Caracterização de desvios sintáticos em um *corpus* de redações O processo de anotação

Renata Ramisch[1], **Ariani Di Felippo**[1]

[1]Departamento de Letras
Universidade Federal de São Carlos (UFSCar)
Núcleo Interinstitucional de Linguística Computacional (NILC)
São Carlos – SP – Brasil

{renata.ramisch,arianidf}@gmail.com

Abstract. *This article describes the annotation of syntactic errors in essays written by High School students in Brazilian Portuguese (BP). Using a typology of syntactic errors based on the formal written style of BP, we annotated a set of sentences from the essays, which reveals that absence of punctuation marks and verbal agreement are the most common errors. Such annotation may contribute to refine the computational treatment of these violations.*

Resumo. *Descreve-se a anotação de desvios sintáticos em redações nos moldes do ENEM, escritas por estudantes do ensino médio. A partir de uma tipologia de desvios da modalidade escrita formal do português brasileiro, um conjunto de sentenças foi anotado, relevando que a ausência de pontuação e a concordância verbal são os desvios mais comuns. Tal anotação pode refinar o tratamento computacional dessas violações.*

1. Introdução

Escrever redações é um processo inerente à trajetória educacional. A redação também é frequentemente utilizada como mecanismo de avaliação dos conhecimentos de português e de produção textual em vestibulares e exames de seleção, como o Exame Nacional do Ensino Médio[1] (ENEM). Assim, um bom desempenho nessa tarefa garante melhores notas e, por consequência, aumenta as chances na disputa pelas vagas mais concorridas para o ensino superior. Porém, textos escritos por estudantes, mesmo na etapa final da educação básica, ainda apresentam diversos desvios de ortografia e gramática quanto à modalidade escrita esperada pelos avaliadores desses exames de seleção [Castaldo 2009].

Nesse sentido, algumas aplicações computacionais podem ser úteis no processo de correção e avaliação dos textos por professores ou avaliadores, assim como no aperfeiçoamento das habilidades de produção textual pelos próprios alunos. Exemplos de ferramentas que podem ser usadas pelos estudantes são os corretores ou revisores gramaticais (isto é, sistemas que detectam desvios gramaticais em um texto e sugerem correções [Soni e Thakur 2018]) e as ferramentas de auxílio à escrita (FAE), que dão suporte a todo o processo de escrita, seja no agrupamento de ideias ou na composição do texto. Um exemplo de FAE que auxilia na composição de textos acadêmicos em português é o SciPo

[1]O ENEM é uma prova realizada pelo INEP/MEC para avaliar a qualidade do ensino médio no país e dar acesso ao ensino superior em universidades públicas brasileiras e em algumas universidades estrangeiras.

[Feltrim 2004], um ambiente na *web* composto por um conjunto de ferramentas integradas para auxiliar estudantes a escreverem resumos e introduções de textos da área da computação[2].

Além disso, como o processo de avaliação e correção manual dessas redações costuma ser longo e caro, tem crescido o interesse pelo desenvolvimento de aplicações computacionais que possam agilizar também a correção e/ou a avaliação humana. Um exemplo de trabalhos nessa área é o de [Santos et al. 2016], que propõem um analisador léxico-sintático para a avaliação automática de atividades escritas em português. No experimento, eles utilizaram 20 textos e identificaram desvios que não haviam sido marcados pelos corretores humanos.

Sistemas que buscam realizar o processamento automático de uma língua natural requerem uma série de ferramentas, como os *part-of-speech (POS) taggers* (etiquetadores morfossintáticos) e os *parsers* (analisadores sintáticos). Para que seja possível utilizar tais ferramentas e aplicações também para analisar redações escritas por estudantes, é necessário que elas sejam capazes de lidar com textos que apresentam desvios de escrita. Essa tarefa constitui um desafio para a área de processamento de língua natural (PLN), uma vez que esses desvios são de vários tipos (p.ex., pontuação, concordância, regência, crase, etc.), e a ocorrência de alguns desses tipos costuma ser pouco frequente.

Para subsidiar o desenvolvimento de tais aplicações, os *corpora* anotados são recursos importantes, pois permitem modelar computacionalmente os fenômenos e/ou as tarefas linguísticas, além de treinar e avaliar tais modelagens. Esses *corpora* podem então ser utilizados como base para estudos linguísticos desses fenômenos, bem como para a construção de ferramentas como *POS taggers* e *parsers* e para aplicações de apoio à escrita e de correção/avaliação automática de textos.

Neste artigo, descreve-se a anotação manual de desvios sintáticos em um *corpus* de redações nos moldes do ENEM, escritas por estudantes do ensino médio. O objetivo é caracterizar os desvios sintáticos desses textos de forma a gerar descrições linguísticas que possam subsidiar o refinamento do tratamento computacional dessas violações.

2. A construção do *corpus* de redações

A construção de *corpora* de aprendizes é útil para a análise da linguagem utilizada, a avaliação de ferramentas de PLN e o desenvolvimento de sistemas de correção de desvios gramaticais [Köhn e Köhn 2018]. Neste estudo, focam-se os desvios sintáticos em redações de falantes nativos de português do Brasil, mas aprendizes da modalidade escrita formal da língua. Para definir o conceito de "desvio sintático", utiliza-se essa mesma modalidade, que também é adotada como critério de avaliação no ENEM, conforme consta na Cartilha do Participante [Brasil 2018]. Segundo a Cartilha, a avaliação das redações se divide em cinco competências, sendo que a Competência 1 avalia o domínio das convenções de escrita e a estrutura sintática, que deve estar adequada às regras gramaticais e à fluidez de leitura.

Da mesma maneira, as ferramentas de PLN também são desenvolvidas com base na modalidade escrita, a partir de uma perspectiva mais tradicional da língua. Logo, as análises linguísticas automáticas se apoiam, em grande medida, na gramática tradicional.

[2]http://www.nilc.icmc.usp.br/scipo/

Como esta tarefa de anotação leva em conta tanto as noções estabelecidas pelo ENEM quanto a abordagem das ferramentas de PLN, um desvio sintático é definido aqui como todo aquele relacionado a problemas na organização das palavras e suas combinações, de acordo com a modalidade escrita formal do português, podendo ser de ordem, de concordância e de dependência entre palavras.

Para analisar os desvios sintáticos presentes em textos de estudantes do ensino médio, construiu-se um *corpus* de 1.045 redações dissertativo-argumentativas (já em formato digital) nos moldes do ENEM, fornecidas pela empresa *Letrus*[3], um centro de tecnologia e letramento que desenvolve ferramentas de escrita e avaliação de textos para escolas. O *corpus* possui 10.653 sentenças, totalizando 325.111 palavras e 184.967 *types* (palavras únicas).

A construção do *corpus* seguiu as etapas de [Aluísio e Almeida 2006]: i) projeto do *corpus* (seleção dos textos); ii) compilação, manipulação, nomeação dos arquivos; iii) anotação. Antes da anotação, procedeu-se à (i) limpeza do *corpus*, (ii) segmentação das redações em sentenças, visto que a anotação se deu em nível sentencial, e (iii) correção ortográfica (via MS Word®), a fim de manter o foco de atenção do anotador nos desvios sintáticos (foram corrigidos apenas os desvios ortográficos identificados automaticamente por essa ferramenta). O arquivo original foi preservado, efetuando-se todas as alterações em arquivo específico. Na sequência, selecionou-se a parcela do *corpus* que seria anotada (6.000 sentenças) e construiu-se uma diretriz de anotação, que engloba uma tipologia de desvios sintáticos, exemplos e orientações gerais e específicas[4].

3. Metodologia de anotação de desvios sintáticos

Observando-se as questões de [Hovy e Lavid 2010] sobre anotação linguística, a tarefa aqui descrita teve início com a adaptação da tipologia de desvios gramaticais de [Pinheiro 2008] aos dados do *corpus*, resultando em 11 categorias e 27 subcategorias, organizadas conforme a Tabela 1.

Para evitar a sobreposição de categorias, estabeleceu-se uma regra hierárquica de anotação. Assim, a categoria crase, por exemplo, é hierarquicamente superior à regência, o que determinou que problemas de crase relacionados a regência fossem anotados apenas em uma das subcategorias de crase.

A escolha das categorias a serem anotadas se deu em função da noção de desvio e das alterações nas estruturas das árvores sintáticas das sentenças causadas pela presença de desvios. Nesse sentido, por exemplo, os desvios que alteram a classe morfossintática das sentenças (ou a interface entre a ortografia e a sintaxe) tornam a etiquetação automática via *POS taggers* difícil, já que essas ferramentas em geral consideram o contexto em que as palavras ocorrem para atribuir etiquetas. Assim, se a classe morfossintática de uma palavra é identificada equivocadamente devido a problemas de ortografia (p. ex. a ausência ou o excesso de acento no par *esta*/*está*), as etiquetação das palavras que a circundam provavelmente também terá problemas.

Da mesma forma, a presença ou ausência de determinadas palavras (tanto grama-

[3]https://www.letrus.com.br/

[4]Tanto o *corpus* como a diretriz de anotação poderão ser disponibilizados mediante contato com as autoras.

Tabela 1. Tipologia de desvios sintáticos

Categoria	Subcategoria	Descrição
01 - Pontuação	Ausência (pont-aus)	Ausência de pontuação em casos obrigatórios (p. ex. em adjuntos adverbiais deslocados).
	Excesso (pont-exc)	Ocorrência de pontuação em lugares não permitidos, como separação de sujeito e verbo com vírgula.
	Uso inadequado (pont-desv)	Uso inadequado de um sinal de pontuação no lugar de outro (p. ex. aglutinações de sentenças por vírgulas).
02 - Crase	Ausência (crase-aus)	Falta de crase quando a sua ocorrência é obrigatória.
	Excesso (crase-exc)	Uso da crase quando ela não é permitida.
03 - Regência	Verbal (rege-verb)	Ausência, excesso ou uso inadequado de preposições quando o termo regente é um verbo.
	Nominal (rege-nom)	Ausência, excesso ou uso inadequado de preposições quando o termo regente é um substantivo, adjetivo ou advérbio.
04 - Concordância	Verbal (concor-verb)	Problemas de concordância entre sujeito e verbo.
	Nominal (concor-nom)	Problemas de concordância entre adjetivo, artigo, etc. e os termos a que se referem (substantivo ou pronome).
	Anafórica (concor-anaf)	Retomada equivocada de elementos citados na sentença, mas cujo retomador não concorda com o retomado.
05 - Pronomes	Colocação (pronom-col)	Colocação irregular dos pronomes em termos de posição na sentença (uso de próclise/ênclise).
	Ausência (pronom-aus)	Casos de ausência de qualquer tipo de pronome quando a sua ocorrência é obrigatória.
	Excesso (pronom-exc)	Ocorrência excessiva de pronome (p. ex. retomada do sujeito por meio de pronome pessoal).
	Uso inadequado (pronom-desv)	Utilização inadequada de pronomes, como uso de *cujo*/*cuja* sem valor de retomada.
06 - Preposições	Ausência (prepo-aus)	Ausência de preposição obrigatória não ligada a regência (p. ex. ausência de preposições em locução).
	Excesso (prepo-exc)	Excesso ou repetição de preposições (p. ex. *mediante a* ou *muitas das vezes*).
	Uso inadequado (determ-desv)	Uso inadequado de preposição ou contração (p. ex. uso de contração quando a estrutura exige a forma não contraída).
07 - Determinantes	Ausência (determ-aus)	Falta de determinante (p. ex. em casos de paralelismo obrigatório).
	Excesso (determ-exc)	Uso duplicado/excessivo de determinantes (p. ex. *cujo o*).
	Uso inadequado (determ-desv)	Ocorrência inadequada de determinantes (pouco frequente).
08 - Conjunções	Uso inadequado (conjunc)	Ausência, excesso ou uso inadequado de conjunções (p. ex. uso inadequado dos porquês, *mas porém*).
09 - Formas verbais	Uso equivocado de formas verbais (verbo-mod)	Ocorrência de desvios de formas, tempos, modos verbais (p. ex. uso de indicativo em vez de subjuntivo).
	Uso equivocado de formas nominais (verbo-nom)	Uso equivocado das formas nominais gerúndio, particípio e infinitivo.
10 - Segmentação	Segmentação inadequada de sentenças (segment)	Sentenças que foram segmentadas, mas deveriam estar ligadas à anterior (p. ex. que comecem por *assim como*).
11 - Outros	Ordem (ordem)	Ordem equivocada, ausência ou excesso de palavras ou grupos de palavras de conteúdo (não abarcadas pelas demais categorias).
	Interface ortografia-sintaxe (orto-sin)	Problemas de ortografia que alterem a classe morfossintática da palavra, influenciando na sintaxe.
	Sem especificação (sem-espec)	Desvios que não se encaixem em nenhuma das categorias anteriores.

ticais quanto de conteúdo) e os desvios ligados às formas verbais geram dificuldades para um *parser* encontrar as relações corretas entre os elementos de uma sentença. Portanto, uma vez que a anotação tem como objetivo gerar subsídios para o desenvolvimento e o aprimoramento de tais ferramentas, é importante que esses fenômenos façam parte do esquema de anotação.

Estabelecida a tipologia, a tarefa de anotação se deu em duas fases: classificação das sentenças em "com desvio" e "sem desvio"; e tipificação dos desvios presentes em parte das sentenças com desvio. Na primeira fase, classificaram-se 6.000 sentenças (56,3% do *corpus*) em "sem" e "com desvio". Em um arquivo *xls* composto por três colunas, a classificação consistiu em apenas identificar se cada sentença possuía (ou não) ao menos um desvio sintático. As duas primeiras colunas do arquivo codificavam o ID de uma sentença e o respectivo texto, e a terceira coluna era a da anotação, que foi realizada por meio da atribuição de uma das *tags*: N (= sem desvio) e D (= com desvio).

Na segunda fase da anotação, 2.500 sentenças classificadas como D tiveram seus desvios categorizados conforme a tipologia, o que foi feito por meio da plataforma de anotação FLAT (*FoLia Linguistic Annotation Tool*) [Gompel e Reynaert 2013]. A caracterização consistiu em delimitar o segmento sentencial relativo ao desvio e anotá-lo com a etiqueta correspondente à sua subcategoria, seguindo as diretrizes de anotação. Os desvios caracterizados pela ausência de um elemento foram ancorados no *token* imediatamente anterior à posição em que o elemento ausente deveria ocorrer, com exceção dos casos de regência, cuja anotação foi associada ao termo regente.

4. Resultados da anotação

A Tabela 2 apresenta os resultados da primeira fase, contendo o número de sentenças com e sem desvio nas 6.000 sentenças anotadas (56,3% do *corpus*), e os respectivos percentuais.

Tabela 2. Número de sentenças com e sem desvio

	Nº sentenças	**Percentual (%)**
Contém desvio	4.409	73,48
Não contém desvio	1.591	26,52

A presença significativa de sentenças com desvio justifica a necessidade de tais descrições linguísticas. A segunda fase identificou os tipos de desvios das categorias e subcategorias mais e menos frequentes, chegando a um total de 7.290 desvios. Os desvios por categoria se distribuem como mostra a Tabela 3, por ordem de frequência.

Os desvios mais frequentes são os de pontuação e de concordância. As categorias menos frequentes são as de uso de conjunções e de determinantes. Em função da estrutura da tipologia, é preciso analisar também as subcategorias para que seja possível estabelecer melhor os padrões de desvios encontrados. A Tabela 4 mostra a distribuição dos desvios por subcategoria, ordenados por frequência.

Analisando as subcategorias, vê-se que há maior ocorrência de desvios de ausência de pontuação do que de excesso ou de uso inadequado desses sinais. Já nas subcategorias de concordância, os desvios de concordância verbal são quase duas vezes mais frequentes

Tabela 3. Distribuição dos desvios sintáticos por tipo: categoria

Categoria	Nº desvios
01 - Pontuação	3.224
04 - Concordância	1.378
09 - Formas verbais	500
05 - Pronomes	422
06 - Preposições	418
02 - Crase	312
10 - Segmentação	303
03 - Regência	250
11 - Outros	168
08 - Conjunções	167
07 - Determinantes	148

Tabela 4. Distribuição dos desvios sintáticos por tipo: subcategoria

Subcategoria	Nº desvios
01.2-pont-exc	726
01.3-pont-desv	614
04.2-concor-nom	455
10.1-segment	303
09.1-verbo-mod	300
05.4-pronom-desv	246
02.1-crase-aus	229
03.1-rege-verb	208
09.2-verbo-nom	200
08.1-conjunc	167
06.2-prepo-exc	160
06.3-prepo-desv	142
11.2-orto-sin	135
06.1-prepo-aus	116
05.3-pronom-exc	106
02.2-crase-exc	83
07.1-determ-aus	81
07.2-determ-exc	61
04.3-concor-anaf	48
03.2-rege-nom	42
05.1-pronom-col	36
05.2-pronom-aus	34
11.1-ordem	33
07.3-determ-desv	6
11.3-sem-espec	-

que as outras subcategorias de concordância. Além disso, como era esperado, a subcategoria de uso inadequado de determinantes foi a menos frequente (excetuando-se a de desvios sem especificação, que deveria ser usada apenas em casos que realmente não se inseriam em nenhuma das outras subcategorias definidas, e que não foi aplicada a nenhum desvio).

A segunda menos frequente foi a subcategoria relacionada à ordem de palavras (isto é, a estruturas cuja ordenação das palavras está equivocada), repetição, ausência ou excesso de palavras de conteúdo. Essa pouca ocorrência de problemas de ordem provavelmente está associada ao fato de as redações terem sido escritas por falantes nativos, que já têm internalizada a ordem adequada de elementos na sentença.

Vê-se que a subcategoria de uso de conjunções é a décima mais frequente, sendo que a categoria referente a ela estava entre as menos frequentes. Isso se dá porque a categoria é composta de uma única subcategoria, isto é, todos os desvios da categoria "08 - Conjunções" também são os que aparecem na respectiva subcategoria. O mesmo ocorre com a categoria "10 - Segmentação".

5. Considerações finais

Neste artigo, apresentou-se a tarefa de anotação de desvios sintáticos presentes em redações de estudantes do ensino médio nos moldes do ENEM. O *corpus* anotado busca servir de subsídio para o desenvolvimento de ferramentas de PLN capazes de lidar com textos que tenham como característica a presença de desvios sintáticos. A partir da tipologia proposta e da metodologia de anotação estabelecida, observou-se que os textos do *corpus* apresentam muitos desvios, sendo que os mais frequentes são de pontuação (principalmente a sua ausência) e de concordância, com ênfase para a concordância verbal.

Agradecimento. À CAPES, pelo suporte financeiro.

Referências

Aluísio, S. M. e Almeida, G. M. d. B. (2006). O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. *Calidoscópio*, 4(3):156–178.

Brasil (2018). *Redação no ENEM 2018: Cartilha do Participante*. INEP/MEC, Brasília.

Castaldo, M. M. (2009). *Redação no vestibular: a língua cindida*. Tese (doutorado em educação), Universidade de São Paulo.

Feltrim, V. D. (2004). *Uma abordagem baseada em corpus e em sistemas de crítica para a construção de ambientes Web de auxilio à escrita acadêmica em português*. Tese (doutorado em computação), Universidade de São Paulo.

Gompel, M. v. e Reynaert, M. (2013). FoLiA: A practical XML format for linguistic annotation – a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81.

Hovy, E. e Lavid, J. (2010). Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation Studies*, 22:13–36.