Atividades unidade 2

TERMINOLOGIA E CONCEITOS

TC.2.1. Sob o ponto de vista linguístico, qual a diferença entre corpus de texto, parágrafo, frase, oração e palavra? Ilustre com um exemplo e indique como diferentes tipos de significado podem estar atrelados a cada um desses elementos.

A diferença está em como cada um desses elementos é analisado, organizado e usado na linguagem. Começando pela **palavra**, ela é a menor unidade que possui significado dentro da linguagem, situada entre dois espaços em branco. Exemplo: Manga. Diferentes significados podem estar atrelados a essa palavra, como à fruta ou à parte de uma blusa, dependendo do contexto.

Oração é uma unidade que contém no mínimo um verbo que transmite ação ou estado. Exemplo: Ela comeu manga. Onde o significado é da ação de comer. Já em "Ela usou a manga da blusa," o significado muda.

A **frase** expressa um sentido completo e pode ter mais de uma oração.
Exemplo: A manga estava madura e ela a usou para fazer uma sobremesa. A frase tem significados relacionados a frutas e ação culinária.

O **parágrafo** é um conjunto de frases que desenvolvem uma ideia central.
Exemplo: Um parágrafo descrevendo uma festa de verão pode mencionar a "manga" como parte de uma bebida refrescante e, em outro momento, mencionar a "manga" como parte de um traje. Os significados variam conforme a ideia principal do parágrafo.

O **corpus de texto** é um conjunto de dados linguísticos de uma língua, usado para estudo, análise e que pode ser utilizado pelo computador. Exemplo: Um corpus pode conter várias obras que usam "manga" em diferentes contextos, em diferentes gêneros (contos, crítica, crônica, poesia, romance).

PRÁTICA DE PROGRAMAÇÃO

PP.2.8. Exemplifique o funcionamento de um corretor ortográfico, aplicável à língua portuguesa, que efetue correção de palavras baseado em um corpus de texto considerado como referência e que utilize métricas de distância e estatísticas de ocorrência de palavras no corpus considerado.
Alterar o corpus pode afetar o comportamento do corretor? Se sim, dê um exemplo prático utilizando dados diferentes para o corpus.
A resposta a esse problema deverá ser um programa que:
a) Leia uma frase digitada pelo usuário.
b) Verifique se há ou não palavras potencialmente incorretas.
c) Informe ao usuário a frase potencialmente corrigida ou então diga que a frase aparenta estar correta.

O corretor ortográfico implementado no código lê o corpus dos contos de Machado de Assis, para criar um dicionário com as palavras únicas e suas frequências. Faz o processamento das palavras deixando-as em minúsculas, faz a tokenização de palavras ao dividir o texto em tokens (palavras), remove caracteres especiais, mantém letras e números, mas se o caractere não for alfanumérico nem espaço, ele é substituído por um espaço.

Quando o usuário insere uma frase, cada palavra é comparada com as palavras do corpus. Se a palavra não for encontrada, o programa calcula a distância de edição (Levenshtein) para sugerir correções com as palavras mais frequentes.

A alteração do corpus pode afetar o comportamento do corretor ortográfico porque as palavras consideradas corretas e suas frequências mudam com diferentes corpora. Nos corpora de Machado de Assis e da Bíblia, a palavra "digitar" não está presente, já que esses textos pertencem a épocas e contextos onde esse termo não era utilizado, resultando sempre na correção de "digitar" para outra palavra.

Se considerarmos a frase "Ele foi ao mercad para comprr maças e laranjas", o comportamento do corretor será diferente dependendo do corpus. Em ambos os casos, o corretor sugere como incorreta, no corpus de Machado de Assis dará a

possível correção como "ele foi ao mercar para comprar moças e laranjas", enquanto na Bíblia "ele foi ao mercado para compra maras e franjas".

Da mesma forma, há sugestões de correções iguais para a frase “O homen caminhou pelá rua”, que será corrigida para “o homem caminhou pela rua”. E para a frase “o hoemem deve sejuir os mandamenttos de deus”, a correção será “o homem deve seguir os mandamentos de Deus”.