño and Søgaard 2018]	MultiTask MLP	**86,62**

Mais recentemente, [Stajner et al. 2017] e [Scarton et al. 2018] contribuíram para a tarefa, avaliando a complexidade com o apoio do córpus Newsela (que possui 550 mil sentenças, três vezes maior que o Wikipedia-SimpleWikipedia) e [Bosco et al. 2018] obtiveram bons resultados para o italiano, utilizando Redes Neurais Recorrentes do tipo *Long Short Term Memory* (LSTM). Finalmente, [Brunato et al. 2018] contribuíram com um trabalho sobre a percepção da complexidade e concordância entre anotadores, enquanto [Timm 2018] investigou simplificações sentenciais automáticas, utilizando rastreamento ocular. Para o português (até onde foi possível verificar), foi encontrado apenas o trabalho de [Leal et al. 2018], em que foi publicado o córpus PorSimplesSent, com foco nesta tarefa.

3. Método de Avaliação da Complexidade Sentencial para o Português

3.1. Córpus

O PorSimplesSent [Leal et al. 2018] é um córpus de sentenças alinhadas, disponível publicamente, que foi compilado a partir do PorSimples [Caseli et al. 2009] e organizado a partir do alinhamento sentencial dos textos, em três níveis: a) **Original:** Sentenças originais; b) **Simplificação Natural:** Textos simplificados de forma livre pelos anotadores e c) **Simplificação Forte:** Textos simplificados seguindo as regras do manual desenvolvido no projeto.

O córpus PorSimplesSent possui três versões, com diferentes abordagens para as sentenças que sofreram operação de divisão. O PSS1 repete a sentença original para cada sentença resultante da divisão; o PSS2 seleciona apenas a maior sentença resultante, que também tenha maior sobreposição de palavras, e o PSS3 contém apenas sentenças que não sofreram divisão, sendo, portanto, o menor dos três. Para este trabalho foi escolhida a versão **PSS2**, que possui 4.962 pares de sentenças, com alinhamentos Original-Natural, Natural-Forte e Original-Forte, obtida no formato TSV (*Tab Separated Values*)[4].

3.2. *Features*

Este trabalho utiliza como *features* as métricas disponibilizadas pelas ferramentas Coh-Metrix-Port, Coh-Metrix-Dementia, LIWC, AIC e as métricas psicolinguísticas disponibilizadas por [dos Santos et al. 2017], que anotaram automaticamente um banco[5]

[4] http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources

[5] A base está disponível em: http://143.107.183.175:21380/portlex/index.php/en/component/content/article/2-uncategorised/23-psycholinguistic

de 26.874 palavras do PB com Imageabilidade, Concretude, Familiaridade e Idade de Aquisição. O conjunto utilizado totaliza 189 features.

O Coh-Metrix-Port[6] [Scarton and Aluísio 2010] é uma adaptação para o PB do Coh-Metrix, desenvolvida dentro do projeto PorSimples, e implementa 48 métricas, divididas nas categorias: contagens básicas, operadores lógicos, frequências, hiperônimos, tokens, constituintes, conectivos, ambiguidade, co-referência e anáforas. O Coh-Metrix-Dementia[7] [Cunha 2015] é uma adaptação do Coh-Metrix-Port para análise automática de distúrbios de linguagem nas demências (como Doença de Alzheimer) ou no Comprometimento Cognitivo Leve (CCL). Ele adiciona 25 novas métricas às 48 do Coh-Metrix-Port, nas categorias: disfluências, análise de semântica latente, diversidade lexical, complexidade sintática e densidade semântica. LIWC (*Linguistic Inquiry and Word Count* - liwc.wpengine.com) é uma ferramenta baseada em dicionários para análise dos vários componentes emocionais, cognitivos e linguísticos em amostras de textos, com categorias como: estatísticas comuns do texto, dimensão linguística, processos psicológicos, relatividade, assuntos pessoais e miscelânea, totalizando aproximadamente 100 métricas [Cunha 2015]. A tradução e adaptação do dicionário para o PB foi realizada em uma colaboração entre NILC, Checon Pesquisa e Unisinos no período de 2010 a 2012 e está disponível no site do projeto PortLex[8]. Também criada dentro do contexto do PorSimples [Maziero et al. 2008], a ferramenta AIC (Análise Automática de Inteligibilidade de Córpus) traz 39 métricas, com o principal diferencial de utilizar o analisador sintático PALAVRAS [Bick 2000] para o cálculo delas. Elas estão organizadas em seis classes: estatísticas do texto, voz passiva, características das orações, densidade, personalização e marcadores discursivos [Cunha 2015].

3.3. Modelo Proposto

O treinamento do modelo foi feito em três fases, ilustradas na Figura 1, na qual é feita uma analogia entre a complexidade e o espectro de cores (vermelho = mais complexo, azul = mais simples). Neste exemplo, uma sentença nunca vista antes (F3), de cor verde mais intensa que o verde da posição 4, tem o valor estimado de complexidade 4,5 (entre 4-verde e 5-amarelo).

Fase 1 - *Pairwise Ranking*

A primeira tarefa consistiu em treinar uma rede neural MLP (*Multi Layer Perceptron*) com 3 camadas, utilizando os 4.962 pares de sentenças do córpus PSS2 e todas as *features* disponíveis. Metade dos pares foi invertida, de forma a balancear melhor as duas classes: o lado complexo foi anotado com 1 e o lado simples com 0.

A camada de entrada da MLP contou com 378 neurônios (189 *features* x 2 sentenças); a camada oculta foi configurada com 30 neurônios, utilizando a função de ativação *ReLU*, e a saída possui apenas um neurônio, com a função *sigmoid*, predizendo 1 quando a sentença A é mais complexa que a B e 0 no inverso. O ranqueador conseguiu uma acurácia de **87,8%**, utilizando *10 fold cross validation*, um pouco superior ao estado da arte da tarefa para a língua inglesa no córpus Wikipedia-SimpleWikipedia. Utilizamos como *baseline* os resultados reportados por [Leal et al. 2018], cujo melhor modelo atin-

[6] http://fw.nilc.icmc.usp.br:22680/
[7] http://fw.nilc.icmc.usp.br:22380/
[8] http://nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc

Figura 1. Diagrama da sequência de passos do modelo

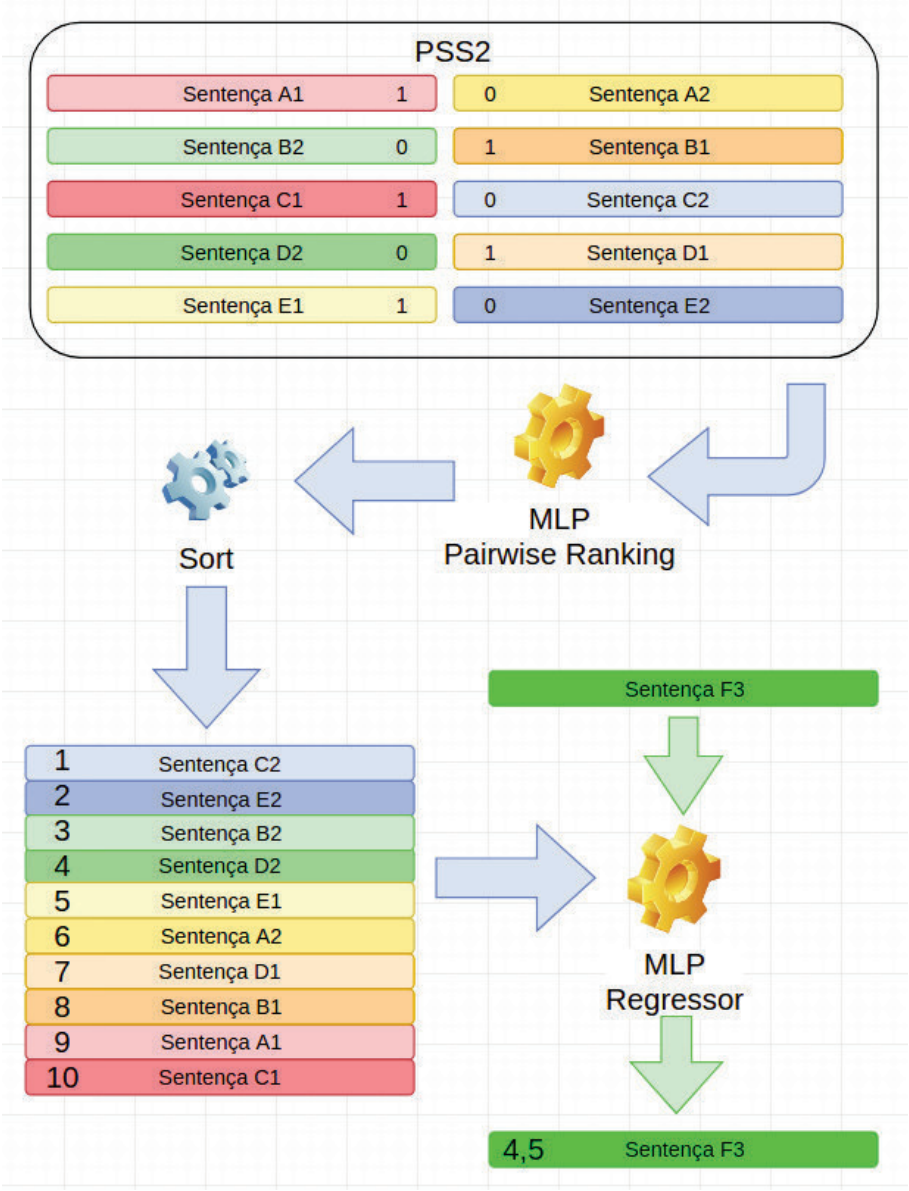


Tabela 3. *Baselines* e resultados do ranqueamento no PSS2

Modelo	Acurácia
Número de orações por sentença	41,28%
Média de sílabas por palavra de conteúdo	50,90%
Total de *tokens* por sentença	69,35%
SVMRank [Leal et al. 2018]	74,20%
MLP Pairwise Ranking	**87,80%**

giu 74,20% no mesmo córpus, conforme tabela 3, que também mostra outras 3 *baselines* simples nas primeiras linhas da tabela.

Fase 2 - *Ranking* global de sentenças

Uma vez obtido um modelo que conseguiu julgar razoavelmente bem a comple-

xidade das sentenças apresentadas em pares, ele foi utilizado para comparar todas as sentenças do PSS2, resultando em um *ranking* ordenado de 1 a 9.924. Algumas sentenças estão repetidas, quando aparecem em lados opostos do par, uma vez como a mais simples e outra vez como a mais complexa. Essas sentenças foram mantidas para uma validação adicional do resultado da ordenação, pois mesmo em lados diferentes, enquanto pares, no *ranking* global elas precisam estar próximas.

Conforme esperado, a maioria das sentenças do nível Original ficaram nas últimas posições do *ranking*, enquanto as primeiras posições foram preenchidas pelos níveis Natural e Forte. O primeiro terço do ranking ficou com 16% das sentenças originais, 30% das naturais e 55% das fortes. O último terço ficou com 52% de originais, 35% de naturais e 12% das fortes.

Fase 3 - Regressor

O *ranking* resultante da fase 2 foi normalizado entre 0 e 1 e utilizado para treinar uma segunda rede neural (MLP), também com 3 camadas. Porém, dessa vez, com apenas 189 neurônios na camada de entrada (apenas uma sentença) e predizendo a complexidade entre 0 e 1 no único neurônio de saída (utilizando ReLU), na forma de um regressor. O *dataset* foi dividido em 80% para treinamento e 20% para testes. Com todas as *features*, obtivemos uma raiz do erro quadrático médio (ou RMSE, em inglês) de 0,04 (MSE: 0,0017). Em seguida, foi aplicado o método de seleção de *features* ***Permutation Importance*** implementado no *eli5.sklearn*[9] para escolher as 50 mais importantes [10] e o regressor foi retreinado, obtendo uma RMSE de **0,06** (MSE: 0,0033). Foi implementada uma interface simples para validação no portal *open source* Simpligo (`https://simpligo.sidle.al`), na qual é possível entrar com um texto e conferir os valores preditos para cada sentença, numa escala de complexidade entre 0 e 100.

4. Teste de Robustez: generalização para outros gêneros de texto

O teste de robustez foi projetado para avaliar o desempenho do preditor de complexidade (em termos de *F-Measure*) em sentenças de outros gêneros diferentes do jornalístico no qual o modelo foi treinado. Foram escolhidas 500 sentenças do domínio rural conforme Tabela 4, de materiais selecionados com ajuda de uma pesquisadora da Embrapa Gado de Leite. Os materiais vieram dos gêneros instrucionais/procedimentais, técnicos e administrativos e foram agrupados para atender os quatro níveis de letramento do INAF 2018, sendo eles: rudimentar[11]; elementar[12]; intermediário[13] e proficiente[14].

[9] `https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html`

[10] A lista completa das *features* pode ser vista junto aos códigos fonte do trabalho em `https://github.com/sidleal/simpligo-ranker`

[11] As cartilhas podem ser acessadas a partir dos seguintes links: `https://www.infoteca.cnptia.embrapa.br/handle/doc/1035506` e `https://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1055203`.

[12] Os materiais de EaD vieram do espaço e-Campo (`https://www.embrapa.br/e-campo/`)

[13] Os comunicados técnicos (COT) podem ser acessados a partir dos links: `https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1034878/1/COT77Teormatseca.pdf` e `https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/594901/1/COT28Ensilagemdomilhoedosorgo.pdf`

[14] O Plano Diretor: `http://www.cnpms.embrapa.br/edital/PDU.pdf`.

As sentenças foram passadas pelo regressor e classificadas em 4 níveis, de acordo com a complexidade predita, sendo 1 o mais simples e 4 o mais complexo: Nível 1 (1-25), Nível 2 (26-50), Nível 3 (51-75) e Nível 4 (76-100). Posteriormente, elas foram avaliadas pela pesquisadora da Embrapa que anotou aquelas em que discordou do nível atribuído pelo regressor, atribuindo o nível que considerava apropriado. A Tabela 5 traz os resultados, sendo a *F-Measure* média obtida de **88,4%**.

Tabela 4. Sentenças selecionadas para teste de robustez do piloto

Publicação	Quantidade de Sentenças
Cartilhas de Ensilagem Milho e Sorgo	98
Curso de EaD de Silagem Capim	97
Curso de EaD de Silagens de milho e sorgo para produção de leite	95
Comunicado Técnico sobre Matéria Seca	61
Comunicado Técnico sobre Milho e Sorgo	91
Plano Diretor para Milho e Sorgo	58
Total	**500**

Quanto aos erros, 59,6% foram contíguos (nos quais o nível correto é imediatamente acima ou inferior ao predito) e 40,3% foram erros distantes. Os erros mais comuns aconteceram pela presença de termos que são simples no domínio rural, mas pouco frequentes nos demais domínios, e nas sentenças com pontuação diferente (por exemplo, as terminadas em dois pontos), o que vai exigir reavaliação das métricas utilizadas, incluindo novas métricas de natureza lexical para o domínio rural. Importante salientar que esses erros não são necessariamente do regressor, mas podem ter sido introduzidos por deficiências das ferramentas das etapas anteriores.

Tabela 5. Resultados do teste robustez

Nível	Precisão (%)	*Recall* (%)	*F-Measure* (%)
Nível 1	98,0	76,2	85,7
Nível 2	86,0	83,2	84,6
Nível 3	89,1	94,6	91,8
Nível 4	84,6	100,0	91,6
Média	89,4	88,5	**88,4**

5. Considerações Finais

Este trabalho apresentou uma evolução significativa para a abordagem da tarefa de Avaliação da Complexidade Sentencial para o português brasileiro, com um incremento de mais de 10% na acurácia sobre o melhor resultado anterior reportado em [Leal et al. 2018]. Também disponibilizou uma aplicação prática para o modelo, permitindo a avaliação das sentenças de um texto, além de provar sua capacidade de generalização para outros domínios. O teste de robustez demonstrou que o modelo desenvolvido pode ser útil no apoio da avaliação e simplificação dos materiais usados pela Embrapa, mesmo esses sendo de outros gêneros textuais. A análise de erros mostrou que novas métricas simples podem ajudar a aumentar o desempenho da tarefa. Como trabalhos futuros, pretendemos investigar a tarefa utilizando métodos de *Deep Learning* e *Transfer Learning*, além da inclusão de mais *features* no modelo, em especial as de rastreamento ocular reportadas no trabalho detentor do estado da arte da tarefa para a língua inglesa.

Referências

Aluisio, S. and Gasperin, C. (2010). Fostering digital inclusion and accessibility: the Porsimples project for simplication of Portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas - Association for Computational Linguistics*, pages 46–53.

Ambati, B. R., Reddy, S., and Steedman, M. (2016). Assessing relative sentence complexity using an incremental CCG parser. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1051–1057.

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

Bosco, G. L., Pilato, G., and Schicchia, D. (2018). A neural network model for the evaluation of text complexity in Italian language: a representation point of view. In *Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2018)*, pages 464–470.

Brunato, D., Mattei, L. D., Dell'Orletta, F., Iavarone, B., and Venturi, G. (2018). Is this sentence difficult? Do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699.

Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A. S., Gasperin, C., and Aluísio, S. M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. *Advances in Computational Linguistics, Research in Computer Science (CICLing-2009)*, vol. 41:59–70.

Cunha, A. L. V. (2015). *Coh-Metrix-Dementia: análise automática de distúrbios de linguagem nas demências utilizando Processamento de Línguas Naturais*. Master's thesis, ICMC - USP, São Carlos - SP - Brasil.

Dell'Orletta, F., Wieling, M., Cimino, A., Venturi, G., and Montemagni, S. (2014). Assessing the readability of sentences: Which corpora and features? *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). Read-it: Assessing readability of Italian texts with a view to text simplification. *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

dos Santos, L. B., Duran, M. S., Hartmann, N. S., Candido, A., Paetzold, G. H., and Aluisio, S. M. (2017). A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In Ekštein, K. and Matoušek, V., editors, *Text, Speech, and Dialogue. TSD 2017. Lecture Notes in Computer Science*, volume 10415, pages 281–289. Springer, Cham.

Dubay, W. H. (2007). *Smart Language: Readers, Readability, and the Grading of Text*. Impact Information, Costa Mesa, CA. ISBN: 1-4196-5439-X.

Gonzalez-Garduño, A. V. and Søgaard, A. (2018). Learning to predict readability using eye-movement data from natives and learners. *Proceedings of the The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5118–5124.

Howcroft, D. M. and Demberg, V. (2017). Psycholinguistic models of sentence processing improve sentence readability ranking. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 958–968.

IPM (2018). *INAF Brasil 2018: Indicador de Alfabetismo Funcional - Resultados Preliminares*. Instituto Paulo Montenegro. Disponível em http://acaoeducativa.org.br/wp-content/uploads/2018/08/Inaf2018_Relat%C3%B3rio-Resultados-Preliminares_v08Ago2018.pdf.

Leal, S. E., Duran, M. S., and Aluísio, S. M. (2018). A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413.

Magalhães, V. M. A., Bernardo, W. F., Diniz, F. H., dos Santos, K. C. L., Fonseca, L. M. G., Aluisio, S. M., and Leal, S. E. (2017). E-rural methodology: Contents elaborated according to the literacy level of the target audience. In *Twelfth Latin American Conference on Learning Technologies (LACLO)*, pages 1–9.

Maziero, E. G., Pardo, T. A. S., and Aluísio, S. M. (2008). *Ferramenta de Análise Automática de Inteligibilidade de Córpus (AIC)*. NILC - ICMC-USP. Disponível em http://www.nilc.icmc.usp.br/nilc/download/NILCTR0808-MazieroPardo.pdf.

Scarton, C. and Aluísio, S. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, pages 45–62.

Scarton, C., Paetzold, G. H., and Specia, L. (2018). Text simplification from professionally produced corpora. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3504–3510.

Singh, A. D., Mehta, P., Husain, S., and Rajkumar, R. (2016). Quantifying sentence complexity based on eye-tracking measures. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 202–212.

Stajner, S., Ponzetto, S. P., and Stuckenschmidt, H. (2017). Automatic assessment of absolute sentence complexity. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4096–4102.

Timm, L. B. (2018). *Looking at text simplification: Using eye tracking to evaluate the readability of automatically simplified sentences*. PhD thesis, Linköping University, Department of Computer and Information Science, Human-Centered systems, Linköping, Sweden.

Vajjala, S. and Meurers, D. (2014). Assessing the relative reading level of sentence pairs for text simplification. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 288–297.

Vajjala, S. and Meurers, D. (2016). Readability-based sentence ranking for evaluating text simplification. *CoRR - Computer Research Repository*, Disponível em http://arxiv.org/abs/1603.06009.

5.3 Estado da arte para o PB com *Transfer Learning*

Título:	***Using Eye-tracking Data to Predict the Readability of Brazilian Portuguese Sentences in Single-task, Multi-task and Sequential Transfer Learning Approaches***
Autores:	**Sidney Evaldo Leal, João Marcos Munguba Vieira, Erica Rodrigues, Elisangela Nogueira Teixeira e Sandra Maria Aluísio**
Ano:	**2020**
Conferência:	**COLING - The 28th International Conference on Computational Linguistics - Espanha (online)**
Situação:	**Publicado**

Depois da coleta dos dados de rastreamento ocular do córpus RastrOS, finalmente foi possível replicar o método do estado da arte para a tarefa na língua inglesa (GONZALEZ-GARDUñO; SØGAARD, 2018) com *Multi-task Learning*.

No entanto, este artigo vai além da replicação dos métodos para o inglês e propõe um novo método utilizando *Sequential Transfer Learning* que atingiu o estado da arte para o PB com 97,5% de acurácia no PorSimplesSent, unindo as métricas de rastreamento ocular às linguísticas e psicolinguísticas.

O artigo foi apresentado no COLING em 2020, que aconteceu de forma *online* por causa da pandemia do covid-19, e foi indicado a *Best Paper*, ficando entre os dezoito melhores artigos longos da conferência.

Using Eye-tracking Data to Predict the Readability of Brazilian Portuguese Sentences in Single-task, Multi-task and Sequential Transfer Learning Approaches

Sidney Evaldo Leal [1] sidleal@gmail.com
João Marcos Munguba Vieira [2,3] joaomvieira@gmail.com
Erica dos Santos Rodrigues [4] ericasr@puc-rio.br
Elisângela Nogueira Teixeira [2,3] elisteixeira@letras.ufc.br
Sandra Maria Aluísio [1] sandra@icmc.usp.br

[1] **Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (USP)**
Av. do Trabalhador Saocarlense, 400, São Carlos - SP - Brazil

[2] **Programa de Pós-graduação em Linguística - Universidade Federal do Ceará (UFC)**
Avenida da Universidade, 2683, BL. 125, 1o andar - Fortaleza - CE - Brazil

[3] **Laboratório de Ciências Cognitivas e Psicolinguística - Universidade Federal do Ceará (UFC)**
Avenida da Universidade, 2683, BL. 125, Sala 4, 1o andar - Fortaleza - CE - Brazil

[4] **Programa de Pós-graduação em Estudos da Linguagem - Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)**
Laboratório de Psicolinguística e Aquisição da Linguagem (LAPAL/ PUC-Rio)
Rua Marquês de São Vicente, 225 - Edifício Pe. Leonel Franca, 3o andar - Gávea - Rio de Janeiro - RJ - Brazil

Abstract

Sentence complexity assessment is a relatively new task in Natural Language Processing. One of its aims is to highlight in a text which sentences are more complex to support the simplification of contents for a target audience (e.g., children, cognitively impaired users, non-native speakers and low-literacy readers (Scarton and Specia, 2018)). This task is evaluated using datasets of pairs of aligned sentences including the complex and simple version of the same sentence. For Brazilian Portuguese, the task was addressed by (Leal et al., 2018), who set up the first dataset to evaluate the task in this language, reaching 87.8% of accuracy with linguistic features. The present work advances these results, using models inspired by (Gonzalez-Garduño and Søgaard, 2018), which hold the state-of-the-art for the English language, with multi-task learning and eye-tracking measures. First-Pass Duration, Total Regression Duration and Total Fixation Duration were used in two moments; first to select a subset of linguistic features and then as an auxiliary task in the multi-task and sequential learning models. The best model proposed here reaches the new state-of-the-art for Portuguese with 97.5% accuracy[1], an increase of almost 10 points compared to the best previous results, in addition to proposing improvements in the public dataset after analysing the errors of our best model.

1 Introduction

Readability is the ease of reading a text, not in its typographical aspects such as font size, but by measures such as its syntactic structure complexity, vocabulary frequency, content, style, and organisation that can be fitted to prior knowledge, reading skill, interest and motivation of the reader (Dubay, 2007).

Tracking the automation of readability back to its origin, the first readability formulas can be found a century ago in the United States, aiming to help teachers, librarians and scholars to select reading material

[1] Accuracy in our task is how close the model is to the true value, when assessing whether a given sentence is simple or complex, in a 10-fold cross-validation test.

for classes (Davison and Green, 1988) (Bohn, 1990). At that time, it was considered that complexity could be inferred by surface-level metrics of words and sentences, based on the frequency and size (number of letters) of the words and on the average number of words per sentence. Since then, readability analysis has become a large area of multidisciplinary research, which has an ever growing body of literature, related tasks (e.g., text simplification task (Vajjala and Meurers, 2014a) and text summarization task (Vodolazova and Lloret, 2019)), and has gained new computational approaches in this century using Natural Language Processing (NLP) and Machine Learning methods (Collins-Thompson, 2014).

Traditionally, the readability assessment task has been applied to the text level, assigning a grade (or level of proficiency ranking) for an entire document. However, in a document classified as simple, complex sentences can occur, just as there are simple sentences in a complex document. A sentence is an important unit that provides, in most cases, enough information to be able to infer and analyse its complexity. Although the same approach can be used to assess the complexity of texts at the sentence level, (Dell'Orletta et al., 2014) demonstrated that a greater number of *features* are needed for readability prediction at the sentence level. A study conducted by (Gonzalez-Garduño and Søgaard, 2018) has achieved state-of-the-art performance in readability prediction for English sentences, using multi-task learning and eye-tracking measures. An example of an application for the sentence level approach is the complexity checker tool, proposed by (Scarton et al., 2017) that analyses all sentences in a text, highlighting the complex ones to help with the simplification process.

This paper presents a thorough evaluation of sentence readability prediction for Brazilian Portuguese (BP), starting by evaluating single-task methods, followed by a replication of the work developed by (Gonzalez-Garduño and Søgaard, 2018). At the end, we propose a new model based on the sequential transfer learning approach (Ruder et al., 2019), which has achieved state-of-the-art performance in readability prediction of BP sentences.

Section 2 presents a literature review of the main works in readability prediction at sentence level (RPSL). Section 3 describes the corpora and metrics used and Section 4 presents the models evaluated and experimental results. Section 5 presents an analysis of the main errors of our best model, followed by a revision of the evaluation dataset and the results of our final best model. Section 6 draws the conclusions and proposes future research.

2 Readability Prediction at Sentence Level

The first studies on RPSL appeared in the last decade. Therefore, we can consider them as a recent research task, which aims to individually analyse and evaluate the sentences of a text, allowing for more accurate information of their complex points. The first study to consider the RPSL task was (Dell'Orletta et al., 2011), who compared its difficulty with readability at text level. However, a proposal of assessing the task was only consolidated by (Vajjala and Meurers, 2016), leading to further studies to improve the results comparatively (see Table 1).

According to (Dell'Orletta et al., 2014), sentence level readability is relevant because approaches to classifying text readability do not bring great advantages to the subsequent application of automatic simplification methods. Furthermore, considering all sentences as complex in a text classified as complex can impair the training of methods, especially when these sentences are used to assess the task of predicting sentence complexity. This was demonstrated by (Vajjala and Meurers, 2014b) when investigating the reasons for the low accuracy obtained from the Wikipedia-SimpleWikipedia corpus, used without any sentence alignment method. (Howcroft and Demberg, 2017) and (Singh et al., 2016) also explored the RPSL task using new metrics; the first study exclusively evaluated psycholinguistic metrics and the second one eye-tracking metrics. (Ambati et al., 2016) improved the results significantly by using a Combinatory Categorial Grammar (CCG) parser, and (Gonzalez-Garduño and Søgaard, 2018) achieved state-of-the-art performance in readability prediction for English sentences, using multi-task learning and eye-tracking measures combined with linguistic and psycholinguistic features. In addition, (Gonzalez-Garduño and Søgaard, 2018) compared the performance of readability models that use eye-tracking data of native speakers with models using data from language learners. There was no significant drop in performance when replacing learners with natives, i.e. language learner difficulties can be efficiently

estimated from native speakers. These findings are important since, in this paper we replicate the results of (Gonzalez-Garduño and Søgaard, 2018), using an eye-tracking data with native speakers of Brazilian Portuguese, that we created to study predictability in reading.

Study	Method	Accuracy
Flesch-Kincaid	Baseline	72.30
(Vajjala and Meurers, 2016)	RankSVM	74.58
(Ambati et al., 2016)	SMO	78.87
(Singh et al., 2016)	Logistic Regression	75.21
(Howcroft and Demberg, 2017)	Rank as Classification	73.22
(Gonzalez-Garduño and Søgaard, 2018)	MultiTask MLP	**86.62**

Table 1: State-of-the-art results for English using Wikipedia-SimpleWikipedia corpus.

Recently, (Stajner et al., 2017) and (Scarton et al., 2018) evaluated the RPSL task with a huge dataset — the Newsela dataset, comprising 550 thousand sentences, three times greater than Wikipedia-SimpleWikipedia. (Brunato et al., 2018) evaluated the perception of complexity and agreement between annotators, while (Timm, 2018) investigated automatic sentence simplifications, using eye-tracking tools.

For Italian, (Bosco et al., 2018) developed a good performance model for the RPSL task, using Long Short-Term Memory units (LSTMs), a well-known subset of Recurrent Neural Networks (RNN), and (Schicchi et al., 2020) evaluated RNN methods with attention-based mechanisms. For Portuguese, there are two studies: (Leal et al., 2018) compiled the PorSimplesSent (PSS) corpus and proposed baseline methods, and (Leal et al., 2019) developed a model using neural networks with 87.80% accuracy on PorSimplesSent2 (PSS2). PorSimplesSent2 is the most challenging version of PorSimplesSent and comprises **4,968** simplification pairs, where for splitting operation only the longest sentences derived are chosen, paired with the original sentence (details in Table 2).

Study	Method	Accuracy
Tokens per sentence	Baseline	69.35
(Leal et al., 2018)	RankSVM	74.20
(Leal et al., 2019)	MLP Pairwise ranking	**87.80**

Table 2: State-of-the-art results for BP using PorSimplesSent2 corpus.

3 Resources

3.1 Data

PSS Corpus PorSimplesSent (Leal et al., 2018) is a publicly available corpus of aligned sentences that was compiled from the PorSimples corpus (Caseli et al., 2009), which is organised into three readability levels: a) **Original:** Original sentences; b) **Natural Simplification:** Texts freely simplified by the annotators and c) **Strong Simplification:** Simplified texts following the rules of the simplification manual developed in the PorSimples project.

The PorSimplesSent corpus has three versions that include different approaches to sentences that have undergone the split operation. PSS1 repeats the original sentence for each sentence resulting from the division; PSS2 selects only the largest resulting sentence, which also has the greatest overlap of words, and PSS3 contains only sentences that have not been divided, and thus is the smallest of the three. For the present study, version **PSS2** was chosen. It has 4,962 pairs of sentences, with Original-Natural, Natural-Strong and Original-Strong alignments, obtained in TSV format[2].

RastrOS Corpus To compare models of readability using data of eye movements during reading, since there was no public corpus of eye movements in BP, we created a Brazilian national project to build an

[2] http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources

eye-tracking corpus[3] with predictability norms for study several parameters, among them readability and syntactic complexity. We followed a very similar methodology as the Provo Corpus (Luke and Christianson, 2018). We compiled a corpus with 50 shorts paragraphs taken from various sources, at a rate of 40% for news, 40% for pop-science and 20% for literary paragraphs. Currently, we have Cloze scores (full-orthographic form, Part-of-Speech and inflectional properties) for all 2,494 words (1,237 types) in 120 sentences distributed among the 50 paragraphs. To build this Brazilian eye-tracking corpus, we created to tasks: a word-by-word Cloze task and a silent reading task to record eye-movements.

For the Cloze task, at the time of writing this paper, 315 undergraduate students from six universities in three different Brazilian regions read 5 paragraphs each, from a pool of 50 paragraphs. Moreover, 30 undergraduate students have their eye movements recorded during silent reading of all 50 paragraphs. We used a high-accuracy eye-tracker with chin and forehead rest — the EyeLink 1000 Hz Desktop from SR Research. All participants were native Brazilian Portuguese speakers. None had participated in the Cloze task. They were asked to read the passages silently for comprehension, one by one, in a random order, preceded by two practice paragraphs. After a gaze trigger, a whole paragraph is presented and after read the paragraph, participants should press a joystick button to continue the experiment and read the forthcoming paragraphs, presented in a random order. Yes-no comprehension questions appeared 20 times, to ensure participants attention.

3.2 Linguistic Features

This study used 156 features, developed for BP, known to affect text complexity which are available in the Coh-Metrix-Port (Scarton et al., 2010) and Coh-Metrix-Dementia (Aluísio et al., 2016) tools, as well as 24 psycholinguistic metrics created from a repository of 26,874 words in BP annotated with Imageability, Concreteness, Familiarity and Age of Acquisition scores (dos Santos et al., 2017).

Moreover, using the parser PALAVRAS (Bick, 2000), 39 syntactic metrics were also developed to extract the passive voice and other sentence and clause information. We also included 5 classical readability formulas, some of them adapted to BP, and several other metrics using lists of words, such as easy-conjunction ratios and PALAVRAS semantic tags, for example, abstract-noun ratio.

The Coh-Metrix-Port is an adaptation for BP of the metrics available in the Coh-Metrix project. It was developed in the scope of the PorSimples project, and implements 48 metrics (Scarton et al., 2010), divided into the following categories: basic counts, logical operators, frequencies, hyperonyms, tokens, constituents, connectives, ambiguity, co-reference and anaphors. Coh-Metrix-Dementia (Aluísio et al., 2016) is an adaptation of Coh-Metrix-Port for automatic analysis of language disorders in dementias (such as Alzheimer's Disease) or Mild Cognitive Impairment. 25 new metrics were added to the Coh-Metrix-Port's 48, in the following categories: disfluencies, latent semantic analysis, lexical diversity, syntactic complexity, semantic density and idea density.

The list of the first 50 features obtained after the feature selection described in Section 4 can be seen in Table 3; to visualise them better, they are grouped into readability formulas, syntactic complexity, morphosyntactic complexity, psycholinguistic metrics and types of clauses.

3.3 Eye-tracking Measures

As stated in (Gonzalez-Garduño and Søgaard, 2018), previous research has demonstrated a correlation between eye-tracking measures and text difficulty (Rayner et al., 2012). This research opened up new possibilities of assessments with the machine learning approach that use both eye-tracking measures and widely known linguistic features.

This study tried to use similar eye-tracking metrics adopted by (Gonzalez-Garduño and Søgaard, 2018). We do not used the same metrics because we decided to use the sum of the times of each word of the sentence, instead of the average used by (Gonzalez-Garduño and Søgaard, 2018). The main reason to not use the average come from the fact that our results with average were poor as the Pearson correlation was below 0.2. After analysing it, we verified that our average values were all in a very close range,

[3]The RastrOS Corpus (`http://www.nilc.icmc.usp.br/nilc/index.php/rastros`) will be described in a forthcoming paper and will be publicly available in the OSF platform.

Metric name	Definition
Readability Formulas	
Brunet's Statistics	Brunet's Statistics is a form of type/token ratio that is less sensitive to text size.
Gunning Fox	Gunning Fog readability index.
Flesch Index adapted to BP	The Flesch Readability Index.
Honore's Statistics	Honore's Statistics takes into account words that are only used once, indicating a higher lexical richness.
Dale–Chall formula adapted to BP	Combines the number of unfamiliar words with the average number of words per sentence.
Syntactic Complexity	
TTR	Type-Token Ratio is the proportion of words without repetition (types) in relation to the total of words (tokens).
Frazier	Bottom-up approach for calculating the syntactic complexity of a sentence, climbing the tree from the word.
Yngve	It measures deviations from the tendency of syntactic trees to branch to the right.
Words	Number of words in the sentence.
Dependence Distance	The dependency distance using a dependency tree.
Punctuation Diversity	Proportion of punctuation mark types in relation to punctuation mark tokens in the text.
Sentences with four clauses	Proportion of sentences containing 4 clauses.
Sentences with five clauses	Proportion of sentences containing 5 clauses.
Sentences with six clauses	Proportion of sentences containing 6 clauses.
Sentences with seven more clauses	Proportion of sentences containing 7 clauses.
Easy conjunctions ratio	Proportion of easy frequent conjunctions in relation to all words in the text.
Adverbs ambiguity	Proportion between the amount of meanings of the text adverbs in the TeP 2.0[4] and the amount of adverbs.
Adjectives ambiguity	Proportion between the amount of meanings of the text adjectives in the TeP 2.0 and the amount of adjectives.
Min-content words freq	Average of the absolute frequencies of the rarest content words in the text sentences.
Morphosyntactic Complexity	
Gerund verbs	Proportion of verbs in the present participle in relation to all verbs in the text.
Verbs	Proportion of verbs in relation to the number of words in the text.
Adjectives min	Minimum proportion of adjectives in relation to the number of words in the sentences.
Adjectives max	Maximum proportion of adjectives in relation to the number of words in the sentences.
Aux-plus-PCP per sentence	Proportion of auxiliary verbs followed by participle in relation to the number of sentences in the text.
Prepositions per sentence	Average prepositions per sentence.
Pronouns max	Maximum proportion of pronouns in relation to the number of words in the sentences.
Pronoun ratio	Proportion of relative pronouns in relation to the number of pronouns in the text.
Adjective ratio	Proportion of adjectives in relation to the number of words in the text.
Subjunctive imperfect ratio	Proportion of verbs in the past imperfect subjunctive in relation to the total inflected verbs.
Preposition diversity	Proportion of different prepositions in relation to the total prepositions of the text.
Indicative imperfect ratio	Proportion of verbs in the past imperfect indicative, in relation to the total number of verbs in the text.
Subjunctive present ratio	Proportion of verbs in the present of the subjunctive in relation to the total number of inflected verbs in the text.
Third person pronouns	Proportion of personal pronouns in third persons in relation to all personal pronouns in the text.
Adverbs diversity ratio	Proportion of adverb types in relation to the number of adverb tokens in the text.
Inflected Verbs	Proportion of inflected verbs in relation to all verbs in the text.
Abstract-nouns ratio	Proportion of abstract nouns in relation to the number of words in the text.
Psycholinguistic Metrics	
Concreteness mean	Average of the concreteness values of the words of the sentence.
AoA 1-2.5 ratio	Proportion of content words with Age of Acquisition (AoA) values between 1 to 2.5, in relation to all content words.
Imageability 2.5-4 ratio	Proportion of content words with Imageability values between 2.5 to 4, in relation to all content words.
AoA std	Standard deviation of the Age of Acquisition values of the sentence content words.
Types of Clauses	
Coordinate conjunctions ratio	Proportion of coordinated conjunctions in relation to the total number of text conjunctions.
Coordinate conjunctions per clauses	Proportion of coordinating conjunctions in relation to the total number of sentences in the text.
Logical operators	Proportion of logical operators in relation to the number of words in the text.
Relative-pronouns ratio	Proportion of relative pronouns in relation to the number of pronouns in the text.
Positive-temporal connectives ratio	Proportion of positive temporal connectives in relation to the number of words in the text.
Subordinating conjunctions ratio	Proportion of subordinating conjunctions in relation to the sum of subordinating and coordinating conjunctions.
Negative-temporal connectives ratio	Proportion of negative temporal connectives in relation to the number of words in the text.
Positive-logical connectives ratio	Proportion of positive logical connectives in relation to the total words of the text.
Apposition per clause	Average number of apposition per clause.
Subordinate clauses	Proportion of subordinating clauses in relation to all clauses in the text.

Table 3: Top 50 linguistic metrics (of 156 obtained after feature selection).

therefore we decided to use the sum of times (or late measures[5]), significantly improving the results with the Pearson correlation to a value above 0.8, as seen in Table 4.1.

It seemed intuitive that to measure complexity, the sum works better than the average, for instance, a single word in a sentence with a fixation over 800 milliseconds can be the cause of the complexity for that sentence, but when using the average, these 800 ms can be diluted in a large sentence in which all other words have a fixation of 250 ms or less.

The eye-tracking metrics are described below[6]:

- **First Pass Reading Time (FirstPass)**: Sum of the duration of the fixations in a given word, it does not consider new fixations in the word after a regression.

[5]There is a distinction between early and late measures in eye tracking experiments. For example, early measures are related with first fixation duration and late measures are related to total fixation duration, that means the sum of all fixations in an interest area like a word or a phrase. Late measures usually provide evidence of difficulties during reading.

[6]Once we exported data from Data Viewer (from SR Research), First Pass Reading Time corresponds a IA_FIRST_RUN_DWELL_TIME; Total Regression Duration is IA_REGRESSION_PATH_DURATION and Total Fixation Duration is IA_DWELL_TIME.

- **Total Regression Duration (Regression)**: Total duration spent looking back at previous words, searching for a context; this movement can indicate difficulty in understanding a passage.
- **Total Fixation Duration (TotalFix)**: Sum of the duration of all fixations in a given word, before and after regressions.

4 Approaches and Results

The models developed and evaluated for the task are presented below. The first step was to validate whether the current linguistic features could predict the eye-tracking measures. Only after proving that, the measures were used as a basis for feature selection, followed by a comparison among the single-task, multi-task and sequential transfer learning approaches. All models were evaluated with 10-fold cross validation and trained with an Adam optimiser, implemented using the Keras (Chollet and others, 2015) and Scikit-Learn packages (Pedregosa et al., 2011) for the Python language. As far as we could investigate, this work is the first to use sequential transfer learning in the RPSL task.

4.1 Predicting the Eye-tracking Measures and Feature Selection

First of all, we validated whether eye-tracking measures could be predicted from linguistic features, as mentioned previously. This was done with a simple regressor, implemented as an MLP with 3 layers, 189 neurons in the input (related to all the metrics evaluated), 100 neurons in the hidden layer and one neuron in the output layer, using ReLU as the activation function in all layers.

The model was trained and tested on the 120 sentences of RastrOS corpus using cross-validation and calculating the Pearson correlation between the predicted and real values. The results can be seen in Table 4. FirstPass alone obtained the best result with a correlation value above 0.9. To predict the three metrics at the same time, the architecture was changed to 3 neurons in the output layer, each predicting one of them. The simultaneous prediction of the 3 metrics reached 0.88 correlation with a p-value of 0.001.

Measure	**RMSE**	**Pearson's p Correlation**	**p-value**
First pass (FirstPass)	0.058	0.92	< 0.001
Regression Duration (Regression)	0.096	0.82	0.005
Total Fixation Duration (TotalFix)	0.094	0.84	0.008
FirstPass+Regression+TotalFix	0.092	0.88	0.001

Table 4: Root Mean Squared Error and Pearson's p correlation between predictions and true values using Single Task MLP.

Once the feasibility of using linguistic features to predict eye movements was validated, the model was used to perform the selection of features to be used in the evaluated models. Using the **Permutation Importance** method implemented in *eli5.sklearn*[7], it was found that from all of the 189 features available, 156 contributed to the prediction with a value above zero (see Section 3.2).

4.2 Single-Task and Multi-Task MLP

The first model developed to predict the complexity in PSS2 was a Single-Task MLP with 3 layers and 100 neurons in the hidden layer, which was similar to the prior state-of-the-art model for Brazilian Portuguese.

We did not use eye-tracking measures and we only increased the number of neurons in the hidden layer from 30 to 100, including a sigmoid activation function on the output and ReLU on the other layers.

The input for this model was the 156 linguistic features for each sentence of the Simple-Complex pair, and the output was just one neuron that tried to predict which sentence was complex and which was simple. The Single-Task MLP model showed no significant improvements (see Table 5) but used less features at the input.

[7] `https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html`

Then, Multi-Task Learning models adapted from (Gonzalez-Garduño and Søgaard, 2018) were tested, where two MLPs were connected by the hidden layer with 100 neurons and trained simultaneously. While the first MLP tried to predict the eye-tracking measures, the second attempted to predict which sentence was more complex, receiving all the linguistic features from the pair and predicting 0 when sentence A was simpler than B or 1 when it was the opposite. Half of the 4,962 pairs were randomly inverted for training and testing, resulting in 50% simple-complex and 50% complex-simple pairs and then split by the 10-fold cross validation method.

The first network had 156 neurons in the input layer and one neuron in the output layer for the individual eye tracking measure or 3 neurons to predict the 3 eye-tracking measures simultaneously: all with the ReLU activation function. The second network had 312 neurons at the input (156 of each sentence of the aligned pair), and one neuron at the output activated by the sigmoid function.

The results of the network that tried to predict one of the eye-tracking measures one at a time did not improve when compared to the single-task approach.

However, an increase of 3 points was observed in the accuracy when using the prediction of the 3 measures at the same time in the first task (see Table 5).

Model	**Accuracy**
Easy Baseline (Tokens per sentence)	0.694
Strong baseline (Previous State-of-the-Art (Leal et al., 2019))	0.878
Single-Task (without eye-tracking measures)	0.884
Multi-Task FirstPass	0.880
Multi-Task Regression	0.858
Multi-Task TotalFix	0.856
Multi-Task FirstPass+Regression+TotalFix	0.908
Sequential FirstPass+Regression+TotalFix	**0.968**
Sequential FirstPass+Regression+TotalFix (After Error Analysis)	0.975*

Table 5: Accuracy for all multi-task and single-task models, including the two step training approach.
* This result value is not directly comparable with others in this table, because the dataset was slightly different after cleaning.

4.3 Sequential Transfer Learning

Finally, we proposed a new model as an evolution of the previous ones, which reached the state-of-the-art for the RPSL task in Brazilian Portuguese, with an improvement of almost 10% over the best previous result.

The model was chosen from several other models implemented to try to improve accuracy, inspired by the models proposed by (Gonzalez-Garduño and Søgaard, 2018) and (Singh et al., 2016). Several architectures were tested, varying the number of layers, number of neurons, training time and how to make better use of eye-tracking measures to predict complexity.

Figure 1 shows the final architecture. In the first phase, a single-task MLP with 2 hidden layers (with 64 and 100 neurons and ReLU activation) was trained with all the RastrOS corpus sentences throughout 100 epochs. Once training was complete, the two hidden layers were transferred to the second MLP and frozen. This second network had two parallel layers at the input, one for each sentence of the pair comprising 156 neurons each. These input layers were then completely connected to the first transferred layer and to the other hidden layer with 64 neurons. The predicted result for the 3 eye-tracking measures for each of the sentences was then concatenated with the 64-neuron layer that fed the final layer with only one neuron using the sigmoid function to return 0 or 1. All the other layers used the ReLU function. This architecture allowed the training to also adjust the weights of the middle layer of the eye-tracking measure prediction and was trained throughout 30 epochs.

The results show that there was a significant improvement in the accuracy of the previous state-of-the-art model, supporting the conclusions of the studies cited for the English language. We also confirmed the usefulness of the eye-tracking measures for the task of evaluating sentence complexity.

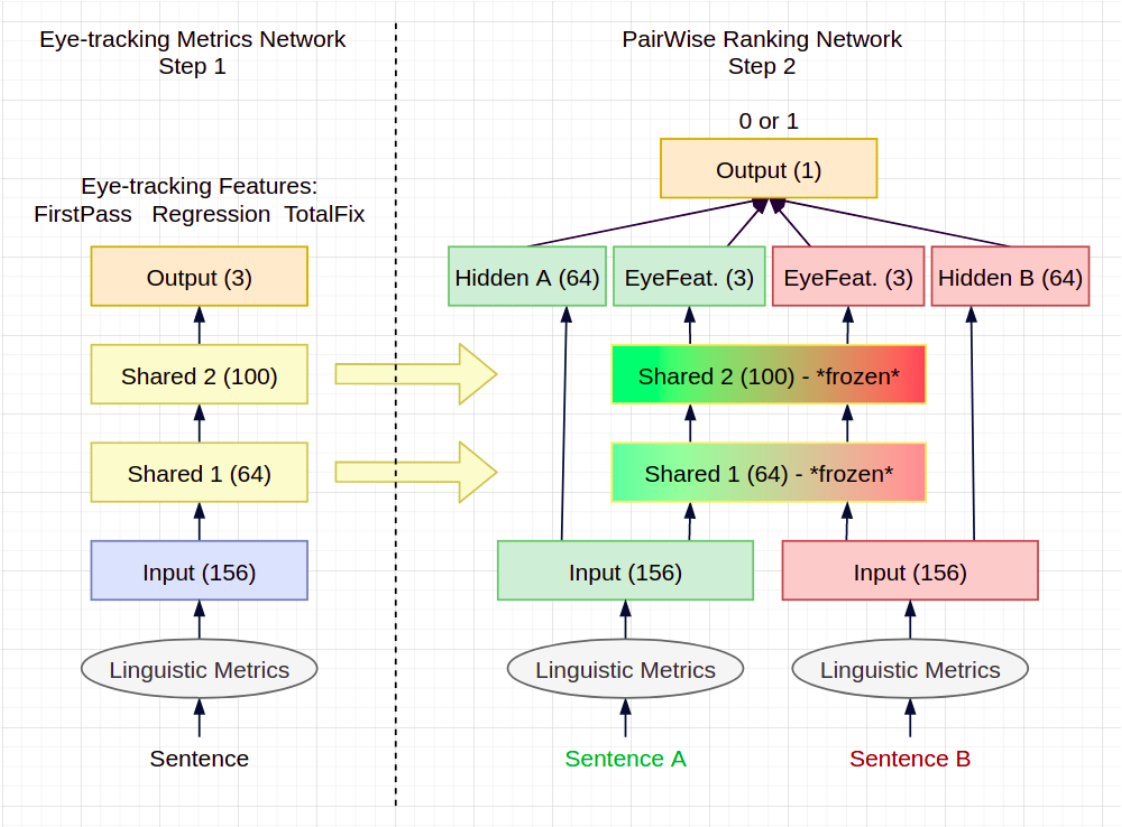


Figure 1: Sequential Transfer Learning

5 Error Analysis

After running the best model with 10-fold cross validation, all 151 pairs in which the model failed the prediction were manually annotated in a thoughtful error analysis, shown in Figure 2.

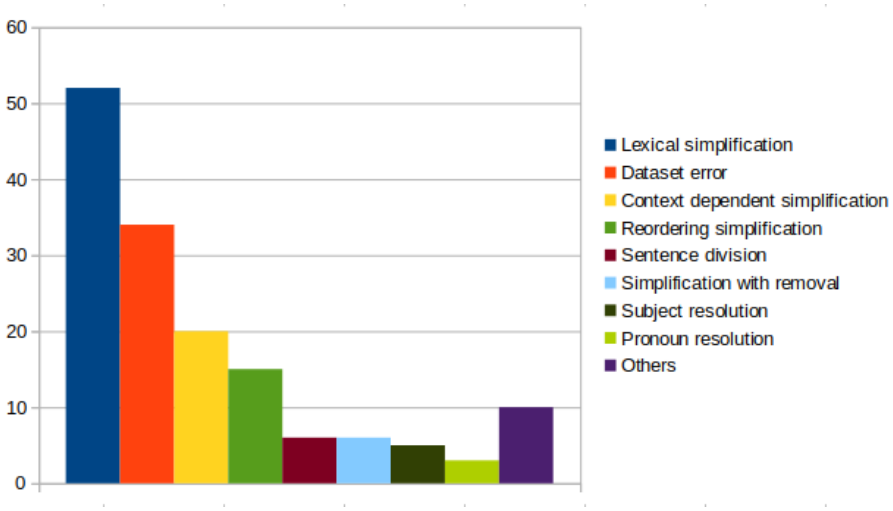


Figure 2: Main error types obtained with manual annotation

This enabled us to verify that the second biggest cause of errors are problems in the dataset, possibly caused by the automatic alignment of sentences. There were problems in 23% of the pairs, such as just a capitalised letter, an added comma or spelling correction on the simplified side. The authors understand that these errors should be removed from the public dataset, as they do not represent real

cases of simplification. The third major cause of errors are simplifications that go beyond the context of the sentence itself and would necessarily need the previous or subsequent sentences so that the model could automatically decide the correct class.

However, it is important to note that the main cause of errors refers to simplifications at the lexical level, indicating the need to refine the metrics at this level and possibly to include new ones. To validate how the model behaves without the 54 pairs of sentences suggested for removal, it was run again in the new dataset, without these pairs, reaching the accuracy of **97.5%** with an improvement of approximately 1 point as expected. The cleaned dataset was sent to the PSS authors recommending them to publish it as a new revised version.

6 Conclusions

This work reinforces the observations of (Gonzalez-Garduño and Søgaard, 2018) on the importance of using eye-tracking measures, as well as models with transfer learning (multi-task and sequential learning) for the task of sentence readability assessment. Moreover, it establishes the new state-of-the-art for the task of assessing sentence complexity in the Brazilian Portuguese language, with a substantial increase of almost 10 points over the best accuracy obtained so far. It also contributes to improving the PSS2 dataset, identifying and proposing the elimination of alignment errors for future evaluations.

The source codes with the implemented models are publicly available at `https://github.com/sidleal/simpligo-ranking`.

Regarding future research, in order to mitigate errors at the lexical level, as mentioned in Section 5, it is worth using a model combining word embeddings in an architecture with multi-view learning, as well as implementing and validating the Recurring Neural Networks and Attention-based architectures. Another question that deserves further investigation is the difference observed when using the average and sum of eye-tracking measures, and why the average did not work well in our scenario. One hypothesis is that this may be related with the text genres of the eye-tracking dataset. We also intend to test the models proposed here in the two other versions of PorSimplesSent corpus.

Acknowledgements

This research project received financial support from The São Paulo Research Foundation (FAPESP) (Fundação de Amparo à Pesquisa do Estado de São Paulo, in Portuguese) process number 2019/09807-0. The second author was financed with a Master's Scholarship from the Ministry of Education of Brazil, through the Agency for the Development of Higher Education Personnel (CAPES) (Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior).

References

Sandra M. Aluísio, Andre Cunha, and Carolina Scarton. 2016. Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In *12th International Conference on Computational Processing of the Portuguese Language (PROPOR 2016)*, volume 9727 of *Lecture Notes in Computer Science*, pages 109–114. Springer Cham.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental CCG parser. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1051–1057.

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

Hilario Inacio Bohn. 1990. Linguistic complexity and text comprehension: Readability lssues reconsidered by davison and green. *Revista Fragmentos*, v.3,n.2.

Giosué Lo Bosco, Giovanni Pilato, and Daniele Schicchia. 2018. A neural network model for the evaluation of text complexity in Italian language: a representation point of view. In *Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2018)*, pages 464–470.

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? Do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699.

Helena Medeiros Caseli, Tiago Freitas Pereira, Lúcia Specia, Thiago A. S. Pardo, Caroline Gasperin, and Sandra Maria Aluísio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. *Advances in Computational Linguistics, Research in Computer Science (CICLing-2009)*, vol. 41:59–70.

François Chollet et al. 2015. Keras. `https://keras.io`.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

Alice Davison and Georgia Green. 1988. *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*. Routledge.

Felice Dell'Orletta, Martijn Wieling, Andrea Cimino, Giulia Venturi, and Simonetta Montemagni. 2014. Assessing the readability of sentences: Which corpora and features? *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of Italian texts with a view to text simplification. *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

Leandro Borges dos Santos, Magali Sanches Duran, Nathan Siegle Hartmann, Arnaldo Candido, Gustavo Henrique Paetzold, and Sandra Maria Aluisio. 2017. A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In K. Ekštein and V. Matoušek, editors, *Text, Speech, and Dialogue. TSD 2017. Lecture Notes in Computer Science*, volume 10415, pages 281–289. Springer, Cham.

William H. Dubay. 2007. *Smart Language: Readers, Readability, and the Grading of Text*. Impact Information, Costa Mesa, CA. ISBN: 1-4196-5439-X.

Ana Valeria Gonzalez-Garduño and Anders Søgaard. 2018. Learning to predict readability using eye-movement data from natives and learners. *Proceedings of the The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5118–5124.

David M. Howcroft and Vera Demberg. 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 958–968.

Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Maria Aluísio. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413. Association for Computational Linguistics, August.

Sidney Evaldo Leal, Vanessa Maia Aguiar de Magalhães, Magali Sanches Duran, and Sandra Maria Aluísio. 2019. Avaliação automática da complexidade de sentenças do português brasileiro para o domínio rural. In *Symposium in Information and Human Language Technology - STIL*. SBC.

Steven G. Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton. 2012. *The Psychology of Reading*. Psychology Press.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, page 712–718.

Carolina Scarton, O. Oliveira-Junior, Arnaldo Candido-Junior, Caroline Gasperin, and Sandra Maria Aluísio. 2010. Simplifica: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 41–44.

Carolina Scarton, Alessio Palmero Aprosio, Sara Tonelli, Tamara Martin Wanton, and Lucia Specia. 2017. Musst: A multilingual syntactic simplification tool. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 25–28, Tapei, Taiwan. Association for Computational Linguistics.

Carolina Scarton, Gustavo Henrique Paetzold, and Lucia Specia. 2018. Text simplification from professionally produced corpora. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3504–3510.

Daniele Schicchi, Giovanni Pilato, and Giosué Lo Bosco. 2020. Deep neural attention-based model for the evaluation of italian sentences complexity. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 253–256, San Diego, CA, USA.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajakrishnan Rajkumar. 2016. Quantifying sentence complexity based on eye-tracking measures. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 202–212.

Sanja Stajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4096–4102.

Linnea Björk Timm. 2018. *Looking at text simplification: Using eye tracking to evaluate the readability of automatically simplified sentences*. Ph.D. thesis, Linköping University, Department of Computer and Information Science, Human-Centered systems, Linköping, Sweden.

Sowmya Vajjala and Detmar Meurers. 2014a. Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification*.

Sowmya Vajjala and Detmar Meurers. 2014b. Assessing the relative reading level of sentence pairs for text simplification. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 288–297.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *CoRR - Computer Research Repository*, Disponível em http://arxiv.org/abs/1603.06009.

Tatiana Vodolazova and Elena Lloret. 2019. Towards adaptive text summarization: How does compression rate affect summary readability of L2 texts? In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1265–1274, Varna, Bulgaria, September. INCOMA Ltd.

CAPÍTULO

6

CONCLUSÃO

O objetivo geral deste trabalho consistiu na avaliação de modelos computacionais de classificação da complexidade das sentenças de um texto, para atender inicialmente o público alvo do domínio rural. Os modelos foram criados para dar apoio à criação de textos informativos, técnicos e procedimentais para leitores com variados níveis de letramento. Até onde foi possível verificar, este trabalho foi o primeiro a abordar a tarefa de avaliação automática da complexidade sentencial para o PB.

Durante a revisão da literatura, ficou clara a necessidade de criar diversos recursos, sendo os principais um córpus com dados de rastreamento ocular e outro com sentenças alinhadas (versão simples/complexa) para avaliação da tarefa. Foram assim criados o córpus RastrOS e o córpus PorSimplesSent, respectivamente. Ambos os objetivos foram atingidos, sendo detalhados na Seção 4.3 e na Seção 5.1, respectivamente.

Outro objetivo importante foi a disponibilização pública do NILC-Metrix, com 200 métricas linguísticas e psicolinguísticas, também atingido e detalhado no Capítulo 3.

Por fim, com os recursos criados e as *baselines* definidas, o objetivo de criar um modelo computacional para a avaliação automática da complexidade sentencial foi atingido, com a investigação da tarefa alvo utilizando diversas abordagens. Culminando no estado da arte para o PB com 97,5% de acurácia via *Transfer Learning* e métricas de rastreamento ocular aliadas às linguísticas e psicolinguísticas (Seção 5.3). Também foi demonstrada a capacidade de generalização do modelo para o domínio rural.

A partir do melhor método desenvolvido, foi criada uma aplicação (<http://fw.nilc.icmc.usp.br:23380/simpligo-ranking>) que permite a avaliação das sentenças de um texto, atribuindo um índice de complexidade individual para cada sentença, variando de 1 (simples) a 100 (complexo) (ver um exemplo na Seção 6.1). Os métodos e recursos criados estão todos disponibilizados publicamente para utilização pela comunidade.

Hipótese

A hipótese da pesquisa estava relacionada com as métricas de rastreamento ocular. Sendo elas representantes da complexidade mensurável durante a leitura das sentenças por humanos, acreditávamos que contribuiriam com um aumento de cerca de 8% na acurácia do melhor modelo de predição da complexidade sentencial para o PB, assim como contribuíram para o inglês.

A hipótese foi confirmada. Após a adição das métricas de rastreamento ocular ao melhor modelo de predição da complexidade sentencial que utiliza *Transfer Learning*, houve um incremento de 87,80% para 97,5% na acurácia do modelo no córpus PorSimplesSent. Isso representou um aumento de **9,7%**, valor superior ao resultado obtido no trabalho de Gonzalez-Garduño e Søgaard (2018) para a língua inglesa.

Questões de pesquisa

Abaixo são retomadas as principais questões levantadas no início desta pesquisa e apresentadas as considerações sobre cada uma, após reflexão na fase final desta etapa.

Qual o impacto de grupos de *features* linguísticas (morfológicas, lexicais, sintáticas), clássicas, psicolinguísticas e do rastreamento ocular para a tarefa de predição de complexidade sentencial?

Com a experiência deste trabalho é possível afirmar que as *features* de rastreamento ocular desempenharam um papel fundamental para o PB, para o aperfeiçoamento dos modelos de predição, assim como desempenharam para a língua inglesa contribuindo com um aumento de 2,4% na acurácia do modelo com *multi-task learning* e 9,1% no modelo com *sequential transfer learning*. Também foi verificado que as fórmulas clássicas continuam entre as medidas com maior peso nos modelos. As métricas morfológicas, lexicais, sintáticas e psicolinguísticas atuaram no refinamento dos modelos.

Qual o impacto de *features* individuais para a tarefa de predição de complexidade sentencial?

A maior surpresa quando foram avaliadas as *features* individuais foi a contribuição das fórmulas clássicas (ver Subseção 2.2.1), principalmente as mais tolerantes ao tamanho do texto, como a Brunet e Honoré. Fora as clássicas, as morfológicas e sintáticas foram as melhores classificadas no *ranking* da seleção de *features*, além das três medidas extraídas do rastreamento ocular: tempo da primeira passagem, tempo total de fixação e tempo total da regressão.

Qual é a influência do tamanho do córpus de treinamento no desempenho dos métodos para a tarefa de predição da complexidade sentencial?

Foram obtidos bons resultados com o córpus PorSimplesSent, com cerca de cinco mil

pares de sentenças alinhadas. Em teoria um córpus maior beneficiaria os métodos, permitindo investigar mais facilmente modelos mais recentes com *Deep Learning* e *Transformers*, mas a tarefa se mostrou relativamente simples de se resolver utilizando AM clássica.

É importante lembrar que o PorSimplesSent possui três versões, de acordo com a estratégia escolhida para as operações de divisão de sentenças. Neste trabalho foi explorada a versão PSS2, que utiliza apenas a maior sentença resultante da divisão. A avaliação da tarefa nas outras duas versões pode contribuir para o melhor entendimento da tarefa e ficará como trabalhos futuros.

Para os dados de rastreamento ocular, o córpus utilizado no treinamento foi ainda menor, com apenas 120 sentenças (cerca de 2.500 tokens), porém lidas por 37 participantes. Esses dados contribuíram bastante para atingir o estado da arte para o PB. Para comparação, o córpus usado por Gonzalez-Garduño e Søgaard (2017) foi o Dundee, com 2.368 sentenças lidas por 10 participantes, e os córpus utilizados por Gonzalez-Garduño e Søgaard (2018) foram o Dundee e o GECO, com 5.000 sentenças lidas por 33 participantes. A avaliação com os dados de rastreamento do projeto RastrOS utilizou uma versão com 30 participantes, ficando com um valor intermediário de participantes entre os dois córpus utilizados em trabalhos relacionados ao desta tese.

Qual método de seleção de *features* fornece um melhor desempenho para a tarefa de predição de complexidade sentencial?

Durante o desenvolvimento dos modelos desta pesquisa, o método escolhido foi o *Permutation Importance*[1], pela simplicidade de acoplamento aos modelos no Scikit-learn e rapidez de execução. De forma bem resumida esse método troca uma das *features* por vez por números aleatórios e compara o resultado em um modelo já treinado. A saída dele é um *ranking* das *features* que mais contribuíram.

Um ponto que merece destaque é que a seleção de *features* para o melhor modelo foi feita utilizando o regressor que estima as métricas de rastreamento ocular no RastrOS e não o classificador final. Em teoria isso melhora a generalização do método, pois o RastrOS possui três gêneros e o PorSimplesSent apenas um.

Quais são os erros sistemáticos que persistem usando o melhor método de predição de complexidade sentencial?

O principal erro verificado foi por conta de simplificações lexicais, o que pareceu indicar que as métricas de frequência de palavras não estavam tão boas. Este fato retroalimentou o processo de pesquisa e motivou a criação de oito novas métricas de frequência no NILC-Metrix. O melhor modelo não foi novamente avaliado após a criação das métricas, ficando como trabalho

[1] <https://scikit-learn.org/stable/modules/permutation_importance.html>

futuro, mas é esperada uma redução nesses erros.

Qual método de AM atinge o melhor resultado para a tarefa de predição de complexidade sentencial?

Os métodos de AM que mais se destacaram durante a revisão bibliográfica e também durante os experimentos desta pesquisa foram o SVM, sobretudo o SVMRank[2] e as MLP's. As grandes vantagens das MLP's foram a flexibilidade e facilidade de customização durante o processo de treinamento com as abordagens de *Transfer Learning*, e terminaram sendo responsáveis pelo estado da arte para o PB.

6.1 Contribuições

As principais contribuições desta pesquisa são apresentadas a seguir, agrupadas em Córpus, Métodos e Aplicações.

Córpus

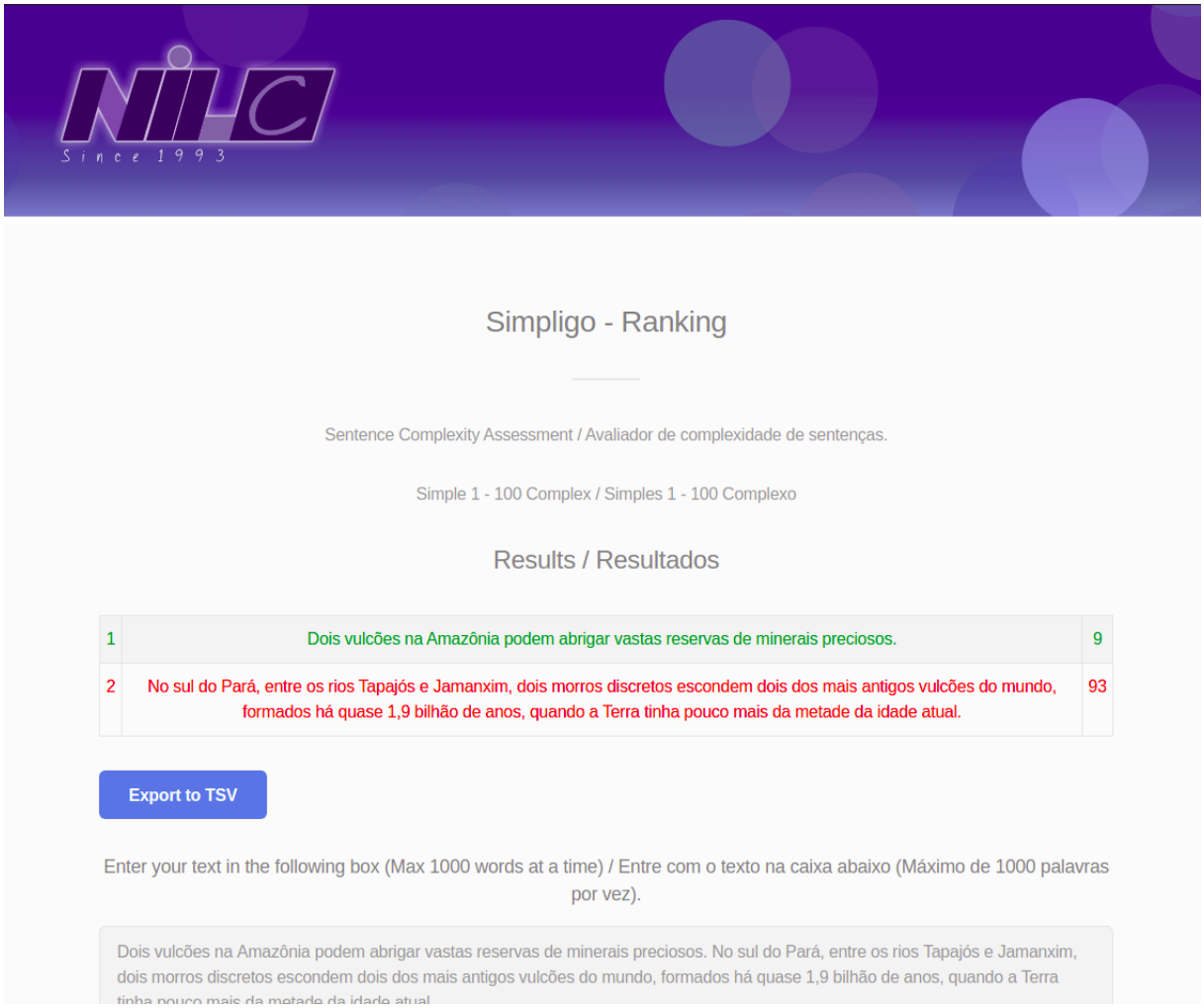
- **Córpus PorSimplesSent**: Córpus de pares de sentenças alinhadas Complexa-Simples para treinamento e avaliação da tarefa de complexidade sentencial em PB (ver Seção 5.1);
- **Córpus RastrOS**: Córpus com dados de rastreamento ocular e normas de previsibilidade, tendo a coleta sido realizada com estudantes do ensino superior de seis universidades brasileiras (ver Seção 4.3);
- **Córpus RastrOS - Sentence Completion**: Córpus para a tarefa de *Sentence Completion*, criado a partir das respostas ao teste Cloze do RastrOS (ver Seção 4.2). Foi utilizado para avaliação de métodos de similaridade semântica para criação das normas de previsibilidade semânticas.

Métodos

- **Método de clusterização para criação de córpus**: Método de aprendizagem não supervisionada para gerar grupos utilizando métricas linguísticas, permitindo a seleção de textos similares ou com pouca similaridade, dependendo do objetivo final (ver Seção 4.1);
- **Método de avaliação de similaridade semântica**: Método híbrido para o cálculo da similaridade semântica entre palavras e contexto ou entre duas palavras, com BERT e *Word Embeddings* (ver Seção 4.2);

[2] vide Seção 2.4.

Figura 14 – Tela da ferramenta Simpligo-Ranking, com o resultado do processamento de um dos parágrafos do córpus RastrOS



Fonte: Elaborada pelo autor.

- **Método de avaliação da complexidade sentencial**: Método para avaliação da complexidade de sentenças, treinado em córpus de sentenças alinhadas, utilizando *ranking* para generalizar para sentenças individuais (ver Seção 5.2 e Seção 5.3).

Aplicações

- **NILC-Metrix**[3]: Ferramenta para extração das métricas de textos em PB, reunindo 200 métricas de coesão, coerência e complexidade textual (ver Capítulo 3);

- **Simpligo-Cloze**[4]: Ferramenta para aplicação do teste Cloze de forma incremental, uma palavra por vez e todas as palavras de um texto, parágrafo ou sentença, com configuração simplificada (ver Seção 4.3);

[3] <http://fw.nilc.icmc.usp.br:23380/nilcmetrix>
[4] <https://simpligo.sidle.al>

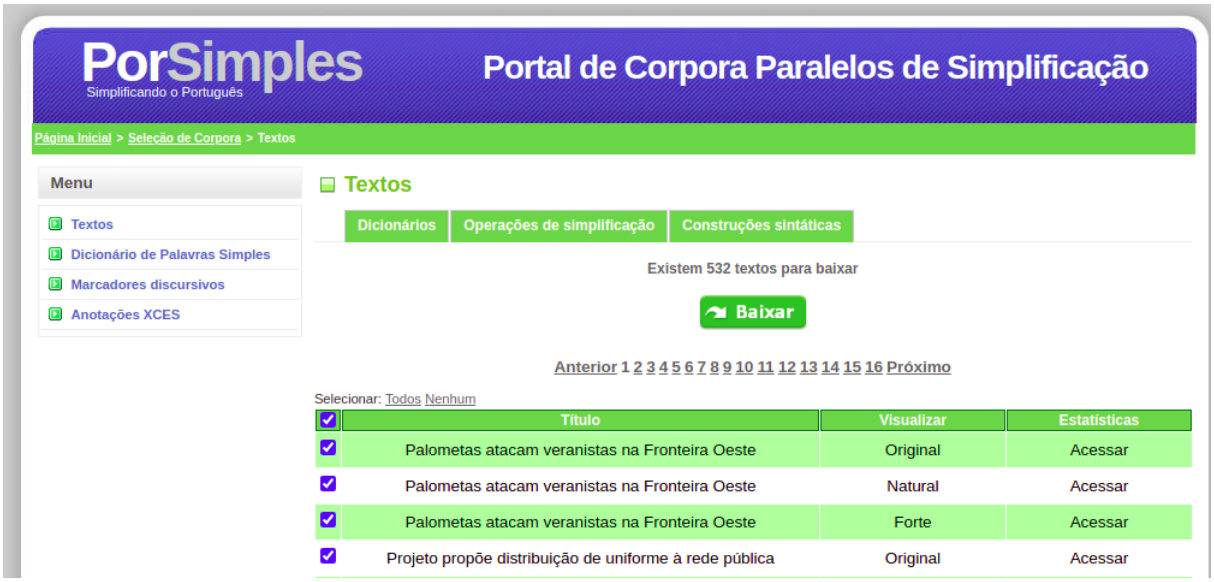
- **Simpligo-Ranking**[5]: Ferramenta beta para demonstração do método de avaliação da complexidade sentencial, a partir do texto submetido. Executa a sentenciação e avalia individualmente a complexidade de cada sentença, atribuindo um valor de 1 a 100, sendo 1 o mais simples e 100 o mais complexo. Na Figura 14 vemos um exemplo da saída da ferramenta, que apresenta também uma classificação em cores de quatro níveis de complexidade: verde (1 a 25), amarelo (26-50), laranja (51-75) e vermelho (76-100).

Além dessas contribuições, é importante também citar o trabalho inicial de recuperação do legado do projeto PorSimples. Como as tecnologias das ferramentas desenvolvidas já estavam desatualizadas, foi necessário utilizar técnicas de virtualização para a restauração. Duas ferramentas merecem destaque:

- **Editor de Anotação de Simplificação**[6]: É o editor original utilizado para simplificar e anotar os textos do PorSimples, ele utiliza o formato "bi-texto" para acompanhamento do texto original e simplificado, permitindo alinhamento das sentenças e anotação das operações de simplificação (vide Figura 8 na Subseção 2.3.2).

- **Portal de Corpora Paralelos de Simplificação**[7]: Essa ferramenta fornece uma interface amigável para navegação entre os textos originais e simplificados do PorSimples, além de uma lista de palavras simples e de marcadores discursivos (vide Figura 15).

Figura 15 – PorSimples: Tela do Portal de Corpora Paralelos de Simplificação



Fonte: Elaborada pelo autor.

5 <http://fw.nilc.icmc.usp.br:23380/simpligo-ranking>
6 <http://fw.nilc.icmc.usp.br:23080>
7 <http://fw.nilc.icmc.usp.br:23081>

A próxima ferramenta a ser resgatada será o SIMPLIFICA (CANDIDO-JUNIOR; OLIVEIRA; ALUÍSIO, 2009), como um trabalho futuro. Porém, esta tarefa será feita com novos métodos, como a avaliação da complexidade no nível textual (Capítulo 3 e sentencial (Seção 5.3), além da simplificação lexical (HARTMANN; ALUíSIO, 2020).

6.2 Limitações

Este trabalho reconhece as seguintes limitações:

- O projeto RastrOS foi bastante impactado por causa da pandemia do Covid-19; a coleta dos dados do eye-tracker dos participantes precisou ser interrompida. A coleta do teste Cloze foi menos impactada, pois permitia a resposta remota. No entanto, teve menos participantes do que os 600 inicialmente planejados. Isso resultou em um córpus com dados de menos participantes do que o previsto, mas ainda assim relevante;
- O córpus PorSimplesSent tem um tamanho razoável, de aproximadamente cinco mil pares de sentenças, mas está bem distante dos tamanhos dos córpus utilizados para avaliação da tarefa na língua inglesa;
- Houve a tentativa de criação de um córpus de textos nos quatro níveis do público alvo para o domínio rural, mas não foi adiante por conta do prazo e política de privacidade dos dados da Embrapa;
- A ferramenta NILC-Metrix usa internamente 3 parsers, sendo um deles um *parser* proprietário. Uma melhoria importante seria substituir essas ferramentas por um único parser robusto de código livre. Outro ponto de melhoria necessária na ferramenta NILC-Metrix é a performance. Textos grandes exigem vários minutos de processamento.

6.3 Trabalhos futuros

Alguns pontos de evolução e sequências naturais deste trabalho são:

- Um dos objetivos iniciais era disponibilizar uma ferramenta de edição *online* que permitisse não apenas a avaliação individual das sentenças dos textos, mas que também sugerisse as simplificações. Esse é o próximo passo bem definido que já pode ser visualizado, juntando o método criado neste trabalho com as técnicas mais recentes que tratam a simplificação como uma tradução intra-língua;
- O Córpus RastrOS deve continuar sendo ampliado, em número de participantes e análises;
- Investigar o melhor modelo deste trabalho nas outras duas versões do PorSimplesSent: PSS1 (com 7.909 pares de sentenças) e PSS3 (com 2.508 pares de sentenças). A PSS1 é

considerada a versão mais simples de todas, com repetição das sentenças originais para cada divisão, e a PSS3 só possui sentenças que não foram divididas, portanto espera-se que seja mais desafiadora;

- Para a tarefa de avaliação da complexidade sentencial, novos datasets maiores são desejáveis. As três versões do PorSimplesSent são relativamente pequenas se comparadas aos datasets disponíveis para a língua inglesa. Outros domínios e gêneros também são importantes;

- Avaliar o melhor método de avaliação de complexidade desenvolvido (LEAL *et al.*, 2020) com as novas métricas disponíveis na última versão do NILC-Metrix;

- No final deste trabalho, foram incluídas duas novas medidas no RastrOS: *Surprisal* e *Entropy Reduction*. Será importante avaliar essas medidas junto às de rastreamento ocular na tarefa de avaliação de complexidade sentencial;

- Durante este trabalho, foram analisadas alternativas mais acessíveis para a captura de dados de rastreamento ocular; uma das melhores apostas de custo-benefício foi o dispositivo Fove[8], que reúne realidade virtual e rastreamento ocular. Porém essa frente foi interrompida por conta da pandemia que impediria uma avaliação. Este tópico é importante e deve ser retomado no futuro.

6.4 Artigos e Publicações

Foram escritos onze artigos durante o doutorado (listados na Tabela 20), sete já publicados, dois aceitos para publicação e dois submetidos (em processo de avaliação por pares), sendo oito artigos para conferências e três para revistas. As premiações e artigos em destaque foram:

- Prêmio de *Best Paper* no *workshop* VI JDP - Jornada de Descrição do Português em 2019, com o artigo: "Métodos de Clusterização para a Criação de Corpus para Rastreamento Ocular durante a Leitura de Parágrafos em Português";

- Prêmio de terceiro *Best Paper* no STIL - *Symposium in Information and Human Language Technology* em 2019, com o artigo: "Avaliação Automática da Complexidade de Sentenças do Português Brasileiro para o Domínio Rural.";

- Indicação para *Best Paper* (ficando entre os 18 melhores artigos longos) no *COLING - 28th International Conference on Computational Linguistics* em 2020, com o artigo: "*Using Eye-tracking Data to Predict the Readability of Brazilian Portuguese Sentences in Single-task, Multi-task and Sequential Transfer Learning Approaches*.".

[8] <https://fove-inc.com/>

Na Tabela 20, todos os artigos nos quais o autor consta como principal estão diretamente relacionados a esta tese e foram reproduzidos integralmente aqui. Os demais trabalhos têm certa relação, pois utilizam as mesmas métricas e recursos descritos, porém foram desenvolvidos em parceria com colegas da Embrapa ou do NILC em suas próprias pesquisas.

Tabela 20 – Lista de artigos desenvolvidos durante esta pesquisa, em ordem cronológica.

Artigos
MAGALHAES, V. M. A. ; BERNARDO, W. F. ; DINIZ, F. H. ; LAGE DOS SANTOS, K. C. ; FONSECA, L. M. G. ; LEAL, S.; ALUISIO, S. M. . **E-rural methodology: Contents elaborated according to the literacy level of the target audience.** In: Twelfth Latin American Conference on Learning Technologies (LACLO), 2017, La Plata. p. 1–9.
LEAL, S.; DURAN, M. S. ; ALUISIO, S. M. . **A Nontrivial Sentence Corpus for the Task of Sentence Readability Assessment in Portuguese.** In: Proceedings of the 27th International Conference on Computational Linguistics, 2018. p. 401-413.
GAZZOLA, M. ; LEAL, S. E. ; ALUISIO, S. M. . **Predição da Complexidade Textualde Recursos Educacionais Abertos em Português.** In: Proceedings of 12th Brazilian Symposium in Information and Human Language Technology, 2019.
LEAL, S. E.; RODRIGUES, E. S. ; VIEIRA, J. M. M. ; TEIXEIRA, E. N. ; ALUISIO, S. M. . **Métodos de Clusterização para a Criação de Corpus para Rastreamento Ocular durante a Leitura de Parágrafos em Português.** In: VI Jornada de Descrição do Português (Workshop), 2019, Salvador - BA. Proceedings of 6th Workshop on Portuguese Description (JDP), 2019.
LEAL, S. E.; MAGALHAES, V. M. A. ; DURAN, M. S. ; ALUISIO, S. M. . **Avaliação Automática da Complexidade de Sentenças do Português Brasileiro para o Domínio Rural.** In: Proceedings of 12th Brazilian Symposium in Information and Human Language Technology, 2019. Salvador - BA
LEAL, S. E.; VIEIRA, J. M. M. ; RODRIGUES, E. S. ; TEIXEIRA, E. N. ; ALUISIO, S. M. . **Using Eye-tracking Data to Predict the Readability of Brazilian Portuguese Sentences in Single-task, Multi-task and Sequential Transfer Learning Approaches.** In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020. p. 5821-5831.
SANTOS, R. L. S. ; WICK-PEDRO, G. ; LEAL, S. ; VALE, O. A. ; PARDO, T. A. S. ; BONTCHEVA, K. ; SCARTON, C. . **Measuring the Impact of Readability Features in Fake News Detection.** In: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). Marseille, France: European Language Resources Association (ELRA), 2020. p. 1397-1406.
GAZZOLA, M.; LEAL, S.; PEDRONI, B.; ROCHA, F.T.; POMPÉIA, S.; ALUÍSIO, S. . **Text Complexity of Open Educational Resources in Portuguese: Mixing Written and Spoken Registers in a Multi-task Approach.** In: Language Resources and Evaluation (LREV), 2021. *Aceito para publicação (11.03.21)*
LEAL, S. E.; CASANOVA, E. ; PAETZOLD, G. H. ; ALUISIO, S. M. . **Evaluating Semantic Similarity Methods to build Semantic Predictability Norms of Reading Data.** In: The twenty-fourth International Conference on Text, Speech and Dialog, 2021. *Aceito para publicação (24.05.21)*
LEAL, S. E.; DURAN, M. S.; SCARTON, C. E.; HARTMANN N. S.; ALUÍSIO, S. M. . **NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese.** In: Language Resources and Evaluation - Special Issue: Computational approaches to Portuguese, 2021. *Submetido - Sob revisão editorial*
LEAL, S. E.; LUKASOVA, K.; CARTHERY-GOULART, M. T.; ALUÍSIO, S. M. . **RastrOS Project: Natural Language Processing contributions to the development of an eye-tracking corpus with predictability norms for Brazilian Portuguese.** In: Language Resources and Evaluation (LREV), 2021. *Submetido - Sob revisão editorial*

Fonte: Elaborada pelo autor.

REFERÊNCIAS

ABREU, K. N. M. de; GARCIA, D. C. de; HORA, K. da; SOUZA, C. R. de. O teste de cloze como instrumento de medida da proficiência em leitura: fatores linguísticos e não linguísticos. **Revista de Estudos da Linguagem**, v. 25, n. 3, p. 1767–1799, 2017. Citado na página 41.

ALUÍSIO, S.; GASPERIN, C. Fostering digital inclusion and accessibility: the porsimples project for simplication of portuguese texts. **Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas - Association for Computational Linguistics**, p. 46–53, 2010. Citado nas páginas 1 e 14.

AMBATI, B. R.; REDDY, S.; STEEDMAN, M. Assessing relative sentence complexity using an incremental ccg parser. **Proceedings of the North American Chapter of the Association for Computational Linguistics on Language technologies, California, USA**, p. 1051–1057, 2016. Citado na página 6.

ARFÉ, B.; MASON, L.; FAJARDO, I. Simplifying informational text structure for struggling readers. **Read Writ (2018) Volume 31, Issue 9**, p. 2191–2210, 2018. Citado na página 12.

BARRETT, M. J.; AGIC, Z.; SØGAARD, A. The dundee treebank. **Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories: TLT14**, p. 242–248, 2015. Citado nas páginas 26 e 35.

BICK, E. The parsing system"palavras": Automatic grammatical analysis of portuguese in a constraint grammar framework. **Aarhus University Press**, 2000. Citado nas páginas 14 e 24.

BOHN, H. I. Linguistic complexity and text comprehension: Readability lssues reconsidered by davison and green. **Revista Fragmentos**, v.3,n.2, 1990. Citado na página 3.

BOITO, M. Z. Simplificação lexical de substantivos e multiword expressions. **Salão de Iniciação Científica (26. : 2014 out. 20-24 : UFRGS, Porto Alegre, RS)**, 2014. Citado na página 13.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. **CoRR abs/1607.04606**, 2016. Citado na página 47.

BOSCO, G. L.; PILATO, G.; SCHICCHIA, D. A neural network model for the evaluation of text complexity in italian language: a representation point of view. **Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2018)**, p. 464–470, 2018. Citado na página 6.

____. A recurrent deep neural network model to measure sentence complexity for the italian language. **Proceedings of the sixth International Workshop on Artificial Intelligence and Cognition**, 2018. Citado nas páginas 31 e 46.

BRUNATO, D.; CIMINO, A.; DELL'ORLETTA, F.; VENTURI, G. Paccss–it: A parallel corpus of complex–simple sentences for automatic text simplification. **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, p. 351–361, 2016. Citado na página 31.

BRUNATO, D.; MATTEI, L. D.; DELL'ORLETTA, F.; IAVARONE, B.; VENTURI, G. Is this sentence difficult? do you agree? **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, p. 2690–2699, 2018. Citado na página 6.

CANDIDO-JUNIOR, A. **Análise bidirecional da língua na simplificação sintática em textos de português voltada à acessibilidade digital**. ICMC - USP São Carlos: Biblioteca Digital USP, 2013. Citado na página 14.

CANDIDO-JUNIOR, A.; OLIVEIRA, M. de; ALUÍSIO, S. M. Simplifica: um sistema web de autoria de textos simplificados. **Simpósio Brasileiro de Sistemas Multimídia e Web (Webmedia 2009) v.2**, p. 55–58, 2009. Citado nas páginas 7, 14 e 195.

CARUANA, R. Multitask learning. **Machine Learning - Special issue on inductive transfer - Volume 28**, p. 41–75, 1997. Citado na página 45.

CASELI, H. de M.; PEREIRA, T. de F.; SPECIA, L.; PARDO, T. A. S.; GASPERIN, C.; ALUÍSIO, S. M. Building a brazilian portuguese parallel corpus of original and simplified texts. **Advances in Computational Linguistics, Research in Computer Science (CICLing-2009)**, v. 41, p. 59–70, 2009. Citado nas páginas 9, 11, 8, 32 e 33.

CHALL, J. S.; DALE, E. **Readability revisited: the new Dale-Chall readability formula**. [S.l.]: Brookline Books, 1995. Citado na página 18.

CHANDRASEKAR, R.; DORAN, C.; SRINIVAS, B. Motivations and methods for text simplification. **Proceedings of the 16th International Conference on Computational Linguistics (COLING)**, p. 1041–1044, 1996. Citado na página 12.

COLEMAN, M.; LIAU, T. L. A computer readability formula designed for machine scoring. **Journal of Applied Psychology**, v. 60, p. 283–284, 1975. Citado na página 18.

COP, U.; DIRIX, N.; DRIEGHE, D.; DUYCK, W. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. **Behavior Research Methods**, v. 49, p. 602–615, 05 2016. Citado na página 36.

CSIKSZENTMIHALYI, M. **Flow: The Psychology of Optimal Experience**. [S.l.]: Harper Perennial, 2008. Citado na página 4.

CUNHA, A. L. V. da. **Coh-Metrix-Dementia: análise automática de distúrbios de linguagem nas demências utilizando Processamento de Línguas Naturais**. ICMC - USP São Carlos: Biblioteca Digital USP, 2015. Citado nas páginas 19, 23 e 24.

DAVISON, A.; GREEN, G. **Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered**. [S.l.]: Routledge, 1988. Citado na página 3.

DELL'ORLETTA, F.; MONTEMAGNI, S.; VENTURI, G. Assessing document and sentence readability in less resourced languages and across textual genres. **International Journal of Applied Linguistics (ITL). Special Issue on Readability and Text Simplification**, 2014a. Citado na página 6.

DELL'ORLETTA, F.; MONTEMAGNI, S.; VENTURI, G. Read-it: Assessing readability of italian texts with a view to text simplification. **Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies**, p. 73–83, 2011. Citado nas páginas 6, 17, 18 e 31.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **Google AI Language**, 2018. Citado na página 47.

DUBAY, W. Robert gunning's fog readability formula. **Plain Language At Work Newsletter**, v. 8, 2014. Disponível em: <http://www.impact-information.com/impactinfo/newsletter/plwork08.htm>. Citado na página 18.

DUBAY, W. H. **Smart Language: Readers, Readability, and the Grading of Text**. Costa Mesa, CA: Impact Information, 2007. Citado nas páginas 3 e 17.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. de. **Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina**. Rio de Janeiro: LTC - Livros Técnicos e Científicos, 2011. Citado nas páginas 42, 43 e 50.

FELLBAUM, C. Wordnet: An electronic lexical database. **MIT Press**, 1998. Citado na página 13.

FETTER, G. L. **Divulgação Tecnológica para Agricultores Familiares: Análise de Terminologias sob a Ótica da Linguística Sistêmico-Funcional**. UFRGS - Porto Alegre - RS: Instituto de Letras - UFRGS, 2017. Citado na página 7.

FILHO, J. A. W.; WILKENS, R.; IDIART, M.; VILLAVICENCIO, A. The brWaC corpus: A new open resource for Brazilian Portuguese. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Disponível em: <https://www.aclweb.org/anthology/L18-1686>. Citado na página 48.

FISCHER, H. Só é acessível se der para entender. In: **Acessibilidade cultural : atravessando fronteiras - Seminário Internacional de Acessibilidade Cultural promovido pela Rede de Museus da Pró-Reitoria de Extensão e Cultura da UFPel**. [S.l.]: Editora UFPel, 2020. p. 244–261. Citado na página 1.

FOVE. **Fove Eye Tracker**. 2018. Disponível em: <https://www.getfove.com/>. Citado na página 26.

GONZALEZ-GARDUñO, A. V.; SØGAARD, A. Using gaze to predict text readability. **Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications**, p. 438–443, 2017. Citado nas páginas 36, 46 e 191.

____. Learning to predict readability using eye-movement data from natives and learners. **Proceedings of the The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)**, p. 5118–5124, 2018. Citado nas páginas 6, 37, 44, 46, 177, 190 e 191.

GOOGLEBLOG. **Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing**. 2018. Disponível em: <https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>. Citado na página 47.

GRAESSER, A. C.; MCNAMARA, D. S.; KULIKOWICH, J. M. Coh-metrix: Providing multilevel analyses of text characteristics. **Educational Researcher Vol. 40, N. 5**, p. 223–234, 2011. Citado nas páginas 16, 20 e 22.

GRAESSER, A. C.; MCNAMARA, D. S.; LOUWERSE, M. M.; CAI, Z. Coh-metrix: Analysis of text on cohesion and language. **Behavior Research Methods, Instruments, n Computer - Springer**, p. 193–202, 2004. Citado na página 19.

HARTMANN, N. S.; ALUíSIO, S. M. Adaptação lexical automática em textos informativos do português brasileiro para o ensino fundamental. **Linguamática**, v. 12, n. 2, p. 3–27, Dez. 2020. Disponível em: <https://linguamatica.com/index.php/linguamatica/article/view/323>. Citado na página 195.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation, MIT Press**, p. 1735–1780, 1997. Citado na página 46.

HUSAIN, S.; VASISHTH, S.; SRINIVASAN, N. Integration and prediction difficulty in hindi sentence comprehension: Evidence from an eye-tracking corpus. **Journal of Eye Movement Research, 8(2)**, 2015. Citado na página 37.

HWANG, W.; HAJISHIRZI, H.; OSTENDORF, M.; WU, W. Aligning sentences from standard wikipedia to simple wikipedia. **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, p. 211–217, 2015. Citado na página 30.

IMOTIONS. **Eye Tracking - The Complete Pocket Guide**. [S.l.]: www.imotions.com, 2017. Citado na página 26.

IPM. Inaf brasil 2015: Indicador de alfabetismo funcional - alfabetismo no mundo do trabalho. **Instituto Paulo Montenegro**, 2016. Disponível em: <http://www.ipm.org.br/pt-br/programas/inaf/relatoriosinafbrasil/Paginas/Inaf-2015---Alfabetismo-no-Mundo-do-Trabalho.aspx>. Citado nas páginas 1 e 7.

____. Inaf brasil 2018: Indicador de alfabetismo funcional - resultados preliminares. **Instituto Paulo Montenegro**, 2018. Disponível em: <http://acaoeducativa.org.br/wp-content/uploads/2018/08/Inaf2018_Relat%C3%B3rio-Resultados-Preliminares_v08Ago2018.pdf>. Citado nas páginas 1 e 3.

JOACHIMS, T. Training linear SVMs in linear time. In: **Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.]: ACM Press, 2006. v. 3, p. 217–226. Citado na página 44.

JONNALAGADDA, S.; GONZALEZ, G. Biosimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction. **AMIA Annual Symposium Proceedings**, p. 351–356, 2010. Citado na página 11.

KENNEDY, A.; HILL, R.; PYNTE, J. The dundee corpus. **Proceedings of the 12th European conference on eye movement**, 2003. Citado na página 35.

KENNEDY, A.; PYNTE, J. Parafoveal-on-foveal effects in normal reading. **Vision Research - Volume 45, Issue 2**, 2005. Citado na página 35.

KENT.EDU. **SPSS TUTORIALS: PEARSON CORRELATION**. 2021. Disponível em: <https://libguides.library.kent.edu/SPSS/PearsonCorr>. Citado na página 50.

KINCAID, J. P.; FISHBURNE, R. P.; ROGERS, R. L.; CHISSOM, B. S. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for navy enlisted personnel. **Research Branch Report**, p. 8–75, 1975. Citado na página 18.

KLIEGL, R.; GRABNER, E.; ROLFS, M.; ENGBERT, R. Length, frequency, and predictability effects of words on eye movements in reading. **European Journal of Cognitive Psychology, 16**, p. 262–284, 2004. Citado na página 36.

KLIEGL, R.; NUTHMANN, A.; ENGBERT, R. Tracking the mind during reading: The influence of past, present, and future words on fixation durations. **Journal of Experimental Psychology: General, v. 135**, p. 12–35, 2006. Citado na página 36.

LANDAUER, T. K.; LAHAM, D.; REHDER, B.; SCHREINER, M. E. How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In: SHAFTO, M. G.; LANGLEY, P. (Ed.). **Proceedings of the 19th annual meeting of the Cognitive Science Society**. [S.l.: s.n.], 1997. p. 412–417. Citado na página 47.

LAURINAVICHYUTE, A. K.; SEKERINA, I. A.; ALEXEEVA, S.; BAGDASARYAN, K.; KLIEG, R. Russian sentence corpus: Benchmark measures of eye movements in reading in russian. **Behavior Research Methods**, p. 1–18, 2018. Citado na página 37.

LEAL, S. E.; VIEIRA, J. M. M.; RODRIGUES, E. dos S.; TEIXEIRA, E. N.; ALUÍSIO, S. Using eye-tracking data to predict the readability of Brazilian Portuguese sentences in single-task, multi-task and sequential transfer learning approaches. In: **Proceedings of the 28th International Conference on Computational Linguistics**. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020. p. 5821–5831. Disponível em: <https://www.aclweb.org/anthology/2020.coling-main.512>. Citado na página 196.

LEITÃO, M. M.; RIBEIRO, A. J. C.; MAIA, M. Penalidade do nome repetido e rastreamento ocular em português brasileiro. **Revista LinguíStica**, v8 n2, 2012. Citado na página 26.

LI, H. **Learning to Rank for Information Retrieval and Natural Language Processing**. 2a. ed. [S.l.]: Morgan & Claypool, 2014. Citado nas páginas 42 e 43.

LORENA, A. C.; CARVALHO, A. C. P. L. F. de. Uma introdução às support vector machines. **RITA - Revista de Informática Teórica e Aplicada**, v. 14, p. 43–67, 2007. Citado na página 44.

LUKE, S. G.; CHRISTIANSON, K. The provo corpus: A large eye-tracking corpus with predictability norms. **Behavior Research Methods**, 2018. Citado nas páginas 37 e 41.

MAGALHÃES, V. M. A. de; BERNARDO, W. F.; DINIZ, F. H.; SANTOS, K. C. L. dos; FONSECA, L. M. G.; ALUISIO, S. M.; LEAL, S. E. E-rural methodology: Contents elaborated according to the literacy level of the target audience. **Twelfth Latin American Conference on Learning Technologies (LACLO)**, 2017. Citado na página 1.

MAIA, M.; LEMLE, M.; FRANçA, A. I. Efeito stroop e rastreamento ocular no processamento de palavras. **Ciências e Cognição 2007**, v. 12, p. 02–17, 2007. Citado na página 26.

MARGARIDO, P. R. A.; PARDO, T. A. S.; ANTONIO, G. M.; FUENTES, V. B.; AIRES, R.; ALUÍSIO, S. M.; FORTES, R. P. M. Automatic summarization for text simplification: Evaluating text understanding by poor readers. In: **Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web**. New York, NY, USA: ACM, 2008. (WebMedia '08), p. 310–315. ISBN 978-85-7669-199-0. Disponível em: <http://doi.acm.org/10.1145/1809980.1810057>. Citado na página 15.

MARTINS, T.; GHIRALDELO, C.; NUNES, M.; JR., O. **Readability Formulas Applied to Textbooks in Brazilian Portuguese**. [S.l.], 1996. Martins, T.B.F.; Ghiraldelo, C.M.; Nunes, M.G.V.; Oliveira Jr., O.N. Readability Formulas Applied to Textbooks in Brazilian Portuguese. Notas do ICMSC-USP, Série Computação, nro. 28, 1996, 11p. Citado na página 17.

MAYER, R. E. Elaboration techniques that increase the meaningfulness of technical text: An experimental test of the learning strategy hypothesis. **Journal of Educational Psychology**, American Psychological Association, v. 72, n. 6, p. 770–784, 1980. Citado nas páginas 11 e 15.

MAZIERO, E. G.; PARDO, T. A. S.; ALUÍSIO, S. M. Ferramenta de análise automática de inteligibilidade de córpus (aic). **NILC - ICMC-USP**, 2008. Citado nas páginas 24 e 25.

MCCRAY, G.; BRUNFAUT, T. Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. **Language Testing**, v. 35, n. 1, p. 51–73, 2018. Citado na página 41.

MCNAMARA, D.; GRAESSER, A.; CAI, Z.; DAI, J. **Coh-Metrix Common Core T.E.R.A. version 1.0.** 2013. [Online; acessado em 2021.03.10. Disponível em: <http://www.commoncoretera.com/>. Citado na página 22.

MCNAMARA, D. S.; GRAESSER, A. C.; MCCARTHY, P. M.; CAI, Z. **Automated Evaluation of Text and Discourse with Coh-Metrix**. 1a. ed. [S.l.]: Cambridge University Press, 2014. Citado na página 19.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **CoRR abs/1301.3781**, 2013. Citado na página 47.

MILLER, G. A. Wordnet: A lexical database for english. **Communications of the ACM**, Vol. 38, No. 11, p. 39–41, 1995. Citado na página 13.

MIN.AGRICULTURA. **Ministério da Agricultura, Pecuária e Abastecimento: Agropecuária puxa o PIB de 2017**. 2017. Disponível em: <http://www.agricultura.gov.br/noticias/agropecuaria-puxa-o-pib-de-2017>. Citado nas páginas 1 e 7.

NEURALMIND. **NeuralMind disponibiliza modelo BERT, Inteligência Artificial do Google, em português**. 2020. Disponível em: <https://neuralmind.ai/2020/01/26/neuralmind-disponibiliza-modelo-bert-inteligencia-artificial-do-google-em-portugues/>. Citado na página 48.

PAETZOLD, G.; SPECIA, L. Inferring psycholinguistic properties of words. In: **NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016**. [s.n.], 2016. p. 435–440. Disponível em: <http://aclweb.org/anthology/N/N16/N16-1050.pdf>. Citado na página 25.

____. Lexical simplification with neural ranking. In: **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**. Valencia, Spain: Association for Computational Linguistics, 2017. p. 34–40. Disponível em: <http://www.aclweb.org/anthology/E17-2006>. Citado na página 13.

PAETZOLD, G. H.; SPECIA, L. Unsupervised lexical simplification for non-native speakers. In: **Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence**. AAAI Press, 2016. (AAAI'16), p. 3761–3767. Disponível em: <http://dl.acm.org/citation.cfm?id=3016387.3016433>. Citado na página 13.

PASQUALINI, B. **Corpop : um corpus de referência do português popular escrito do Brasil**. UFRGS - Porto Alegre - RS: Instituto de Letras - UFRGS, 2018. Citado nas páginas 34 e 35.

PENNEBAKER, J. W.; BOYD, R. L.; JORDAN, K.; BLACKBURN, K. The development and psychometric properties of liwc2015. **The University of Texas at Austin**, 2015. Citado na página 24.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. **Empirical Methods in Natural Language Processing (EMNLP)**, p. 1532–1543, 2013. Citado na página 47.

RAYNER, K. Eye movements in reading and information processing: 20 years of research. **Psychological Bulletin - APA, vol. 124 n. 3**, p. 372–422, 1998. Citado na página 26.

REIS, G. B. Predição da complexidade textual de notícias jornalísticas usando uma plataforma crowdsourcing. **Monografia Conclusão Curso - USP**, 2017. Citado na página 24.

ROSENBERG, A.; HIRSCHBERG, J. V-measure: A conditional entropy-based external cluster evaluation measure. In: **Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**. [S.l.]: Association for Computational Linguistics, 2007. p. 410–420. Citado na página 50.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1987. Citado na página 51.

RUDER, S. **An Overview of Multi-Task Learning in Deep Neural Networks**. 2017. [Online; acessado em 2021.03.11]. Disponível em: <https://ruder.io/multi-task/>. Citado na página 45.

RUDER, S. **Neural Transfer Learning for Natural Language Processing**. Tese (Doutorado) — NATIONAL UNIVERSITY OF IRELAND, GALWAY, 2019. Citado na página 45.

RUDER, S.; PETERS, M. E.; SWAYAMDIPTA, S.; WOLF, T. Transfer learning in natural language processing. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 15–18. Disponível em: <https://www.aclweb.org/anthology/N19-5004>. Citado na página 45.

RUMELHART, D. E.; GROUP, J. L. M. an P. R. Parallel distributed processing: Explorations in the microstructures of cognition. **The MIT Press**, v. 1 - Foundations, 1986. Citado na página 44.

SANTOS, L. B. dos; DURAN, M. S.; HARTMANN, N. S.; CANDIDO, G. H. P. A.; ALUISIO, S. M. A lightweight regression method to infer psycholinguistic properties for brazilian portuguese. **International Conference on Text, Speech, and Dialogue**, p. 281–289, 2017. Citado nas páginas 24 e 25.

SCARTON, C.; ALUÍSIO, S. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. **Linguamática**, p. 45–62, 2010. Citado nas páginas 7 e 23.

SCARTON, C.; OLIVEIRA-JUNIOR, O.; CANDIDO-JUNIOR, A.; GASPERIN, C.; ALUÍSIO, S. M. Simplifica: a tool for authoring simplified texts in brazilian portuguese guided by readability assessments. **Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies**, Los Angeles, CA, p. 41–44, 2010. Citado nas páginas 7, 14 e 23.

SCARTON, C.; PAETZOLD, G. H.; SPECIA, L. Simpa: A sentence-level simplification corpus for the public administration domain. **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, p. 4333–4338, 2018. Citado na página 31.

____. Text simplification from professionally produced corpora. **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, p. 3504–3510, 2018. Citado nas páginas 6 e 30.

SCARTON, C.; SPECIA, L. Learning simplifications for specific target audiences. **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)**, p. 712–718, 2018. Citado nas páginas 14 e 30.

SEAD. **Secretaria Especial de Agricultura Familiar e do Desenvolvimento Agrário: Agricultura familiar do Brasil é 8ª maior produtora de alimentos do mundo**. 2018. Disponível em: <http://www.mda.gov.br/sitemda/noticias/agricultura-familiar-do-brasil-%C3%A9-8%C2%AA-maior-produtora-de-alimentos-do-mundo>. Citado nas páginas 1 e 7.

SIDDHARTHAN, A. Syntactic simplifcation and text cohesion. **Research on Language and Computation - Springer**, 2006. Citado na página 11.

SINGH, A. D.; MEHTA, P.; HUSAIN, S.; RAJKUMAR, R. Quantifying sentence complexity based on eye-tracking measures. **Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity**, p. 202–212, 2016. Citado na página 36.

SJÖHOLM, J. **Probability as readability: A new machine learning approach to readability assessment for written Swedish**. [S.l.]: LiU Electronic Press, 2012. Citado nas páginas 6, 17, 18 e 31.

SOARES, M. O que é letramento? **Presença Pedagógica Volume 2, n. 10**, p. 15–25, 1996. Citado na página 4.

STAJNER, S.; PONZETTO, S. P.; STUCKENSCHMIDT, H. Automatic assessment of absolute sentence complexity. **Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)**, 2017. Citado nas páginas 6 e 30.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Pearson Education, 2006. Citado na página 49.

TEIXEIRA, E. N.; FONSECA, M. C. M.; SOARES, M. E. Resolução do pronome nulo em português brasileiro: Evidência de movimentação ocular. **VEREDAS: Sintaxe das Línguas Brasileiras**, v. 18, 2014. Citado na página 26.

THOMAS, C.; KEŠELJ, V.; CERCONE, N.; ROCKWOOD, K.; ASP, E. Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. **Proceedings of the IEEE International Conference on Mechatronics and Automation**, p. 1569–1574, 2005. Citado na página 19.

TIMM, L. B. **Looking at text simplification: Using eye tracking to evaluate the readability of automatically simplified sentences**. Linköping, Sweden: Linköping University, Department of Computer and Information Science, Human-Centered systems, 2018. Citado na página 6.

TREVISO, M. V. **Segmentação de sentenças e detecção de disfluências em narrativas transcritas de testes neuropsicológicos**. ICMC - USP São Carlos: Biblioteca Digital USP, 2017. Citado na página 24.

URANO, K. Lexical simplification and elaboration: An experiment in sentence comprehension and incidental vocabulary acquisition. **Hokkai-Gakuen University - Thesis Defense**, 2000. Citado na página 15.

VAJJALA, S.; LUČIĆ, I. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. **Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications**, p. 297–304, 2018. Citado na página 30.

VAJJALA, S.; MEURERS, D. Assessing the relative reading level of sentence pairs for text simplification. **Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)**, p. 288–297, 2014a. Citado nas páginas 6, 12 e 29.

____. Readability-based sentence ranking for evaluating text simplification. **CoRR**, abs/1603.06009, 2016. Citado nas páginas 6, 16, 29, 30, 44 e 152.

VAPNIK, V. N. The nature of statistical learning theory. **Springer-Verlag**, 1995. Citado na página 44.

____. Statistical learning theory. **John Wiley and Sons**, 1998. Citado na página 44.

VAPNIK, V. N.; CHERVONENKIS, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. **Theory of Probability and its Applications, 16(2)**, p. 283–305, 1971. Citado na página 44.

VIEIRA, J. M. M. **The Brazilian Portuguese eye tracking corpus with a predictability study focusing on lexical and partial prediction**. 113 p. Dissertação (Mestrado) — Federal University of Ceará (UFC), Universidade Federal do Ceará, Biblioteca Universitária, 2020. Disponível em: <http://www.repositorio.ufc.br/handle/riufc/55798>. Citado na página 115.

VIEIRA, R.; SANTOS, J. **Tutorial Modelos de Linguagem (Word Embeddings)**. 2019. STIL 2019 - XII Symposium in Information and Human Language Technology. Citado na página 47.

WATANABE, W. M.; FORTES, R. P. de M.; PARDO, T. A. S.; ALUÍSIO, S. M. Facilita: helping the reading of texts available on the web. In: **XV Brazilian Symposium on Multimedia and the Web, WebMedia '09, Fortaleza, Ceará, Brazil, October 5-7, 2009**. [s.n.], 2009. p. 39. Disponível em: <http://doi.acm.org/10.1145/1858477.1858516>. Citado na página 15.

WATANABE, W. M.; JUNIOR, A. C.; UZÊDA, V. R. de; FORTES, R. P. de M.; PARDO, T. A. S.; ALUÍSIO, S. M. Facilita: reading assistance for low-literacy readers. In: **Proceedings of the 27th Annual International Conference on Design of Communication, SIGDOC 2009, Bloomington, Indiana, USA, October 5-7, 2009**. [s.n.], 2009. p. 29–36. Disponível em: <http://doi.acm.org/10.1145/1621995.1622002>. Citado na página 15.

XU, C.; TAO, D.; XU, C. A survey on multi-view learning. **arXiv:1304.5634v1**, 2013. Citado na página 44.

YAN, M.; KLIEGL, R.; RICHTER, E. M.; NUTHMANN, A.; SHU, H. Flexible saccade-target selection in chinese reading. **The Quarterly Journal of Experimental Psychology, 63(4)**, p. 705–725, 2010. Citado na página 37.

ZELENINA, M. **Eye Tracking for NLP**. 2015. Disponível em: <https://www.slideshare.net/mariezelenina/presentation-2-47610828>. Citado nas páginas 27 e 36.

ZHU, Z.; BERNHARD, D.; GUREVYCH, I. A monolingual tree-based translation model for sentence simplification. **Proceedings of The 23rd International Conference on Computational Linguistics (COLING), August 2010. Beijing, China**, p. 1353–1361, 2010. Citado na página 29.


ICMC
USP
SÃO CARLOS