



Complexidade textual e suas tarefas relacionadas

Sidney Evaldo Leal
Cientista de Dados e Pesquisador
Venturus / ICMC-USP
sidney.leal@venturus.org.br
sidleal@gmail.com



ACESSO AO LIVRO

PROCESSAMENTO DE LINGUAGEM NATURAL

Conceitos, Técnicas
e Aplicações em
Português

ISBN: 978-65-00-95750-1
Licença: CC BY-NC-ND 3.0 BR

Organizado por:
Helena de Medeiros Caseli
Maria das Graças Volpe Nunes

2ª Edição | 2024

<https://brasileiraspln.com/livro-pln>

Sumário

- ❑ O que é complexidade textual?
- ❑ Complexo para quem?
- ❑ Tarefas de PLN relacionadas
- ❑ Avaliação da complexidade e métricas
- ❑ Recursos para o português brasileiro
- ❑ Grandes modelos de linguagem

O que é
complexidade
textual?

Complexidade x inteligibilidade

Readability:

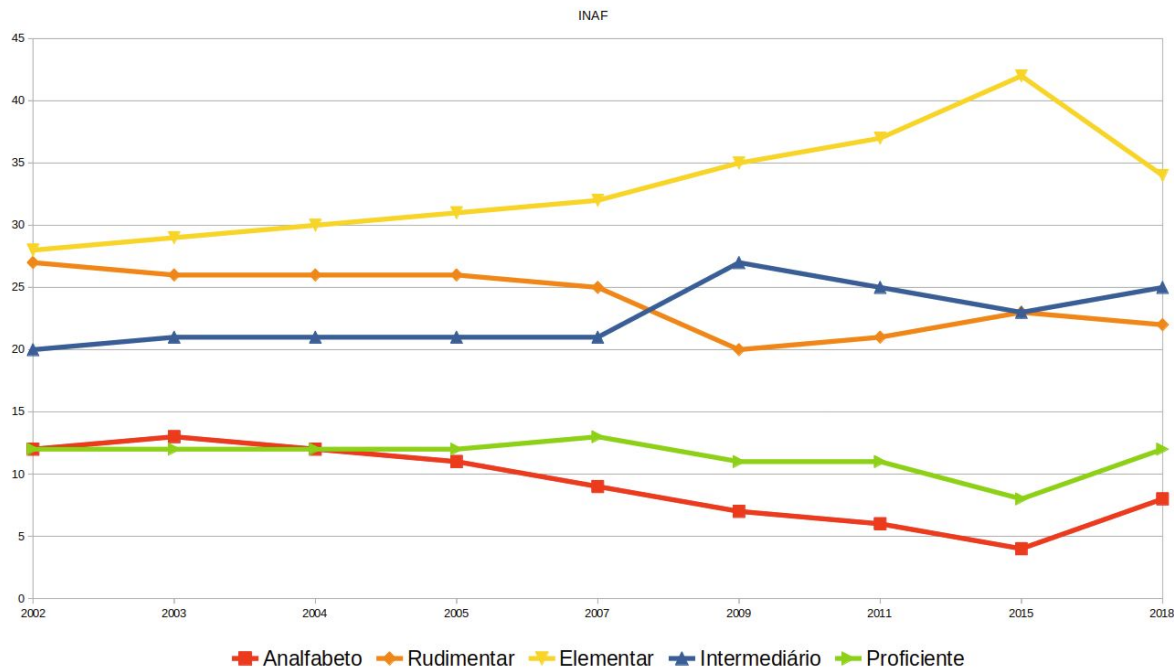
- Inteligibilidade
- Leiturabilidade
- ~~— Legibilidade~~

“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)



Complexo
para quem?

Letramento no Brasil



(IPM, 2018)

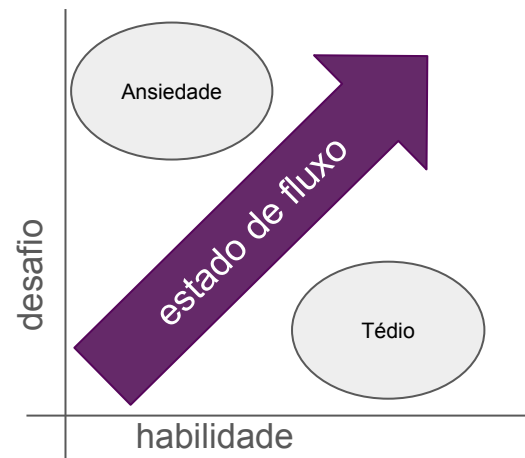
Complexo para quem?



Por que não ensinar a ler?

Tempo versus acesso: Ensinar a ler exige tempo, enquanto simplificar pode permitir o acesso à informação no momento presente.

Nível ideal de complexidade: A evolução do processo de ensino-aprendizagem pode ser muito mais eficiente se os textos fornecidos aos estudantes apresentarem um nível de desafio adequado.

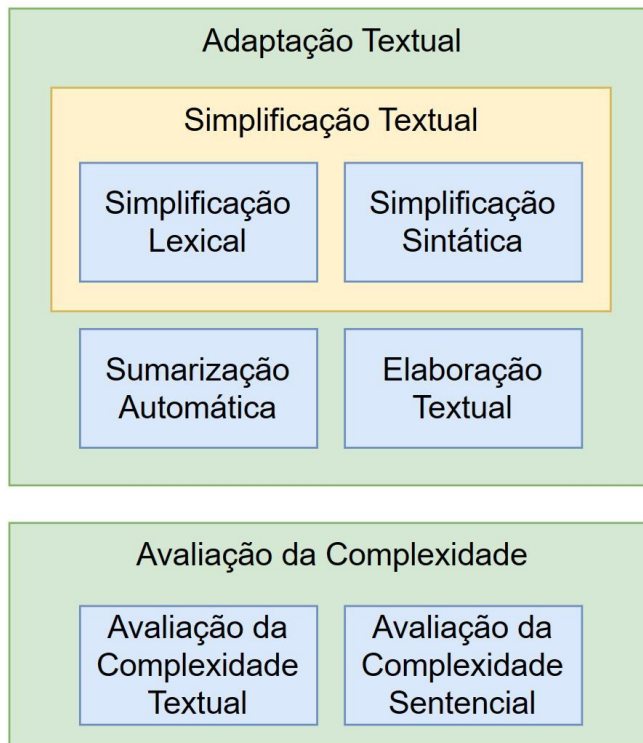


Tarefas de PLN* relacionadas



*Processamento de
Linguagem Natural

Principais tarefas PLN



Simplificação

Complexo

Simple

Simplificação lexical

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

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

Simplificação sintática, via divisão de sentença

O uso de forragem conservada, **cujas** formas mais comuns são: ensilagem e fenação, é uma solução para alimentar o rebanho.

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

Elaboração via definição curta

A ensilagem é o processo de conservação do alimento por fermentação **anaeróbia**.

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



Sumarização

<https://brasileiraspln.com/livro-pln/2a-edicao/parte-aplicacoes/cap-as/cap-as.html>

Avaliação da complexidade sentencial

>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur ut uma eu turpis mollis maximus e u ac tortor. Fusce non pellentesque enim. Proin a faucibus neque. Nam iaculis non odio vitae blandit. Cras pretium porttitor ligula, et placerat metus blandit ut. Donec condimentum commodo urna, id luctus felis finibus vel. Donec pulvinar, lacus vitae ullamcorper vulputate, elit arcu pharetra lorem, ut tristique nisl libero quis nisl. Interdum et malesuada fames ac ante ipsum primis in faucibus. Vivamus pharetra vel neque at molestie.

Etiam laoreet nulla ut nulla auctor sagittis. Cras ullamcorper convallis neque molestie aliquet. Integer placerat nec nisl eget fringilla. Aenean nisl mauris, condimentum rutrum neque vitae, mattis volutpat elit. Etiam sit amet elit nec nisl ultrices congue vitae in augue. Sed venenatis pretium risus. Integer tempus elementum ultricies. Ut id lobortis mauris. Integer tincidunt nunc nec bibendum accumsan. Cras semper hendrerit ex quis luctus. Mauris felis sem, mollis sed uma id, pharetra feugiat nunc.

Vivamus eu euismod augue. Sed eu velit placerat, rutrum velit sed, malesuada diam. Aliquam eu ipsum vestibulum non lacus eu aliquet. Duis eget sapien tristique, auctor lorem ut, volutpat mi. Nulla blandit ipsum imperdiet magna laoreet, ac mollis lorem tristique. Maecenas et velit felis. Cras nec fermentum ligula, bibendum commodo lacus. Interdum et malesuada fames ac ante ipsum primis in faucibus. Phasellus mollis suscipit molestie. Morbi imperdiet sem euismod lorem euismod vehicula. Duis in erat eu justo ornare pretium id et odio. Ut at arcu luctus, gravida odio eu, condimentum ligula. Maecenas orci tellus, vestibulum sit amet iaculis mattis, feugiat nec tellus. Ut dignissim, ligula id condimentum consequat, diam tortor ultricies arcu, nec tincidunt nibh massa sodales orci.

Donec vitae euismod libero. Etiam neque nunc, rhoncus ac tortor eget, gravida pellentesque magna. Etiam pulvinar hendrerit eros non elementum. Praesent convallis leo sagittis odio porttitor, at rutrum tortor scelerisque. Fusce a pharetra ligula, ac sollicitudin eros. Aliquam ultricies hendrerit turpis vitae tempus. Suspendisse potenti. Nunc dignissim nunc magna. Praesent interdum ex sit amet dui consectetur, id rutrum mauris fringilla. Pellentesque in ultricies dolor. Donec eu ipsum congue, viverra odio a, euismod lectus. Aliquam tempor tortor augue, maximus pellentesque turpis ullamcorper molestie. Proin velit elit, euismod id diam vel, congue scelerisque augue.

Mauris laoreet mauris eget sem gravida consequat. Fusce in ante lacus. Praesent non vehicula massa. Quisque blandit pulvinar vulputate. Praesent congue, augue vitae feugiat tincidunt, tortor sapien faucibus purus, egetas porta leo risus et nibh. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Nulla blandit lectus risus, sed pellentesque ante hendrerit vitae. Nulla nunc elit, interdum ut erat non, mattis bibendum magna. Nullam vitae metus nunc. Duis luctus mi a mauris vulputate, ut blandit mi congue. Ut molestie erat id convallis aliquet. Duis a metus orci. Donec vehicula justo a luctus porttitor. Phasellus ligula nibh, vehicula vel sapien vel, congue elementum velit. Vivamus vel felis eu dui elementum pulvinar sit amet vel turpis.

Aliquam convallis eu nisi non mattis. Maecenas magna sem, consectetur fringilla convallis efficitur, placerat in tellus. Etiam quis laoreet ex. Ut augue orci, condimentum ac justo quis, sollicitudin varius velit. Phasellus ultrices velit nulla, at tempor lectus pulvinar nec. In ut nunc nunc. Donec aliquam nisl eget consectetur venenatis. Integer consequat at orci nec viverra. Duis laoreet sagittis odio, eget maximus est dapibus in. Mauris consequat a dui et hendrerit.

- Primeiras fórmulas de inteligibilidade (~1920)
- Primeiros estudos no nível sentencial (~2010)
- **Objetivo:** analisar e avaliar as sentenças de um texto a partir de suas métricas, permitindo identificar seus pontos mais complexos.

<http://fw.nilc.icmc.usp.br:23380/simpligo-ranking>

Simple  Complex 



Avaliação da complexidade e suas métricas

Fórmulas clássicas

Flesch

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

Brunét

$$W = N^{V-0.165}$$

Gunning Fog Index

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

Honoré

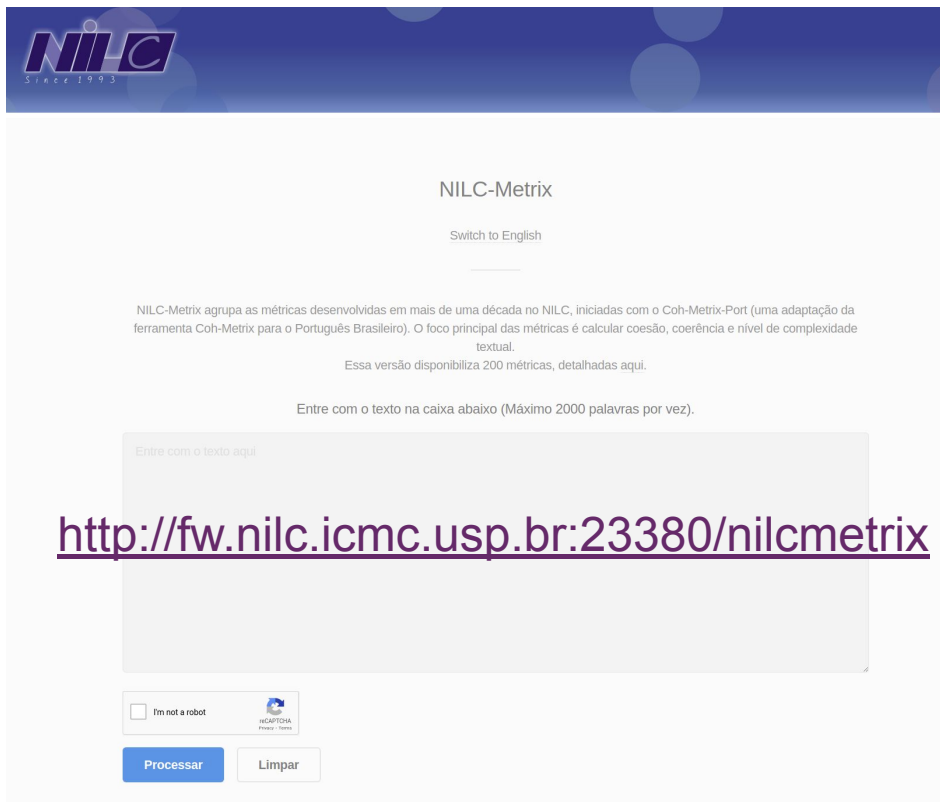
$$R = \frac{100 \log N}{1 - \frac{V_1}{V}}$$

Dale-Chall

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



Métricas Linguísticas



NILC
Since 1993

NILC-Metrix

[Switch to English](#)

NILC-Metrix agrupa as métricas desenvolvidas em mais de uma década no NILC, iniciadas com o Coh-Metrix-Port (uma adaptação da ferramenta Coh-Metrix para o Português Brasileiro). O foco principal das métricas é calcular coesão, coerência e nível de complexidade textual.

Essa versão disponibiliza 200 métricas, detalhadas aqui.

Entre com o texto na caixa abaixo (Máximo 2000 palavras por vez).

Entre com o texto aqui

☐ I'm not a robot 

Processar Limpar

<http://fw.nilc.icmc.usp.br:23380/nilcmatrix>

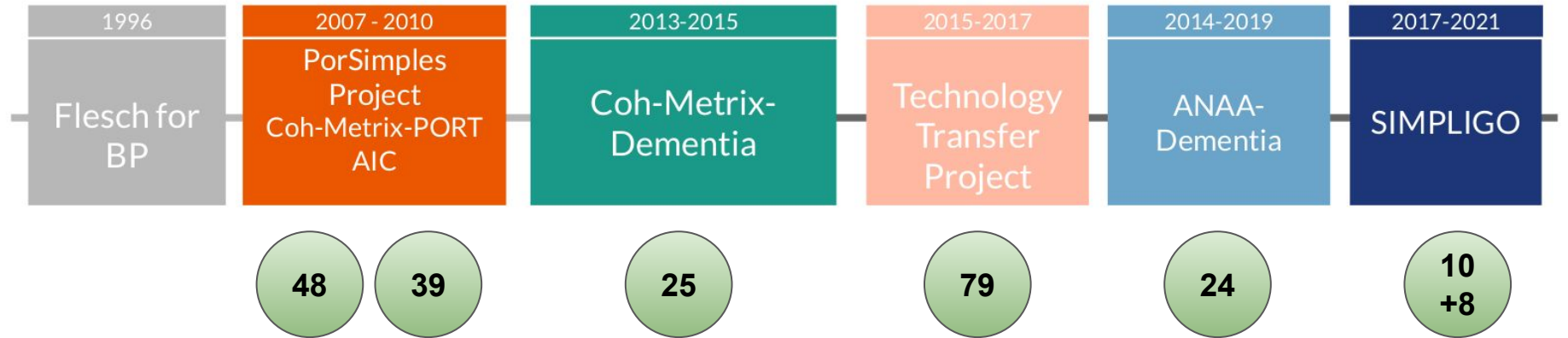


Métricas Linguísticas

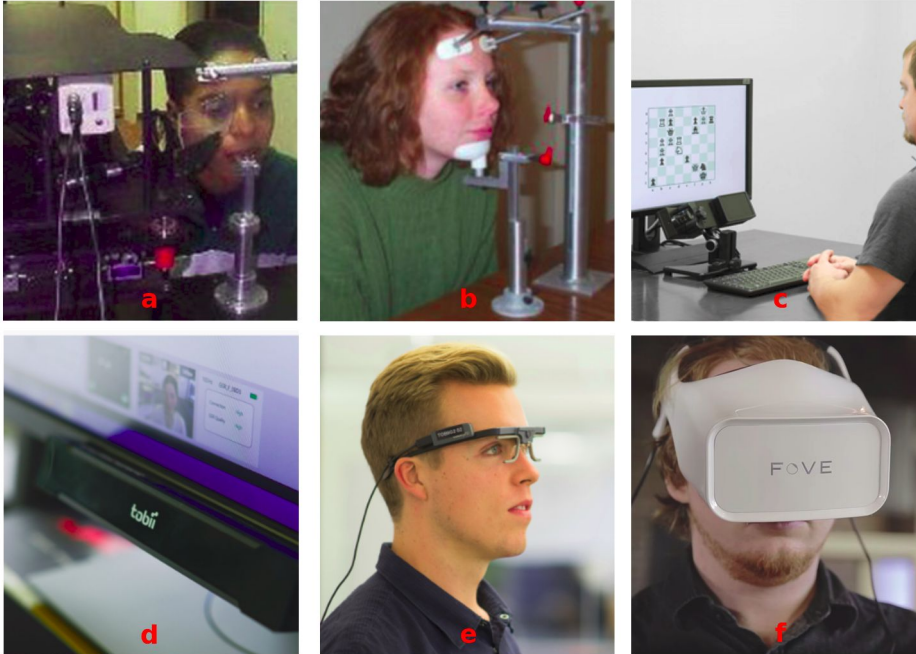
Grupo	Qtde	Descrição
Medidas Descritivas	10	Estatísticas básicas: número de palavras, sentenças, ...
Simplicidade Textual	8	Tamanhos das sentenças, proporção de palavras simples, ...
Coesão Referencial	9	Presença de elementos para cadeias de correferência.
LSA - Coesão Semântica	11	Sobreposição de palavras semanticamente relacionadas.
Medidas Psicolinguísticas	24	Idade de aquisição, concretudo, imageabilidade, familiaridade.
Diversidade Lexical	15	Type-token ratio (TTR).
Conectivos	12	Mede coesão pelas conexões entre as partes do texto.
Léxico Temporal	12	Tempos e modos verbais.
Complexidade Sintática	27	Métricas extraídas de parsers de dependência e constituintes.
Densidade de padrões sintáticos	4	Palavras por sintagma nominal, orações no gerúndio.
Informações Morfossintáticas de Palavras	42	Adjetivos, advérbios, verbos, substantivos, preposições, ...
Informações Semânticas de Palavras	11	Ambiguidade, positivo/negativo, hiperônimos, ...
Frequência de Palavras	10	Frequência BrWaC e Brasileiro, cruas e normalizadas.
Índices de Leiturabilidade	5	Fórmulas clássicas.



Métricas Linguísticas



Rastreamento Ocular



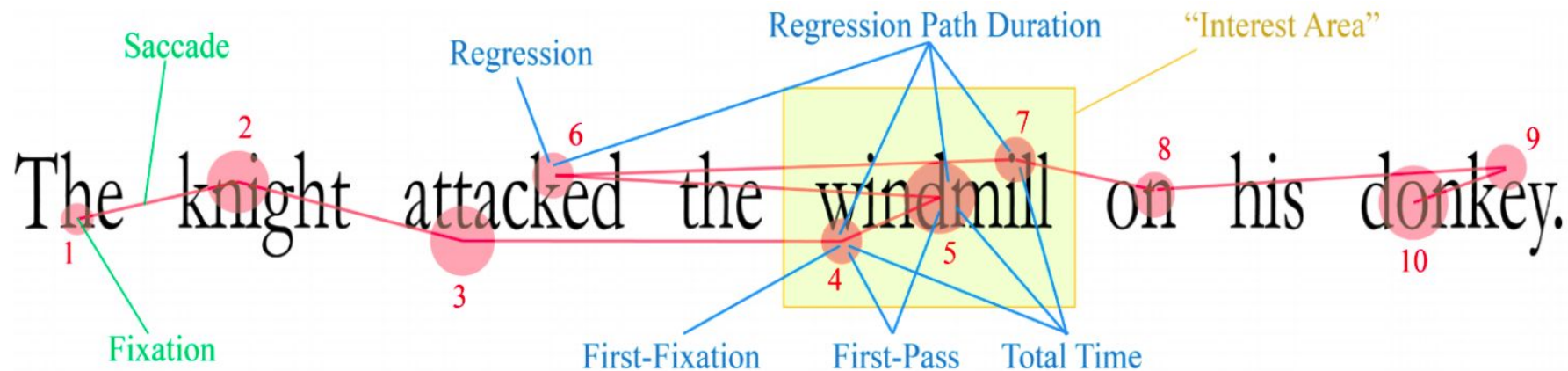
Ward Technical Consulting (2018), SR Research (2018), Imotions (2017), FOVE (2018)

Medidas extraídas a partir do monitoramento dos movimentos dos olhos durante a execução de uma tarefa (por ex: leitura).

Os movimentos dos olhos podem ser interpretados como uma janela para o processamento do cérebro, refletindo os tempos cognitivos envolvidos em determinada tarefa.

(BARRETT; AGIC; SØGAARD, 2015)

Rastreamento Ocular



(ZELENINA, 2015)

Recursos para o português brasileiro

Recursos para as tarefas de Complexidade

Corpus de Complexidade Textual para Estágios Escolares do Sistema Educacional

Brasileiro: Grande corpus com textos utilizados em diferentes etapas de ensino do Sistema Educacional Brasileiro

CorPop e PorPopular: Reúnem textos de diversas fontes, como jornais populares, livros para leitores de baixo letramento e publicações de pessoas em situação de rua, com o objetivo de representar a linguagem do brasileiro médio.

MedSimples: Ferramenta que simplifica textos sobre temas de saúde.



PorSimples

154 textos simplificados em dois níveis
(104 jornalísticos e 50 divulgação científica)

Original:

Uma parcela critica o uniforme, porque acredita que ele ameaçaria a individualidade de cada um.

Natural:

Alguns criticam o uniforme, porque acreditam que ele ameaça a individualidade de cada um.

Forte:

Alguns acreditam que o uniforme ameaça a individualidade de cada um.

The screenshot displays the PorSimples web application interface, which is designed for text simplification. It features a top navigation bar with tabs for 'Texto Natural', 'Sintático', 'Léxico', and 'Sintático'. The main content area is divided into two columns. The left column, titled 'Texto Natural', contains a list of simplification operations such as 'Não simplifica', 'Operações de reescrita', 'Fazer reescrita forte', 'Fazer reescrita simples', 'Reescrever em SVO', 'Mudança de voz', 'Operações de movimentação', 'Inverter ordem da sentença', 'Dividir sentença', 'Unir sentença', 'Operações de remoção', 'Remover sentença', and 'Remover parte da sentença'. The right column, titled 'Texto Simplificado', shows the result of applying these operations to a sample text. The sample text is about technological resources in cars and the benefits of DVD players. The simplified text is shown in a box on the right, with a 'Sintático' tab selected. Below the simplified text, there are buttons for 'Remover Operação' and 'Detalhar Operação'. The interface also includes a search bar and a list of operations on the right side.

Distribuição das sentenças originais de acordo com o nível de simplificação

Application of Simplification	Number of Sentences
Operations in PorSimples Sentences	
NOT simplified in any level	372
Simplified in two levels	1141
Simplified only in Natural Level	1289
Simplified only in Strong Level	178
TOTAL	2980

Distribuição dos pares por nível nas três versões do cópús

Types of Pairs	PorSimplesSent1	PorSimplesSent2	PorSimplesSent3
Original-Natural	3535	2372	1543
Natural-Strong	4976	1501	782
Original-Strong	2105	1095	275
TOTAL	10616	4968	2600

<http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources>

RastrOS



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

Um grande corp us com medidas de RASTReamento Ocular e normas de previsibilidade durante a leitura de estudantes do ensino Superior do Brasil



PONTIF CIA UNIVERSIDADE CAT LICA DO RIO DE JANEIRO



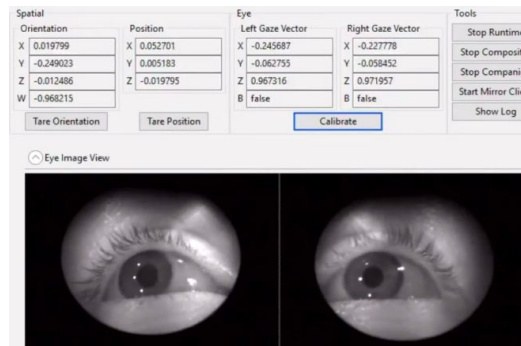
- 50 par grafos curtos ( g neros jornal stico, divulga  o cient fica e liter rio)
- Coleta de dados em 6 universidades
- **37 participantes no estudo do rastreamento ocular**
- **414 participantes no teste Cloze**

Par grafo 1 de 5 - Palavra 5 de 40.

(Digite a pr xima palavra e depois tecla enter, ou clique no bot o.)

A chuva  cida  

Pr xima



Recursos [corpora] para o portugu s brasileiro



Grandes Modelos de linguagem

Prompt

Simplifique o seguinte texto para que um aluno do quarto ano do ensino fundamental consiga compreender, após a simplificação, forneça passo a passo os detalhes das mudanças e motivos para fazer as adaptações no texto:

O ano era 1978. As salas de cinema de todo o mundo exibiam uma produção do diretor Joe Dante em que um cardume de piranhas escapava de um laboratório militar e atacava participantes de um festival aquático.

Quase 30 anos depois, banhistas assustados estão se afastando do principal balneário de Uruguaiana, na Fronteira Oeste.

Mais de 20 pessoas foram mordidas por palometas (*Serrasalmus spilopleura*, espécie de piranha) que vivem nas águas da barragem Sanchuri, na margem da Br-472, a 40 quilômetros da cidade.

- Os ataques se tornaram mais freqüentes. Por isso, aconselhamos mais cautela- diz o subprefeito Nei Pinto. As mordidas em pés e canelas de banhistas não são novidade. O que chamou a atenção das autoridades foi o aumento no número de ataques em relação aos outros anos.



Em 1978, os cinemas exibiam um filme do diretor Joe Dante, onde piranhas escapavam de um laboratório e atacavam pessoas em um festival aquático.

Quase 30 anos depois, em Uruguaiana, uma cidade na Fronteira Oeste, banhistas ficaram com medo e saíram da água. Mais de 20 pessoas foram mordidas por palometas, um tipo de peixe parecido com a piranha, que vive na barragem Sanchuri, a 40 quilômetros da cidade.

- Os ataques estão acontecendo com mais frequência. Por isso, pedimos mais cuidado - disse o subprefeito Nei Pinto. As mordidas em pés e pernas já aconteciam antes, mas agora estão acontecendo mais vezes do que nos outros anos.

Passo a passo das mudanças:

1. Remoção de detalhes desnecessários para o público-alvo:
 - No texto original, há muitas informações específicas (como o nome científico da palometa, "Serrasalmus spilopleura", e o nome da rodovia "BR-472") que não são essenciais para uma criança. Retirei essas informações para deixar o texto mais claro.
1. Uso de palavras mais simples:
 - Troquei "salas de cinema de todo o mundo exibiam uma produção" por "os cinemas exibiam um filme", porque é uma forma mais simples e direta de dizer a mesma coisa.
 - "Participantes de um festival aquático" virou "pessoas em um festival aquático", já que "participantes" é um termo mais formal.
 - "Balneário" foi trocado por "água", pois balneário é uma palavra mais difícil e menos conhecida por crianças dessa idade.
2. Redução de frases complexas:
 - No texto original, há frases longas e com muitas informações, como "Mais de 20 pessoas foram mordidas por palometas (Serrasalmus spilopleura, espécie de piranha) que vivem nas águas da barragem Sanchuri, na margem da Br-472, a 40 quilômetros da cidade". Simplifiquei isso dividindo a frase e removendo partes menos essenciais.
1. Explicações curtas e claras:
 - Mantive o foco em uma narrativa direta, que é mais acessível para crianças, sempre mantendo a ideia original do texto, mas de forma fácil de entender.



GPT-4o1-preview

Em 1978, os cinemas mostravam um filme do diretor Joe Dante. Nesse filme, piranhas escapavam de um laboratório militar e atacavam pessoas em um festival na água.

Quase 30 anos depois, em Uruguaiana, uma cidade na Fronteira Oeste, pessoas que nadavam ficaram com medo e saíram da principal área de banho.

Mais de 20 pessoas foram mordidas por palometas, que são peixes parecidos com piranhas. Elas vivem nas águas da barragem Sanchuri, a 40 quilômetros da cidade.

“Os ataques estão acontecendo mais vezes. Por isso, pedimos mais cuidado”, disse o subprefeito Nei Pinto. Mordidas em pés e pernas já aconteciam antes, mas o que chamou a atenção das autoridades foi o aumento no número de ataques em comparação com outros anos.



GPT-4o1-preview

Passo a passo das mudanças e motivos:

Vocabulário Simplificado:

- "As salas de cinema de todo o mundo exibiam uma produção" foi alterado para "os cinemas mostravam um filme" para usar palavras mais simples e comuns.
- "Atacava participantes de um festival aquático" mudou para "atacavam pessoas em um festival na água", facilitando a compreensão.

Explicação de Termos Difíceis:

- "Palometas (*Serrasalmus spilopleura*, espécie de piranha)" foi simplificado para "palometas, que são peixes parecidos com piranhas", ajudando o aluno a entender do que se trata sem termos científicos complicados.

Remoção de Detalhes Desnecessários:

- "Na margem da BR-472" foi removido porque o nome da rodovia não é essencial para entender a história e pode confundir o aluno.
- "A 40 quilômetros da cidade" foi mantido para dar uma noção de distância, mas sem sobrecarregar com detalhes.

Frases Mais Curtas e Diretas:

- Frases longas foram divididas em sentenças menores para facilitar a leitura e compreensão.
- Por exemplo, a frase sobre as pessoas sendo mordidas foi separada para destacar a informação importante.

Uso de Palavras Familiares:

- "Banhistas assustados estão se afastando do principal balneário" foi alterado para "pessoas que nadavam ficaram com medo e saíram da principal área de banho", usando palavras que o aluno provavelmente já conhece.

Citação Simplificada:

- A fala do subprefeito foi mantida, mas ajustada para linguagem mais simples: "Os ataques estão acontecendo mais vezes. Por isso, pedimos mais cuidado".

Explicitação de Conceitos:

- Esclareci que "mordidas em pés e pernas já aconteciam antes", para contextualizar a situação de forma clara.



Gemini 1.5-pro

Texto Simplificado:

Era 1978 e nos cinemas do mundo todo, as pessoas assistiam a um filme onde um monte de piranhas escapavam de um laboratório e atacavam as pessoas em um festival na água.

Agora, quase 30 anos depois, na cidade de Uruguaiana, na fronteira com a Argentina, as pessoas estão com medo de entrar na água.

Mais de 20 pessoas foram mordidas por palometas, que são um tipo de piranha, que vivem em um rio chamado Sanchuri. Esse rio fica perto da cidade, a 40 quilômetros de distância.

O subprefeito Nei Pinto disse que as mordidas estão acontecendo com mais frequência. Por isso, ele pediu para as pessoas tomarem mais cuidado ao entrar na água.

As pessoas já tinham sido mordidas por palometas antes, mas este ano está tendo muito mais gente sendo mordida.



Gemini 1.5-pro

Passos e Motivos das Adaptações:

1. **Frases curtas e simples:** O texto original usa frases longas e complexas. Para facilitar a leitura de um aluno do quarto ano, as frases foram divididas em partes menores e mais simples. Exemplo: "O ano era 1978. As salas de cinema de todo o mundo exibiam uma produção do diretor Joe Dante em que um cardume de piranhas escapava de um laboratório militar e atacava participantes de um festival aquático." foi transformado em "Era 1978 e nos cinemas do mundo todo, as pessoas assistiam a um filme onde um monte de piranhas escapavam de um laboratório e atacavam as pessoas em um festival na água."
1. **Vocabulário simplificado:** Palavras como "exibiam", "produção", "cardume", "laboratório militar" foram substituídas por termos mais simples como "assistiam", "filme", "monte", "laboratório".
1. **Explicação de termos desconhecidos:** "Palometas" foi explicado como "um tipo de piranha" para melhor compreensão.
1. **Informações essenciais:** O texto original continha informações como o nome científico da palometa. Essas informações foram retiradas, pois não são relevantes para a compreensão do aluno do quarto ano.
1. **Organização em parágrafos:** O texto foi dividido em parágrafos para facilitar a leitura e compreensão.
1. **Uso de linguagem informal:** A linguagem foi adaptada para um tom mais informal, com frases como "O subprefeito Nei Pinto disse que as mordidas estão acontecendo com mais frequência", em vez de "Os ataques se tornaram mais frequentes".
1. **Foco na informação principal:** O texto original focava em diversos aspectos, como o filme de 1978. O texto simplificado foca na informação principal, que é o aumento de ataques de palometas em Uruguiana.



GPT-4o1-preview

ChatGPT o1-preview ▾

 Compartilhar 

simplifique a sentença: "o jovem mancebo golpeou fortemente o brinquedo esférico com a destra inferior."



Pensou por alguns segundos ^

Analisando termos complexos
Estou estabelecendo o significado de "destra inferior" como "pé direito", já que "destra" geralmente se refere à mão direita.

O rapaz chutou a bola.

1º Ciclo de Encontros: 4xPLN

Outubro, 2024

09/10, 10h: Mariza Ferro
Ética em IA e PLN

14/10, 14h: Sidney Leal
Complexidade Textual

24/10, 17h: Elisa Terumi
PLN na Saúde

28/10, 14h: Renata Vieira,
Fernanda Olival e
Helena Cameron
PLN e Humanidades Digitais

EVENTOS VIRTUAIS

Para participar, acesse:
meet.google.com/puj-fxrh-vfd

<https://brasileiraspln.com/livro-pln>

<https://brasileiraspln.com/livro-pln/2a-edicao/parte-aplicacoes/cap-complexidade-textual/cap-complexidade-textual.html>



Referências

Caseli, H. M.; Nunes, M. G. V. Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português. 2a. edição. São Carlos: Brasileiras em Processamento de Linguagem Natural, 2024. Disponível em: brasileiraspln.com/livro-pln

DUBAY, W. H. Smart Language: Readers, Readability, and the Grading of Text. Costa Mesa, CA: Impact Information, 2007.

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. Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular. tese de doutorado—[s.l.] Universidade de São Paulo, 2021.

ZELENINA, M. "Eye Tracking for NLP". 2015. Available at: <https://www.slideshare.net/mariezelenina/presentation-2-47610828>.