Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

FERRAMENTA DE ANÁLISE AUTOMÁTICA DE INTELIGIBILIDADE DE CÓRPUS (AIC)

Erick Galani Maziero
Thiago Alexandre Salgueiro Pardo
Sandra Maria Aluísio

NILC-TR-08-08

Julho, 2008

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Apresenta-se, neste relatório, uma ferramenta de análise automática de inteligibilidade de córpus chamada AIC. A ferramenta extrai automaticamente diversos atributos de textos, baseando-se na anotação sintática dos textos produzida pelo sistema PALAVRAS. A AIC foi desenvolvida para o projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital), mas pode ser usada para os mais variados fins.

Este trabalho contou com o apoio das agências de fomento à pesquisa FAPESP, CAPES e CNPq.

ÍNDICE

1. Introdução

No contexto do projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital), uma ferramenta para automatização de análise de córpus, mais especificamente análise de inteligibilidade, foi desenvolvida com vistas a facilitar o trabalho de verificar uma lista de características que possivelmente são diferenciais em textos simples quando comparados com textos elaborados, os quais não têm a preocupação com simplicidade textual.

Os resultados da análise de inteligibilidade de córpus podem ajudar a guiar a tarefa de simplificação textual, fornecendo quais características realmente tornam um texto mais simples de ser entendido por pessoas dos mais diversos níveis de letramento. Apesar de ter sido criada para este fim, a ferramenta pode ser utilizada para quaisquer fins que necessitem de tais informações.

Na próxima seção, discute-se a questão de análise de córpus, mais especificamente, análise de inteligibilidade. Na seção 3, mostra-se o papel dos conhecimentos morfossintático e sintático na referida análise. Na Seção 4, é apresentada a interface web, que constitui a ferramenta de análise em si.

2. Análise de córpus

Córpus designa um conjunto de textos de uma língua, geralmente pertencentes a um mesmo estilo de escrita ou domínio de conhecimento. Uma análise de um conjunto de textos pode ter os mais variados objetivos, que vão desde um nível morfológico a uma análise pragmático-discursiva, para obtenção de informações úteis a outras tarefas, sejam elas de PLN (Processamento de Linguagem Natural) ou não.

Como será tratado abaixo, este documento trata da análise de características de inteligibilidade de textos feita por uma ferramenta web.

Inteligibilidade diz respeito à qualidade de algo ser claramente entendido. No caso de uma análise de córpus, inteligibilidade refere-se às características que tornam um texto facilmente inteligível. Por exemplo, o fato de um texto ter muitas construções utilizando a voz passiva pode ser uma característica que dificulta o entendimento do texto.

Na ferramenta de análise de inteligibilidade, delineada mais à frente, faz-se a verificação de diversas características do texto. Tais características estão listadas na Figura 1.

1. Número de caracteres
2. Número de palavras
3. Número médio de caracteres por palavra
4. Número médio de palavras por sentença
5. Número de sentenças
6. Número de palavras simples
7. Número de sentenças na voz passiva
8. Número de orações
9. Número de orações que iniciam com conjunções subordinadas
10. Número de orações que iniciam com conjunções coordenadas
11. Sentenças que contenham número definido de orações.
12. Moda do cálculo anterior
13. Número médio de orações por sentença
14. Número de conjunções subordinadas
15. Número de conjunções coordenadas
16. Número de verbos no gerúndio
17. Número de verbos no particípio
18. Número de verbos no infinitivo
19. Soma dos verbos no gerúndio, particípio e infinitivo
20. Número de objetos preposicionais
21. Número médio de objetos preposicionais por sentença
22. Número médio de objetos preposicionais por oração
23. Número de orações relativas
24. Número de apostos especificadores
25. Número de adjuntos adverbiais
26. Número de advérbio
27. Número de adjetivos
28. Número de pronomes pessoais
29. Número de pronomes pessoais na segunda pessoa do singular
30. Número de pronomes pessoais na primeira pessoa do plural
31. Número de pronomes pessoais na segunda pessoa do plural
32. Número de pronomes pessoais na terceira pessoa do plural
33. Número de pronomes possessivos
34. Número de pronomes possessivos na segunda pessoa (Teu, Tua, Teus, Tuas)
35. Número de pronomes possessivos na segunda pessoa (Vosso, Vossa, Vossos, Vossas)
36. Número de pronomes possessivos na primeira pessoa
37. Número de pronomes possessivos na terceira pessoa
38. Número de marcadores discursivos
39. Número de marcadores discursivos ambíguos

Figura 1: Lista das características calculadas pela ferramenta de análise

Algumas características não refletem o quanto um texto está inteligível, mas servem como operandos dos cálculos de outras características. Por exemplo, um texto pode conter um

número de caracteres proporcional ao seu tamanho, mas este número não indica se o texto é mais ou menos fácil de entender. Entretanto, a contagem de caracteres é utilizada no cálculo do número médio de caracteres por palavra; o tamanho médio das palavras pode ser um indicativo de inteligibilidade de um texto.

3. Utilização de conhecimento morfossintático e sintático

A maioria das características que se analisa na ferramenta advém de informações morfossintáticas e sintáticas obtidas automaticamente do texto. Portanto, a análise morfossintática e sintática prévia é indispensável para que a ferramenta funcione corretamente. A ferramenta utilizada para isso é o sistema PALAVRAS (Bick, 2000).

O PALAVRAS é um dos melhores analisadores sintáticos automáticos (ou parsers) para o português do Brasil. O texto submetido a este parser recebe uma marcação palavra a palavra, que indica, dentre outras informações, a categoria gramatical, a função sintática e algumas informações semânticas.

Este parser utiliza a Gramática de Dependências, uma representação formal. Na representação usada no PALAVRAS, por exemplo, o símbolo @ é utilizado para introduzir as etiquetas (ou marcações) de nível sintático e os marcadores < e > indicam a direção do núcleo sintático de que os constituintes são dependentes, com exceção do verbo principal, que não exibe marcadores de dependência.

O parser tem como entrada um texto simples, sem anotação prévia, tal como um arquivo de texto sem formatações. Sua saída pode ser seguindo diversos formatos; por padrão, tem-se a saída no formato vislCG como no Figura 2, que contém um exemplo de análise feita pelo parser.

```
Esta            [este] <dem> DET F S @>N
frase-exemplo   [frase-exemplo] <act-s> N F S @SUBJ>
será            [ser] <fmc> <aux> V FUT 3S IND VFIN @FS-STA
verificada      [verificar] <mv> V PCP F S @ICL-AUX<
para            [para] PRP @<ADVL
demonstração    [demonstração] <act> N F S @P<
de              [de] <sam-> <np-close> PRP @N<
as              [o] <artd> <-sam> DET F P @>N
etiquetas       [etiqueta] <ac> N F P @P<
$.
```

Figura 2: exemplo de frase analisada pelo parser PALAVRAS

A saída do PALAVRAS, na Figura 2, contém as palavras do texto analisado, uma a uma, na primeira coluna. Nos casos de palavras contraídas, como *das*, que é a contração da preposição *de* com o artigo *as*, o PALAVRAS separa e coloca cada palavra em uma linha, para facilitar a análise.

Na segunda coluna, além da palavra lematizada (entre os sinais [ ]), encontram-se etiquetas para a classe morfossintática da palavra (V para verbo, PROP para substantivos ou ADJ para adjetivos, por exemplo) e suas inflexões (F indicando feminino ou S para singular, como exemplos). Iniciada pelo símbolo @ está a etiqueta para indicar a função sintática (@ SUBJ> ou @ <SUBJ para indicar sujeito da oração, em que os sinais < ou > são usados para indicar a direção na sentença onde está o verbo principal da oração com relação à palavra que recebe esta etiqueta).

Algumas etiquetas auxiliares, ou secundárias, são usadas para desambiguar alguma etiqueta morfossintática ou de função sintática. Estas etiquetas encontram-se entre os sinais <>. Ainda entre <> podem ser encontradas etiquetas semânticas para substantivos, verbos e alguns adjetivos. Essas etiquetas semânticas agrupam as palavras em classes de significado, tais como “grupo de animais” ou “veículo”.

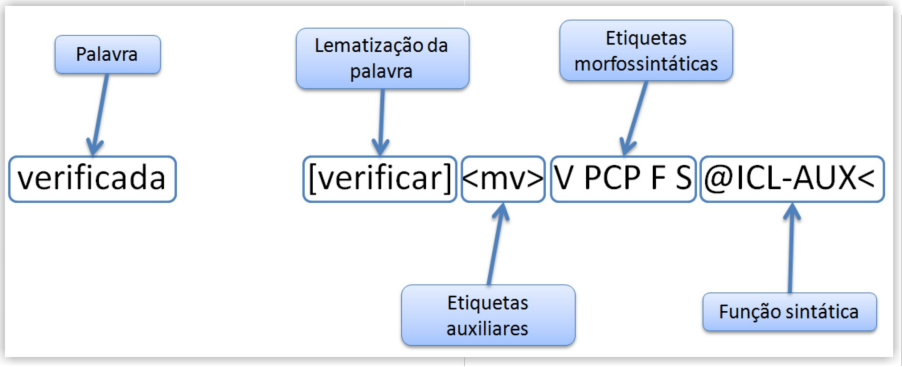

Figura 3: organização das etiquetas do PALAVRAS

Na Figura 3 pode ser observada a organização das etiquetas pelo parser, em que, na segunda coluna, separada por um *tab*, estão as etiquetas da palavra encontrada na primeira coluna.

Embora se baseie no formalismo de Gramáticas de Restrições (de Helsinki) para resolver ambigüidades, o parser apresenta algumas dificuldades quando diante de alguns fenômenos lingüísticos como adjetivos acidentais, como o caso da palavra *nervoso* no trecho:

"Ele bebeu um copo nervoso", em que o parser não relaciona o adjetivo *nervoso* com o sujeito da oração. Além de suas limitações, e como em qualquer sistema automático de PLN, suas análises não têm uma precisão de 100%. Portanto seus erros podem se propagr na análise.

4. A ferramenta AIC

A ferramenta consiste em uma interface web, responsável por obter as informações a serem processadas e repassá-las aos scripts e outras ferramentas que farão a análise de inteligibilidade do texto submetido. Após a análise, a interface exibe os resultados.

A interface web pode ser acessada no seguinte endereço: www.nilc.icmc.usp.br/AIC. Muito simples de ser manipulada, contém uma área para inserção do texto a ser analisado e um botão para iniciar a análise (veja Figura 4).

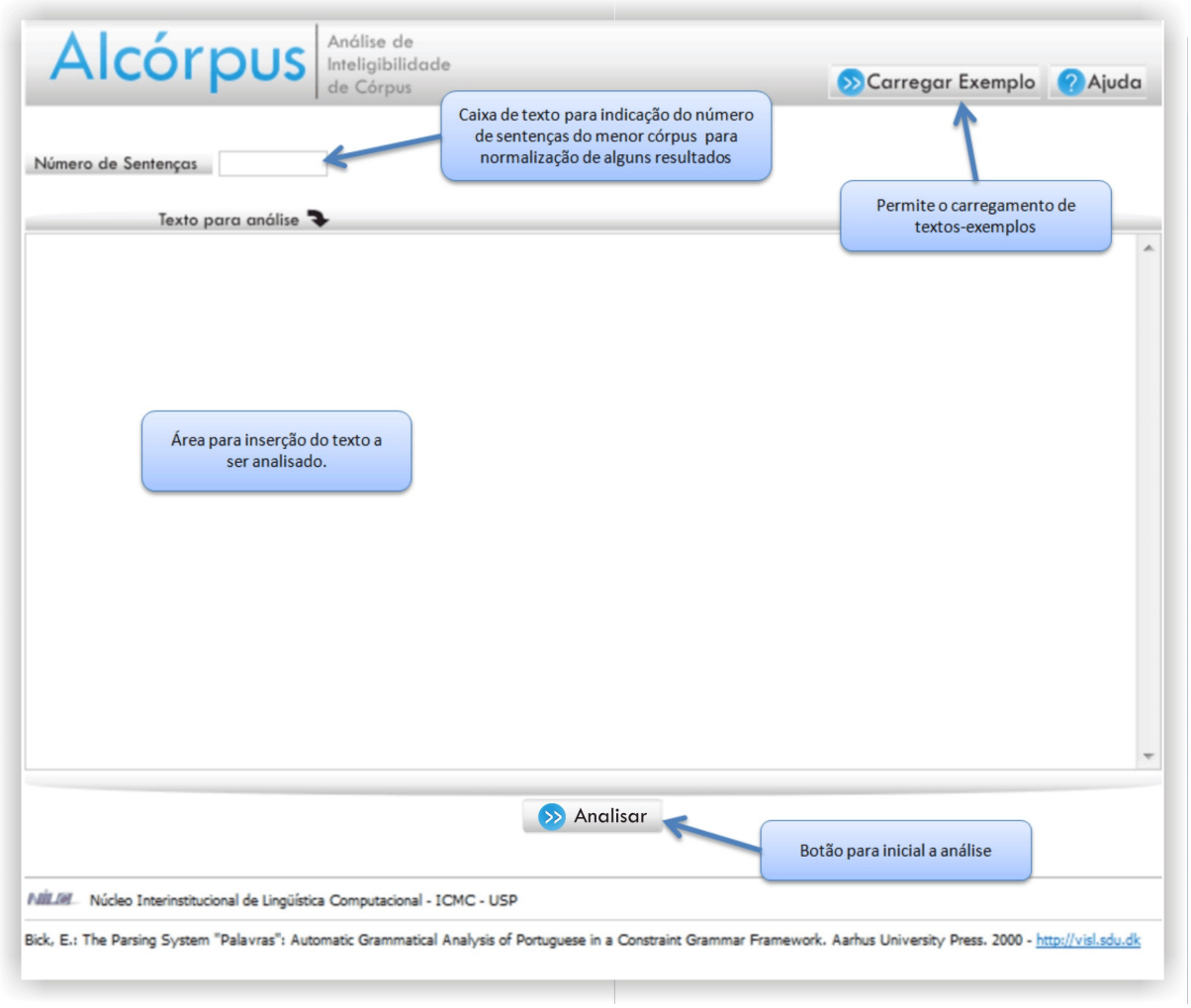

Figura 4: Tela inicial da ferramenta de análise

Como a análise envolve diversos textos a serem processados, muitos de tamanhos diferentes, algumas das características foram calculadas para o texto inteiro submetido, e calculadas, também, para as x primeiras sentenças do mesmo texto, sendo x o número informado na interface da ferramenta, equivalente ao número de sentenças do menor texto a ser analisado, para fins de normalização de alguns cálculos no conjunto de textos sob análise. Este número é inserido em seu local indicado na Figura 4.

Ao iniciar a análise, clicando no botão "Analisar", o usuário deve esperar até que a análise seja completada e os resultados sejam exibidos. Um quadro, com o texto "Aguarde" é apresentado e mantido durante o processamento do texto. Quando a análise for finalizada, os resultados serão exibidos pela interface, como mostrado na Figura 5.

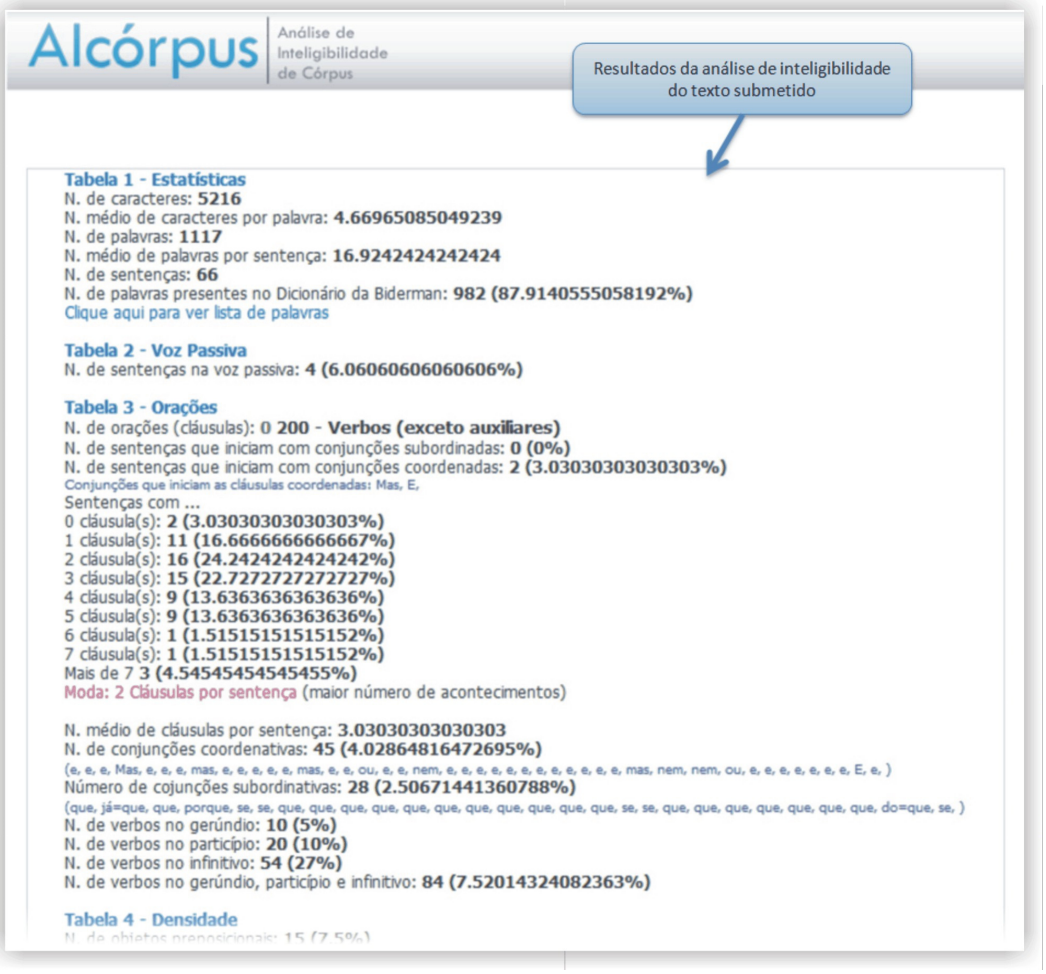

Figura 5: Exibição dos resultados da análise

Além dos dados exibidos, há alguns links na página para outras informações provenientes da análise. Por exemplo, logo abaixo da característica "N. de palavras presentes no Dicionário da Biderman", há um link para se visualizar quais foram as palavras simples do texto que estão presentes em uma lista de palavras comprovadamente simples (Biderman, 2005).

No fim dos resultados, há dois links (veja Figura 6): um para um arquivo de texto contendo o resultado exibido na presente página da ferramenta – *"Save as...(Right click -> save as..."*; outro contendo para o arquivo contendo o resultado da análise do parser – *"Save as...(PALAVRAS)(Right click -> Save as..."*

Figura 6: links para arquivos da análise

No Apêndice A, mostra-se um exemplo de texto completamente processado pela ferrametna.

Agradecimentos

Aos caros colegas Tiago de Freitas Pereira e Paulo Rodrigues Alves Margarido pelas contribuições na confecção da ferramenta.

Referências

Bick, E. (2000). *The Parsing System PALAVRAS: Automatic Gramatical Analysis of Porutugese in a Constraint Grammar Framework.* Aarhus University Press.

Biderman, M.T.C. (2005). *Dicionário Ilustrado de Português*. Editora Ática.

Apêndice A: Exemplo de análise da ferramenta

Como exemplo da análise descrita neste documento, abaixo se tem uma coletânea de textos de uma seção infantil de textos simples: Para Seu Filho Ler do Jornal Zero Hora

Texto de entrada para a ferramenta:

Os créditos de carbono são como uma moeda. Imagine o seguinte: as empresas que poluem mais a natureza, ou seja, soltam fumaça e jogam restos de lixo nos rios, têm que repassar este "dinheiro" para aquelas que não fazem mal ao meio ambiente.
Trata-se de algo novo, que existe há pouco tempo no mundo e também no Brasil. Foi a forma que se encontrou de tentar reduzir a destruição da natureza, já que as empresas que poluem vão tentar mudar, para não ter que gastar seu dinheiro.
Uma das peças mais conhecidas de William Shakespeare é Romeu e Julieta. Mas o dramaturgo inglês, que viveu há 400 anos, escrevia vários textos que falavam da história do seu tempo.
Ricardo III é um deles. Ele conta a trajetória de um nobre coxo e corcunda que não hesita em matar amigos e parentes para conquistar o trono da Inglaterra.
É uma trama misteriosa cheia de intrigas, e que mostra o mal que a ambição pode fazer.
Existem pelo menos dois filmes muito legais sobre o assunto: o mais clássico é um lançado em 1955, com Lawrence Olivier no papel-título, mas o documentário À Procura de Ricardo III (1996), de Al Pacino, é muito bacana também.
Os bombeiros tentam encontrar pessoas embaixo da terra em São Paulo. Na sextafeira, um obra do metrô desabou, aumentando um buraco que já existia, fazendo carros, caminhões e pessoas caírem dentro dele.
O trabalho é difícil, porque tem muita lama, pedaço de cimento, carros e caminhões lá dentro,
e os bombeiros têm de tirar tudo isso de cima para conseguir chegar até as pessoas e ver se elas estão vivas.
Ninguém sabe dizer ainda por que a terra desmoronou e fez o chão perto do buraco afundar, mas já tem gente estudando o que aconteceu para dizer se alguém fez alguma coisa errada.
Os Estados Unidos acham que o Irã quer construir bombas atômicas, que podem matar milhares de pessoas. O Irã diz que não é verdade e que só pretende produzir eletricidade.
Os americanos ameaçam aprovar medidas contra os iranianos na Organização das Nações Unidas, como proibir que o Irã compre produtos de outros países. Os Estados Unidos também podem começar uma guerra contra o Irã para impedir a fabricação das bombas.
Ontem, o presidente iraniano desafiou seus inimigos e disse que não acredita em guerra ou castigos contra seu país. O presidente do Irã não gosta de Israel e também fez críticas aos israelenses.
Numa ditadura, quem manda no país não é eleito e nem permite que os meios de comunicação contem a verdade sobre o que acontece. Os ditadores usam a força e a violência para fazer
que todos aceitem suas regras.
O Brasil também teve ditaduras. A última acabou há mais de 20 anos, e algumas pessoas que combateram aquele governo ainda acham importante lembrar as histórias de quem sofreu na época.

Por isso, lançaram mais um livro sobre como as pessoas eram torturadas na ditadura. Os militares que governaram o Brasil de 1964 e 1985 acham que o passado já foi explicado.

Dizem que, na época, queriam evitar que o Brasil se tornasse comunista, um sistema onde o governo não é eleito e é dono de quase tudo. Hoje, os militares fazem parte da democracia, regime no qual se elege o presidente, os deputados, os vereadores, os prefeitos e se pode protestar quando não se está contente.

Quer saber se é maneiro ser escoteiro? Então aparece neste domingo no Cais do Porto de Porto Alegre, na Área Infantil e Juvenil da Feira. A partir das nove da manhã e até as três da tarde, a União dos Escoteiros do Brasil promove bate-papos, palestras, exibições de vídeo, jogos, brincadeiras e canções - tudo para mostrar como são as atividades dos escoteiros. - Vai ser bem interativo - promete Carlos Chaise, presidente da Seção Rio Grande do Sul da União.

Diz ele que, no mundo, há hoje cerca de 30 milhões de escoteiros - 70 mil deles no Brasil, 7,2 mil aqui no Estado.

Para chegar lá sem fazer feio, vá decorando o cumprimento deles: três dedos estendidos para cima, o polegar abraçando o mindinho.

O ex-governador do Rio Anthony Garotinho é candidato a presidente na eleição marcada para outubro e quer chamar a atenção. Desde domingo, ele não come e só bebe água para evitar que os jornais, rádios e TVs divulguem problemas relacionados a sua candidatura.

Ele quer impedir a divulgação de notícias sobre doações de dinheiro a sua campanha, que podem ter sido feitas fora da lei. Sem conseguir explicar o que aconteceu, o candidato diz que está sendo perseguido pela imprensa. Jornais, rádios e TVs dizem que não vão dar bola para a birra de Garotinho.

Todos os anos, a Organização das Nações Unidas (ONU) lança um livro em que mostra em quais países as pessoas estão vivendo melhor e em quais a vida é mais difícil. O Brasil vem melhorando aos poucos, mas continua no meio da lista, nem entre os melhores, nem entre os piores.

O livro lançado este ano mostra ainda que muita gente no mundo não tem água pura para beber ou um banheiro limpo para usar. Por isso, pede a ajuda dos países ricos para ajudar os mais pobres.

Nos últimos anos, o mundo começou a perceber mudanças no clima, como chuva forte e calor intenso, que provocam enchentes e derretem grandes camadas de gelo em vários países.

Esse clima maluco é conseqüência do aquecimento da Terra, o chamado efeito estufa. Esse aquecimento é provocado por excesso de gases que são ruins para a atmosfera (a camada de ar que envolve o planeta). Indústrias e automóveis, por exemplo, liberam esses gases. Os desmatamentos e as queimadas de florestas também. Preocupados com isso tudo, representantes de 84 países se comprometeram, em 1997, a passar a lançar menos gases.

Esse acordo foi feito na cidade de Kyoto, no Japão. Por isso o nome se chama Protocolo de Kyoto.

Muitos países não estão cumprindo o que prometeram. O aquecimento da Terra continua aumentando, e o clima é cada vez mais maluco.

Seus brinquedos são de plástico, né? Ele é muito usado na indústria por ser barato, leve e fácil de amassar e esticar. O problema é que aquele joguinho velho, esquecido no armário, vai durar mais do que você. O plástico demora séculos para sumir da natureza.

E se as pecinhas forem levadas pela chuva, então, podem matar animais e entupir canos. Um saco.

Muitos pais, quando eram crianças, não sabiam disso. Então, cresceram pedindo um brinquedo atrás do outro (pergunte pelo bambolê).

Nesta página, há brinquedos feitos de outros materiais. Ajude seu pai a montar um, afinal ele é meio atrapalhado.

Saída da Ferramenta:

Tabela 1 - Estatísticas

N. de caracteres: **5216**
N. médio de caracteres por palavra: **4.66965085049239**
N. de palavras: **1117**
N. médio de palavras por sentença: **16.9242424242424**
N. de sentenças: **66**
N. de palavras presentes no Dicionário da Biderman: **982 (87.9140555058192%)**

Tabela 2 - Voz Passiva

N. de sentenças na voz passiva: **4 (6.06060606060606%)**

Tabela 3 - Orações

N. de orações (cláusulas): **200 - Verbos (exceto auxiliares)**
N. de sentenças que iniciam com conjunções subordinadas: **0 (0%)**
N. de sentenças que iniciam com conjunções coordenadas: **2 (3.03030303030303%)**
Conjunções que iniciam as cláusulas coordenadas: Mas, E,
Sentenças com ...
0 cláusula(s): **2 (3.03030303030303%)**
1 cláusula(s): **11 (16.6666666666667%)**
2 cláusula(s): **16 (24.2424242424242%)**
3 cláusula(s): **15 (22.7272727272727%)**
4 cláusula(s): **9 (13.6363636363636%)**
5 cláusula(s): **9 (13.6363636363636%)**
6 cláusula(s): **1 (1.51515151515152%)**
7 cláusula(s): **1 (1.51515151515152%)**
Mais de 7 **3 (4.54545454545455%)**
Moda: 2 Cláusulas por sentença (maior número de acontecimentos)

N. médio de cláusulas por sentença: **3.03030303030303**
N. de conjunções coordenativas: **45 (4.02864816472695%)**
(e, e, e, Mas, e, e, e, mas, e, e, e, e, e, mas, e, e, ou, e, e, nem, e, e, e, e, e, e, e, e, e, e, e, mas, nem, nem, ou, e, e, e, e, e, e, e, E, e,)
Número de cojunções subordinativas: **28 (2.50671441360788%)**
(que, já=que, que, porque, se, se, que, que, que, que, que, que, que, que, que, que, que, que, se, se, que, que, que, que, que, que, que, do=que, se,)
N. de verbos no gerúndio: **10 (5%)**
N. de verbos no particípio: **20 (10%)**
N. de verbos no infinitivo: **54 (27%)**
N. de verbos no gerúndio, particípio e infinitivo: **84 (7.52014324082363%)**

Tabela 4 - Densidade

N. de objetos preposicionais: **15 (7.5%)**
N. médio de preposicionais por sentença: **0.227272727272727**
N. médio de preposicionais por cláusula: **0.075**
N. de cláusulas relativas: **26 (13%)**

Iniciadas por: que, que, que, que, que, que, que, que, que, que, que, que, como, quem, que, como, que, onde, o=qual, que, quais, que, que, que, que, que, quando,

N. de apostos especificadores: **1**

N. adjuntos adverbiais: **131**

N. de adjetivos: **43 (3.84959713518353%)**

N. de advérbios: **150 (13.4288272157565%)**

Tabela 6 - Personalização

N. de pronomes pessoais: **6 (0.53715308863026%)**, dado total de palavras

2PS	1PP	2PP	3PP
Tu, Você, Te, Ti, Contigo	Nós, Nos, Conosco	Vós, Vocês, Vos, Convosco	Eles, Elas, Os, As, Lhes, Se, Si, Consigo
1 (16.6666666666667%)	0 (0%)	0 (0%)	5 (83.3333333333333%)

N. de pronomes possessivos: **9 (0.805729632945389%)**, dado total de palavras

Teu, Tua, Teus, Tuas: **0 (0%)**

Nosso, Nossa, Nossos, Nossas: **0 (0%)**

Vosso, Vossa, Vossos, Vossas: **0 (0%)**

Seu, Sua, Seus, Suas: **9 (100%)**

Tabela 5 - Marcadores Discursivos

também : **1 (100%)**

Marcadores Ambíguos:

também

N. total de marcadores: **1 (0.0895255147717099%)**

N. total de marcadores ambíguos: **1 (100%)**