Proposta recebida em Fevereiro 2023 e aceite para publicação em Maio 2023 .

# A compilação e a análise de métricas textuais de um corpus de redações

**Compilation and analysis of textual metrics of an essay's corpus**

Átila Augusto Soares Vital
Faculdade de Letras da Universidade Federal de Minas Gerais

**Resumo**

A prova de redação do Exame Nacional do Ensino Médio (Enem) é decisiva para a garantia da vaga em instituições de ensino superior no Brasil. De 2010 a 2020, foi observado que a quantidade de redações avaliadas em nota máxima (mil pontos) caiu de maneira drástica e abrupta: de 3.694 redações nota máxima em 2011 para apenas 28 em 2020. O objetivo deste trabalho é apresentar um corpus de redações nota máxima avaliadas pela banca do Enem, descrevê-las e tecer breves considerações a partir da análise de métricas textuais na série histórica de 2010 a 2020. A compilação foi feita de forma manual, pela internet. Para as descrições, foram utilizados o programa Orange: Data Mining e o analisador de complexidade textual NILC-Metrix (Leal et al., 2022). Os resultados sugerem que houve aumento expressivo no número de palavras e diminuição da razão type/token ao longo dos anos. Além disso, foram feitas medidas sintáticas que constataram o aumento da complexidade dos textos.

**Palavras chave**

redações; linguística de corpus; complexidade textual

**Abstract**

The writing test of the National High School Exam (Enem) is very important to guarantee a place for students in undergraduate institutions in Brazil. From 2010 to 2020, the number of texts evaluated in maximum grade (one thousand points) dropped abruptly: in 2011, 3,694 texts gained 1,000 points, and in 2020, only 28 texts were evaluated with the same grade. The objective of this research is to present a corpus of texts graded one thousand points by Enem's team, to describe them and to make brief considerations about their characteristics during the historical series from 2010 to 2020. The compilation was made manually, using the internet. We used Orange: Data Mining and the NILC-Metrix (Leal et al., 2022) textual complexity analyzer. The results suggest an expressive increase in the number of words and a decrease in the type/token ratio during the period. Finally, syntactic metrics were measured and confirmed the increase in textual complexity.

**Keywords**

essays; corpus linguistics; textual complexity

## 1. Introdução

O Exame Nacional do Ensino Médio (Enem) ocorre anualmente no Brasil desde 1998, tendo sido criado com a intenção de avaliar os estudantes da educação básica. A partir de 2009, com a aderência da maior parte das universidades brasileiras ao exame, houve busca crescente pela prova, acompanhada pelas revisões nos critérios de correção e nos assuntos a serem discutidos nas questões de múltipla escolha. No caso da redação, há, uma vez a cada ano, a disponibilização de materiais para os alunos por parte do Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) — a Cartilha de Redação — e para os corretores — o Manual do Corretor — onde são arrolados os critérios para a correção dos textos dissertativo-argumentativos.

Boa parte do trabalho dos docentes, seja em cursinhos pré-vestibulares ou em escolas preparatórias, se relaciona com a análise detida dos critérios em cada um dos manuais prescritivos, de modo que se possa ter uma visão sucinta dos elementos necessários para a escrita da redação, que deve pertencer ao tipo textual dissertativo-argumentativo. As correções devem ser pautadas em cinco competências, que procuram tornar objetiva a análise (i) da norma padrão da Língua Portuguesa; (ii) da correspondência ao tipo textual; (iii) da consistência da argumentação; (iv) da coesão e (v) da apresentação de uma proposta de intervenção. Esses elementos e suas explicações oficiais podem ser visualizados nos tópicos abaixo, que relacionam cada

DOI: 10.21814/lm.15.1.393

competência com seus objetivos de avaliação, segundo a Cartilha do Participante. As notas de cada competência variam de 0 a 200 pontos e, somadas, constituem a nota final da prova de redação (Brasil, 2020).

- *Competência I*: Demonstrar domínio da modalidade escrita formal da Língua Portuguesa;
- *Competência II*: Compreender a proposta de redação e aplicar conceitos das áreas de conhecimento, dentro dos limites do texto dissertativo-argumentativo em prosa;
- *Competência III*: Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista;
- *Competência IV*: Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação;
- *Competência V*: Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos;

Além disso, na etapa do Sistema de Seleção Unificada (SISU), principal meio de alocação dos candidatos em vagas nas universidades públicas, a nota da prova de redação é inserida na média simples em relação às outras áreas do conhecimento. Em muitos casos, como salientam Cançado et al. (2020, pp. 64), essa nota "é responsável por uma grande parte da classificação de um candidato, fechando ou abrindo as portas de entrada em nossas universidades."

Curiosamente, dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) mostram que, embora o número de participantes nas edições do ENEM tenha aumentado consideravelmente desde 1998, houve, a partir de 2011, queda acentuada na quantidade de textos avaliados em nota máxima — 1000 pontos. Partindo de um pico de 3.694 avaliações nota mil em 2011, em 2020, apenas 28 textos foram avaliados como nota máxima no exame, evidenciando a diminuição de textos exemplares (Figura 1).

O objetivo deste trabalho, portanto, é, a partir da ferramenta NILC-Metrix (Leal et al., 2022), desenvolvida pelo Núcleo Interinstitucional de Linguística Computacional (NILC) da Universidade de São Paulo (USP), descrever, de formas quantitativa e qualitativa, algumas das características linguísticas das redações nota mil ao longo dos anos de 2010 a 2020, período crítico em que houve diminuição expressiva da quantidade de textos avaliados em nota máxima. Dessa forma, evidenciar o perfil dos textos exemplares

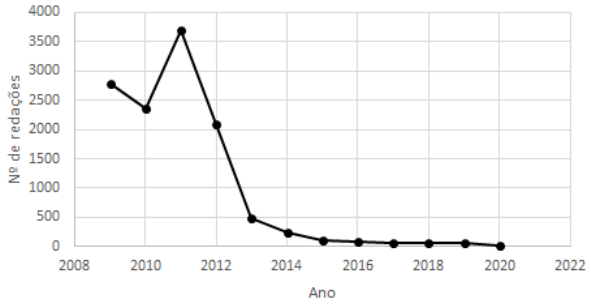


**Figura 1:** Quantidade de redações avaliadas em nota 1000 ao longo dos anos. Fonte: Sinopses Estatísticas do INEP.

no decorrer do período pode ajudar a reunir características linguísticas para futuros estudos em redações do Enem, além de revelar em que medida os textos da série histórica se aproximam e se distanciam. Para o caso de diferenças significativas entre os períodos, é possível pensar na mudança da estrutura dos textos coligada à mudança de critérios de correção. Este trabalho, portanto, oferece subsídios para se pensar a mudança estrutural para, posteriormente, se fazer a análise dos critérios e incrementar o corpus com textos de variados níveis.

Para isso, compilamos, sob o viés metodológico da Linguística de Corpus e Computacional, um corpus de 96 redações nota mil que passaram pela correção oficial do exame ao longo do período analisado. Como o corpus conta apenas com textos nota máxima, foi possível observar mudanças estruturais, ao longo do tempo, nas redações que foram avaliadas, a princípio, num mesmo patamar. Os dados foram arquitetados de modo que pudéssemos ter pelo menos um texto de cada um dos anos, totalizando, até o momento, 37.459 palavras, em uma razão type/token de 0,1606.

Para o estudo linguístico a partir da análise do corpus, lançamos mão da consagrada conceptualização histórica de Berber Sardinha (2000), que salienta que a descrição de um conjunto de dados está diretamente relacionada à representatividade e, portanto, ao caráter probabilístico do uso linguístico. Desse modo, as conclusões a serem observadas no estudo poderão assumir valores de verdade apenas em relação aos dados elencados na compilação. É importante salientar o fato de que, até o momento da confecção deste trabalho, não havia disponíveis corpora expressivos de redações nota mil corrigidas pela banca oficial do Enem.

Com o auxílio de métodos computacionais, há poucos trabalhos disponíveis que procuram correlacionar notas ou níveis de proficiência escrita em uma dada língua com métricas textuais. Esse é o caso de Crossley & McNamara (2012), que, através da ferramenta Coh-Metrix (Graesser et al., 2004), examinaram a relação entre as estratégias de coesão e de sofisticação linguística com o nível de proficiência conferido a aprendizes de inglês como segunda língua (L2). Como medidas de coesão, os autores observaram o uso de operadores lógicos, retomada lexical, coesão sequencial, referenciação semântica, causalidade, conectivos e diversidade lexical. Por sofisticação, os autores consideram como sendo a capacidade de produção de estruturas linguísticas menos frequentes e tomam como principais métricas medidas psicolinguísticas, frequência de hiperônimos, frequência de palavras e complexidade sintática. Como resultados, o estudo mostrou que as medidas de sofisticação se correlacionam com o grau de proficiência dos alunos.

Outro trabalho que parte de textos de aprendizes de inglês como L2 é o de Alexopoulou et al. (2017), que investigou os efeitos linguísticos nos textos dos alunos que podem ser gerados pela complexidade das tarefas e pelo enfoque instrucional dado por elas. Nesse sentido, diferentes instruções para o desenvolvimento de textos gera particularidades em suas estruturas, como diferentes distribuições de tempos verbais, complexidade sintática e uso de verbos irregulares. Para aprendizes de português como língua estrangeira, está disponível a plataforma LX-CEFR, apresentada por Branco et al. (2014), que calcula o grau de proficiência linguística a partir do Quadro Comum Europeu de Referência para Línguas (CEFR). O processamento do texto a partir do LX-CEFR enquadra nos níveis A1, A2, B1, B2 ou C1 as seguintes métricas textuais: índice Flesch, tamanho médio das frases (em palavras), tamanho médio das palavras (em sílabas) e proporção de nomes. Com as possibilidades de se compilar e de se treinar modelos de linguagem natural, é cada vez mais frequente o surgimento de ferramentas para avaliação automática. Como bem salientam Branco et al. (2014), é importante que esses recursos sejam utilizados de modo a complementarem a avaliação do aprendizado de línguas estrangeiras, já que, em muitos casos, os modelos são desenvolvidos com um modesto número de textos, fator limitante da acurácia.

Westerlund (2019), após compilar 30 ensaios produzidos em inglês para o Swedish National Exam por estudantes do segundo grau na Suécia, se utilizou de uma metodologia similar à de Crossley & McNamara (2012), processando os textos na ferramenta Coh-Metrix 3.0 e confirmando resultados semelhantes: textos com maior sofisticação, isto é, com uso de palavras menos frequentes, hiperônimos, voz passiva e alta densidade de sintagmas adverbiais foram aqueles que receberam notas mais altas.

Para análises que não lançam mão das métricas propostas por ferramentas de tratamento de textos como o Coh-Metrix, há, no Brasil, o desenvolvimento incipiente de trabalhos voltados para a avaliação automática de redações (Automatic Essay Scoring), como é o caso de Amorim & Veloso (2017). No estudo, os autores propõem um sistema de análise de múltiplos aspectos (multi-aspect analysis) para a correção automática de redações, considerando um dataset de 1840 textos corrigidos pelos corretores do website *Educação UOL* e métricas desenvolvidas para a língua inglesa. De forma similar, com objetivo de fornecer conjunto de dados para análises de redações do Enem, Marinho et al. (2021) criam um corpus com 4.570 redações coletadas das plataformas *Vestibular UOL* e *Educação UOL*. Com um conjunto de textos ainda maior, de 56.644 redações, Fonseca et al. (2018) criaram outro sistema de avaliações automático baseado em redes neurais. Nesse caso, o modelo considerou métricas de 5 grupos principais:

1. *Count metrics*: contagem de estatísticas básicas dos textos, como número de vírgulas, número de caracteres, número de parágrafos, tamanho médio das sentenças, dentre outras;
2. *Specific expressions*: contagem de grupos de palavras esperados em redações, como agentes sociais, conectores discursivos, palavras propositivas e marcas de oralidade;
3. *Token n-grams*: verifica a correlação entre a ocorrência de n-gramas e as notas altas;
4. *POS n-grams*: verifica a classe de palavras dos n-gramas;
5. *POS counts*: conta o número de classes de palavras no texto.

É importante evidenciar que os trabalhos elencados acima para redações em português brasileiro i) coletaram textos que não necessariamente passaram pela correção oficial do Enem e ii) e que possuem pontuações variadas - não apenas nota mil. Para este artigo, optamos pela coleta de redações que restritivamente passaram pela correção oficial e que receberam nota máxima (mil pontos) ao longo do período entre 2010 e 2020. Com isso, ao obtermos diferenças significativas em métricas textuais ao longo da série

estudada, teremos indícios que denotarão diferentes perfis linguísticos para textos avaliados num mesmo nível pelos corretores da banca oficial.

Nesse sentido, para o refinamento metodológico, foram utilizados outros artigos que se relacionam com a temática da linguística das redações do ENEM, dentre eles, citamos Buzato et al. (2021) e Cançado et al. (2020), que se assentam diretamente na Linguística de Corpus, Bertucci et al. (2020) e Cruz et al. (2021), que tratam, respectivamente, da ocorrência de anáforas encapsuladoras em textos do Exame e da argumentação em propostas de intervenções de redações nota mil. Na esteira da Linguística Computacional, lançamos mão do trabalho de Westerlund (2019), que correlaciona métricas textuais com as notas de avaliações do Exame Nacional Sueco (Swedish National Exam) a partir das métricas da ferramenta Coh-Metrix, da qual o NILC-Metrix (Leal et al., 2022) é uma das derivações. Além disso, as investigações de avaliação automática de textos para aprendizado de segunda língua, como Alexopoulou et al. (2017), para o inglês, e Branco et al. (2014), para o português foram importantes para a composição metodológica.

## 2. Metodologia

Diferentemente de boa parte dos trabalhos sobre redações do Enem, nosso objetivo foi levar em consideração apenas os textos que foram corrigidos pela banca oficial de corretores da prova. Nesse sentido, os passos metodológicos adotados passeiam pelas fases de:

1. pesquisa e compilação dos textos;
2. caracterização do corpus coletado;
3. escolha dos programas e das métricas de análise;
4. apresentação dos resultados e da análise linguística.

Na fase de pesquisa e compilação, foram coletados, de diferentes domínios da internet, textos de redações nota mil. A maior parte delas é proveniente de sites que veiculam boas práticas de escrita e que fornecem, de forma gratuita, instruções e cursos para a realização da prova, além de textos nota mil publicados em reconhecidos sites de notícias e reportagens.

Os textos foram retirados dos domínios na internet e dispostos em arquivos `.txt` com codificação UTF-8, nomeados conforme o ano em que foram escritos e o número associado a cada redação específica daquele ano. O exemplo 1 ilustra um trecho do texto 8, escrito e corrigido no ano de 2014.

> Exemplo 1
>
> *Durante o século XX, o estímulo à produção industrial, por Getúlio Vargas, e o incentivo à integração nacional, de Juscelino Kubitschek, foram fatores que possibilitaram a popularização dos meios de comunicação no Brasil. Com isso, cresceu também a publicidade infantil, que busca introduzir nas crianças, desde cedo, o princípio capitalista de consumo. No entanto, essa visão negativa pode ser significativamente minimizada, desde que acompanhada de uma forte base educacional que auxilia as crianças a discernir por meio do desenvolvimento de senso crítico próprio.*

Para a análise automática do corpus, foram selecionadas três vias principais, cada uma com suas métricas e bibliotecas particulares: (i) script em *Python*, para visualização das palavras mais frequentes; (ii) *Orange: data mining*, para aferição da razão type/token; (iii) NILC-Metrix, para o cálculo da complexidade textual.

Além de (i) e (ii), mecanismos extremamente difundidos para análise de dados em geral, a interface do NILC-Metrix foi desenvolvida para que os textos possam ser processados em relação a 200 métricas de complexidade textual, coerência e coesão. As métricas são divididas em 14 grandes grupos, considerando a natureza das análises e as técnicas de PLN empregadas, conforme Leal et al. (2022). Para este trabalho, foram escolhidas métricas de 7 grupos, a saber: duas métricas descritivas, uma morfossintática, uma de densidade de padrões sintáticos, uma de complexidade sintática, uma de conectivos e uma de leiturabilidade. Os nomes das métricas escolhidas e seus respectivos objetivos estão arrolados a seguir.

- *words per sentences*: métrica descritiva que calcula a quantidade média de palavras por sentença. Quanto maior a métrica, maior é o tamanho das sentenças no texto, e, portanto, a complexidade textual. É importante pontuar que a noção de sentença programada para o NILC-Metrix não leva em consideração a máxima projeção do sintagma verbal Jackendoff (1982), mas sim a unidade iniciada por letra maiúscula e finalizada por ponto final, ponto de exclamação, ponto de interrogação ou reticências;

- *sentences per paragraph*: métrica descritiva que calcula a quantidade média de sentenças por parágrafo. Para esta métrica, a relação com a complexidade textual é menos evidente, já que esta última dependerá tanto do tamanho do parágrafo quanto do tamanho das sentenças que o compõem. De toda forma, sua aferição será importante para a análise da série histórica de textos, já que, a princípio, redações exemplares mais divulgadas possuem quantidades próximas (e, em alguma medida, padronizada) de frases nos parágrafos.
- *content words*: métrica morfossintática que calcula a proporção de palavras de conteúdo (substantivos, verbos, adjetivos, advérbios e palavras denotativas) em relação ao número de palavras total. Quanto maior a métrica, maior é o vocabulário exigido para a compreensão do texto e, portanto, maior a complexidade;
- *mean noun phrase*: métrica de densidade de padrões sintáticos que calcula a quantidade média de palavras que compõem os sintagmas nominais. Para o cálculo desta métrica, o texto é processado pelo LX-Parser, que identifica os constituintes sintáticos, dentre eles, os sintagmas nominais (SNs). Quanto maior a métrica, maior é o tamanho médio dos SNs, e, portanto, maior a complexidade textual;
- *words before main verb*: métrica de complexidade sintática que calcula a quantidade média de palavras antes do verbo principal. Quanto maior a métrica, mais informações precisam ser armazenadas na memória de trabalho, aumentando a complexidade textual;
- *conn ratio*: métrica de conectivos que calcula a quantidade de conectivos em relação à quantidade de palavras total. Por meio de uma lista de palavras pré-determinada, a ferramenta procura por conectivos aditivos, temporais, causais e lógicos. Quanto maior o uso de conectivos, mais simples tende a se tornar o entendimento do texto, diminuindo a complexidade. O uso da métrica se justifica pelo fato de haver uma competência específica para a avaliação da coesão (competência IV), que se dá, dentre outros fatores, através da análise dos conectivos intra e interparagrafais. Quanto maior a métrica, maior a coesão entre as partes do texto;
- *flesch*: métrica de leiturabilidade que considera o tamanho médio das palavras (calculado em número médio de sílabas) e sentenças (calculado em número médio de palavras). Quanto maior a métrica, menor a complexidade do texto.

Embora o índice flesch possa representar uma medida grosseira da leiturabilidade do texto — já que depende da acurácia do segmentador silábico e do tokenizador de palavras e sentenças — o índice mostrou resultados intrigantes ao longo da série. Para o cálculo, são utilizadas a média de palavras por sentenças e a média do número de sílabas por palavra no texto. A fórmula empregada é ajustada empiricamente.

$$F = 248,835 - [1,015 \times MPS] - [84,6 \times MSP] \quad (1)$$

Sendo $F$ o índice flesch, $MPS$ a média de palavras por sentença e $MSP$ a média de sentenças por parágrafo.

Por conta de problemas para o processamento de longas sequências de arquivos de texto na ferramenta, optamos por não calcular o Índice de Honoré, que também é uma medida de leiturabilidade, mas que leva em consideração o número de tokens e de hápax legomena, isto é, palavras que são utilizadas uma única vez no texto.

O fato de a quantidade de palavras compiladas a cada ano ser diferente ao longo da série histórica implica na impossibilidade de comparação direta entre os valores das métricas. Por conta disso, foi feita a normalização de cada um dos resultados para cada 100 (cem) palavras. Isso assegura que possamos comparar os valores de uma métrica relativa a um mesmo número de palavras entre dois ou mais períodos diferentes.

## 3. Resultados e discussões

O número de palavras coletadas no corpus, até o momento, é de 37.395, com razão type/token de 0,160602. A partir do ano de 2012, embora a quantidade de textos nota máxima diminua, boa parte deles se encontram disponíveis na internet e em compilações feitas por plataformas de vestibulares. A coleta manual e a dificuldade de se encontrarem redações de determinados períodos explicam os diferentes números de textos compilados para cada edição do exame na Tabela 1.

Os textos e seus respectivos temas estão organizados a seguir. O corpus está disponível publicamente.[1]

Os dados do NILC-Metrix, do script e do Orange foram dispostos em gráficos, com o objetivo de facilitar a visualização da série histórica das notas. O primeiro gráfico a ser notado é o do número médio de palavras por ano, calculado pelo script, e que quase duplicou, saindo

[1] https://github.com/atilavital/corpus-redacao

| Ano | Tema | Quantidade |
|---|---|---|
| 2010 | O trabalho na construção da dignidade humana | 4 |
| 2011 | Viver em rede no século XXI: os limites entre o público e o privado | 3 |
| 2012 | O movimento imigratório para o Brasil no século XXI | 9 |
| 2013 | Efeitos da implantação da lei seca no Brasil | 10 |
| 2014 | Publicidade infantil em questão no Brasil | 10 |
| 2015 | A persistência da violência contra a mulher | 10 |
| 2016 | Caminhos para combater a intolerância religiosa | 10 |
| 2017 | Desafio para a formação educacional de surdos no Brasil | 10 |
| 2018 | Manipulação do comportamento do usuário pelo controle de dados | 10 |
| 2019 | Democratização do acesso ao cinema | 10 |
| 2020 | O estigma associado às doenças mentais na sociedade brasileira | 10 |

**Tabela 1:** Quantidade de textos coletados para cada tema.

de 226,5, em 2010, para 441,5, em 2020, conforme a Figura 2. Como o espaço para a escrita das redações não foi alterado ao longo dos anos — mantendo-se um limite de 30 (trinta) linhas em uma área sem alterações significativas —, é de se esperar que, ao se incrementar gradativamente a quantidade de palavras, mudanças graduais devam ser observadas em outras métricas linguísticas.

A média da razão type/token (Figura 3), por outro lado, como métrica importante para indicação da riqueza lexical ao longo dos textos, foi calculada pelo Orange e sugeriu uma diminuição ao longo dos anos, embora com uma taxa não tão evidente quanto a da média de palavras, conforme o valor de $R^2$. A diminuição da razão type/token pode ser o reflexo do aumento gradual da quantidade de palavras (tokens), conforme a Figura 2, sem acréscimos, na mesma proporção, de tipos diferentes de palavras (types). Sob esse ponto de vista, a riqueza lexical tende a diminuir ao longo da série estudada.

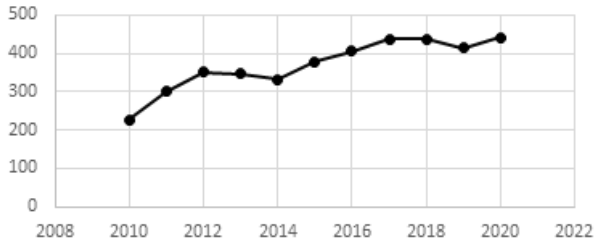


**Figura 2:** Média de palavras por ano. Fonte: elaborado pelo autor.

Como já salientado, a ferramenta NILC-Metrix compreende a noção de sentença como todos os elementos linguísticos posicionados entre uma letra maiúscula e um sinal de pontuação (ponto final, ponto de interrogação, exclamação ou reticências). As Figuras 4 e 5 nos mostram que tanto a média de sentenças por parágrafo

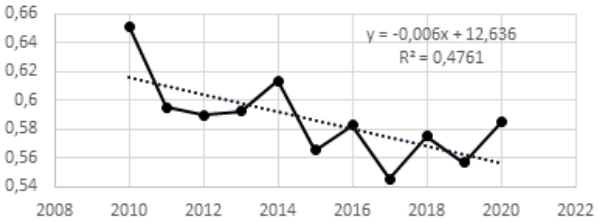


**Figura 3:** Evolução da média da razão type/token. Fonte: elaborado pelo autor.

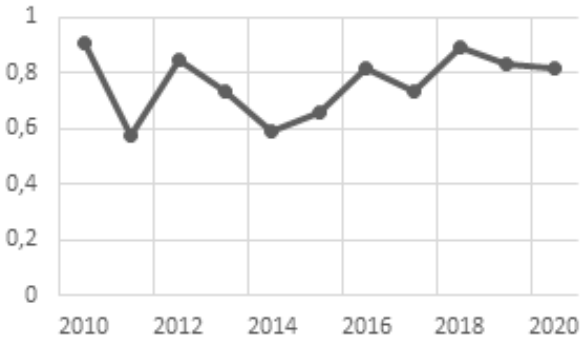


**Figura 4:** Média de sentenças por parágrafo.

quanto a média de palavras por sentenças não exibiram mudanças tão evidentes. A princípio, com um número de parágrafos inalterado e o mesmo espaço para a escrita do texto, o aumento do número de sentenças poderia indicar uma menor complexidade textual, já que a tendência seria a de haver mais sentenças, ainda que menores, no interior dos parágrafos; no entanto, essa conclusão só seria verdadeira caso a quantidade de palavras por sentença diminuísse, o que não se observa na Figura 5.

O tamanho médio dos sintagmas nominais, evidente na Figura 6, depende acurácia da anotação sintática do LX-Parser, ferramenta utilizada para a anotação sintática do NILC-Metrix.

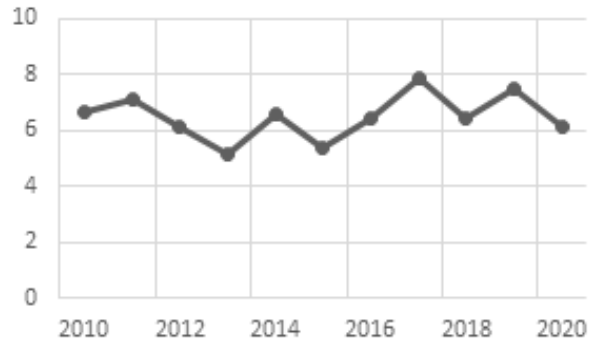


**Figura 5:** Média de palavras por sentença.

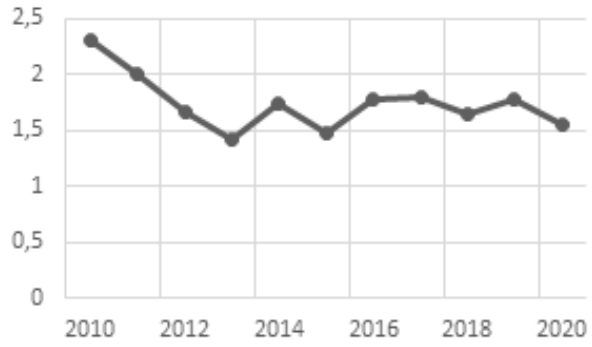


**Figura 6:** Tamanho médio dos SNs.

Pelo gráfico, podemos perceber que, a partir de 2015, o tamanho médio dos SNs tende a um descrescimento de 2010 a 2013, mas o padrão não se mantém até 2020.

Tanto a média de palavras de conteúdo (Figura 7) — medida pela biblioteca nlpnet (Fonseca & Rosa, 2013), do Python, modelo baseado em redes neurais e que faz etiquetagens semânticas e de classes de palavras — quanto a média de palavras antes do verbo principal (Figura 8) — que inclui medidas de anotação do LX-Parser (Silva et al., 2010) e de tokenização do Parser Palavras (Bick, 2000) — demonstraram um decrescimento.

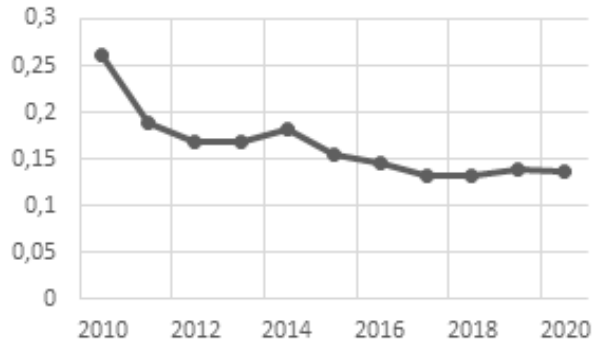


**Figura 7:** Média de palavras de conteúdo / palavras total.

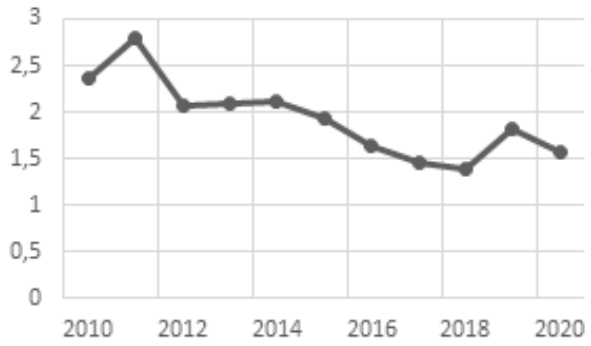


**Figura 8:** Média do nº de palavras antes do verbo principal.

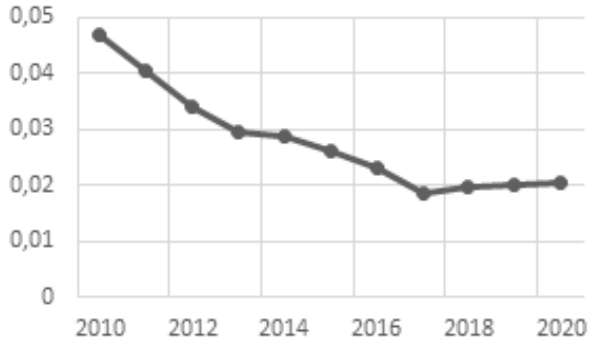


**Figura 9:** Média de conectivos em relação ao número de palavras.

No caso das palavras de conteúdo, é possível afirmar que o aumento do número de palavras total ao longo da série (Figura 2) desencadeia uma diminuição na proporção entre palavras significativas e gramaticais. As palavras acrescidas, por sua vez, parecem ocupar as posições sintáticas após o verbo principal, já que a média de palavras antes do verbo principal diminui ao longo dos anos (Figura 8). Neste ponto, é possível pensar em variadas explicações para a não variação no tamanho dos SNs: uma hipótese seria que o aumento de palavras no texto implicaria um maior número de sintagmas nominais com tamanhos próximos, o que evitaria um simples incremento da complexidade dentro dos SNs, cuja quantidade já parecia estar estável em relação ao tamanho e à informatividade do texto. É importante destacar que, para trabalhos futuros, o processamento de outros tipos de sintagmas seria valioso, de modo a esclarecer em quais constituintes, precisamente, as novas palavras tendem a se encaixar.

A Figura 9, por sua vez, representa a média de conectivos em relação ao número de palavras no texto, que diminuiu ao longo dos anos analisados, fazendo com que haja mais palavras para cada conectivo. Este dado não sugere que os textos passam a ser menos coesos, mas que o número de

elementos coesivos não aumenta na mesma taxa do número de palavras.

Além dos gráficos de complexidade textual que medem parâmetros sintáticos, semânticos e paragrafais, foi calculado o índice Flesch de leiturabilidade. Calculado automaticamente pelo NILC-Metrix, o índice tem o objetivo de verificar a correlação média entre tamanhos médios de palavras e sentenças. O comportamento da métrica ao longo dos anos pode ser visualizado na Figura 10.

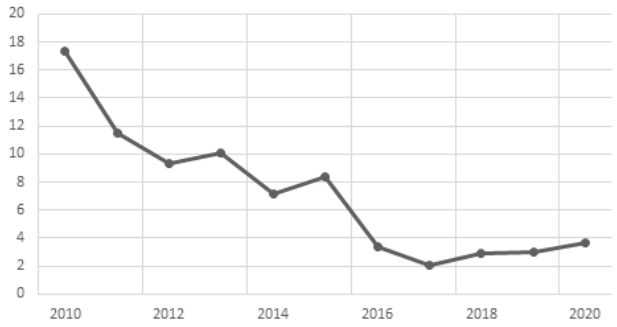


**Figura 10:** Evolução do índice de leiturabilidade Flesch.

Houve queda acentuada na média de leiturabilidade dos textos entre os anos de 2010 a 2017, o que implica numa maior complexidade textual e, portanto, numa menor facilidade de leitura. Embora o índice tenha apresentado resultados intrigantes e compatíveis com as métricas anteriores, é importante notar que os resultados dependem do desempenho do tokenizador, do sentenciador e do segmentador silábico, sendo que este último produz uma medida grosseira em relação ao tamanho médio das palavras do texto.

Ainda que incipientes, os resultados encontrados permitem que teçamos alguns comentários a respeito da série histórica de textos nota mil. O primeiro deles diz respeito ao aumento contínuo do número de palavras, que, em média, quase dobrou durante o período analisado, embora o espaço para a escrita do texto tenha se mantido constante. Os dados do aumento expressivo de tokens são enriquecidos com os dados de palavras de conteúdo, que, por sua vez, diminuem progressivamente, sugerindo que a relação entre as palavras significativas e o número total de palavras se dê em taxas diferentes: se o número de palavras total aumenta em taxa $x$ , o número de palavras de conteúdo diminui em taxa $y$, tal que $x$ é maior que $y$.

É importante notar que, aparentemente, nas condições em que há um limite rigoroso de linhas, como é o caso das redações, um aumento de palavras implica numa reorganização textual, impactando a estrutura das sentenças e o encadeamento coesivo. É precisamente nesse sentido, portanto, que seria esperado um aumento no número de palavras nos SNs, que, a princípio, enriqueceriam a descrição e os assuntos introduzidos nas sentenças. Por outro lado, a média de palavras antes do verbo principal se mostrou descrescente — mesmo que saibamos que a quantidade de palavras nas sentenças aumentou, em média. Esse fato pode sugerir que as sentenças se desenvolveram mais à direita do verbo, dentro do sintagma verbal e dos complementos.

Outro reflexo do aumento de palavras é a quantidade média de palavras para cada conectivo, que diminui na série histórica. Isso nos indica que o número de conectivos não cresce na mesma proporção do número de tokens, encaixando maiores quantidades de informação entre uma estratégia de conexão e outra, e, portanto, aumentando a complexidade coesiva. Buzato et al. (2021), ao analisar a ocorrência de operadores argumentativos em um corpus de redações de notas variadas, constatou alta frequência de uso de poucos operadores. Uma futura comparação poderia levar em conta os resultados de Buzato et al. (2021), na tentativa de encontrar aproximações e distanciamentos entre os dois corpora.

O índice de leiturabilidade Flesch, por sua vez, foi a métrica utilizada para verificar, sob um ponto de vista amplo, a complexidade de cada um dos textos. Como mostramos acima, acreditamos que boa parte dos comportamentos das métricas podem ser explicados, numa primeira análise, pelo aumento do número de palavras, embora esse fenômeno possa ter criado outras particularidades linguísticas que não foram captadas pelas medidas realizadas neste trabalho. Isso não é diferente no caso da leiturabilidade, já que, ao considerar os valores de *MPS* e *MSP* — dois valores que tendem a crescer com o aumento de tokens no texto —, é esperado que seu índice diminua ao longo dos anos analisados.

No sentido de tecer possíveis explicações para a diminuição drástica no número de redações nota mil entre os anos de 2011 e 2013, seria necessária, também, i) uma análise das métricas em conjunto com a evolução dos critérios de correção ao longo dos anos e ii) uma maior compilação de textos englobando, inclusive, redações de diferentes níveis. Sob o ponto de vista da logística das correções, as discrepâncias aceitas entre as notas dos dois primeiros corretores passaram de até 300 (trezentos) pontos, em 2011, para 200 (duzentos) pontos, em 2012, e, por fim, 100 (cem) pontos em 2013, valor considerado até os dias de hoje. Dessa forma, é possível que, com a diminuição

progressiva da tolerância, os textos passem a ser mais re-corrigidos, o que diminuiria a probabilidade de haver um elevado número de redações nota máxima. Embora a causalidade entre as duas variáveis não possa ser comprovada neste trabalho, é possível que a diminuição das notas mil não esteja relacionada com fatores puramente linguísticos, mas com elementos relativos à metodologia de correção.

## 4. Considerações finais

Este trabalho teve como objetivo fazer uma breve descrição de um corpus de redações nota mil a partir de métricas textuais, procurando tendências linguísticas ao longo do período de 2010 a 2020, em que houve a drástica diminuição na atribuição de notas máximas bela banca de correções do Enem. Neste momento, ainda não é possível afirmar, com precisão, os motivos de tal diminuição, mas são observáveis diferenças quantitativas entre textos considerados exemplares ao longo da série. Para isso, lançamos mão de ferramentas computacionais, dentre elas, o analisador de complexidade textual NILC-Metrix (Leal et al., 2022), disponível gratuitamente na internet. Como resultados encontrados, pudemos ressaltar o aumento de palavras e, de um modo geral, o aumento da complexidade textual, reportada pelas métricas de conectivos e de leiturabilidade.

Na esteira contemporânea da Linguística de Corpus, objetivamos, em momento oportuno, a disponibilização dos textos, que constituirão o primeiro corpus de redações nota mil avaliadas pela banca de corretores do Enem.

## Agradecimentos

Este trabalho foi desenvolvido durante a disciplina de Linguística de Corpus e Computacional, da pós-graduação em Linguística da Universidade Federal de Minas Gerais. Agradeço à equipe do NILC pela disponibilização da plataforma NILC-Metrix, à Dr[a]. Heliana Mello pela oferta da disciplina e à revisão da Linguamática.

## Referências

Alexopoulou, Theodora, Marije Michel, Akira Murakami & Detmar Meurers. 2017. Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning* 67(S1). 180–208. 10.1111/lang.12232.

Amorim, Evelin & Adriano Veloso. 2017. A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. Em *Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 94–102.

Berber Sardinha, Tony. 2000. Linguística de corpus: histórico e problemática. *D.E.L.T.A* 16(2). 323–367. 10.1590/S0102-44502000000200005.

Bertucci, Roberlei Alves, Andréa Jacqueline Malheiros & Wanderlei de Souza Lopes. 2020. Ocorrências de anáforas encapsuladoras em redações do enem. *Filologia e Linguística Portuguesa* 22(1). 81–102. 10.11606/issn.2176-9419.v22i1p81-102.

Bick, Eckhard. 2000. *The parsing system PALAVRAS: Automatic grammatical analysis of Portuguese in a constraint grammar framework*. Aarhus University Press.

Branco, António, João Rodrigues, Francisco Costa, João Silva & Rui Vaz. 2014. Rolling out text categorization for language learning assessment supported by language technology. Em *Computational Processing of the Portuguese Language (PROPOR)*, 256–261. 10.1007/978-3-319-09761-9_29.

Brasil. 2020. *A redação do Enem 2020: cartilha do participante*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Brasília, DF.

Buzato, Dalmo, Elias Victor de Jesus Cardoso Machado, Priscilla Tulipa da Costa & Suélen Érica Costa Silva. 2021. Operadores argumentativos em redações modelo enem: uma análise baseada em corpus. *Scientia Prima* 7(1). e97.

Cançado, Márcia, Luana Amaral, Amorin Evelin, Veloso Adriana & Heliana Mello. 2020. Subjetividade em correções de redações: detecção automática através de léxico de operadores de viés linguístico. *Linguamática* 12(1). 63–79. 10.21814/lm.12.1.313.

Crossley, Scoot A. & Danielle S. McNamara. 2012. Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35(2). 115–135. 10.1111/j.1467-9817.2010.01449.x.

Cruz, Daniel Ribeiro da, Rafaela Gonçalves Ulian, Robson Faleiros Ribeiro & Sheila Fernandes Pimenta e Oliveira. 2021. REDAÇÃO NOTA 1000: argumentos de propostas de intervenções em redações do ENEM 2009–2018. *Revista Eletrônica de Letras* 14(14). on–line.

Fonseca, Erick, Ivo Medeiros, Dayse Kamikawachi & Alessandro Bokan. 2018. Automatically grading Brazilian student essays. Em *Computational Processing of the Portuguese Language (PROPOR)*, 170–179.

Fonseca, Erick Rocha & João Luís G. Rosa. 2013. Mac-Morpho revisited: Towards robust part-of-speech tagging. Em *9th Brazilian Symposium in Information and Human Language Technology (STIL)*, 98–107.

Graesser, Arthur C., Danielle S. McNamara, Max M. Lowerse & Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36. 193–202. 10.3758/BF03195564.

Jackendoff, Ray. 1982. X syntax: A study of phrase structure. *Journal of Linguistics* 18. 409–497.

Leal, Sidney Evaldo, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann & Sandra Maria Aluísio. 2022. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. ArXiv cs.CL.

Marinho, Jeziel, Rafael Anchiêta & Raimundo Moura. 2021. Essay-BR: a brazilian corpus of essays. Em *Anais do III Dataset Showcase Workshop*, 53–64. 10.5753/dsw.2021.17414.

Silva, João, António Branco, Sérgio Castro & Ruben Reis. 2010. Out-of-the-box robust parsing of Portuguese. Em *Computational Processing of the Portuguese Language (PROPOR)*, 75–85.

Westerlund, Marcus. 2019. *Correlations between textual features and grades on the Swedish national exam in English: A Coh-Metrix analysis*: Stockholms Universitet. Tese de Mestrado.