



EACH

Escola de Artes, Ciências e Humanidades
da Universidade de São Paulo

Relatório Técnico PPgSI

Análise Estatística entre Corpora

Norton Trevisan Roman
norton@usp.br

Marcos Martins de Oliveira Pacheco
mpachecooliveira@usp.br

Outubro - 2024

Série de Relatórios Técnicos

PPgSI-EACH-USP
Rua Arlindo Béttio, 1000 – Ermelino Matarazzo
03828-000 – São Paulo, SP
TEL: (11) 3091-8197 <http://www.each.usp.br/ppgsi>

Resumo

Apresenta-se, neste documento, um projeto na área de Processamento de Língua Natural, ramo de pesquisa da Inteligência Artificial. De forma a subsidiar as pesquisas para o processamento computacional do Português, propõe-se uma comparação da análise estatística da anotação morfossintática atualmente presente em dois corpus de tweets do mercado financeiro em português do Brasil, sendo um anotado de forma manual, e outro de forma automática, conforme o modelo de representação Universal Dependencies. A mesma análise foi conduzida em um terceiro corpus, com um escopo mais amplo, também focado no mercado financeiro.

1 Introdução

O Processamento de Língua Natural (PLN) é o ramo da Inteligência Artificial que visa a habilitar os computadores a lidar com a língua humana (Jurafsky e Martin [2008]), estudando as relações entre a linguística e a informática, de modo a tornar possível a construção de sistemas capazes de reconhecer e produzir informação apresentada em uma língua natural (Vieira e Lima [2001]). Prover a inteligência linguística para a máquina a torna capaz de realizar tarefas rotineiras que um ser humano realiza, como tradução, sumarização, reconhecimento e produção de fala, entre muitas outras, sendo de grande apoio para as tarefas diárias e também para outras frentes de pesquisa, como mineração de textos, recuperação de informação e ciência de dados, por exemplo.

A área tem avançado significativamente na última década. Esse avanço é explicado em função principalmente de duas razões. Por um lado, as redes neurais artificiais, com seus modelos profundos (Goodfellow et al. [2016]), bem como técnicas recentes, como a modelagem de atenção (Vaswani et al. [2017]), têm revolucionado toda a área de Inteligência Artificial, apoiadas pelos avanços no poder computacional e pela grande quantidade de dados disponíveis. Por outro lado, os modelos distribucionais, responsáveis pelos vetores/embeddings de palavras, treinados a partir de dados originalmente não rotulados, têm permitido que se alcancem resultados melhores em diversas tarefas, tendo como marcos os modelos word2vec e BERT, já disponíveis para o português.

Apesar do progresso recente significativo, ainda há um longo caminho para que muitas tarefas de PLN atinjam um dia resultados comparáveis ao que o humano produz. Ainda se mostram necessários criar recursos e investigar e definir modelagens e métodos linguístico-computacionais mais robustos e apropriados. Evidência disso é a iniciativa do projeto Universal Dependencies (ou UD, simplesmente), por exemplo, que busca desenvolver um modelo de representação sintática “universal”, capaz de ser utilizado por diferentes línguas, representando suas características importantes e possibilitando a investigação de métodos de uso geral para processamento linguístico e estudos de língua cruzada, assim como produzir e disponibilizar corpora anotados para o treinamento de sistemas de análise linguística, como etiquetadores morfossintáticos (taggers) e analisadores sintáticos (parsers).

Para além da anotação linguística de corpora, outro desafio bastante atual trata da criação de métodos e recursos para a identificação de outros fenômenos presentes na língua, de modo a fornecer uma visão mais abrangente do contexto, quando da execução de certos tipos de análise, como é o caso da identificação de padrões comuns a diversos gêneros literários ou domínios, ou mesmo pontos característicos de divergência, no que diz respeito a como informação linguística e, mais especificamente sintática, é utilizada. Nesse sentido, uma análise comparativa poderia fornecer informações de grande valia quando da definição de métodos e ferramentas que melhor se adequassem a essas variáveis.

Nesse contexto, o presente projeto de pesquisa visou a caracterização de três corpora já anotados com etiquetas morfossintáticas conforme o modelo UD (Lopes et al. [2022]), com relação à distribuição dessas etiquetas ao longo do corpus. Esses corpora são todos escritos em português do Brasil. Após isso, foram analisados os padrões e correlações/associações existentes entre a distribuição dessas etiquetas nos corpora, buscando assim identificar possíveis semelhanças e diferenças entre eles.

2 Sobre os Corpora

Abaixo está contido algumas informações importantes sobre os corpora utilizados no trabalho.

Os corpus DanteStocks Large e DanteStocks Automático foram etiquetados automaticamente com o Porttager.

huggingface.co/spaces/Emanuel/porttagger

2.1 DanteStocks Manual

Descrição: Corpus com 4048 Tweets do mercado financeiro que foram anotados manualmente.

Onde encontrar o corpus: [DanteStocks](#)

Versão utilizada: DANTEStocks (V2.1)

2.2 DanteStocks Automático

Descrição: Corpus com 4048 Tweets do mercado financeiro que foram extraídos e anotados automaticamente do DantesStocks anotado manualmente.

Onde encontrar o corpus: [DanteStocks Automático](#)

Versão utilizada: DANTEStocks (V2.1)

2.3 DanteStocks Large

Descrição: Corpus com 126.579 Tweets do mercado financeiro que foram anotados automaticamente com informações morfológicas e morfossintáticas.

Onde encontrar o corpus: [DanteStocks Large](#)

Versão utilizada: DanteStocks Large (10aug2022)

3 Análise de Frequência das Tags POS

As etiquetas POS (Part-of-Speech) são rótulos que identificam as características linguísticas de cada palavra em um texto. Esta análise compara a distribuição dessas classes gramaticais entre os corpora.

Índice de Tags POS

As principais tags utilizadas na análise são:

- **ADJ** - Adjetivos
- **ADP** - Preposições
- **ADV** - Advérbios
- **AUX** - Verbos Auxiliares
- **CCONJ** - Conjunções Coordenativas
- **DET** - Determinantes
- **INTJ** - Interjeições
- **NOUN** - Substantivos
- **NUM** - Numerais
- **PART** - Partículas
- **PRON** - Pronomes
- **PROPN** - Nomes Próprios
- **PUNCT** - Pontuação
- **SCONJ** - Conjunções Subordinativas
- **SYM** - Símbolos
- **VERB** - Verbos
- **X** - Outros
- **_** - Não Classificado

Tag	DanteStocks Manual	DanteStocks Automatizado	DanteStocks Large
ADJ	2,917	2,223	136,684
ADP	8,755	5,805	482,554
ADV	2,686	2,303	91,867
AUX	1,318	1,048	43,186
CCONJ	1,696	1,315	71,888
DET	6,726	3,946	326,737
INTJ	146	112	1,526
NOUN	12,001	8,533	621,879
NUM	5,021	4,139	221,990
PART	0	0	1
PRON	1,297	879	44,279
PROPN	11,762	11,131	544,737
PUNCT	13,124	12,578	540,876
SCONJ	754	598	16,078
SYM	4,456	19,502	197,719
VERB	6,583	5,029	277,381
X	1,755	1,413	114,880
	3,420	3,561	198,342

Tabela 1: Frequência das Tags POS por Corpus

Tag POS	% DanteStocks Manual	% DanteStocks Automatizado	% DanteStocks Large
ADJ	3.46%	2.64%	3.48%
ADP	10.37%	6.90%	12.27%
ADV	3.18%	2.74%	2.34%
AUX	1.56%	1.25%	1.10%
CCONJ	2.01%	1.56%	1.83%
DET	7.97%	4.69%	8.31%
INTJ	0.17%	0.13%	0.04%
NOUN	14.22%	10.14%	15.81%
NUM	5.95%	4.92%	5.64%
PART	0.00%	0.00%	0.00%
PRON	1.54%	1.04%	1.13%
PROPN	13.93%	13.23%	13.85%
PUNCT	15.55%	14.95%	13.75%
SCONJ	0.89%	0.71%	0.41%
SYM	5.28%	23.18%	5.03%
VERB	7.80%	5.98%	7.05%
X	2.08%	1.68%	2.92%
	4.05%	4.23%	5.04%

Tabela 2: Porcentagem das Tags POS

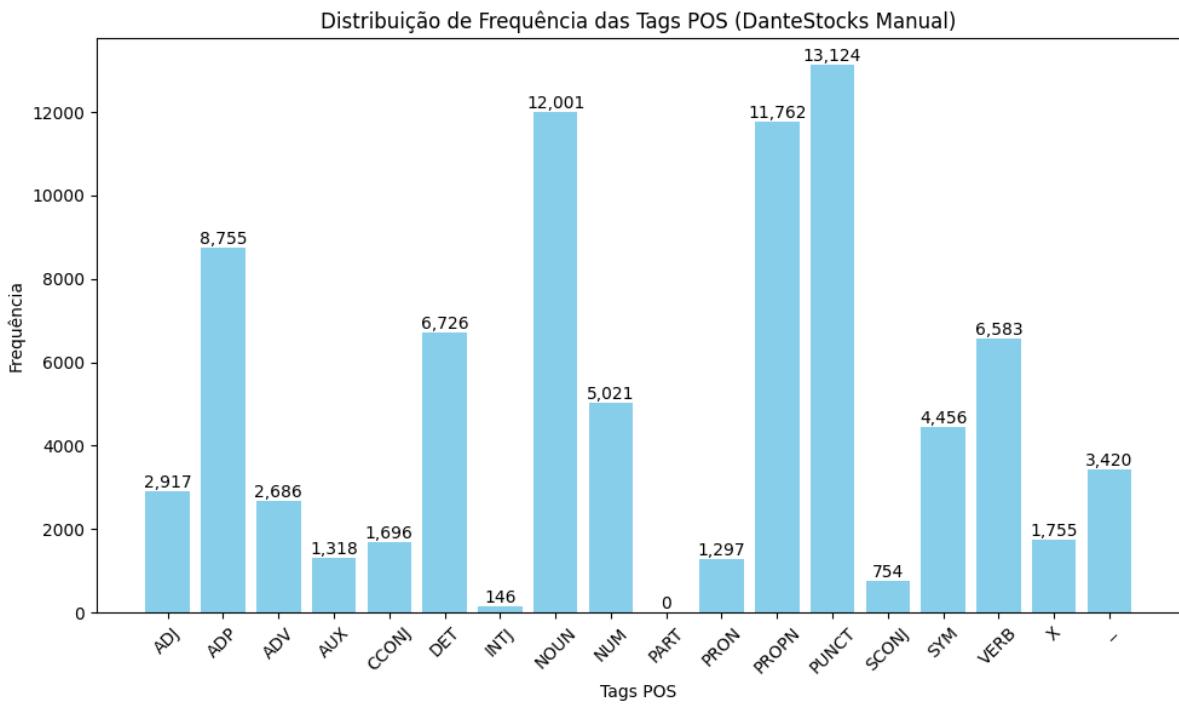


Figura 1: Distribuição de Frequência das Tags POS (DanteStocks Manual)

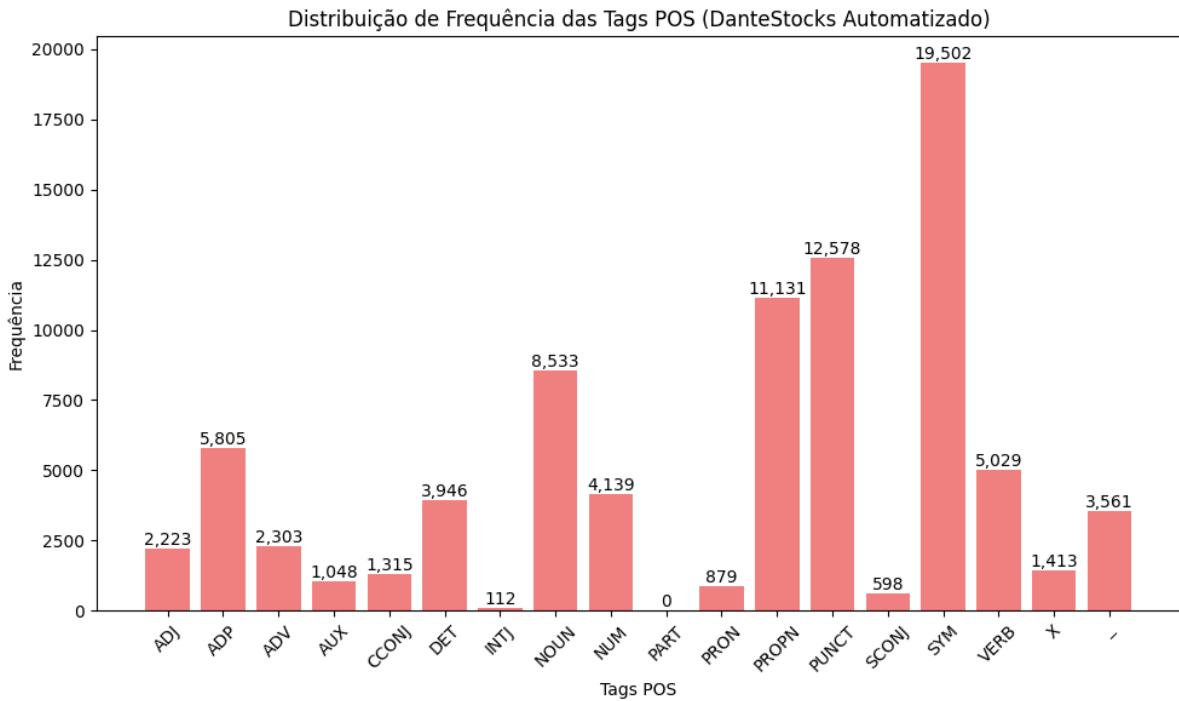


Figura 2: Distribuição de Frequência das Tags POS (DanteStocks Automatizado)

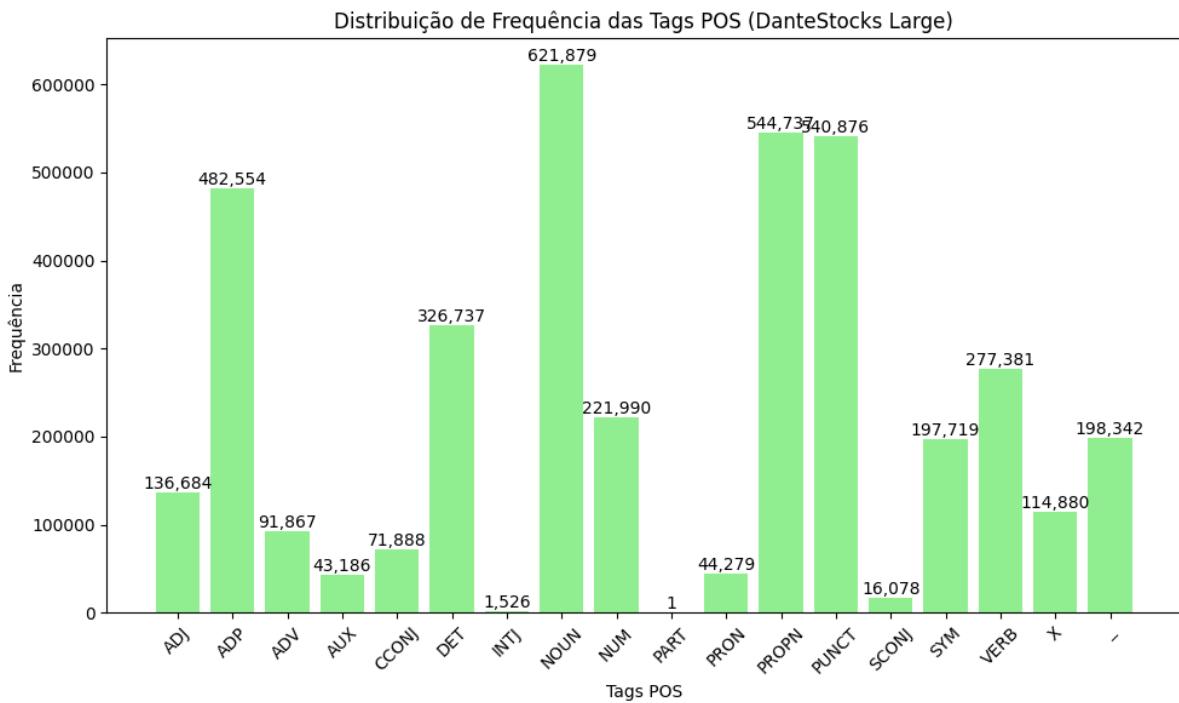


Figura 3: Distribuição de Frequência das Tags POS (DanteStocks Large)

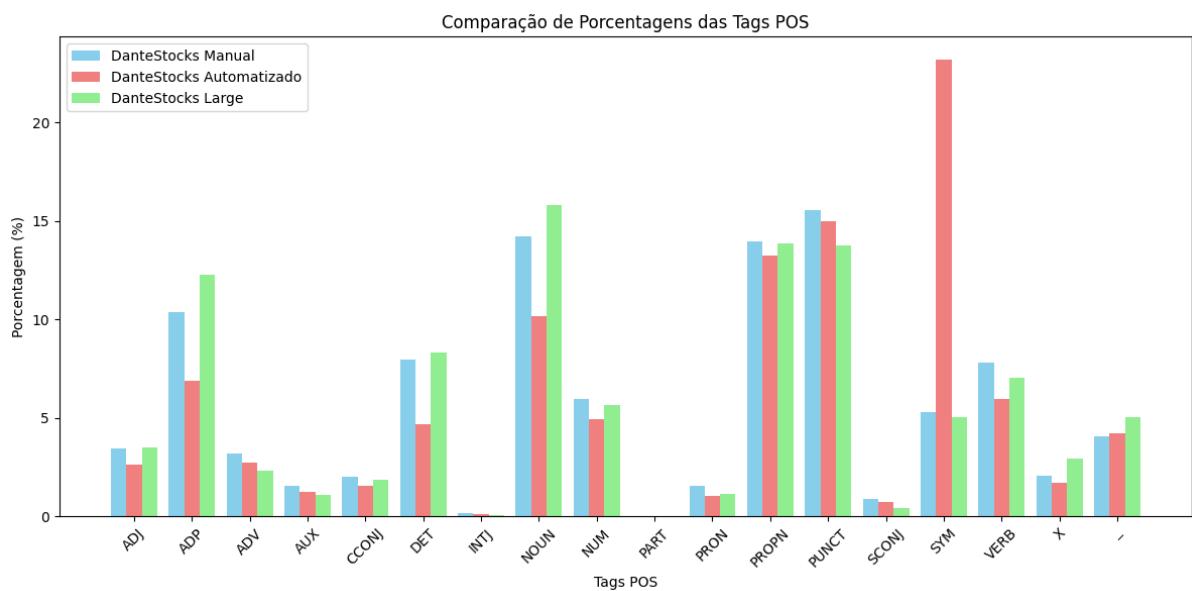


Figura 4: Comparação de Porcentagens das Tags POS

4 Análise de Comprimento dos Tokens

O comprimento dos tokens é uma medida importante na análise linguística, pois nos permite entender a distribuição do tamanho das palavras em um corpus. Esta análise pode revelar padrões no uso de palavras curtas versus longas, além de ajudar a identificar possíveis diferenças estilísticas entre os corpora.

Métrica	DanteStocks Manual	DanteStocks Automatizado	DanteStocks Large
Comprimento Médio	4.31	4.33	4.01
Mediana	3.00	3.00	3.00
Desvio Padrão	3.88	3.92	3.09

Tabela 3: Comparação do Comprimento dos Tokens

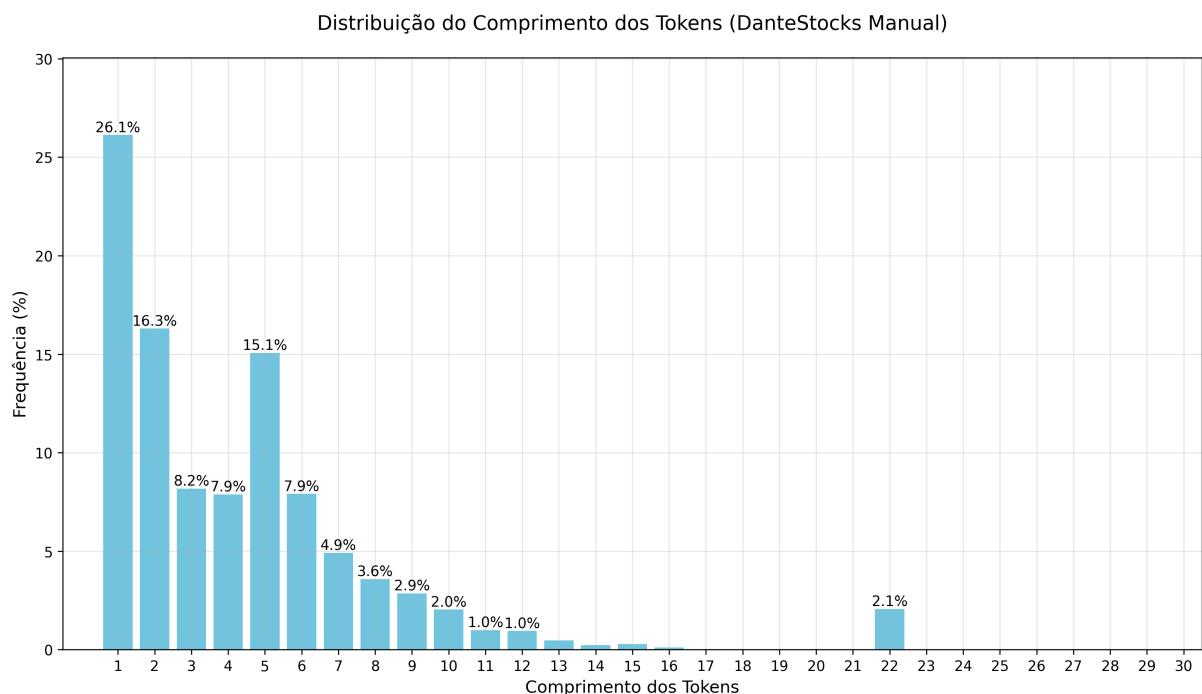


Figura 5: Distribuição de Comprimento dos Tokens (DanteStocks Manual)

Distribuição do Comprimento dos Tokens (DanteStocks Automatizado)

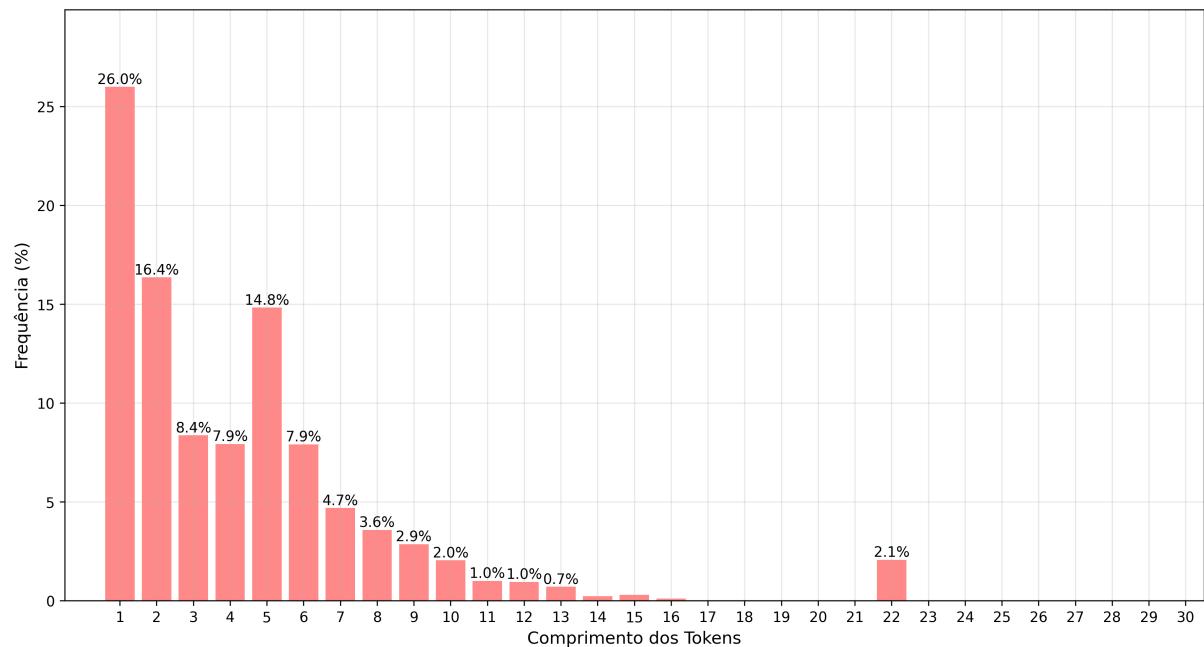


Figura 6: Distribuição de Comprimento dos Tokens (DanteStocks Automatizado)

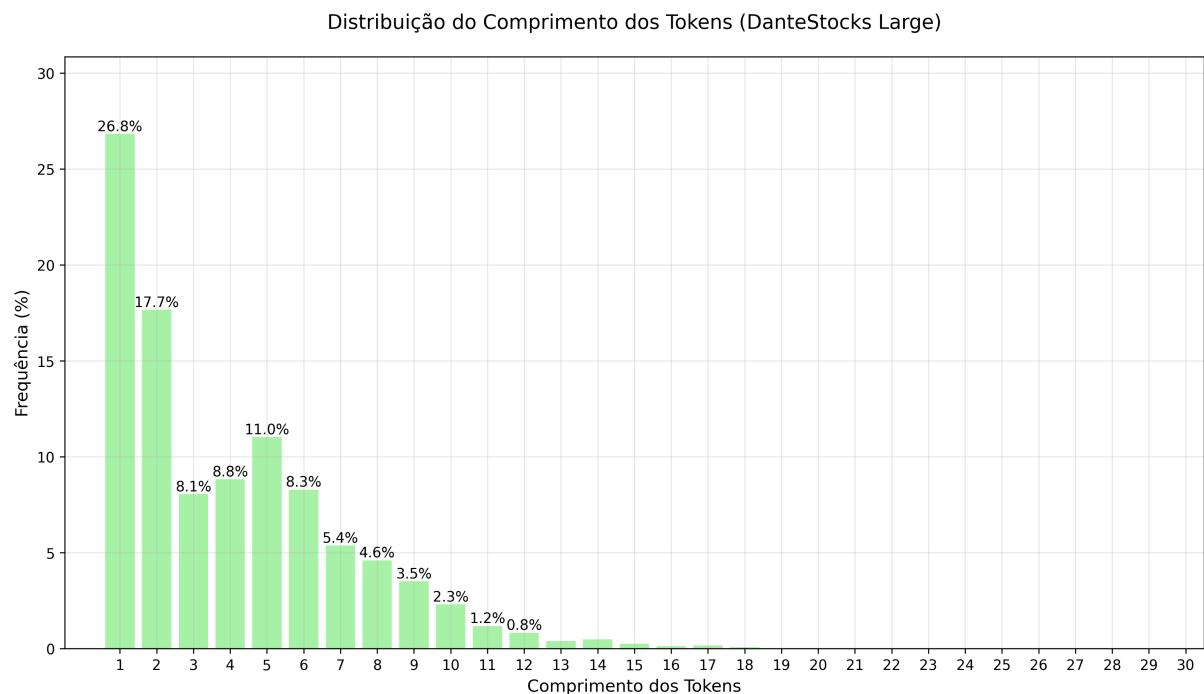


Figura 7: Distribuição de Comprimento dos Tokens (DanteStocks Large)

Comparação de Comprimento dos Tokens

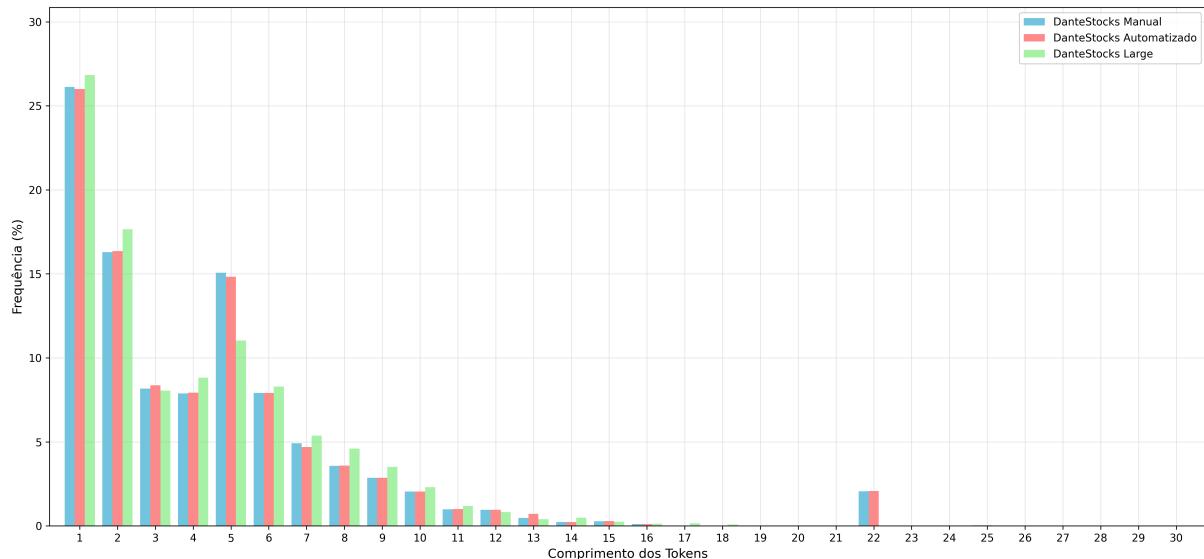


Figura 8: Comparação de Comprimento dos Tokens

5 Análise de Tags por Sentença

Esta seção apresenta uma análise da distribuição do número de tags por sentença em cada corpus, permitindo comparar a complexidade estrutural das sentenças.

Métrica	DanteStocks Manual	DanteStocks Automatizado	DanteStocks Large
Total de Sentenças	4,042	4,042	126,653
Média	20.88	20.81	31.05
Mediana	21.00	21.00	25.00
Desvio Padrão	8.29	8.24	19.65
Primeiro Quartil (Q1)	15.00	15.00	18.00
Terceiro Quartil (Q3)	27.00	27.00	41.00

Tabela 4: Estatísticas do Número de Tags por Sentença

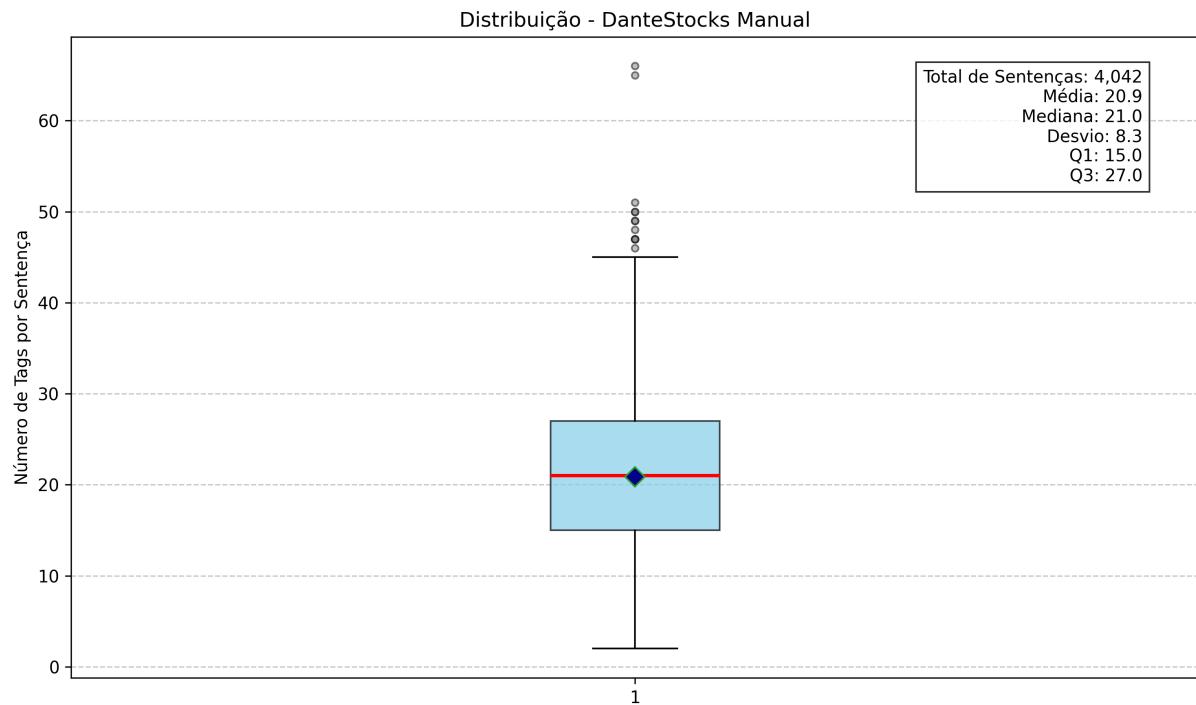


Figura 9: Distribuição do Número de Tags por Sentença - DanteStocks Manual

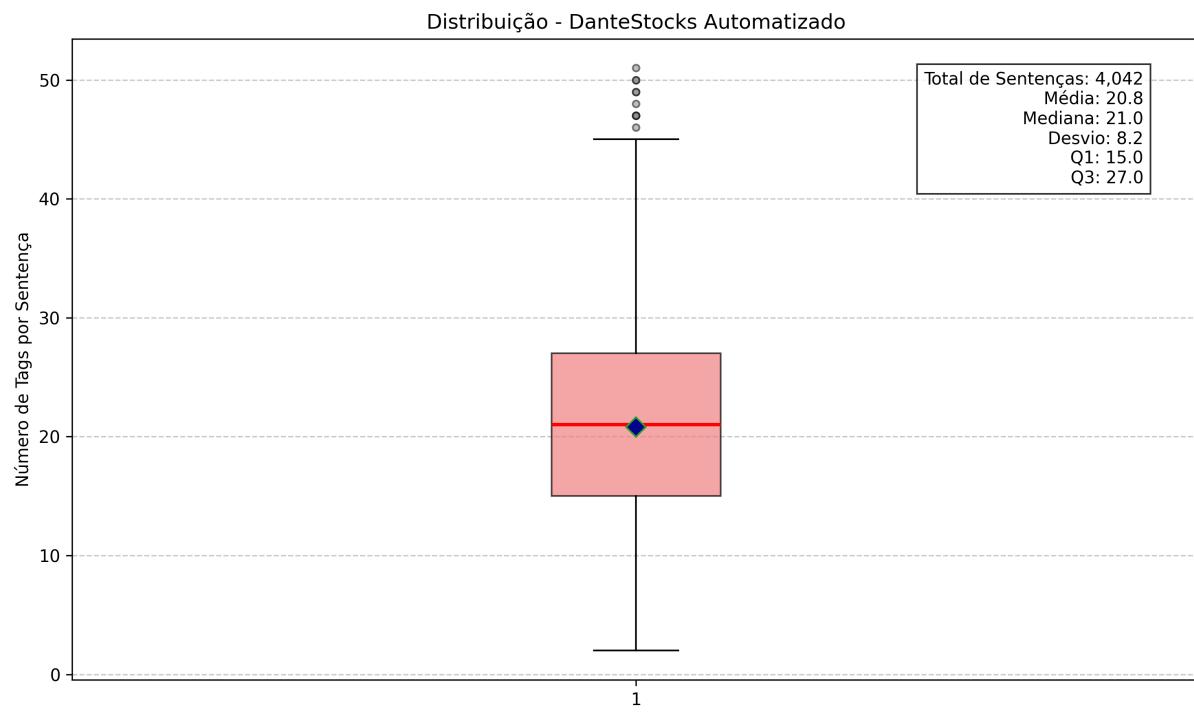


Figura 10: Distribuição do Número de Tags por Sentença - DanteStocks Automatizado

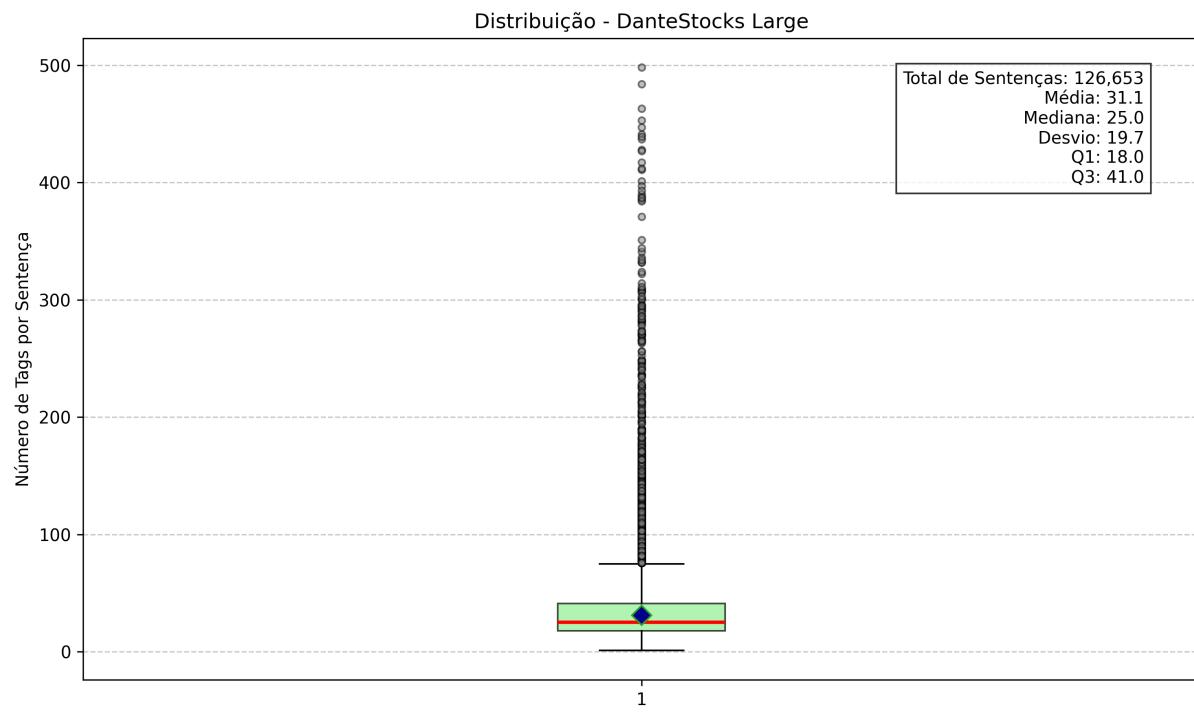


Figura 11: Distribuição do Número de Tags por Sentença - DanteStocks Large

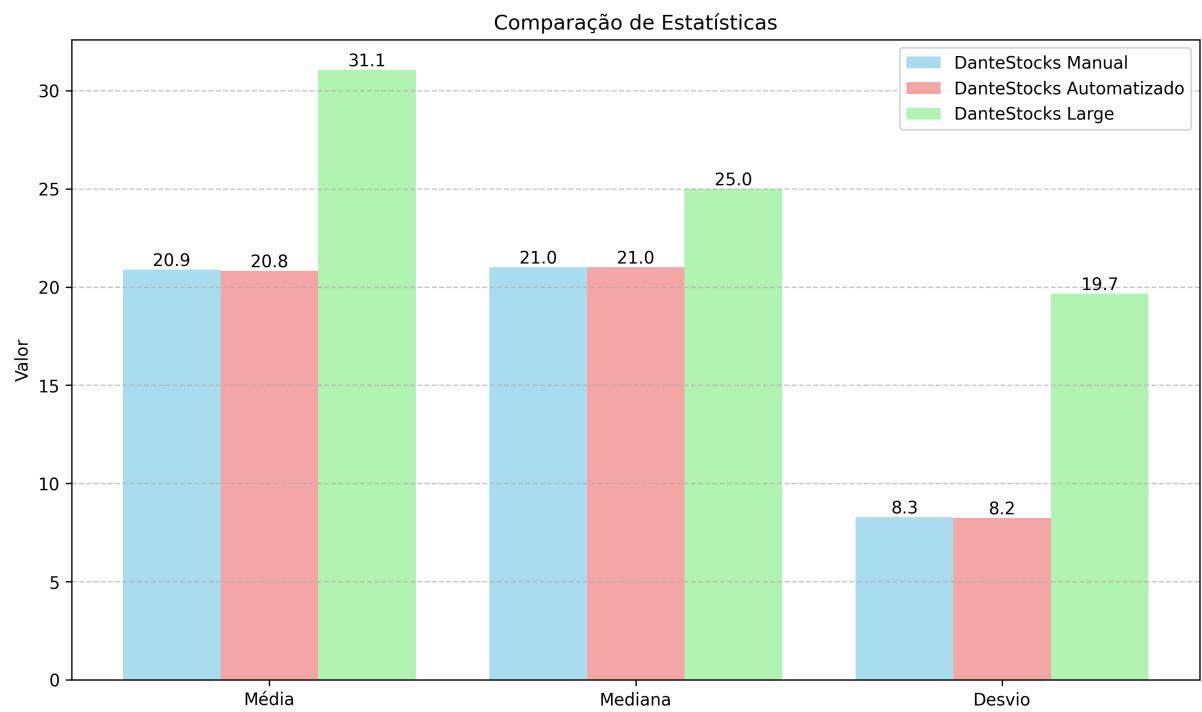


Figura 12: Comparação de Estatísticas entre os Corpora

6 Análise Comparativa de Diversidade Lexical

Esta seção apresenta uma análise detalhada da diversidade lexical dos corpora, comparando diferentes métricas e distribuições entre os textos.

6.1 Métricas Principais

As métricas a seguir nos permitem quantificar diferentes aspectos do uso do vocabulário:

DanteStocks Manual:

- Total de Tokens: 84,417
- Total de Types: 10,731
- Hapax Legomena: 6,154 (7.3%)
- Dis Legomena: 1,702 (2.0%)
- Type-Token Ratio (TTR): 0.127

DanteStocks Automatizado:

- Total de Tokens: 84,115
- Total de Types: 10,752
- Hapax Legomena: 6,163 (7.3%)
- Dis Legomena: 1,713 (2.0%)
- Type-Token Ratio (TTR): 0.128

DanteStocks Large:

- Total de Tokens: 3,932,604
- Total de Types: 79,275
- Hapax Legomena: 38,481 (1.0%)
- Dis Legomena: 10,603 (0.3%)
- Type-Token Ratio (TTR): 0.020

6.2 Análise Comparativa

Comparando os corpora, observamos:

- DanteStocks Automatizado apresenta a maior diversidade lexical ($TTR = 0.128$)
- Tamanhos dos corpora: 84,417, 84,115, 3,932,604 tokens

6.3 Análise dos Gráficos

1. Comparação de Métricas de Diversidade: Os gráficos de pizza mostram a distribuição de:

- Hapax Legomena: Palavras que aparecem uma única vez
- Dis Legomena: Palavras que aparecem duas vezes
- Outras palavras únicas: Types que aparecem três ou mais vezes

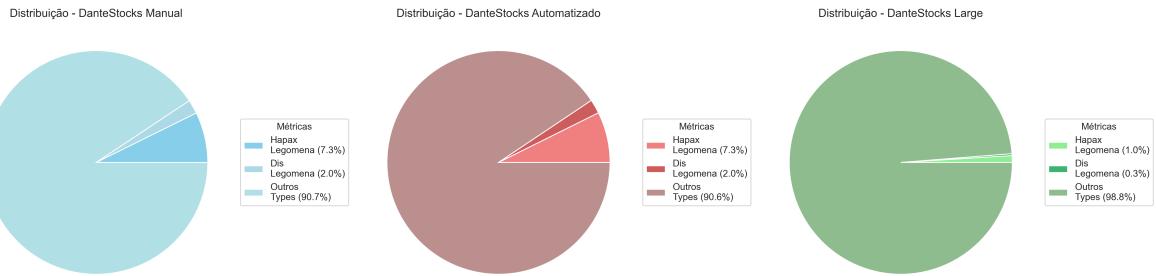


Figura 13: Comparação de Métricas de Diversidade

2. Type-Token Ratio (TTR): O gráfico de barras compara o TTR entre os corpora, onde valores mais altos indicam maior diversidade lexical.

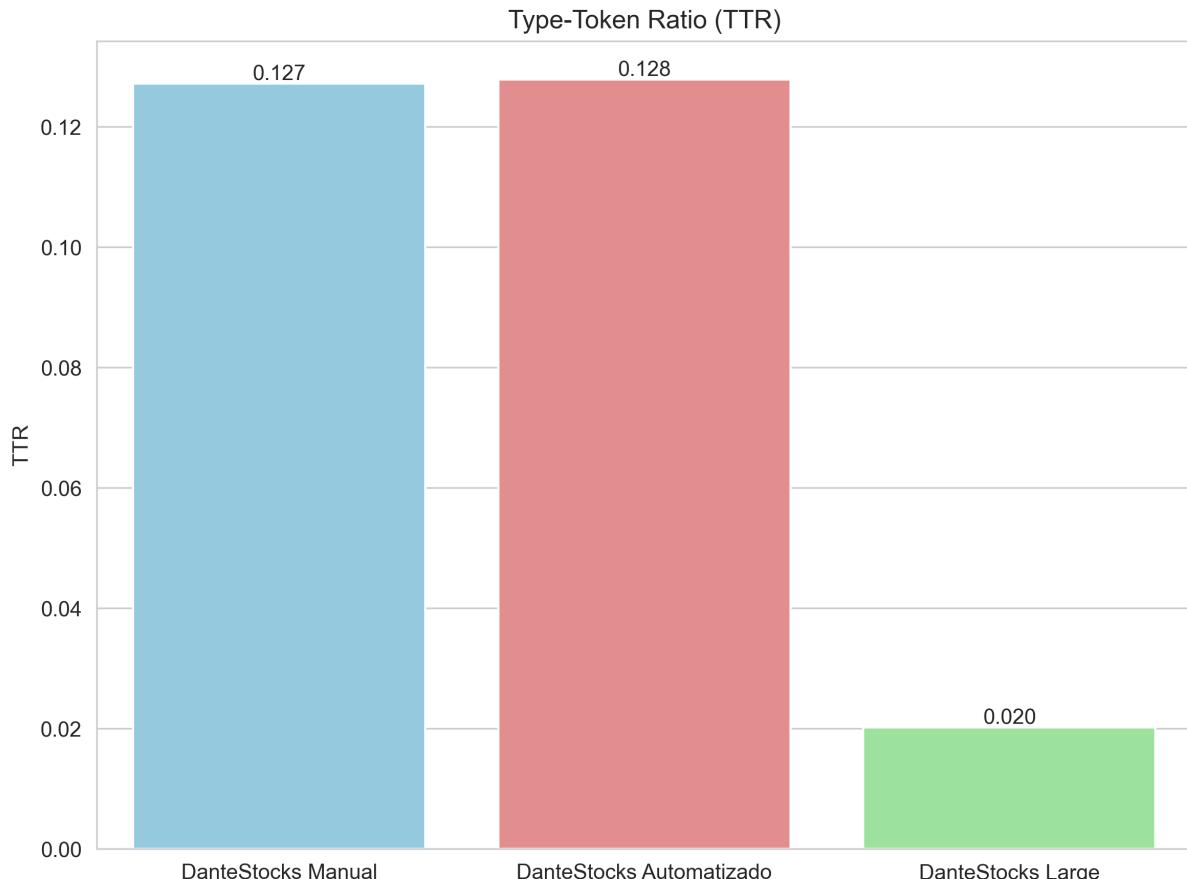


Figura 14: Type-Token Ratio (TTR)

7 Análise de Similaridade entre Distribuições

7.1 Introdução à Análise

Esta seção apresenta uma análise comparativa detalhada entre os corpora, utilizando métricas estatísticas avançadas para quantificar suas semelhanças e diferenças. O foco principal está na distribuição das Etiquetas POS (Part of Speech) em cada corpus.

7.2 Metodologia: Distância de Jensen-Shannon

A análise utiliza a Distância de Jensen-Shannon (JS), uma métrica estatística sofisticada que:

- Mede o grau de similaridade entre duas distribuições de probabilidade
- Produz valores entre 0 e 1, onde:
 - 0 indica distribuições idênticas
 - 1 indica distribuições completamente diferentes
- É mais robusta que medidas simples como diferença percentual

7.3 Resultados das Comparações

Corpus 1	Corpus 2	Distância JS	Divergência JS
DanteStocks Manual	DanteStocks Automatizado	0.1947	0.0379
DanteStocks Manual	DanteStocks Large	0.0553	0.0031
DanteStocks Automatizado	DanteStocks Large	0.2085	0.0435

Tabela 5: Resultados da Análise de Jensen-Shannon

7.4 Interpretação dos Resultados

DanteStocks Manual vs DanteStocks Automatizado: Os corpora são moderadamente similares, com algumas variações sutis no uso de classes gramaticais.

DanteStocks Manual vs DanteStocks Large: Os corpora são muito similares, sugerindo uma forte consistência no uso de classes gramaticais.

DanteStocks Automatizado vs DanteStocks Large: Os corpora são moderadamente diferentes, indicando variações significativas no uso de classes gramaticais.

7.5 Visualização dos Resultados

O gráfico abaixo apresenta uma comparação visual das distribuições de classes gramaticais entre todos os corpora analisados.

Distribuição de Classes Gramaticais e Divergência de Jensen-Shannon

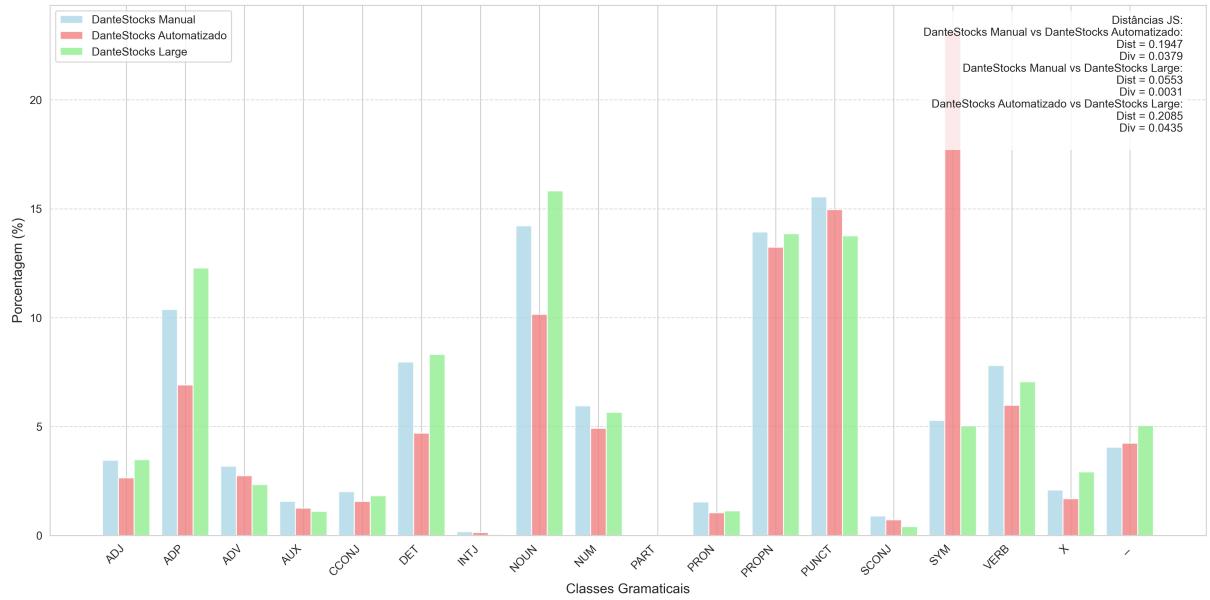


Figura 15: Distribuição de Classes Gramaticais e Divergência de Jensen-Shannon

8 Teste de Kolmogorov-Smirnov (KS)

8.1 Introdução ao Teste KS

O teste de Kolmogorov-Smirnov é um teste estatístico não-paramétrico que avalia se duas amostras provêm da mesma distribuição. Este teste é particularmente útil para comparar distribuições de frequência de classes gramaticais entre diferentes corpora.

8.2 Resultados das Comparações

Comparação	Estatística KS	P-valor	Interpretação
DanteStocks Manual vs DanteStocks Automatizado	0.1111	0.99997	Estatisticamente similares
DanteStocks Manual vs DanteStocks Large	0.1111	0.99997	Estatisticamente similares
DanteStocks Automatizado vs DanteStocks Large	0.1667	0.97154	Estatisticamente similares

Tabela 6: Resultados dos Testes de Kolmogorov-Smirnov

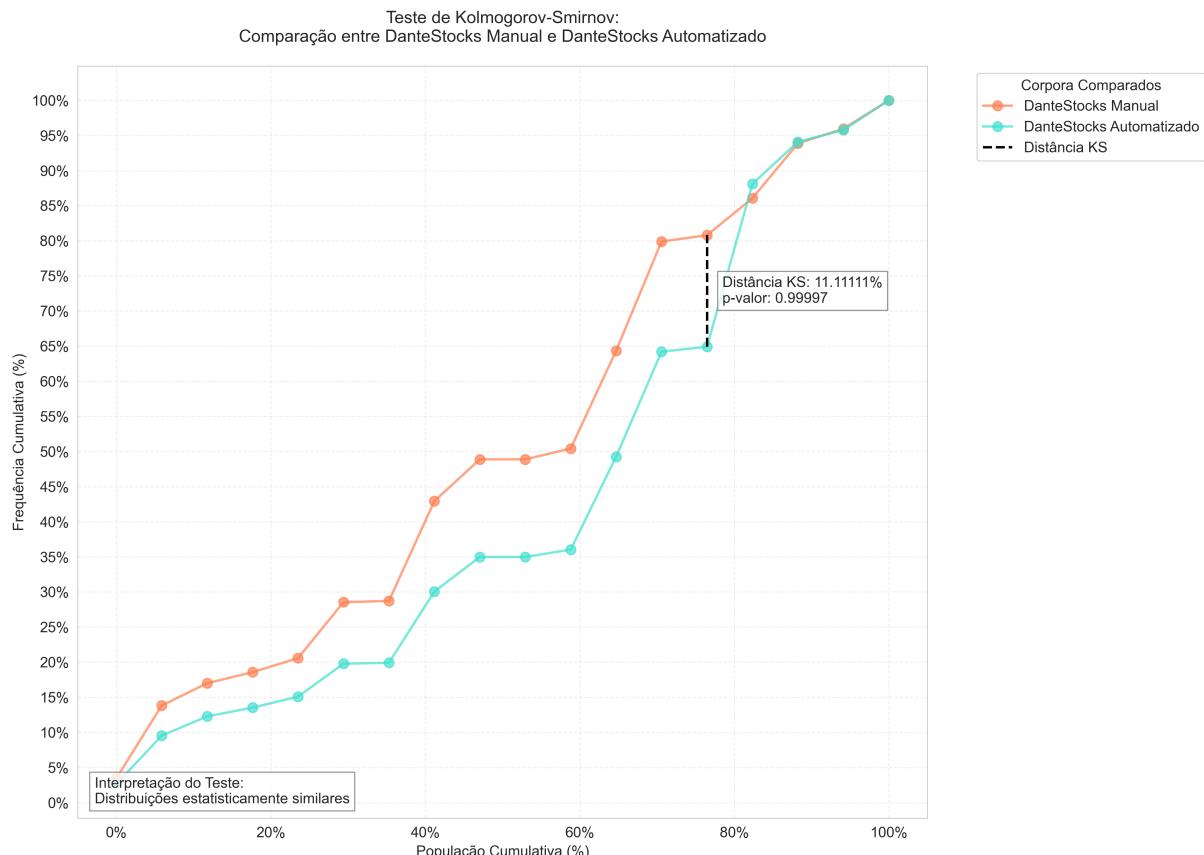


Figura 16: Comparação KS: DanteStocks Manual vs DanteStocks Automatizado

Teste de Kolmogorov-Smirnov:
Comparação entre DanteStocks Manual e DanteStocks Large

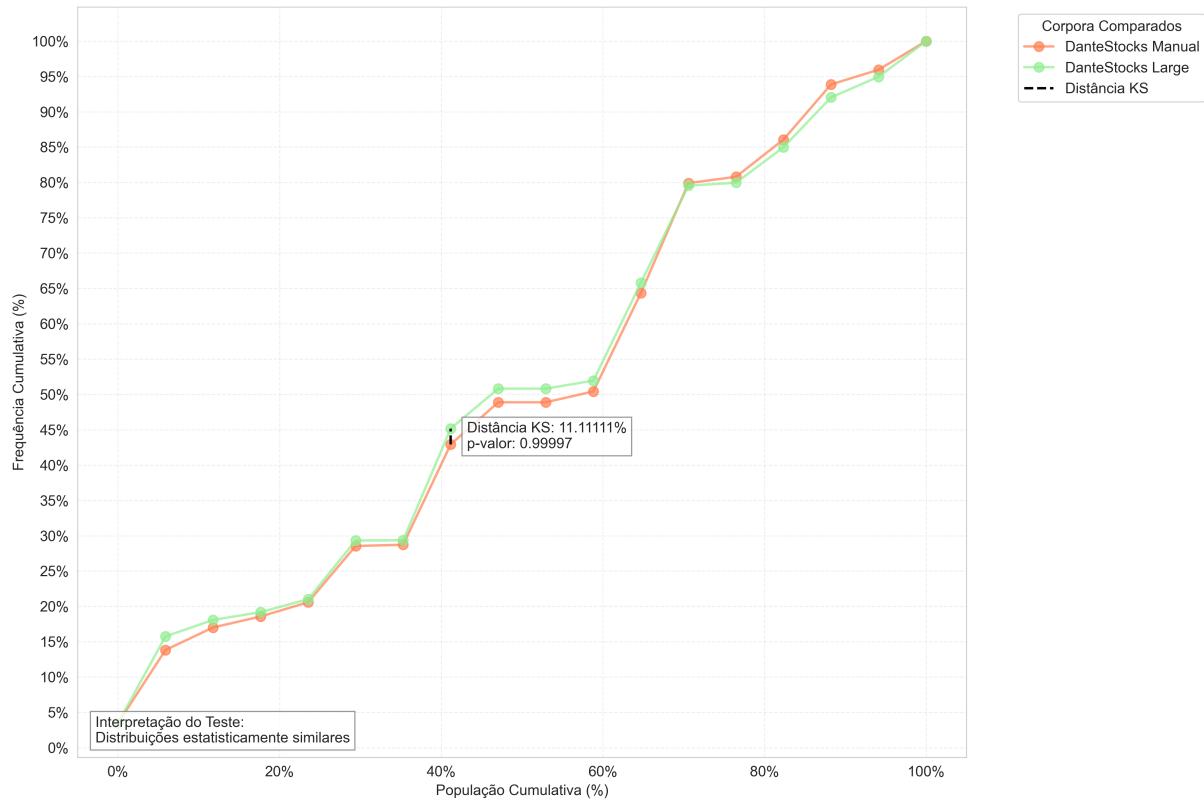


Figura 17: Comparação KS: DanteStocks Manual vs DanteStocks Large

Teste de Kolmogorov-Smirnov:
Comparação entre DanteStocks Automatizado e DanteStocks Large

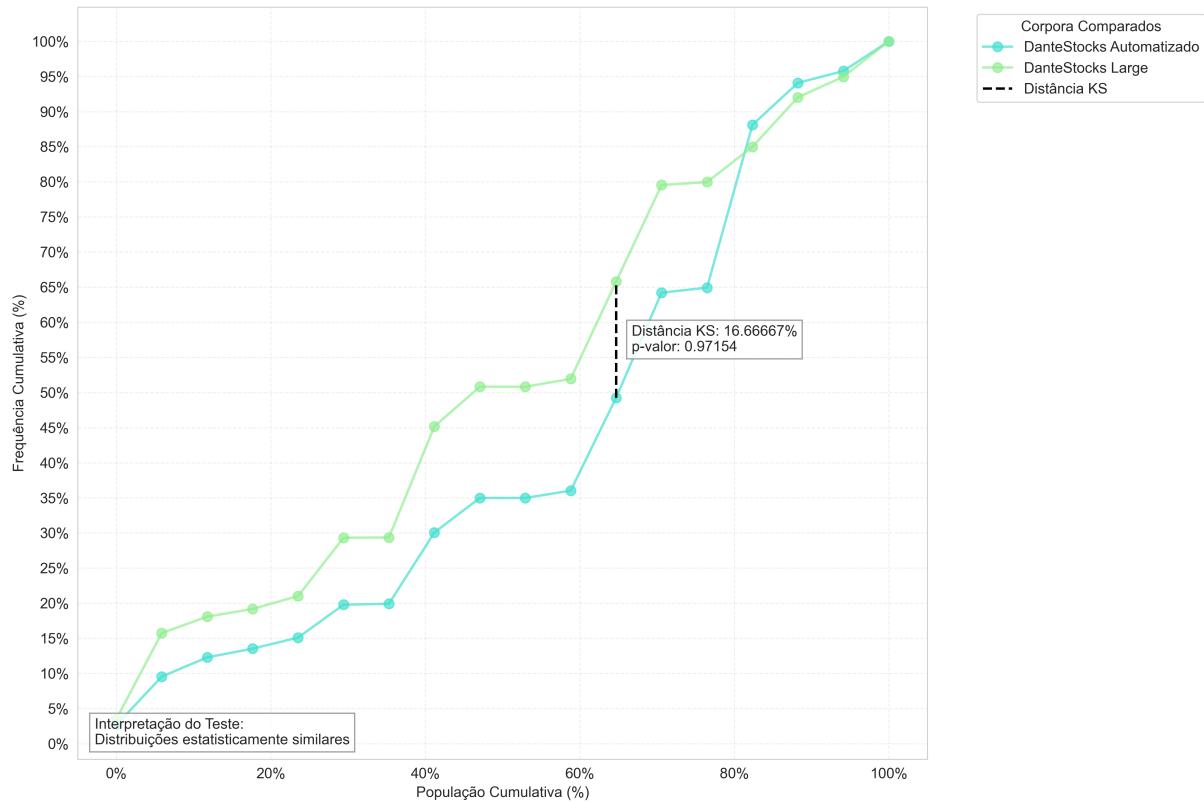


Figura 18: Comparação KS: DanteStocks Automatizado vs DanteStocks Large

9 Análise Comparativa do Tamanho das Sentenças

Esta seção apresenta uma análise detalhada da distribuição do número de tokens por sentença nos corpora. Esta métrica nos ajuda a entender a complexidade estrutural das sentenças e as diferenças de estilo entre os textos.

9.1 Estatísticas Descritivas

Métrica	DanteStocks Manual	DanteStocks Automatizado	DanteStocks Large
Número de Sentenças	4,042	4,042	126,653
Média de Tokens	20.88	20.81	31.05
Mediana	21.00	21.00	25.00
Desvio Padrão	8.29	8.24	19.65
Mínimo	2	2	1
Máximo	66	51	498

Tabela 7: Estatísticas do Tamanho das Sentenças

9.2 Análise Comparativa

DanteStocks Large apresenta, em média, as sentenças mais longas (31.1 tokens por sentença).

DanteStocks Large mostra maior variabilidade no tamanho das sentenças, com desvio padrão de 19.7 tokens.

9.3 Interpretação dos Histogramas

Os histogramas mostram a distribuição do número de tokens por sentença em cada corpus. As linhas verticais indicam a média e a mediana, permitindo visualizar:

- A forma geral da distribuição e sua simetria
- A concentração de sentenças em determinados tamanhos
- A presença de outliers (sentenças muito curtas ou muito longas)
- A relação entre média e mediana, indicando possível assimetria

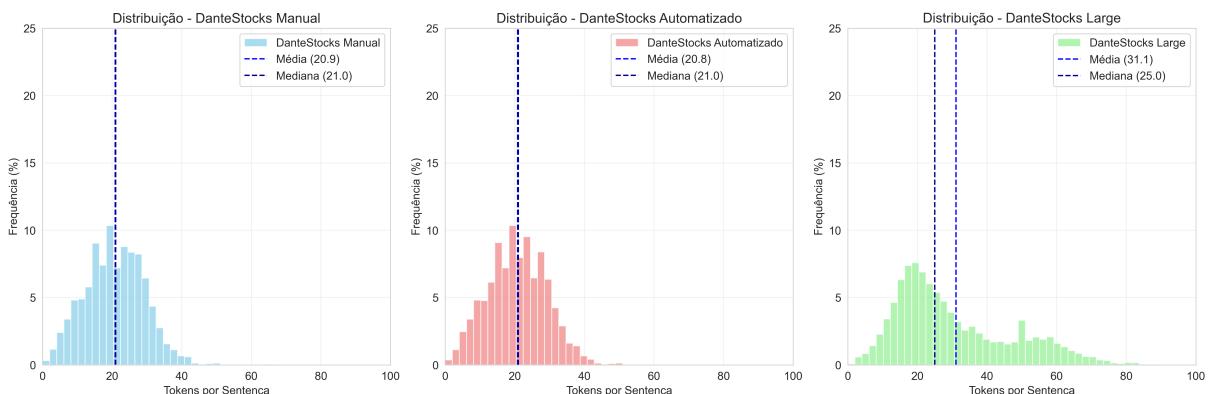


Figura 19: Distribuição Comparativa do Tamanho das Sentenças