Programa de Pós-Graduação em Linguística

PPGL

CARACTERIZAÇÃO DE DESVIOS SINTÁTICOS EM REDAÇÕES DE ESTUDANTES DO ENSINO MÉDIO: subsídios para o Processamento Automático das Línguas Naturais

Renata Ramisch

SÃO CARLOS

2020

Universidade Federal de São Carlos

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

CARACTERIZAÇÃO DE DESVIOS SINTÁTICOS EM REDAÇÕES DE ESTUDANTES DO ENSINO MÉDIO: subsídios para o Processamento Automático das Línguas Naturais

Renata Ramisch
Bolsista CAPES

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Linguística.

Área de concentração: Descrição, análise e processamento automático das línguas naturais

Orientadora: Profa. Dra. Ariani Di Felippo

São Carlos – São Paulo – Brasil
2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Linguística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado da candidata Renata Ramisch, realizada em 27/03/2020:

Profa. Dra. Ariani Di Felippo
UFSCar

Profa. Dra. Sandra Maria Aluísio
USP

Profa. Dra. Livy Maria Real Coelho
B2W Digital

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Sandra Maria Aluísio, Livy Maria Real Coelho e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Profa. Dra. Ariani Di Felippo

*Dedico esta dissertação à minha avó, Gisela Ponath (*in memoriam*), que foi a principal responsável por eu ter aprendido a falar alemão antes mesmo do português, o que abriu as portas para toda a minha trajetória acadêmica e profissional.*

AGRADECIMENTOS

Durante a minha formação acadêmica, aprendi que textos científicos precisam ser impessoais, descrevendo fatos com rigor e evitando linguagem subjetiva. Lemos um texto acadêmico como se o caminho que levou até ele fosse sereno. Não ficam evidentes as dores, as alegrias, os sofrimentos e as conquistas, porque sentimentos precisam ser ocultados. Mas nesse espaço de agradecimentos, olho para trás e tento traduzir em palavras os sentimentos de gratidão com as pessoas que me ajudaram a trilhar esse caminho.

Pai e mãe, vocês muitas vezes disseram que eu era inteligente, incentivaram o meu estudo, falaram da importância da educação. Na minha opinião, mãe, inteligente mesmo é quem sabe fazer blusão de tricô, quem dedica o seu tempo a fazer sapatos e cobertas de bebê para doação, quem se doa de corpo e alma para cuidar de outras pessoas. Inteligente mesmo, pai, é quem sabe consertar um registro, entende de construção civil como ninguém, tem o jardim mais lindo da cidade, arranja soluções criativas para qualquer problema. Para mim, inteligente é quem apoia e incentiva, mesmo que às vezes discorde das decisões tomadas ou não entenda o porquê dos caminhos escolhidos. Palavras não são suficientes para demonstrar a minha gratidão, então fica aqui o meu imenso muito obrigada pelo amor incondicional.

Carlos, meu querido maninho, essa jornada não teria sido possível sem você. Nas nossas semelhanças, descobrimos singularidades. Nas nossas discordâncias, concordamos no respeito mútuo à trajetória, às escolhas, aos valores de cada um. Obrigada pela presença que os mais de 8 mil quilômetros e um oceano inteiro nunca conseguiram distanciar. Obrigada pelo incentivo, pelo apoio, pelas caminhadas em conjunto (literais e metafóricas), pelas ajudas, consultorias, ideias, revisões. Todos os léxicos de todas as línguas conhecidas e desconhecidas nesse planeta (e em outros, quem sabe) não são suficientes para expressar o tamanho do amor.

Guilherme, obrigada pelas pontes que construímos ao longo dos anos. Agradeço por confiar em mim e me incentivar, por respeitar as minhas escolhas, por me fazer relembrar que há vida para além do mestrado, por chamar sempre a minha atenção para os detalhes belos desse mundo. Teria sido muito mais difícil sem a conexão tão intensa que estabelecemos nas várias horas de conversas no Skype, telefone, WhatsApp, e na intensidade dos momentos de presença.

Oma e Opa, vocês infelizmente não puderam estar presentes entre nós até o fim dessa jornada. Opa, como eu prometi na nossa última conversa, estou trazendo mais um diploma para casa. Oma, cada incentivo, biscoito, doce e comida que eu gostava nas voltas ao Sul, cada ligação no meio da semana para perguntar como eu estava me deram forças para concluir essa

etapa. Onde quer que vocês dois estejam, saibam que vocês foram fundamentais em todo o caminho. O meu eterno muito obrigada!

Um grande agradecimento à Nice, ao Marcos e aos meus dois sobrinhos lindos, Ana e Mateus, que me trazem uma alegria imensa e uma saudade enorme. À Rebeca, que me apoiou e me ouviu em tantas horas de áudios e conversas no Skype, mesmo estando literalmente do outro lado do mundo. À Romara e ao Augusto, por terem me recebido em sua casa tantas vezes durante esses dois anos, mantendo as conexões estabelecidas há muito tempo e facilitando imensamente a minha mudança para este estado.

Muito obrigada também aos queridos amigos que eu conheci em São Carlos e que me acolheram, ajudaram, deram muitas e muitas caronas, explicaram a cidade, mostraram diversos restaurantes e cafés, zoaram o meu sotaque e me ajudaram a identificar o que era "gauchês". Aos meus queridos "irmãos de orientação" e amigos do PPGL, muito obrigada mesmo por essa amizade, pelas trocas, pelo apoio, e por tornarem a chegada a uma cidade nova e a vida longe da família muito menos dolorosas.

Aos amigos do NILC, com quem eu aprendi muito e partilhei momentos de descontração, obrigada por serem essas pessoas tão divertidas. Teria sido bem mais difícil sem a amizade de todos vocês! Um obrigada mais que especial pelas horas gastas me explicando como usar ferramentas computacionais, o que são APIs ou VMs, como usar o Linux, como programar em Python, como fazer os *scripts* de que eu precisava — e quando acabava a paciência, por fazerem os códigos para mim em tempo recorde.

Ariani, orientadora deste percurso, um agradecimento primeiramente por ter aceitado o pedido de orientação vindo de última hora, em um tema que de início não lhe era familiar. Agradeço pela confiança que você depositou em mim ao me dar autonomia para tomar decisões, e pela condução de todo o processo, oferecendo suporte e relembrando a necessidade de limitar o escopo da pesquisa sempre que eu me empolgava demais e queria abraçar o mundo.

Agradeço pelos inúmeros aprendizados valiosos, dicas, sugestões, questionamentos dos professores do PPGL e do ICMC, especialmente aos Professores Oto Vale e Thiago Pardo (que é um dos grandes responsáveis pela minha vinda a Sanca, ao me indicar a Ariani), às Professoras Graça Nunes e Helena Caseli. Um obrigada mais que especial ainda à Professora Sandra Aluísio, por aceitar fazer parte da banca de qualificação e de defesa deste mestrado. Agradeço também à Amanda pelas contribuições valiosas na banca de qualificação, e à Livy, por aceitar compor a banca de defesa.

Por último, agradeço à CAPES pelo suporte financeiro e por ter honrado o seu compromisso num tempo de política nada favorável à ciência e à universidade pública.

RESUMO

Escrever redações é um processo inerente à trajetória educacional, do qual depende o bom desempenho em exames de admissão no Ensino Superior. No entanto, desvios da modalidade escrita padrão do português são bastante frequentes, indo de questões de ortografia e gramática até a estrutura textual e discursiva. A presente pesquisa investigou especificamente a recorrência de desvios de natureza sintática e as suas eventuais correlações com determinados atributos linguísticos das sentenças. Para isso, construiu-se um *corpus* composto por 1.045 redações nos moldes do ENEM, escritas por estudantes do Ensino Médio, o qual foi segmentado em um *subcorpus* de 10.652 sentenças. Esse *subcorpus* foi dividido novamente em *corpus* de treino (8.654 sentenças) e *corpus* de teste (1.998 sentenças). Estabeleceu-se um esquema de anotação manual em duas fases: classificação de sentenças em contendo ou não desvio sintático, e tipificação de desvios em 2.500 sentenças com base em uma tipologia de 11 categorias e 27 subcategorias. A anotação revelou que 73,34% das sentenças anotadas contêm desvios (6.347 sentenças do *corpus* de treino, e 1.425 do *corpus* de teste), e o restante não contém desvio (2.307 sentenças do *corpus* de treino, e 573 do *corpus* de teste). As categorias mais frequentes entre os 7.290 desvios identificados são as de pontuação (44%) e concordância (18,9%). Na sequência, realizou-se a análise linguística qualitativa abrangente dos fenômenos nos quais os desvios ocorrem. Essa análise foi dividida entre fenômenos específicos da sintaxe, como inversões da ordem canônica, coordenação, subordinação, entre outros; e fenômenos de outros níveis linguísticos, como desvios de acentuação, estruturas com verbo-suporte e problemas com o uso do verbo *haver*. O *corpus* também foi anotado automaticamente com o *parser UDPipe* e, a partir dos arquivos de saída, foram extraídos 17 atributos linguísticos, os quais foram correlacionados com a presença de desvios via Aprendizado de Máquina Supervisionado, utilizando o *software Weka*. O melhor resultado obtido no *corpus* de teste foi com o algoritmo *Logistic Regression* (75,62% de acurácia), e os atributos mais fortemente correlacionados com a presença de desvios, indicados pelos algoritmos de seleção de informações, foram o tamanho da sentença e a profundidade da árvore sintática. Como resultado adicional, construiu-se um recurso linguístico-computacional que pode ser útil para sistemas de Processamento Automático das Línguas Naturais. O potencial objetivo dessa parceria é o desenvolvimento de ferramentas de auxílio à escrita que podem facilitar a identificação e a correção de desvios pelos próprios autores de redações.

Palavras-chave: Processamento Automático das Línguas Naturais. Redação escolar. Desvio sintático.

ABSTRACT

Writing essays is a common task for students during school education, and a good performance in this task guarantees better grades to compete for places in the best universities. However, deviations from the standard written Portuguese are quite frequent, ranging from spelling and grammar to textual and discursive structure. This research specifically investigated the recurrence of syntactic errors and their possible correlations with certain linguistic attributes of the sentences. For this purpose, we built a corpus of 1,045 essays following ENEM specifications, that were written by high school students and segmented into a subcorpus of 10,652 sentences. This subcorpus was again segmented into train corpus (8,654 sentences) and test corpus (1,998 sentences). We established a manual annotation scheme in two phases: classification of sentences in containing or not syntactic errors, and categorization of the errors in 2,500 sentences based on a typology of 11 categories and 27 subcategories. The annotation showed that 73.34% of the annotated sentences contain syntactic errors (6,347 sentences from train corpus and 1,425 from test corpus), and the rest of the sentences do not contain syntactic errors (2,307 sentences from train corpus and 573 sentences from the test corpus). The most frequent categories among the 7,290 errors are those of punctuation (44%) and agreement (18.9%). We also carried out an extensive qualitative linguistic analysis of the phenomena in which the errors occur. This analysis looked at specific syntactic phenomena such as inversions of the canonical word order, coordination, subordination, etc., and at the phenomena that stem from further linguistic levels, such as missing accents, light-verb constructions and the use of specific verbs. In addition, the corpus was automatically annotated with the parser *UDPipe*, and we extracted from its output 17 linguistic features, which we correlated with the presence of errors via Supervised Machine Learning, using the software *Weka*. We obtained the best result in the test corpus with the algorithm Logistic Regression (75.62% accuracy). The features that were most strongly correlated with the presence of errors, indicated by feature engineering algorithms, were the sentence size and the depth of the syntactic tree. As an additional result, we built a computational-linguistic resource that can be useful to Natural Language Processing systems. The potential goal of such partnership is the development of writing assistance tools that can facilitate the process of identifying and correcting errors made by the authors of the essays themselves.

Keywords: Natural Language Processing. Students essays. Syntactic errors.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

SUMÁRIO

1 INTRODUÇÃO

Escrever redações é uma tarefa comumente realizada pelos estudantes em sala de aula. No entanto, desenvolver essa habilidade mostra-se como um grande desafio, uma vez que a taxa de pessoas que saem da etapa final da Educação Básica sem terem cumprido os objetivos de aprendizagem esperados é significativa, como mostra o relatório sobre o alfabetismo no Brasil divulgado pelo Instituto Paulo Montenegro, em parceria com a Ação Educativa[1]. Segundo a pesquisa, "apenas 8% dos respondentes estão no último grupo de alfabetismo, revelando domínio de habilidades que praticamente não mais impõem restrições para compreender e interpretar textos em situações usuais" (LIMA; MASAGÃO; CATELLI JR., 2016, p. 7).

Além disso, a redação também é alvo de avaliação de diversos vestibulares e concursos públicos, a exemplo do Exame Nacional do Ensino Médio (ENEM). Um bom desempenho na redação ajuda a garantir melhores resultados e, com isso, permite o acesso às vagas mais concorridas de Ensino Superior nas melhores universidades. Porém, textos escritos por estudantes, mesmo na etapa final da Educação Básica, apresentam diversas formas de desvios, tanto de ortografia como de gramática e de estruturação de sentenças (CASTALDO, 2009).

Nesse sentido, considerando a presença massiva de tecnologias no cotidiano de aprendizes de todas as idades, desenvolver ferramentas que possam, de forma (semi)automática, avaliar redações, identificar desvios e oferecer propostas de reescrita de sentenças no texto caracteriza-se como um desafio para a área de estudos do Processamento Automático das Línguas Naturais (PLN). Porém, as ferramentas automáticas existentes, como etiquetadores morfossintáticos (em inglês, *part-of-speech* ou *POS taggers*) e analisadores sintáticos (em inglês, *parsers*), podem ter dificuldades de lidar com textos que contenham desvios. Assim, justifica-se a importância de conduzir pesquisas a partir de *corpora* constituídos por textos de aprendizes, de modo a compreender as principais dificuldades apresentadas pelos estudantes. Com base nisso, é possível propor recursos e ferramentas que auxiliem tanto professores quanto alunos na tarefa de escrever textos melhores.

De forma a oferecer subsídios aos sistemas de PLN existentes e ao desenvolvimento de novas ferramentas computacionais para o processamento de textos com essas características, é fundamental classificar e analisar os desvios. Logo, a construção e análise de *corpora* de aprendizes é útil em diversas tarefas (KÖHN; KÖHN, 2018), como a análise da linguagem de aprendizes, o desenvolvimento e a avaliação de ferramentas de PLN referentes a essa

[1] Ação Educativa é uma organização civil sem fins lucrativos que atua em prol da educação.

linguagem, e o desenvolvimento de sistemas de correção de desvios gramaticais. Assim, unindo as abordagens baseadas em *corpus* e o PLN, este trabalho se propõe a fazer uma descrição linguística dos desvios sintáticos a partir da produção textual de estudantes em formação, especificamente em nível de Ensino Médio.

1.1 *Hipóteses e objetivos*

Esta pesquisa parte da hipótese principal de que existe correlação entre determinadas construções sintáticas e a presença de desvios sintáticos em redações escritas por estudantes do Ensino Médio. A partir da experiência da pesquisadora com correção e avaliação de redações, verificou-se que os estudantes parecem ter dificuldades para construir sintaticamente tipos específicos de sentenças, por exemplo, aquelas com estruturas de subordinação. Focalizam-se apenas os desvios sintáticos porque a correção ortográfica automática (em inglês, *spell checking*) é uma aplicação mais investigada e que apresenta resultados satisfatórios, como se pode ver nos corretores ortográficos dos editores de texto, de celulares e das ferramentas da *web*. A pesquisa baseia-se nas seguintes hipóteses específicas:

- estruturas de coordenação e subordinação estão positivamente associadas à ocorrência de desvios;
- a voz passiva está positivamente associada à presença de desvios;
- sentenças com maior número de *tokens* têm maior probabilidade de conterem desvios sintáticos do que aquelas que têm menos *tokens*;
- sentenças cujas dependências sintáticas entre elementos são mais distantes (isto é, em que há vários *tokens* intercalados entre dois elementos dependentes diretamente um do outro) apresentam maior probabilidade de conterem desvios sintáticos do que aquelas em que os dependentes diretos são adjacentes entre si.

Para validar tais hipóteses, o objetivo principal desta pesquisa é investigar a recorrência de desvios sintáticos presentes em redações produzidas por estudantes do Ensino Médio, nos moldes exigidos pelo ENEM, bem como caracterizá-los. Além disso, pretende-se investigar a eventual correlação entre a presença de tais desvios e a ocorrência de determinadas construções sintáticas, representadas na forma de atributos linguísticos das sentenças. As descrições linguísticas geradas podem servir como recurso linguístico-computacional para subsidiar o desenvolvimento de ferramentas de PLN, especialmente as de *parsing*, que podem ser

implementadas em ferramentas de auxílio à escrita e sistemas de correção/avaliação automática de redações. Os objetivos específicos do estudo são os seguintes:

- identificar e caracterizar os desvios sintáticos mais frequentes em textos de estudantes;
- expandir a tipologia de desvios gramaticais proposta por Pinheiro (2008), que também analisou desvios em redações de estudantes, e propor uma tipologia de desvios sintáticos abrangente, que possa ser útil para outras tarefas de PLN;
- disponibilizar um *corpus* de redações nos moldes da redação do ENEM, anotado tanto manual quanto automaticamente (via *parsing*) no nível da sentença;
- analisar os fenômenos nos quais tipicamente ocorrem os desvios sintáticos encontrados durante a anotação;
- identificar e extrair atributos linguísticos das sentenças do *corpus*;
- analisar as correlações entre tais atributos linguísticos e a presença de desvios sintáticos, utilizando Aprendizado de Máquina (AM) e algoritmos de seleção de informações.

1.2 *Justificativa*

Esta pesquisa tem como principal motivação a experiência da autora na avaliação e revisão de textos escolares e acadêmicos. Durante a atuação profissional, percebeu-se uma dificuldade por parte dos alunos em realizar determinadas construções sintáticas de acordo com a modalidade escrita formal, assim como uma limitação dos corretores gramaticais disponíveis nos editores de textos, livres e comerciais (como *Libre Office* e *Microsoft® Word*), de identificar tais inadequações sintáticas. A dificuldade dos alunos pode indicar que a competência escrita não é completamente alcançada no Ensino Básico, sendo levada às produções textuais nos níveis superiores, quando não há mais o ensino regular de língua portuguesa (exceto nos cursos específicos da área das Linguagens).

No âmbito da pesquisa, diversas áreas da Linguística e do Ensino e Aprendizagem de Línguas têm como tema a redação em língua portuguesa e as suas questões em sala de aula (HERREIRA, 2000; OLIVEIRA; LOURA, 2017; SANTOS; MOTTA, 2017). No entanto, as possibilidades do uso do PLN e de suas ferramentas no ensino de redação ainda parecem ser pouco exploradas. Trabalhos como o de Pinheiro (2008) buscam analisar desvios textuais com base em abordagens da Linguística de *Corpus* (LC) e do PLN. Porém, parece haver espaço para aprofundar tais estudos partindo da produção textual dos estudantes em formação. Ademais, a descrição e a análise de desvios são fundamentais para a compreensão desse fenômeno,

contribuindo tanto para a pesquisa linguística quanto para a reflexão sobre os processos de ensino de produção textual.

Nesse sentido, a tecnologia surge como aliada na busca por alternativas para desenvolver a capacidade de produzir textos. Assim, o PLN vem como proposta para automatizar (ou semiautomatizar) a análise, a avaliação e as sugestões de ajustes dos textos, por meio do desenvolvimento de diversas ferramentas, como *parsers*, corretores ortográficos e gramaticais, ferramentas de auxílio à escrita, sistemas de correção/avaliação automática de redações, entre outras[2]. O estudo se justifica ainda como uma proposta inicial para tornar mais evidentes as relações entre estruturas sintáticas e presença ou ausência de desvios. Espera-se gerar recursos que sejam úteis como descrição linguística e como subsídios para a formação de produtores de texto e para o desenvolvimento de ferramentas de processamento de textos em língua natural.

1.3 *Metodologia*

Situa-se este trabalho no âmbito da descrição linguística, tomando como pontos de partida os pressupostos da análise linguística baseada em *corpus* e o ferramental do PLN. Trata-se de uma pesquisa quantitativa, porque quantifica os desvios que ocorrem nas redações dos estudantes. No entanto, para além da contagem de desvios, a pesquisa também tem cunho qualitativo, pois analisa e descreve os desvios e alguns dos fenômenos associados à sua ocorrência, buscando relacioná-los com características específicas das sentenças.

Assim, a partir de um *corpus* de 1.045 redações de simulados do ENEM fornecidas pela empresa Letrus[3], verificam-se as hipóteses de pesquisa nas seguintes etapas:

a) **construção e pré-processamento do *corpus*:** seleção dos textos, caracterização do *corpus* e levantamento das estatísticas básicas, como número de *types*, número de *tokens*, número de sentenças, bem como correção de problemas de espaços antes de sinais de pontuação, extração dos textos, conversões de codificação, segmentação das sentenças, *parsing* e preparação dos arquivos para a etapa de anotação.

[2] Dada a natureza linguística do estudo, o desenvolvimento dessas ferramentas é motivação para esta pesquisa, mas está fora do seu escopo.

[3] A pesquisa utilizou um *corpus* de redações de simulados para o ENEM previamente compilado pela empresa Letrus porque tais redações já estavam em formato digital, o que agilizou muito o processo de análise, e porque a utilização de redações oficiais do ENEM implica um processo longo e burocrático de autorização de uso junto ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). A pesquisadora não tem nenhum vínculo com a empresa, que autorizou a disponibilização das anotações e do *corpus* em si (com a condição de não serem disponibilizados os metadados fornecidos). Os resultados desta pesquisa estarão disponíveis publicamente após a publicação da dissertação. Com isso, afirma-se não haver qualquer conflito de interesses nesta pesquisa.

b) **anotação manual de desvios sintáticos:** identificação e etiquetação manual dos desvios em duas fases — classificação das sentenças em contendo ou não desvio sintático; e tipificação desses desvios conforme a tipologia proposta.

c) **extração de atributos linguísticos e correlação entre atributos e desvios com Aprendizado de Máquina:** identificação e extração automática dos atributos observados na anotação como relevantes, seguida da correlação via AM entre os atributos e a ocorrência de desvios sintáticos nas sentenças do *corpus*, e identificação dos atributos considerados mais decisivos para os algoritmos testados.

O primeiro capítulo desta dissertação apresentou os aspectos que servem como pano de fundo para o desenvolvimento da pesquisa. Identificaram-se o contexto e a problemática nos quais o trabalho se insere, elencaram-se as hipóteses e os objetivos, citou-se a justificativa para o desenvolvimento do trabalho, e apresentaram-se brevemente as etapas que constituem a abordagem metodológica proposta. No Capítulo 2, trazem-se os conceitos teóricos principais para que seja possível compreender e situar a pesquisa nos campos nos quais ela se insere: a descrição linguística e o PLN. No Capítulo 3, apresentam-se os trabalhos relacionados que deram subsídio para o desenvolvimento da presente investigação.

O Capítulo 4 dedica-se à descrição detalhada das etapas que compõem a metodologia proposta: a construção e o pré-processamento do *corpus*, as duas fases da anotação de desvios sintáticos, bem como as análises quantitativas dos desvios, a identificação de atributos linguísticos e as respectivas correlações via AM. O Capítulo 5 apresenta as análises linguísticas, que buscam sistematizar e descrever os fenômenos nos quais os desvios sintáticos ocorrem. Já o Capítulo 6 traz os resultados obtidos na extração dos atributos e no Aprendizado de Máquina. Por fim, o Capítulo 7 traz as conclusões que a pesquisa permitiu elaborar, assim como as principais limitações encontradas durante o desenvolvimento do trabalho. Esse capítulo cita ainda algumas das possibilidades de investigações futuras, de modo que novas questões possam ser exploradas a partir deste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Durante quase toda a sua existência, a análise linguística tem se baseado (em maior ou menor grau, dependendo da época e das correntes teóricas em evidência) em conjuntos de textos de uso real da língua para corroborar hipóteses formuladas previamente e/ou buscar por elas (ALUÍSIO; ALMEIDA, 2006). Porém, com a popularização massiva dos computadores e, mais recentemente, com a possibilidade de se obter na internet quantidades enormes de dados em língua natural, a maneira de fazer análise linguística mudou significativamente.

Assim, tornaram-se também cada vez mais populares as análises e as descrições de fenômenos com base em conjuntos de textos autênticos em língua natural tratáveis por computador: os *corpora*. Com eles, as áreas de estudos que lançavam mão do *corpus* como abordagem metodológica ou como pressuposto teórico ganharam espaço. A Linguística de *Corpus* é a disciplina que se ocupa mais especificamente desse objeto (BERBER-SARDINHA, 2004), mas os *corpora* têm sido usados também em pesquisas na área da tradução, na aprendizagem de língua estrangeira e materna, nos estudos terminológicos, entre outros.

Outra área que ganhou evidência com o surgimento, a popularização e a massificação do uso de computadores e da internet foi o PLN. Da mesma forma, hoje o uso de *corpora* por essa área multidisciplinar é muito frequente no desenvolvimento de diversas ferramentas e aplicações (HOVY; LAVID, 2010). *Corpora* bem construídos têm se mostrado valiosos, especialmente quando acompanhados de anotações em diversos níveis linguísticos. Alguns dos aspectos importantes na compilação de um *corpus* dizem respeito a características como gênero dos textos, modalidade da língua, tipologia e finalidade de uso.

Nesse sentido, visto que esta pesquisa busca descrever fenômenos linguísticos a partir de um *corpus* para gerar subsídios ao desenvolvimento de aplicações em PLN, apresentam-se os principais pressupostos envolvidos na descrição linguística baseada em *corpus* e o ferramental do PLN, assim como aspectos que concernem à anotação linguística. Entretanto, antes de iniciar a discussão sobre os pressupostos e os ferramentais utilizados, definem-se dois dos objetos que constituem a base desta pesquisa: o gênero *redação escolar* (especificamente a redação do ENEM) e a modalidade escrita formal da língua portuguesa, que é a modalidade esperada para as redações e que também definirá a noção de desvio sintático utilizada.

2.1 *O gênero textual redação escolar*

Esta pesquisa estuda estruturas linguísticas de um tipo de texto caracterizado como *redação escolar*, especialmente nos moldes da redação exigida pelo ENEM. Porém, antes de adentrar a discussão sobre a análise de tais estruturas, alguns apontamentos sobre esse gênero são necessários, em função das suas particularidades. A relevância dessa discussão se dá porque "os gêneros textuais são fenômenos históricos, profundamente vinculados à vida cultural e social. Fruto de trabalho coletivo, os gêneros contribuem para ordenar e estabilizar as atividades comunicativas do dia-a-dia" (MARCUSCHI, 2007, p. 19). Assim, o conceito de gênero textual é uma noção propositalmente pouco formal que se refere à materialização de textos cotidianos, com determinadas especificidades sociocomunicativas, como romance, notícia, receita culinária, artigo, relatório, redação. Marcuschi (2007) defende que "é impossível se comunicar verbalmente a não ser por algum *gênero*" (p. 22 – grifo do autor).

Observa-se nas atividades que envolvem produção textual, especialmente a partir do Ensino Médio, o direcionamento das produções dos alunos para um gênero chamado de *redação escolar*, que frequentemente deve seguir os moldes solicitados pelo ENEM ou por algum outro processo seletivo ou vestibular. Assim, esse gênero tem a peculiaridade de estar praticamente restrito a um contexto de ensino-aprendizagem:

> A redação, criada **pela e para** a escola, tem sentido e circula quase exclusivamente nesse contexto. Nas raras vezes em que extrapola esse espaço, ela continua servindo a determinados propósitos pedagógicos (...). Assim, se observamos as condições de produção de redações na escola, constatamos que a função sociocomunicativa dessa atividade está estreitamente inter-relacionada ao tratamento dispensado à redação enquanto objeto de ensino. (MARCUSCHI, 2005, p. 142 – grifo da autora)

A redação escrita em sala de aula tem propósitos que se limitam ao contexto da própria escola; logo, o leitor presumido das produções desse gênero é o professor (ou a banca avaliadora). Ainda que a discussão sobre a validade do gênero *redação escolar* esteja fora do escopo desta pesquisa, considerar esse aspecto é relevante para ter em mente que os textos que compõem um *corpus* de redações escolares nos moldes do ENEM não representam, necessariamente, a forma de escrita cotidiana dos alunos, visto que tais produções têm o objetivo específico de obter boas notas em exames de seleção.

Considerando as características específicas da redação do ENEM, a Cartilha do Participante do ENEM 2018 traz orientações sobre o texto esperado:

> A prova de redação exigirá de você a produção de um texto em prosa, do *tipo dissertativo-argumentativo*, sobre um tema de ordem social, científica, cultural ou política. (...) Nessa redação, você deverá defender uma tese – uma opinião a respeito do tema proposto –, apoiada em argumentos consistentes, estruturados com coerência e coesão, formando uma unidade textual. Seu texto deverá ser redigido de acordo com a *modalidade escrita formal da língua portuguesa*. (BRASIL, 2018, p. 7 – grifo nosso)

Tradicionalmente, redações dissertativo-argumentativas são compostas por um parágrafo inicial que introduz o tema; um ou dois parágrafos de desenvolvimento contendo a tese e os argumentos; e um parágrafo final, com a conclusão sobre o tema. Especificamente no caso do ENEM, a redação inclui uma proposta de intervenção, exigida pelo exame (BRASIL, 2018, p. 7). A definição de texto dissertativo-argumentativo que consta na Cartilha do Participante é a seguinte:

> O texto dissertativo-argumentativo se organiza na defesa de um ponto de vista sobre determinado assunto. É fundamentado com argumentos, para influenciar a opinião do leitor, tentando convencê-lo de que a ideia defendida está correta. É preciso, portanto, expor e explicar ideias. Daí a sua dupla natureza: é argumentativo porque defende uma tese, uma opinião, e é dissertativo porque se utiliza de explicações para justificá-la. (BRASIL, 2018, p. 15–16)

Vale lembrar que o não atendimento ao gênero textual exigido resulta em anulação da redação do candidato. Definido o gênero das redações que compõem o *corpus* da pesquisa, destaca-se outro aspecto citado em vários pontos da Cartilha do Participante: a adequação do texto à modalidade escrita formal da língua portuguesa. Assim, na próxima seção, discutem-se brevemente a definição e a veiculação dessa modalidade na Base Nacional Comum Curricular (BNCC)[4] do Ensino Médio, e a sua importância para o PLN. Além disso, define-se a noção de desvio sintático, que é fundamental para a anotação e análise do *corpus*.

2.2 *A modalidade escrita formal do português e a noção de desvio sintático*

É inegável a importância do domínio da modalidade escrita da língua enquanto forma de expressão. Com relação às capacidades de escrita a serem desenvolvidas por um estudante do Ensino Médio, as definições da BNCC para esse nível da Educação Básica são as seguintes:

[4] A BNCC é um documento que regulamenta as aprendizagens fundamentais a serem trabalhadas pelas escolas (públicas e particulares) da Educação Básica. Esse documento tem como objetivo garantir o direito à aprendizagem para todos os estudantes brasileiros, promovendo a igualdade no sistema de ensino.

> (EM13LP15) Planejar, produzir, revisar, editar, reescrever e avaliar textos escritos e multissemióticos, considerando sua adequação às condições de produção do texto, no que diz respeito (…) ao gênero textual em questão e suas regularidades, *à variedade linguística apropriada a esse contexto e ao uso do conhecimento dos aspectos notacionais* (ortografia padrão, pontuação adequada, mecanismos de concordância nominal e verbal, regência verbal etc.), sempre que o contexto o exigir. (BRASIL, 2017, p. 509 – grifo nosso)

Logo, vê-se a necessidade de o estudante entender a utilização da variedade linguística apropriada a cada contexto de uso. No ENEM, os aspectos ligados à adequação a essa modalidade linguística são avaliados na Competência 1[5], na qual o candidato deve "demonstrar domínio da modalidade escrita formal da língua portuguesa" (BRASIL, 2018, p. 8), o que inclui as convenções de escrita. O detalhamento da Competência 1 ainda dá especial atenção à estrutura sintática, ao afirmar que o texto deve estar adequado às regras gramaticais e à fluidez de leitura, para a qual a construção sintática escolhida pode ser benéfica ou prejudicial.

Com relação à sintaxe e à noção de desvio sintático, o conceito empregado nesta pesquisa precisa levar em conta as exigências estabelecidas pelo ENEM. Portanto, a base para as decisões relacionadas à sintaxe é, em grande medida, a gramática padrão, que é a ensinada nas escolas e que guia também a modalidade avaliada. Isso fica claro quando se analisa, na Cartilha do Participante, a maneira como é explicada a forma de avaliação da Competência 1, a partir da qual também se justifica a escolha do termo "desvio"[6]:

> (...) o avaliador corrigirá sua redação, nessa Competência, considerando os possíveis problemas de construção sintática e a *presença de desvios* (gramaticais, de convenções da escrita, de escolha de registro e de escolha vocabular).
> Em relação à construção sintática, você deve estruturar as orações e os períodos de seu texto sempre buscando garantir que eles estejam completos e contribuam para a fluidez da leitura.
> Quanto aos desvios, você deve estar atento aos seguintes aspectos:
> - Convenções da escrita: acentuação, ortografia, separação silábica, uso do hífen e uso de letras maiúsculas e minúsculas.
> - Gramaticais: concordância verbal e nominal, flexão de nomes e verbos, pontuação, regência verbal e nominal, colocação pronominal, pontuação e paralelismo.
> - Escolha de registro: adequação à modalidade escrita formal, isto é, ausência de uso de registro informal e/ou de marcas de oralidade. (BRASIL, 2018, p. 12)

[5] A avaliação do ENEM se baseia em cinco competências, sendo cada uma delas referente a um aspecto específico e avaliada individualmente. Nesta pesquisa, considera-se apenas a Competência 1, mas a descrição das demais pode ser verificada na Cartilha do Participante do ENEM 2018, disponível no endereço: http://download.inep.gov.br/educacao_basica/enem/guia_participante/2018/manual_de_redacao_do_enem_2018.pdf

[6] Esse termo foi utilizado para evitar julgamentos contidos na palavra "erro" e por ser o termo também utilizado na Cartilha do Participante do ENEM.

Portanto, o conceito de desvio sintático[7] empregado se apoia nas noções estabelecidas pelo ENEM. Não se pretende levantar aqui as discussões muito pertinentes sobre as questões de norma padrão, norma culta e as variantes mais ou menos prestigiadas, tendo como razão única as limitações de tempo e de espaço impostas pelo contexto da pesquisa[8]. Para os objetivos deste estudo, um *desvio sintático* é um desvio, ou seja, uma construção linguística idiossincrática com relação à escrita formal, relacionada a problemas na organização das palavras e suas combinações, de acordo com a modalidade escrita formal, podendo ser de ordem, de concordância e de dependência entre palavras.

O estabelecimento dos limites entre o que pertence à sintaxe e o que faz parte de outros níveis se dá de forma bastante instrumental, uma vez que tais limites não são totalmente claros ou bem-estabelecidos. Assim, definem-se os desvios ortográficos como aqueles identificados pela ferramenta de correção ortográfica do editor de textos *Microsoft® Word*, os quais foram corrigidos na etapa de pré-processamento do *corpus*, como se descreverá no capítulo metodológico desta dissertação. Os demais desvios que fogem à modalidade escrita padrão e que não sejam claramente ligados à semântica ou a questões pragmático-discursivas são considerados sintáticos. Ainda assim, quando tal definição se mostrou insuficiente, considerou-se o desvio como sendo sintático.

Essa noção de desvio sintático também é usada pelos sistemas de PLN, o que apoia a decisão por uma visão mais "tradicional" quanto ao uso das estruturas sintáticas. Conforme Pinheiro (2008), ao descrever a tarefa dos sistemas de correção ortográfica e gramatical:

> Fala-se, contudo, em usuários para os quais um revisor automático deve garantir sucesso e esses são, objetivamente, aqueles que desejam se comunicar na *modalidade culta* da língua escrita. Algo que poderia soar como obviedade, mas que expressa um princípio norteador desses programas: a modalidade culta, baseada na norma padrão, se configurará como aquela para a qual esse sistema deve ser preparado para agir. (PINHEIRO, 2008, p. 22)

Em termos de sistemas de PLN existentes, tanto o corretor ortográfico e gramatical *ReGra*[9] (MARTINS *et al*., 1998) quanto o *parser* CURUPIRA (NUNES *et al*., 2005), por exemplo, desenvolvidos para a língua portuguesa pelo Núcleo Interinstitucional de Linguística

[7] A caracterização dos desvios sintáticos tem como objetivo a descrição linguística e a geração de subsídios a ferramentas tecnológicas em sala de aula e em PLN, não contendo em si preconceito linguístico ou juízo de valor.
[8] Uma discussão mais aprofundada sobre esses aspectos no contexto da análise de redações é trazida por Pinheiro (2008), no seu Capítulo 2 (*Desvios da norma padrão e revisão linguística automática*).
[9] *ReGra* (Revisor Gramatical Automático do Português Brasileiro Escrito) é o nome do corretor gramatical desenvolvido por um grupo de pesquisadores do NILC, em parceria com a empresa Itautec-Philco, com o objetivo de criar uma ferramenta que pudesse ser utilizada junto a editores de textos comerciais. Mais tarde, a Itautec foi adquirida pela Microsoft, que comprou a licença e incorporou o *ReGra* ao seu editor de textos *MS Word*.

Computacional (NILC)[10], deixam claro que a modalidade utilizada para a construção desses sistemas foi a "norma culta padrão da língua portuguesa". Martins *et al*. (1998) explicitam que um corretor gramatical tem uma abordagem padrão:

> Também é preciso destacar que está implícita no uso do corretor gramatical a noção de correção linguística, mas não no sentido da pesquisa linguística contemporânea. Enquanto os linguistas tratam a correção em termos de aceitabilidade por qualquer falante nativo de uma língua, os corretores gramaticais normalmente tendem a definir a correção de acordo com uma classe muito específica de falantes nativos: aqueles que se considera que estejam cientes da história da língua e de estruturas linguísticas da literatura tradicional. Nesse sentido, o *ReGra* está de acordo com a visão padrão ao avisar o usuário de que algumas formas linguísticas que possuem significado não são consideradas corretas, e o seu uso frequentemente gera repressão e preconceito. (MARTINS *et al*., 1998, p. 289 – tradução nossa)[11]

Em resumo, esta pesquisa procura unir a descrição das estruturas sintáticas e a ideia de desvio sintático considerando algumas "noções teóricas de aceitação mais ou menos geral" (PERINI, 2017, p. 42). Porém, quando isso não for possível, usam-se conceitos tradicionais, que servem como base para as avaliações das redações no ENEM e para os sistemas de PLN. Estabelecidos os objetos principais que guiarão o estudo, revisam-se os pressupostos teórico-metodológicos nos quais a pesquisa se apoia, e o ferramental do qual pretende lançar mão.

2.3 *Descrição linguística baseada em* corpus *e os* corpora *de aprendizes*

Desde a compilação do primeiro *corpus* linguístico tratável por computador (o *corpus* Brown, composto por textos em língua inglesa e publicado em 1964, auge das abordagens linguísticas introspectivas), a área que se ocupa mais diretamente desse objeto, a Linguística de *Corpus*, ganhou relevância (BERBER-SARDINHA, 2000). Assim, é importante descrever alguns dos pressupostos que deram origem às demais abordagens baseadas em *corpus* e, sobretudo, com base nos quais se compilou o conjunto de textos que constitui o *corpus* desta pesquisa. Segundo Berber-Sardinha (2000):

> [a] Linguística de *Corpus* ocupa-se da coleta e exploração de *corpora*, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-

[10] Disponível em: http://www.nilc.icmc.usp.br/nilc/index.php

[11] "It must also be stressed that implicit in the use of the 'grammar checker' is the notion of linguistic correctness, but not in the sense of contemporary linguistic research. While linguists treat correctness in terms of acceptability by any native speaker of a language, grammar checkers typically tend to define correctness according to a very specific class of native speaker: those presumably aware of language history and linguistic structures of traditional literature. In this sense, ReGra conforms to the standard view by warning the user that some meaningful linguistic forms are not considered correct, and their use often attracts repression and prejudice."

> se à exploração da linguagem através de *evidências empíricas*, extraídas por meio de computador. (p. 325 – grifo nosso)

Porém, nem todo conjunto de elementos em língua natural pode ser considerado um *corpus*. Para que seja denominado como tal, um conjunto de textos autênticos precisa ser "construído a partir de um desenho explícito, com *objetivos específicos*" (BERBER-SARDINHA, 2000, p. 335 – grifo nosso). Para Sinclair (2005 *apud* ALUÍSIO; ALMEIDA, 2006), "um *corpus* é uma coleção de textos em língua natural em formato eletrônico, selecionados de acordo com critérios externos de forma a representar ao máximo possível uma língua ou variedade linguística como fonte de dados para pesquisas linguísticas" (tradução nossa)[12]. Santos (2008) define *corpus* como uma coleção de objetos linguísticos que possa ser utilizada pelas áreas do PLN[13] ou da Linguística. O uso dessa coleção pode englobar estudo, teste e avaliação. Já o objeto linguístico pode se constituir de textos, frases, palavras, desvios ortográficos, traduções, etc. Para Santos (2008), um *corpus* precisa ser finito e concreto, servindo como ferramenta para estudar a língua (e não o objeto de estudo em si). A autora destaca ainda uma questão relevante: "o mais importante num corpo [*corpus*] é saber o que fazer com ele, como usá-lo, e para que tarefas ele é útil" (SANTOS, 2008, p. 46).

Outro aspecto relevante na área refere-se à tipologia do *corpus*. Nesse tema, há diversas tipologias possíveis, mas alguns dos critérios são apresentados no Quadro 1. Em termos de representatividade, é comum considerar que quanto maior o *corpus*, maior a representatividade do fenômeno ou da língua/variação linguística que ele contém. No entanto, o *corpus* tem uma função representativa em si, isto é, ainda que seja considerado pequeno, ele apresenta fenômenos (léxicos, gramaticais, discursivos, etc.) mais ou menos frequentes. Nesse sentido, Sinclair (1991) afirma que uma maneira de garantir uma maior representatividade é construir um *corpus* que seja o maior possível em termos de números de instâncias (palavras, sentenças, textos) daquilo que pretenda representar, e que preferencialmente continue crescendo. Todavia, Evers (2018) salienta que amostras pequenas podem, por exemplo, ser indicativas de padrões lexicais e terminológicos. Para a autora, "um *corpus* pequeno pode ter a mesma validade que um *corpus* grande, desde que consideradas e explicitadas as suas finalidades" (p. 67).

---

[12] "[a] corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research"

[13] A autora utiliza o termo *Linguística Computacional* como sinônimo de PLN.

Quadro 1 – Possibilidades de classificação de um *corpus*.

Modo	Falado	Porções de fala transcritas.
	Escrito	Textos escritos, impressos ou não.
Tempo	Sincrônico	Um período de tempo.
	Diacrônico	Vários períodos de tempo.
	Contemporâneo	Período de tempo corrente.
	Histórico	Período de tempo passado.
Seleção	De amostragem (*sample corpus*)	Porções de textos ou variedades textuais, planejado como amostra finita da linguagem.
	Monitor	A composição reflete o estado atual de uma língua. Opõe-se a *corpora* de amostragem.
	Dinâmico ou orgânico	O crescimento e diminuição são permitidos, qualifica o *corpus* monitor.
	Estático	Oposto de dinâmico, caracteriza o *corpus* de amostragem.
	Equilibrado (*balanced*)	Os componentes distribuem-se em quantidades semelhantes.
Conteúdo	Especializado	Os textos são de tipos específicos (em geral gêneros ou registros definidos).
	Regional ou dialetal	Os textos são provenientes de uma ou mais variedades sociolinguísticas específicas.
	Multilíngue	Inclui idiomas diferentes.
Autoria	De aprendiz	Os autores dos textos não são falantes nativos.
	De língua nativa	Os autores são falantes nativos.
Disposição interna	Paralelo	Textos comparáveis (p.ex. original e tradução).
	Alinhado	Traduções abaixo de cada linha do original.
Finalidade	De estudo	O *corpus* que se pretende descrever.
	De referência	Usado para contraste com o *corpus* de estudo.
	De treinamento ou teste	Construído para permitir o desenvolvimento de aplicações e ferramentas de análise.

Fonte: BERBER-SARDINHA, 2000.

Para a LC, a língua, assim como o *corpus*, tem caráter probabilístico, ou seja, determinados fenômenos têm maior ou menor probabilidade de ocorrência. Da mesma forma, as descrições linguísticas que se baseiam em *corpus* levam em conta que, ainda que jamais poderá representar toda a língua, ele geralmente (se for bem construído e representativo) vai conter os fenômenos mais frequentes e poderá trazer à luz aqueles menos frequentes, mas que também podem ser interessantes para fins de descrição. A questão da representatividade e do tamanho do *corpus* dependerá, novamente, da finalidade para a qual ele foi construído.

Em termos de finalidade, é válido considerar a adequação do *corpus* aos objetivos da pesquisa e aos interesses do pesquisador. Segundo Berber-Sardinha, "[…] em vez de se dizer, 'eu tenho este *corpus*, então agora vou descrevê-lo', deve-se pensar 'eu desejo investigar esta questão, então eu necessito de um *corpus* com estas características'" (2000, p. 349). O *corpus* também permite direcionar o olhar aos padrões lexicais ou léxico-gramaticais que a linguagem apresenta, e que não variam de maneira aleatória, sendo mais ou menos recorrentes e/ou sistemáticos em determinados gêneros ou variações linguísticas. Assim, o uso de *corpus* se justifica mais uma vez aos propósitos deste estudo, que busca padrões sintáticos que propiciem uma maior ou menor ocorrência de desvios sintáticos em textos escritos por estudantes.

Há outro aspecto ligado ao *corpus* que é muito importante: a sua disponibilidade. Considera-se que o esforço empenhado em compilar, limpar, caracterizar e anotar um *corpus* deva servir para outras pesquisas que se interessem por fenômenos representados nele ou para o desenvolvimento de sistemas diversos. Segundo Aluísio e Almeida (2006), a "disponibilização de *corpus* compilado para futuras pesquisas é uma característica inerente ao *corpus*, de forma que todo o esforço empreendido para a sua construção não seja útil apenas para uma pesquisa, uma vez que se tem uma referência padrão de língua ou de variedade de língua que pode ser utilizada por outros pesquisadores" (p. 158).

Partindo para as etapas que constituem a construção do *corpus* em si, toma-se por base a metodologia descrita por Aluísio e Almeida (2006). De acordo com as autoras, a construção do *corpus* deve ser feita nas seguintes etapas: (i) projeto (seleção dos textos que vão compor o *corpus*), (ii) compilação, manipulação, nomeação dos arquivos e autorização para uso, e (iii) anotação. Na primeira etapa, é importante definir com cuidado o tamanho do *corpus*, a sua composição no que diz respeito aos textos, e os gêneros dos elementos textuais. A segunda etapa corresponde ao armazenamento dos arquivos e todas as manipulações necessárias, como conversões, limpeza de "ruídos" (figuras, legendas, datas, etc.), correção de eventuais problemas (p. ex., dados corrompidos pela conversão de formatos), entre outras. Essa segunda etapa ainda envolve a obtenção da licença de uso dos textos para que o *corpus* possa ser disponibilizado publicamente. Na última etapa, realiza-se a anotação do *corpus* de acordo com a tarefa pretendida e em qualquer dos níveis linguísticos de interesse, podendo esta ser feita de forma manual (ou humana), automática (anotação por meio de ferramentas de PLN, como *POS taggers* e *parsers*) ou semiautomática (anotação automática com posterior correção humana). A etapa de anotação é discutida em mais detalhes na próxima seção, uma vez que corresponde a uma parte significativa do trabalho desenvolvido para esta pesquisa.

A fim de classificar o *corpus* do estudo, destaca-se inicialmente o tipo dos textos que o compõem. Sabendo-se que esse *corpus* é constituído por textos escritos por estudantes, é importante definir mais especificamente a tipologia utilizada. Assim, salientam-se alguns aspectos relevantes do que se decidiu chamar aqui de *corpora de aprendizes*[14], ainda que essa definição fuja da sua acepção tradicional.

Como dito anteriormente, trata-se de redações escritas por falantes nativos de português, em fase de formação na Educação Básica. Uma das definições de *corpora* de aprendizes é a de conjuntos de textos que não passaram por nenhum tipo de revisão, tendo sido produzidos por

[14] Essa mesma nomenclatura é utilizada por Pinheiro (2008), cujo *corpus* também é composto por redações de estudantes.

indivíduos que estão em processo de formação. Em geral, *corpora* de aprendizes são constituídos por produções escritas de falantes não nativos de uma língua, ou seja, de aprendizes dessa língua como L2 (BERBER-SARDINHA, 2004). Todavia, falantes nativos também passam pelo processo de aprendizagem — nesse caso, da modalidade escrita formal. Logo, a aquisição da modalidade escrita pode ser entendida da mesma maneira que a aquisição de uma língua estrangeira (ainda que os aspectos particulares dessas duas aquisições sejam diferentes). Segundo Evers (2018), *corpora* de aprendizes são "especiais por serem ricos em inadequações de usos de recursos linguísticos, em marcas de aprendizado de língua e em novos padrões que indicam variação da língua e possíveis mudanças da norma" (p. 60).

2.4 *Anotação de* corpus*: interface entre LC e PLN*

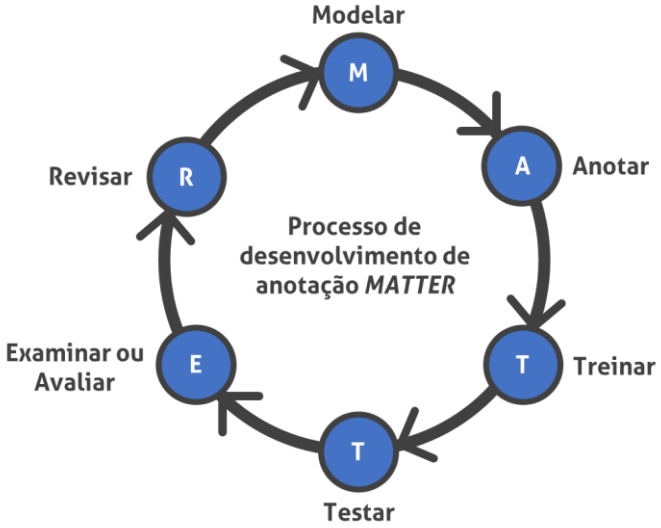
Anotação linguística consiste no processo de enriquecimento de um *corpus* por meio da inserção de informações linguísticas de forma manual ou automática, ou ainda combinando as duas anteriores, com um objetivo teórico ou prático (HOVY; LAVID, 2010). Para Pustejovky e Stubbs (2013), essa tarefa tem como fim melhorar a capacidade de um computador de executar o processamento automático de uma língua natural, pois dados anotados são o principal recurso de sistemas de PLN baseados em Aprendizado de Máquina.

Nesse sentido, *corpora* anotados, manual ou automaticamente, são recursos valiosos de várias maneiras. A anotação é uma tarefa muito frequente na LC e no PLN e, embora as abordagens e os objetivos em ambas as áreas sejam diferentes em alguns pontos, há aspectos importantes a considerar em uma proposta de anotação de *corpus*, os quais são comuns às duas. Do ponto de vista da Linguística, os *corpora* anotados apresentam, de forma estruturada e sistemática, informações linguísticas em larga escala que podem fornecer indícios para os mais diversos tipos de investigações e descrições. A própria elaboração da tarefa de anotação, por meio da instanciação da teoria, anotação piloto, treinamento de anotadores, já pode revelar casos complexos, estruturas que fogem ao comportamento esperado, entre outras questões. Para o PLN, os *corpora* anotados (consistentes e confiáveis) podem ser utilizados para o treinamento de algoritmos e para o desenvolvimento de ferramentas. Porém, independentemente da área (em projetos exclusivos de uma delas e nos multidisciplinares), é fundamental que a tarefa de anotação seja planejada cuidadosamente, considerando uma série de fatores.

Pustejovsky e Stubbs (2013) sugerem que o desenvolvimento de tarefas de AM envolvendo línguas naturais (como a proposta por esta pesquisa) sigam um processo que eles denominam como *MATTER*, um acrônimo em inglês das tarefas de modelar um problema,

anotar os dados, treinar o algoritmo, testá-lo em um conjunto de dados de teste, avaliar a saída e revisar o processo, quantas vezes for necessário, até que se obtenham resultados satisfatórios (em inglês, *model*, *annotate*, *train*, *test*, *evaluate*, *revise*). A Figura 1 ilustra o processo.

Figura 1 – Processo *MATTER*.



Esta seção foca especificamente os dois primeiros pontos: a modelagem do problema e a tarefa de anotação em si, visto que as demais etapas são descritas na Seção 2.6 (p. 38), que trata sobre o Aprendizado de Máquina. No que se refere à primeira etapa, modelar o problema significa caracterizar um fenômeno ou uma tarefa-alvo de maneira abstrata, descrevendo de forma teórica o que se pretende anotar, assim como as etiquetas que serão aplicadas. Essa etapa servirá como base para definir o processo prático de anotação, escrever as diretrizes, escolher as ferramentas a serem utilizadas. Nessa etapa, também é definido o tipo de anotação em relação ao armazenamento dos dados. Segundo Pustejovsky e Stubbs (2013), há dois tipos de anotações possíveis, e cada um deles apresenta vantagens e desvantagens.

Anotação *in-line* – As anotações são inseridas diretamente no texto e são armazenadas junto ao arquivo original. Esse tipo de anotação torna fácil identificar a localização exata da anotação, mas isso faz o texto ser difícil de ler, além de alterar o material original. Também se impõe o desafio de sobrepor etiquetas, caso se decida estender o nível da anotação.

Anotação *stand-off* – Em anotações *stand-off*, que é o tipo de anotação usado por este estudo, as marcações são removidas do texto original e mantidas paralelamente, em colunas ou arquivos separados (ou outras possibilidades). Tais anotações podem ser feitas por *token*, em que cada etiqueta é associada a um *token* específico, previamente identificado no processo de *tokenização*, ou por localização de caracteres, em que se utiliza o local dos caracteres no texto

para definir a qual parte (parte de um *token*, o *token* todo ou um conjunto de *tokens*) se referem as etiquetas. Há ainda a anotação por extensão vinculada, em que etiquetas de vinculação representam um vínculo que se estabelece entre dois elementos anotados (não ao *token* ou ao trecho do texto em si, mas às anotações contidas neles).

Outro fator a ser considerado é a forma como a anotação será realizada: manual, automática ou semiautomática. Na anotação manual, o *corpus* costuma exigir pouca preparação, sendo possível anotar fenômenos complexos sem grandes investimentos, exceto em recursos humanos especializados e tempo. Nesse cenário, os anotadores precisam ter à disposição diretrizes de anotação robustas, de forma a se obter um resultado de maior qualidade e confiabilidade. A anotação automática, por sua vez, ainda que seja muito mais rápida, exige maior preparação do *corpus* e programação/treinamento da ferramenta de anotação, além de possivelmente gerar resultados de menor qualidade (HOVY; LAVID, 2010). Assim, uma abordagem intermediária é a anotação automática com posterior validação ou correção por especialistas. No cenário semiautomático, a anotação é menos cara e demorada que a manual, e os resultados são mais corretos que a automática (ALUÍSIO; ALMEIDA, 2006). Esta pesquisa realiza a anotação automática por meio de *parsing*, descrita na Seção 4.2 (p. 61), e a anotação manual nas duas fases de anotação descritas na Seção 4.3 (p. 65).

Já no segundo ponto, que envolve a tarefa de anotação em si, a partir do que se considerou relevante na etapa de modelagem, deve-se escrever as diretrizes que guiarão a tarefa, escolher e treinar os anotadores, preparar os dados. É comum que, nessa etapa, sejam identificadas questões que não foram bem-definidas, etiquetas específicas ou genéricas demais, ou particularidades nos dados que não foram levadas em conta inicialmente. Nesse caso, deve-se voltar à etapa de modelagem e repetir o processo até que se obtenha o resultado desejado.

Outro aspecto ligado a esse ponto é a confiabilidade da anotação. Mesmo que as diretrizes que guiam a tarefa tenham sido elaboradas de forma a definir claramente o problema e a elencar objetivamente as orientações (diminuindo a subjetividade), é preciso garantir que os anotadores (i) tenham entendido as diretrizes (isso envolve treinamento adequado e suficiente), bem como objetivos/importância da tarefa, (ii) abram mão das suas próprias crenças e tomem as decisões de anotação com base na diretriz, (iii) registrem e discutam os casos duvidosos, problemáticos ou que não são abarcados nas orientações recebidas, e (iv) estejam comprometidos com a tarefa e a executem com seriedade.

Segundo Hovy e Lavid (2010), o PLN adotou uma forma de avaliar não apenas o trabalho dos anotadores, mas também a dificuldade da tarefa e a precisão das diretrizes: os especialistas anotam a mesma porção do *corpus*, e as suas anotações são comparadas em um

processo denominado concordância entre anotadores. Algumas métricas bem-estabelecidas podem ser usadas nesse processo, como a concordância simples, o alfa de Krippendorff e diversas variações da medida *kappa*, dependendo do tipo de tarefa. Assim, cabe aos organizadores estabelecer as métricas e os índices aceitáveis de concordância, dependendo da complexidade dos fenômenos e dos objetivos da anotação. Para Pustejovsky e Stubbs (2013), quando se obtêm valores considerados adequados de concordância, passa-se para a etapa de adjudicação, a partir da qual será criado o *gold standard*, que se refere ao conjunto de dados anotados, avaliados e corrigidos, isto é, o *corpus* com as anotações finais. Na adjudicação, a partir de discordâncias entre anotadores, um ou mais especialistas fazem o papel de "juízes" e tomam decisões para os casos problemáticos.

Para a Linguística de *Corpus*, a anotação de *corpus* pode ser usada para diversos propósitos, como investigações teóricas, descrições linguísticas, criação de dicionários e léxicos, entre outros. Nessa área, há três aspectos fundamentais no que se refere a essa tarefa: reutilização, estabilidade e reprodutibilidade. Em termos de *reutilização*, um *corpus* anotado é interessante porque pode ser utilizado para outras pesquisas e investigações. Com relação à *estabilidade*, ele oferece um padrão de referência para consultas e uma base estável de análise linguística para fins de contraste e comparação. Por fim, a *reprodutibilidade* permite o registro explícito de uma análise linguística, deixando clara a teoria utilizada.

Ainda para Hovy e Lavid (2010), a anotação de *corpus* deve ser alvo de pesquisas da LC porque tem impactos teóricos e práticos no fazer linguístico dessa abordagem. Entre os *impactos teóricos*, os autores citam a formação, a redefinição e o enriquecimento de uma teoria ou um modelo de comportamento para determinados fenômenos linguísticos. Com relação aos *impactos práticos*, a utilização de *corpora* pode ir muito além da descrição linguística ou da aplicação a sistemas de PLN. Na área do ensino de língua, os autores afirmam que há inúmeras possibilidades de exploração de *corpora* anotados a partir de atividades realizadas com alunos. Os exemplos anotados podem ser discutidos em sala de aula, e a anotação pode ser reproduzida com os alunos para o aprendizado de determinados fenômenos linguísticos, por exemplo.

Hovy e Lavid (2010) propõem sete questões relevantes sobre a anotação linguística: (i) seleção do *corpus*, (ii) instanciação da teoria, (iii) seleção e treinamento de anotadores, (iv) especificação do procedimento de anotação, (v) projeto da interface de anotação, (vi) escolha e aplicação das medidas de avaliação e (vii) disponibilização e manutenção do produto. Dependendo do tamanho, dos objetivos e das características da tarefa, cada questão se colocará com maior ou menor relevância. No entanto, todas precisam, no mínimo, ser consideradas. Tais questões serão brevemente descritas a seguir.

Seleção do *corpus* – A questão mais importante desse item é o objetivo da tarefa. Os gêneros textuais, os períodos de tempo e as variações de registro dependem disso, visto que a ocorrência dos fenômenos linguísticos alvo é mais ou menos frequente, dependendo desses aspectos. Assim, o *corpus* deve ser representativo desse fenômeno. O balanceamento do fenômeno nos diferentes gêneros/tempos/variantes também é importante.

Instanciação da teoria – A teoria que sustenta a anotação é a base da definição das categorias e das diretrizes de anotação. Quanto mais complexo é o fenômeno a ser investigado, mais complexa também será a teoria subjacente e, portanto, as diretrizes. Assim, a construção das diretrizes exige reflexão e discussões entre especialistas, anotação piloto, cálculo de concordância entre anotadores, análise e avaliação de dúvidas e casos difíceis, revalidação da teoria e das categorias, inserção de exemplos e aperfeiçoamento das diretrizes.

Seleção e treinamento de anotadores – Deve-se considerar se os anotadores precisam ser especialistas ou se o treinamento fornecido será suficiente mesmo para não especialistas. Se não houver treinamento ou ele for insuficiente, é pouco provável que os resultados da anotação sejam confiáveis, e corre-se o risco de ter decisões subjetivas e não registradas nas diretrizes. O treinamento excessivo, por sua vez, pode indicar que talvez fosse suficiente inserir regras em uma ferramenta automática. O ideal é selecionar anotadores com um nível de formação similar e prover diretrizes e treinamento claros e precisos; no entanto, conhecimentos prévios e intuições dos anotadores são inevitáveis.

Especificação do procedimento de anotação – Inclui anotação piloto e discussões, pois as discordâncias fazem parte desse processo. Deve haver uma equipe ou um especialista responsável por organizar etapas de reconciliação (quando há total discordância e é preciso decidir sobre o que fazer) e adjudicação. Em geral, a fase inicial da anotação serve para que os anotadores se acostumem à tarefa e aos dados a partir de exemplos mais "fáceis", antes de começarem o trabalho de fato.

Projeto da interface de anotação – O objetivo da interface é agilizar e facilitar o trabalho do anotador, e evitar um possível viés na anotação. A escolha da interface/ferramenta a ser utilizada depende da tarefa, mas algumas boas práticas incluem uso do teclado em vez de mouse; limite recomendável de oito categorias; criação de tarefas simples e em mais etapas; apresentação das opções em ordem variável; apresentação de todas as opções numa mesma página (evitando-se a barra de rolagem).

Escolha e aplicação das medidas de avaliação – A avaliação é muito importante no processo de anotação, pois ela garantirá a confiabilidade dos resultados. Assim, deve-se medir

constantemente a concordância entre os anotadores e de um anotador consigo mesmo, de modo a garantir a consistência das decisões e do uso das diretrizes.

Disponibilização e manutenção do produto – Compilar um *corpus* e desenvolver uma tarefa de anotação torna-se muito mais útil se o *corpus* e as anotações puderem ser disponibilizados publicamente, evitando o retrabalho a cada nova investigação de fenômenos semelhantes. Assim, é importante fazer o possível para permitir o acesso aos dados por outros pesquisadores.

Hovy e Lavid (2010) encerram destacando a necessidade de um trabalho colaborativo:

> Portanto, *há uma grande necessidade de esforços colaborativos para avançar em ambos os campos*: enriquecendo os estudos computacionais com teorias e enriquecendo os estudos de *corpus* com metodologias experimentais que assegurem a "confiabilidade" no processo de anotação. (HOVY; LAVID, 2010, p. 32 – tradução nossa, grifo nosso)[15]

A próxima seção traz os aspectos ligados à pesquisa em PLN. Descrevem-se os pressupostos e objetivos dessa área, bem como algumas das ferramentas importantes para este estudo.

2.5 *Ferramentas de PLN para a descrição linguística*

O PLN, também chamado de Linguística Computacional[16], é uma área de estudos que se encontra na intersecção entre Ciência da Computação, Inteligência Artificial e Linguística. Ela busca a sistematização da língua natural por meio de ferramentas computacionais, para que ela possa ser compreendida, aprendida e gerada por computador (MANNING; SCHÜTZE, 1999). Para Ferreira e Lopes (2017), essa disciplina volta-se para a compreensão de línguas naturais para que a comunicação mediada por computador seja possível e, "Como parte da linguística, tem cada vez mais importância nas atividades de descrição e análise, inclusive com a possibilidade de testar hipóteses usando informações sobre as línguas naturais contidas em bases crescentes de dados empíricos" (p. 195). Tais bases de dados são os *corpora*.

Di-Felippo e Dias-da-Silva (2009) fazem um apanhado histórico do surgimento do PLN desde os anos 1940, quando se iniciou a tentativa de fazer o computador entender instruções em língua natural. Os autores afirmam que o PLN tem como objeto a língua escrita (ou, no

[15] "There is, therefore, ample need for collaborative efforts to advance in both camps: enriching computational studies with theory, and enriching corpus studies with experimental methodologies, which ensure the 'reliability' in the annotation process."

[16] Para uma discussão mais sólida sobre semelhanças e diferenças entre esses dois campos de estudo, pode-se consultar Dias-da-Silva (2006). Neste texto, ambos serão considerados sinônimos.

máximo, a língua falada que foi transcrita em forma de texto). Ainda segundo os autores, os papéis dos linguistas e dos cientistas da computação são diferentes nesse campo multidisciplinar. O linguista usa o computador para desenvolver e validar teorias e dados linguísticos, ou fornece conhecimentos para o desenvolvimento de *recursos*; já o cientista da computação implementa *ferramentas* para validar as teorias linguísticas propostas pelos linguistas, e desenvolve *sistemas* (ou aplicações) com base nos conhecimentos linguísticos fornecidos. Logo, trata-se de um campo de estudos heterogêneo.

Nas próximas seções, serão abordados dois tipos de ferramentas que servirão de base para a extração de atributos linguísticos proposta nesta pesquisa: os *POS taggers* e os *parsers*[17].

2.5.1 *POS tagging*: conceitos essenciais

Classes morfossintáticas, classes gramaticais ou *parts-of-speech* (*POS*), termo mais utilizado no PLN, são úteis porque descrevem o comportamento sintático local de uma palavra e das palavras que a cercam. Por exemplo, saber se uma palavra é um *verbo* ou um *nome* permite que se possa predizer quais são os seus possíveis vizinhos (JURAFSKY; MARTIN, 2017). Além disso, tais informações permitem que se identifiquem diversas questões relacionadas à estrutura sintática na qual essa palavra se encontra. Por isso, a tarefa de atribuir etiquetas que explicitem as classes morfossintáticas das palavras é fundamental também para realizar a análise sintática. No PLN, essa tarefa é chamada de etiquetação morfossintática automática (em inglês, *POS tagging*), que é executada pelos *POS taggers*.

Porém, antes de saber como funciona essa ferramenta, é importante compreender quais são as classes morfossintáticas que as palavras podem ter. Tradicionalmente, essas classes são definidas a partir de características morfológicas e da função que uma palavra exerce na frase. Diferentemente das funções sintáticas, atribuídas a pares ou a conjuntos de palavras em termos das relações estabelecidas entre elas, as classes morfossintáticas se referem a cada palavra individualmente e estão relacionadas ao seu sentido em contexto, à função exercida e/ou às variações permitidas (p. ex., flexão, derivação, invariabilidade), levando em conta as suas propriedades distribucionais e morfológicas. A divisão das classes morfossintáticas depende em grande medida da teoria linguística utilizada como base e da língua em questão.

Em geral, considera-se que há dois tipos de classes: as fechadas, que englobam uma lista finita de palavras, a qual dificilmente aumenta (como as preposições); e as abertas, que crescem

[17] Demais ferramentas de PLN estão fora do escopo deste trabalho.

constantemente conforme a evolução da língua (como substantivos e verbos). Os elementos das classes fechadas também recebem o nome de palavras funcionais, enquanto aqueles que pertencem a classes abertas são chamados de palavras lexicais ou de conteúdo. Tradicionalmente, em português, consideram-se as seguintes classes: substantivos/nomes, adjetivos, verbos, advérbios, pronomes, artigos/determinantes, preposições, conjunções, numerais e interjeições. Algumas delas ainda podem ser subdivididas, como substantivos comuns e próprios; advérbios de tempo, modo, local; pronomes pessoais, demonstrativos, possessivos, entre outras divisões possíveis. Há inúmeras críticas a essa divisão, bem como diversas propostas de categorias, mas essa importante discussão não será aprofundada aqui, uma vez que o foco desta seção são as definições utilizadas pelas ferramentas computacionais.

Voltando à tarefa de etiquetação morfossintática automática, segundo Jurafsky e Martin (2017), *POS tagging* é o processo de atribuir automaticamente etiquetas de *POS* a cada palavra em um texto de entrada. Como as etiquetas em geral também são atribuídas aos sinais de pontuação, tais textos precisam passar por uma etapa de *tokenização*, que significa identificar e separar cada *token* (isto é, cada palavra, sinal de pontuação, contração, abreviação, sigla), identificando, por exemplo, se um ponto indica final de sentença ou faz parte de uma abreviação ou um numeral. Como uma mesma palavra pode fazer parte de classes morfossintáticas diferentes, dependendo do seu contexto e do seu sentido (por exemplo, a palavra *se*, que pode ser um pronome reflexivo ou uma conjunção), uma importante questão no desenvolvimento de *POS taggers* é a desambiguação, isto é, identificar a etiqueta correta a ser atribuída a uma palavra ambígua (que tem mais de uma possibilidade de *POS*). A ferramenta precisa, então, escolher a etiqueta adequada de cada palavra, considerando o seu contexto de uso. Ainda de acordo com Jurafsky e Martin (2017), no *corpus* Brown, por exemplo, cujo conjunto de etiquetas é composto por 45 *POS tags* do inglês, 15% dos *types* (palavras diferentes) e 67% dos *tokens* (todas as palavras) são ambíguos. Logo, tais ferramentas precisam ser projetadas para levar em conta o contexto das palavras e a frequência de ocorrência de determinada *POS* em uma língua para que se obtenham bons resultados na tarefa de *POS tagging*.

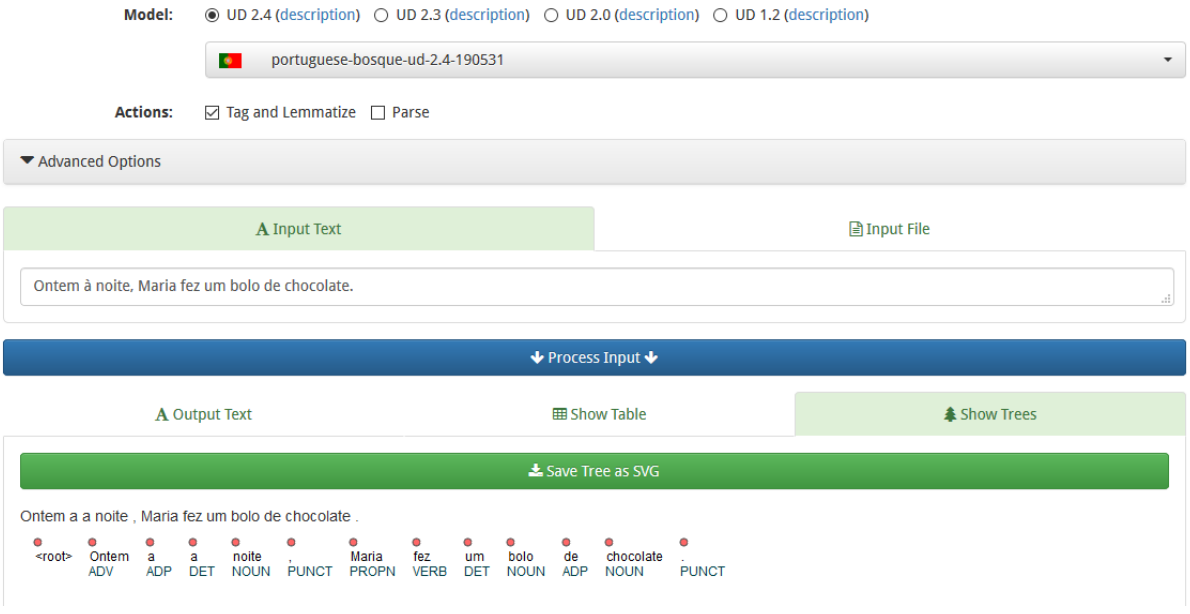
No que se refere ao conjunto de etiquetas (as classes morfossintáticas ou *POS*), o desenvolvimento de um *POS tagger* considera uma teoria linguística específica ou os objetivos de determinada tarefa para defini-lo. Esse conjunto pode conter mais ou menos classes, dependendo da abordagem, e deverá ser disponibilizado junto à documentação da ferramenta. Às vezes, também, são os projetos para a construção de *corpora* anotados que definem as etiquetas, como é o caso do Penn Treebank II, do inglês, com 41 etiquetas, e dos *corpora* do *Universal Dependencies* (UD), com 17 etiquetas de *POS* (NIVRE *et al*., 2016) em mais de 90

línguas (incluindo o português). Esse último projeto busca desenvolver um padrão multilíngue de anotação gramatical consistente e construir bancos de árvores sintáticas (*treebanks*).

Em função da sua importância para a tarefa de *parsing*, muitos dos *parsers* já têm embutido no seu sistema um *POS tagger*. Os *parsers* desenvolvidos no contexto do UD, como o *parser* por dependências *UDPipe* (STRAKA; HAJIČ; STRAKOVÁ, 2016), em geral já contam com um *POS tagger* e um analisador morfológico, que traz informações como gênero, número, pessoa, tempo, modo e subdivisões como tipo de pronome, de forma verbal, de determinante, entre outras questões específicas de cada classe. Essas informações são essenciais para que se possa fazer a análise sintática automática de uma sentença. Atualmente, *taggers* multilíngues embutidos em *parsers* têm demonstrado excelente desempenho, com acurácia de 97–98%, errando principalmente nas etiquetas que também geram dúvidas e discussões em análises humanas.

A Figura 2 mostra um exemplo de saída do *POS tagging* embutido no *parser UDPipe*, obtida na plataforma *on-line* da ferramenta[18].

Figura 2 – Plataforma *on-line* do *parser UDPipe* com a saída do *POS tagger*.

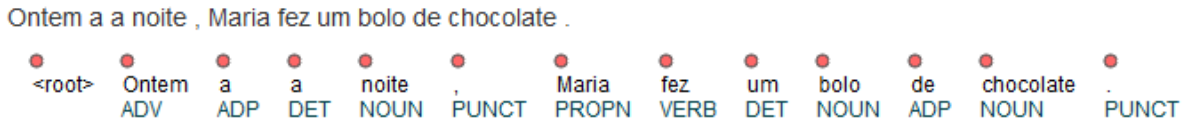


A plataforma *on-line* permite a seleção da versão do modelo que se pretende usar, da língua e do *corpus* de treinamento do modelo. Além disso, pode-se escolher entre utilizar apenas a ferramenta de *tokenização* e *POS tagging*, ou se se quer acrescentar o *parsing*. A entrada pode ser uma sentença escrita diretamente na plataforma, ou um arquivo de texto. Para a saída, pode-

[18] Disponível em: http://lindat.mff.cuni.cz/services/udpipe/.

se escolher a visualização no formato CONLL-U, em tabela ou diretamente na representação visual, como na Figura 2. Para facilitar a visualização das *POS tags* utilizadas para a sentença de entrada, o recorte da Figura 3 mostra apenas a saída do *POS tagger*.

Figura 3 – Saída do *POS tagger* embutido no *UDPipe*.



Vê-se que o *tokenizador* identificou cada um dos elementos, separando a contração *à* em preposição (ADP) e determinante (DET), e etiquetou também os sinais de pontuação. Além disso, o substantivo *Maria* foi etiquetado como nome próprio (PROPN). A seção seguinte explicita a tarefa de *parsing*, que toma como base as etiquetas atribuídas pelo *POS tagger*.

2.5.2 *Parsing* sintático: conceitos essenciais

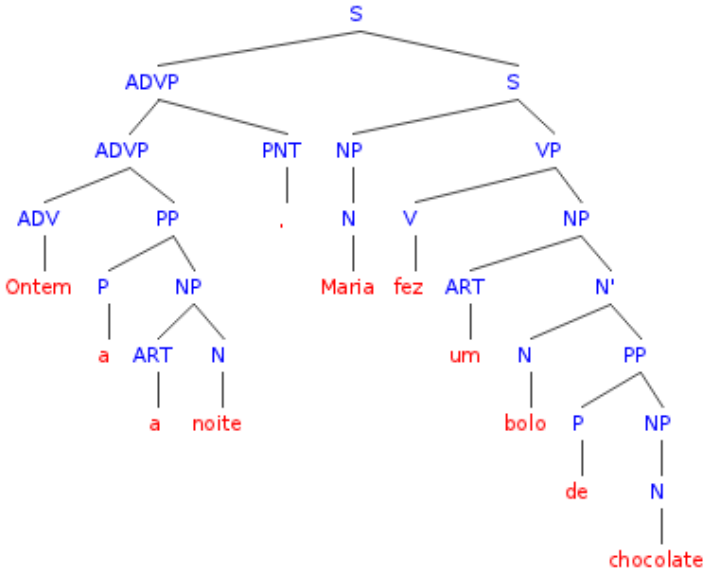
Segundo Jurafsky e Martin (2017), *parsing* sintático é uma tarefa que consiste em reconhecer uma sentença e atribuir a ela uma estrutura sintática. As ferramentas que executam o *parsing* são chamadas de analisadores sintáticos automáticos ou *parsers*[19]. Os autores ressaltam que as árvores sintáticas (isto é, a saída de um *parser*, representada pela estrutura sintática em formato de árvore) são úteis em aplicações como corretores gramaticais em editores de texto. Segundo os autores, uma sentença para a qual o *parser* não consegue gerar uma estrutura pode conter desvios gramaticais ou pode no mínimo ser difícil de ler (JURAFSKY; MARTIN, 2017).

Atualmente há diversos *parsers* disponíveis (gratuitamente ou não) para línguas específicas (p. ex., inglês) ou multilíngues (cujas representações são independentes de língua). Tais ferramentas em geral seguem um dos dois formalismos mais comuns: (i) gramáticas por constituintes ou (ii) gramáticas por dependências. Um *parser* geralmente é desenvolvido com base em grandes bancos de sentenças na(s) língua(s) de trabalho da ferramenta, as quais foram anotadas sintaticamente de forma manual ou (semi)automática. Esses bancos são chamados de *treebanks* (ou bancos de árvores). Uma vez que as estruturas sintáticas das sentenças desses bancos foram explicitadas com base em um dos formalismos citados, estas podem ser usadas especificamente para treinar os *parsers* usando algoritmos de AM.

[19] Nesta dissertação, utiliza-se o termo *parser* como sinônimo de *parser* sintático, isto é, a ferramenta computacional que realiza a análise sintática automática.

No formalismo por constituintes, considera-se que as palavras se agrupam formando constituintes (ou sintagmas), que tipicamente recebem o nome das classes morfossintáticas que estão no topo ou são o núcleo (em inglês, *head*) do constituinte: sintagma verbal (núcleo é um verbo), sintagma nominal (núcleo é um nome), sintagma preposicional (núcleo é uma preposição), sintagma adverbial (núcleo é um advérbio). Segundo Manning e Schütze (1999 – tradução nossa), para o PLN, "uma ideia fundamental é a de que certos agrupamentos de palavras se comportam como constituintes. Constituintes podem ser identificados porque podem ocorrer em várias posições, e porque mostram possibilidades sintáticas regulares de expansão"[20]. A Figura 4 mostra um exemplo de uma árvore por constituintes gerada por meio da plataforma *on-line* de teste[21] do *parser* LX Parser (SILVA *et al*., 2010).

Figura 4 – Exemplo de árvore por constituintes.



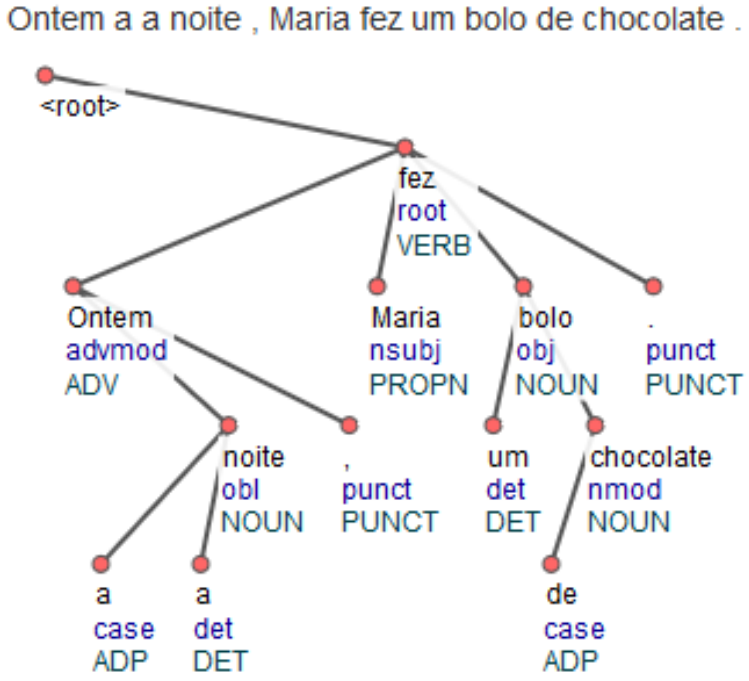
Já o formalismo das gramáticas por dependências, que é o utilizado nesta pesquisa, vem ganhando destaque em diversos sistemas de PLN (JURAFSKY; MARTIN, 2017), e o desenvolvimento de *parsers* que utilizam essa representação tem sido fomentado em campanhas de avaliação (em inglês, *shared tasks*)[22] de projetos internacionais, com destaque

[20] "one fundamental idea is that certain groupings of words behave as constituents. Constituents can be detected by their being able to occur in various positions, and showing uniform syntactic possibilities for expansion"
[21] Disponível em: http://lxcenter.di.fc.ul.pt/services/pt/LXParserPT.html.
[22] Esse modelo tem sido utilizado de forma a unir um conjunto de pessoas vindas de diversas comunidades para o desenvolvimento de sistemas focados na tarefa pretendida. É comum haver um *ranking* para as ferramentas desenvolvidas, com premiações às de melhor desempenho (nem sempre em termos monetários). Esse é o modelo utilizado, por exemplo, pela conferência internacional CONLL, que lança campanhas de avaliação anuais com os mais diversos objetivos, as quais ocorrem desde 1999. Disponível em: https://www.conll.org/2019-shared-task.

para o projeto *Universal Dependencies* (NIVRE *et al*., 2016). Nesse formalismo, a estrutura sintática é representada em termos de associações (ou dependências) binárias entre as palavras (JURAFSKY; MARTIN, 2017). Assim, há uma palavra considerada raiz (em inglês, *root*) da estrutura, que geralmente é o termo predicador (p. ex., o verbo em sentenças com verbos plenos). A partir da raiz, estabelecem-se os dependentes diretos (os núcleos), e os seus demais dependentes (os nodos). Logo, a raiz de cada sentença é sempre o núcleo da estrutura sentencial inteira. As relações de dependência são rotuladas conforme a sua função na sentença (p. ex., sujeito e objeto), e o conjunto de etiquetas é específico de cada ferramenta. Uma das maiores vantagens desse formalismo, segundo Jurafsky e Martin (2017), é que ela permite representar sentenças de línguas em que a ordem das palavras é livre, isto é, em que as funções sintáticas se dão por meio de marcas morfológicas. No formalismo por dependências, pode-se considerar como núcleos as mesmas classes morfossintáticas consideradas núcleos no formalismo por constituintes, ou pode-se dar prioridade às palavras lexicais, como é o caso do *Universal Dependencies*. Esta pesquisa utiliza a segunda forma de representação, com os núcleos sendo as palavras lexicais, e as gramaticais ocupando apenas a posição de nodos.

Para ilustrar esse formalismo, mostra-se, na Figura 5, uma representação por dependências gerada na plataforma *on-line* do *parser UDPipe* (a mesma utilizada para gerar a saída do *POS tagger*, mostrada na Figura 2).

Figura 5 – Exemplo de árvore por dependências.



A ferramenta utilizou as *POS tags* para identificar o verbo *fazer* como a raiz, o nome próprio *Maria* como o sujeito, o substantivo *bolo* como o objeto. Além disso, o advérbio permitiu a identificação da estrutura como tendo a função de modificador do verbo. Na plataforma *on-line*, é possível obter as informações morfológicas e o lema clicando em cada elemento da árvore.

Entre os *parsers* por constituintes existentes para o português do Brasil, destaca-se o já citado LX Parser, desenvolvido pelo grupo NLX[23], de Portugal. Já entre os *parsers* por dependências, há uma variedade de opções, geradas principalmente pelos esforços despendidos no âmbito do referido projeto UD. Entre as opções, estão o *UDPipe* (STRAKA; HAJIČ; STRAKOVÁ, 2016), o *Turbo Parser* (MARTINS *et al*., 2010) e o *MaltParser* (NIVRE *et al*., 2007). O *treebank* mais comumente utilizado para treinar ferramentas para o português é o Bosque, sub*corpus* do projeto Floresta Sintá(c)tica cujas árvores sintáticas foram originalmente revisadas por especialistas humanos. O referido projeto engloba bancos de árvores sintáticas revisadas, não revisadas (Floresta Virgem e Amazônia) e semirrevisadas (Selva) (SANTOS *et al*., 2001). Há ainda um *parser* desenvolvido especialmente para o português, que não utiliza nenhum dos dois formalismos descritos: o PALAVRAS (BICK, 1996), que emprega o formalismo *constraint grammar* (ou gramática constritiva). Esse *parser* não é de uso livre e exige aquisição da licença.

A próxima seção descreve o Aprendizado de Máquina, que é uma técnica comumente utilizada em pesquisas em PLN e em outras áreas que trabalham com grandes conjuntos de dados.

2.6 *O Aprendizado de Máquina em tarefas de PLN*

Segundo Souza e Di-Felippo (2018), "Aprendizado de Máquina é uma área da Inteligência Artificial cujo objetivo é o desenvolvimento de técnicas computacionais para o aprendizado bem como a construção de sistemas (ou algoritmos) capazes de adquirir conhecimento de forma automática a partir de exemplos". De acordo com Marsland (2009), Aprendizado de Máquina é "fazer com que computadores modifiquem ou adaptem as suas ações (...) de forma que elas se tornem mais precisas, e essa precisão é medida com base em quão bem as ações escolhidas

[23] Disponível em: http://nlx.di.fc.ul.pt/

refletem as ações corretas" (p. 5 – tradução nossa)[24]. Logo, a ideia contida no AM é a de que, a partir de um treinamento, o computador (mais precisamente, os algoritmos de AM) possa "aprender" uma tarefa: a máquina seria capaz de, a partir de um conjunto de dados de entrada, "aprimorar a si mesma no desempenho de sua tarefa" (FERREIRA; LOPES, 2017, p. 204).

Há mais de uma possibilidade de execução de AM: Supervisionado, Não Supervisionado e por Reforço. No primeiro, que é o que interessa aos fins desta pesquisa, oferece-se ao algoritmo um conjunto de entradas com informações (os atributos) e exemplos de saídas esperadas, que servirá para que o algoritmo aprenda a correspondência entre atributos da entrada e as respectivas saídas. Em outras palavras, a partir de uma entrada e de uma saída conhecidas, o algoritmo deve ser capaz de aprender a tarefa com tal precisão que consiga prever a saída desconhecida de uma entrada totalmente nova. No Aprendizado Não Supervisionado, não se fornece nenhum tipo de informação sobre a saída esperada, apenas sobre as entradas, e o algoritmo precisa encontrar sozinho regularidades nos dados de entrada. O Aprendizado por Reforço é estabelecido de forma que haja interação e *feedback* conforme o desenvolvimento da tarefa, "premiando" saídas corretas e "punindo as incorretas" (MARSLAND, 2009). A tarefa desta pesquisa consistirá em AM Supervisionado, pois interessa correlacionar atributos identificados por análises linguísticas com a presença de desvios sintáticos. Além disso, pretende-se identificar e analisar os atributos mais relevantes para as decisões tomadas pelos algoritmos, o que só é possível no Aprendizado Supervisionado. Assim, esse tipo será o foco desta seção e, a partir daqui, todas as menções a AM se referem a ele.

Uma das questões primordiais do processo de AM consiste na seleção dos atributos (em inglês, *features*) relevantes a partir dos quais o algoritmo aprenderá. Tais atributos são os dados de entrada do algoritmo (FERREIRA; LOPES, 2017). Assim, em uma tarefa de classificação, a partir da inserção de um conjunto de exemplos de treinamento (por exemplo, o *corpus*) com os respectivos atributos, e da classe desses exemplos (a classificação pretendida, resultante do processo de anotação do *corpus*), que deve ser conhecida para o treinamento, o algoritmo deve ser capaz de aprender as relações existentes entre atributos e classe (SOUZA; DI-FELIPPO, 2018). O processo no qual o algoritmo de AM acessa os dados conhecidos para aprender uma tarefa é chamado de *treinamento do modelo* (FERREIRA; LOPES, 2017), e a parcela do *corpus* utilizada para essa tarefa é chamada de *corpus de treinamento*.

Os resultados obtidos com essa técnica dependem em grande medida da tarefa e do algoritmo de AM selecionado. Em termos de tarefa, há dois tipos de problemas: regressão e

[24] "Machine Learning, then, is about making computers modify or adapt their actions (…) so that these actions get more accurate, where accuracy is measured by how well the chosen actions reflect the correct ones."

classificação. A tarefa proposta neste estudo consiste em um problema de classificação. Entre as possibilidades de algoritmos, há três paradigmas nos quais eles costumam ser classificados: simbólico, conexionista e matemático (ou probabilístico). Em geral, os algoritmos simbólicos são facilmente interpretáveis, uma vez que "descrevem os padrões aprendidos em uma linguagem de fácil compreensão para humanos" (SOUZA; DI-FELIPPO, 2018, p. 143). Como exemplos de algoritmos de cada um dos paradigmas, citam-se: (i) no paradigma conexionista, as redes neurais, (ii) no paradigma matemático, os algoritmos *Naïve Bayes* e *Support Vector Machines* e, (iii) no paradigma simbólico, as árvores de decisão. A evolução do AM tem sido intensa nos últimos anos, com o surgimento do aprendizado profundo (*deep learning*) e as possibilidades utilizando redes neurais. Não cabe aqui uma discussão ou apresentação detalhada desses conceitos, uma vez que a pesquisa pretende utilizar apenas os algoritmos que permitem uma interpretação facilitada por pessoas sem formação em computação. O AM, para este estudo, servirá como uma ferramenta com o potencial de explicitar correlações de forma mais ágil do que as análises manuais.

Nesse sentido, há diversas possibilidades de execução de tarefas de AM, tanto por meio de *scripts*, utilizando bibliotecas específicas para as linguagens de programação, quanto por meio de *softwares* com interfaces gráficas que unem diversos algoritmos. Um deles é o *Weka* (*Waikato Environment for Knowledge Analysis*)[25], um *software* de código aberto desenvolvido pela Universidade de Waikato, na Nova Zelândia (HOLMES; DONKIN; WITTEN, 1994). Ele agrega diversos algoritmos de AM de diferentes paradigmas para mineração de dados, em uma interface gráfica amigável e de utilização fácil por usuários sem experiência em Computação. Em função dessas facilidades e da disponibilidade livre do *software*, optou-se pela sua utilização na etapa de AM desta pesquisa.

Outro aspecto que merece especial atenção em tarefas de AM é a avaliação, isto é, a confirmação de que o algoritmo realmente aprendeu a realizar determinada operação, e não está simplesmente gerando resultados aleatórios. Para isso, faz-se necessário um *corpus de teste*, composto por um conjunto de elementos anotados, mas cujas anotações não são disponibilizadas ao algoritmo durante a fase de treinamento (FERREIRA; LOPES, 2017). Assim, como dito anteriormente, a partir dos dados fornecidos no treinamento, o algoritmo deve ser capaz de classificar elementos novos da forma mais correta possível. Logo, é fundamental que o *corpus* de teste seja selecionado *antes do treinamento* do algoritmo. Ignorar essa etapa

[25] Sabe-se que as técnicas de AM evoluem muito rapidamente, e há diversas ferramentas com resultados melhores do que a que foi escolhida. Porém, escolheu-se um percurso mais facilmente executável e interpretável por pessoas sem formação em disciplinas ligadas à Ciência da Computação.

provavelmente gerará resultados muito bons na avaliação, mas esses resultados são enganosos, pois não indicam o desempenho real do algoritmo frente a dados para os quais ele não foi treinado.

Para medir os resultados do AM, há uma série de métricas disponíveis, mas três delas merecem destaque: a precisão (em inglês, *precision*), a cobertura (em inglês, *recall*) e a medida-f (em inglês, *f-measure*) (MITCHELL, 1997). A precisão consiste no número de ocorrências em que o algoritmo acertou a classificação em relação ao número total de ocorrências. A cobertura indica o número de ocorrências classificadas corretamente em relação às ocorrências que *deveriam ter sido identificadas*. A medida-f é a média harmônica entre as anteriores. Em geral, conforme a cobertura aumenta, a tendência da precisão é diminuir, mas o ideal é atingir o equilíbrio, com resultados satisfatórios nas três métricas (FERREIRA; LOPES, 2017).

Por fim, ressalta-se que, ainda que se faça uma descrição de fenômenos linguísticos, esta pesquisa se insere na área do PLN porque faz uso de ferramentas de PLN (como *POS tagger* e *parser*) e porque busca fornecer subsídios para o desenvolvimento de tais ferramentas. Além disso, utilizam-se técnicas de AM para explicitar, em um conjunto de dados relativamente grande, possíveis correlações entre os atributos encontrados nas análises manuais e a presença ou ausência de desvios sintáticos. No próximo capítulo, apresentam-se alguns dos estudos anteriores que se relacionam mais diretamente com o objeto de estudo desta pesquisa.

3 TRABALHOS RELACIONADOS

Este capítulo traz estudos e pesquisas que antecederam este trabalho, especialmente no que diz respeito à análise de aspectos linguísticos em *corpora* de redações escritas por estudantes em formação e ao interesse da área de PLN nas tarefas de identificação, classificação, avaliação e correção automática ou semiautomática de redações. Inicia-se com os estudos que analisam redações do ponto de vista linguístico, trazendo na sequência aqueles voltados ao desenvolvimento de sistemas de PLN.

3.1 *Os estudos de* corpora *de redações do ponto de vista da Linguística*

O primeiro trabalho a ser citado indica que analisar *corpora* de redações é uma tarefa que desperta interesse científico já há algum tempo. Vale lembrar que a popularização dos computadores e do acesso à internet ocorreu no Brasil por volta dos anos 1990, o que permitiu a compilação de bases de textos eletrônicas, dando início à popularização das pesquisas em LC. Nesse trabalho, Nascimento e Isquerdo (2003) compilaram um *corpus* de 450 redações de vestibular de duas universidades do interior de São Paulo, escritas entre 1999 e 2000, para realizar uma análise léxico-estatística da frequência de palavras desses textos. Na análise, as autoras compararam os dados obtidos no *corpus* com um dicionário de frequências do português (BIDERMAN, 1978) e identificaram um conjunto lexical cuja frequência indicaria a ocorrência de um núcleo de palavras que têm mais chance de aparecer em qualquer tipo de texto.

Ainda no âmbito das análises lexicais, Grama (2016) realizou um estudo piloto sobre o uso inadequado de elementos coesivos sequenciais em 237 redações dissertativo-argumentativas referentes ao ano de 2014. No estudo, a autora utilizou a abordagem da LC e usou como ferramenta o *WordSmith Tools* (SCOTT, 1998), um *software* de análise lexical. A intenção era identificar os elementos coesivos usados de maneira equivocada em termos de sentido e, a partir disso, propor um tratamento lexicográfico desses elementos. Alguns dos coesivos que apresentaram usos inadequados foram *contudo, para tal efeito* e *ou seja*.

Também numa abordagem lexical, Evers (2018) analisou o léxico de redações do vestibular da Universidade Federal do Rio Grande do Sul, com base na LC, na Linguística Textual e nos estudos do léxico. A partir de um *corpus* de 341 redações, a autora analisou padrões léxico-sintáticos correspondentes a três faixas de notas, identificando a repetição de determinadas estruturas com o objetivo de obter a aprovação no vestibular com um bom desempenho na redação (fenômeno chamado de *engaiolamento*). Exemplos desse fenômeno

são as características semelhantes que a autora encontrou nas redações da maior faixa de notas, como número de parágrafos (entre cinco e seis), número de sentenças por parágrafo (de três a quatro) e menor presença de adjetivos, em comparação com as faixas de notas menores.

Outros trabalhos também compilaram *corpora* de redações, mas com intuitos diversos, sem focar atenções em particularidades lexicais. Numa perspectiva mais distante da LC, mas ainda relevante aos estudos relacionados à redação do ENEM, Luna (2009) enfocou o processo de avaliação das redações do ENEM, investigando os atores responsáveis por essa tarefa: os avaliadores. Para isso, tomou como base a Planilha de Critérios de Correção da Redação do ENEM, assim como entrevistas com os profissionais que atuam na avaliação de redações. Essa pesquisa teve um caráter interdisciplinar ao unir uma abordagem linguística e uma pedagógica.

A partir das questões ligadas à política de cotas, Francescon e Fernandes (2010) analisaram as redações produzidas no âmbito do processo seletivo para ingresso na Unioeste em 2009. Com base em um *corpus* de 122 redações, as autoras pretendiam identificar diferenças nos textos produzidos por candidatos cotistas e não cotistas, levando em conta as notas que eles obtiveram na redação. O objetivo era verificar o pressuposto de que alunos de escolas públicas teriam maior dificuldade de ingressar na faculdade, obtendo notas menores nas redações. No entanto, no *corpus* de pesquisa, esse pressuposto não se mostrou verdadeiro, pois as autoras não encontraram diferenças significativas entre as redações de alunos cotistas e não cotistas.

Já a pesquisa de Sousa (2016) se destaca em função do tamanho do *corpus* compilado. No seu estudo, a autora analisou a representação de papéis temáticos (isto é, as relações semânticas entre verbos e seus sujeitos e complementos) dos atores sociais, especificamente aqueles que se referiam a pessoas e a lugares, como *sociedade, pessoas, cidadão, Brasil, mundo*, em redações nos moldes do ENEM. Para isso, ela compilou um *corpus* de 1.405 redações produzidas entre 2009 e 2014 e extraídas do Banco de Redações do Portal UOL Educação, buscando identificar e descrever os principais papéis temáticos desempenhados pelos atores sociais especificados, a partir de pressupostos da LC, da Linguística Cognitiva, da Gramática de Papéis, da Análise Crítica do Discurso e da Linguística Sistêmico-Funcional. A pesquisa identificou que um ator social pode estar representado por papéis temáticos diferentes, e que apresenta circularidade (isto é, desempenha mais de um papel, em que um pode ser o contrário do outro, como no caso de *pessoas*, que pode ter papel tanto de agente quanto de paciente).

Em termos de descrição de desvios, a pesquisa de Sandoval e Zandomênico (2016) direcionou a atenção a um dos aspectos que também é analisado na presente dissertação: a concordância verbal nas redações do ENEM. O objetivo das autoras era verificar se havia diferenças nas marcações de concordância verbal em textos escritos por alunos egressos da

modalidade de Educação de Jovens e Adultos (EJA), em comparação com os egressos do Ensino Regular. Elas também queriam investigar se as marcas de plural ocorriam de forma diferente em estados com Índice de Desenvolvimento Humano (IDH) mais alto (Distrito Federal) ou mais baixo (Acre). Para isso, compilaram dois *corpora*: um composto por 100 redações do ENEM de 2013 de alunos do EJA dos dois estados, e outro composto por 100 redações do ENEM de 2012 de alunos do Ensino Regular desses mesmos estados.

Para a análise, elas coletaram todos os contextos de verbos flexionados no plural, de sujeitos compostos (p. ex., *[álcool e volante] não combinam*) ou formados por nomes coletivos (como *população*, *povo*) e de sujeitos simples no plural (como *os motoristas*). Em seguida, separaram os dados em nove categorias: sujeito nominal; sujeito pronominal (núcleo como pronome ou sujeito oculto); oração adjetiva (explicativa ou restritiva); sintagma nominal pesado (substantivo como núcleo estendido por locução adjetiva, locução adverbial e/ou oração); verbo inacusativo na ordem sujeito-verbo (como *existir*); verbo inacusativo na ordem verbo-sujeito; orações com ordem verbo-sujeito; sujeito coletivo. A hipótese da marcação diferente de plurais em estados distintos não se confirmou, pois as diferenças foram quase insignificantes. Porém, elas identificaram que, ainda que o plural tenha sido marcado na maioria dos casos, os textos do Ensino Regular tinham 89% de concordâncias, enquanto os do EJA tinham 72% de concordâncias adequadas. Alguns dos resultados de Sandoval e Zandomênico (2016) também se revelaram nos fenômenos linguísticos desta pesquisa, descritos no Capítulo 5, como a tendência da falta de concordância na inversão verbo-sujeito.

Outro trabalho que apresenta alguns resultados similares aos encontrados pelo presente estudo foi realizado no contexto do português europeu. Oliveira (2013) propôs um estudo de caso comparando desvios em textos formais escritos por professores e por alunos da Educação Básica da Região Autônoma da Madeira. Ele focou a análise nos desvios de ortografia, pontuação e coesão sintática, que afirma representarem cerca de 80% dos desvios observados, identificando os principais desvios desse tipo em cada grupo estudado. Como *corpus* de análise, o autor utilizou 185 textos formais produzidos por professores (atas, conselhos de turma, etc.) e 64 testes de redação produzidos por alunos do 12º ano de cursos científico-humanísticos.

Entre os desvios encontrados nos textos dos alunos, Oliveira (2013) destaca a quantidade de desvios ortográficos, considerando que “uma percentagem significativa desses erros parece dever-se à desatenção e à ligeireza com que abordam a escrita, podendo ser, pelo menos em parte, resolvida com uma revisão textual cuidada, que, na maioria dos casos, não existe”. Alguns dos desvios também foram identificados nesta pesquisa: confusão entre palavras homônimas e parônimas; aglutinação ou separação indevida de palavras; confusão

entre formas verbais; violação das regras de acentuação. No tocante aos desvios de pontuação, o principal foi o uso da vírgula (caso mais frequente entre os professores e segundo mais frequente entre os alunos, atrás apenas dos desvios ortográficos), e a sua supressão foi mais frequente do que o seu uso indevido. Já entre os desvios de coesão sintática, o autor descreve problemas de concordância; dificuldades com pronomes pessoais átonos; regências problemáticas (principalmente verbais); sintaxe do verbo *haver* e dificuldades na coesão de tempo e aspecto. Alguns desses fenômenos também foram observados nos desvios mapeados nesta dissertação, conforme se verá no Capítulo 5. Ainda que seja um estudo realizado no contexto europeu, é válido para mostrar que a questão dos desvios não se restringe ao Brasil, e que os desvios dos estudantes brasileiros e portugueses apresentam-se de forma similar.

Em uma abordagem que se coloca na interface entre as análises linguísticas e o PLN, destaca-se como mais relevante e mais similar a esta pesquisa o trabalho de Pinheiro (2008), especialmente porque foi a partir da tipologia de desvios gramaticais proposta pela autora que se construiu a tipologia utilizada neste estudo, que é apresentada na Seção 4.3.1 (p. 65). No seu trabalho, ela compilou um *corpus* de 249 redações do ENEM[26] (o *corpus* CORVO), obtidas do banco de redações do INEP, do ano de 2002. O seu objetivo era mapear os desvios gramaticais mais frequentes nos textos e as suas implicações para o desenvolvimento do revisor gramatical automático *ReGra* (MARTINS *et al*., 1998). Assim, as redações selecionadas para o *corpus* foram aquelas das faixas de notas mais baixas de rendimento na Competência 1, porque a autora partiu do pressuposto de que era justamente nessas redações que haveria uma maior quantidade de desvios. O interessante nesse trabalho é justamente a abordagem que utiliza a LC com foco na revisão automática e os interesses específicos do PLN, o que também motiva o presente estudo.

Para realizar a anotação dos desvios, Pinheiro (2008) adaptou a tipologia utilizada na identificação dos desvios mais frequentes executada durante a construção do *ReGra*, sendo composta por 17 categorias: (i) uso de conjunções, (ii) concordância entre modos e tempos verbais, (iii) concordância nominal, (iv) concordância pronominal, (v) concordância verbal, (vi) uso de crase, (vii) uso de artigos e determinantes, (viii) uso de *mal*/*mau*, (ix) uso de *onde*/*aonde*, (x) uso de preposições, (xi) uso de pronomes, (xii) uso dos porquês, (xiii) pontuação, (xiv) uso de particípio, (xv) regência verbal, (xvi) regência nominal e (xvii) uso dos verbos. A pesquisadora previu ainda uma categoria genérica para designar desvios que não se encaixassem em nenhuma das outras. Para definir o que era desvio gramatical e o que era de

[26] As redações eram originalmente arquivos de imagem, que precisaram ser transpostos para registro digitado.

outras ordens, ela se baseou naquilo que o *ReGra* havia sido projetado para identificar como desvio gramatical, uma vez que a ferramenta de correção ortográfica já existia previamente[27]. A autora afirma que a tipologia é consistente para classificar os desvios, mas é incompleta.

Para a anotação, foi desenvolvida uma ferramenta computacional de detecção e etiquetagem dos desvios, que utilizava o *ReGra* como base: o Identificador de Desvios. Tal ferramenta foi criada com diversas funcionalidades e interface atraente, de modo agilizar a classificação de desvios inicialmente como mecânicos, ortográficos, gramaticais ou de estilo. Na sequência, os gramaticais seriam classificados conforme a tipologia. Porém, Pinheiro (2008) afirma que a detecção dos desvios pela ferramenta se mostrou insatisfatória, porque o corretor gramatical usado como base (o *ReGra*) não pôde ser implementado com todas as suas funcionalidades, por questões de licença de uso. Uma das funcionalidades indisponíveis foi o corretor ortográfico, que, segundo a autora, impactou significativamente o desempenho da ferramenta. Nesse sentido, foi grande a quantidade de desvios não detectados.

Assim, para obter parâmetros consistentes de análise, Pinheiro realizou o registro em planilha de dados de todos os desvios de ordem mecânica, ortográfica, gramatical e de estilo no *corpus*. Tal detecção foi feita diretamente no editor de textos *Microsoft® Word*, já que a ferramenta de correção ortográfica do *ReGra* estava implementada nele. Nessa etapa, também foram registrados os indicadores de desempenho do *ReGra*: verdadeiros positivos para cada um dos quatro tipos de desvios e falsos negativos somente da análise gramatical e de estilo.

A partir das anotações e do levantamento de dados, Pinheiro (2008) identificou que o *ReGra* interveio em 98% dos textos do *corpus*; em 75% das vezes, houve mais de uma intervenção. A análise também ressaltou o fato de que houve muitos desvios de ortografia, sendo a maioria entre as intervenções. Em termos de gramática, o *ReGra* fez intervenções devidas em 59% dos textos, mas intervenções indevidas (falsos positivos) em 55% deles. Algumas das questões identificadas são similares ao que foi encontrado no presente estudo, como ficará demonstrado no Capítulo 5. Por exemplo, a autora cita que a acentuação foi "fortemente negligenciada pelos redatores" (p. 110):

> A palavra "país", por exemplo, de alta freqüência no corpus, em virtude do tema da redação, não foi acentuada muitas vezes, o que, por sua vez, acarretou falsos negativos gramaticais em segmentos como "nosso pais", em que se pede a concordância. Mas esses casos foram detectados em sua totalidade pelo revisor; o problema está na não detecção de desvios de acentuação também freqüentes (ausência do acento em "é" e "está", p.ex.) (…) (PINHEIRO, 2008, p. 110)

[27] A presente pesquisa utilizou uma abordagem similar para definir a diferença entre desvio ortográfico e sintático, como explicado anteriormente.

No que se refere aos desvios gramaticais, os mais frequentes identificados pelo *ReGra* foram de concordância verbal, concordância nominal, uso de pontuação e uso de crase. Um aspecto levantado como muito frequentes foi a ausência de concordância do plural de *ter*, o que também ocorreu na presente pesquisa. Em termos de pontuação, Pinheiro (2008) identificou ser este um desvio dos mais frequentes tanto entre os identificados pelo corretor gramatical quanto entre os não identificados (os falsos negativos). A autora destaca que a vírgula foi o sinal de pontuação no qual mais ocorreram desvios.

3.2 *Sistemas de PLN voltados às redações escolares*

Passando aos trabalhos que se inserem mais diretamente no campo de PLN, Litman (2016) destaca que o interesse dessa área por desenvolver aplicações para a educação data dos anos 1960, quando os primeiros trabalhos começaram a focar a avaliação automática de textos escritos por estudantes. Atualmente, identifica-se um grande interesse comercial e acadêmico pelo uso de ferramentas e aplicações de PLN no contexto educacional.

Uma grande quantidade de projetos desenvolvidos na área de avaliação e correção automática de textos foca em aprendizes de inglês (DALE; KILGARRIFF, 2011; LEACOCK *et al.*, 2014; DROLIA *et al.*, 2017). Nesse contexto, vale destacar o trabalho desenvolvido pelo ETS (*English Test Service*) (BURSTEIN; CHODOROW, 2010; HEILMAN *et al.*, 2014; MACARTHUR; GRAHAM; FITZGERALD, 2016) também em termos de pesquisa científica. Duas das plataformas comerciais *on-line* do ETS para a avaliação de redações em inglês são o *Criterion Online Writing Evaluation* (BURSTEIN; CHODOROW; LEACOCK, 2003) e o *e-rater Scoring Engine* (ATTALI; BURSTEIN, 2006). Além disso, a campanha de avaliação anual da conferência internacional CONLL (*The SIGNLL Conference on Computational Natural Language Learning*)[28] teve como objetivo a criação de sistemas de correção gramatical automática nas suas edições de 2013 e 2014 (NG *et al.*, 2014, 2015), incentivando a criação de ferramentas capazes de agilizar a avaliação/correção de textos de aprendizes.

De acordo com Litman (2016), há dois caminhos possíveis no campo da avaliação automática de redações: o primeiro é avaliar a proficiência linguística do estudante para fins que se restringem à atribuição de notas; o segundo é avaliar o seu trabalho e oferecer *feedbacks* a fim de melhorar o seu desempenho ou aumentar a proficiência em termos de escrita. Como

[28] Disponível em https://www.conll.org/.

as ferramentas de correção gramatical e ortográfica não enfocam desvios particularmente importantes para aprendizes de uma língua (materna ou estrangeira), as pesquisas passaram a se interessar pela detecção de desvios gramaticais (LEACOCK *et al.*, 2010).

Segundo Litman (2016), tem crescido o interesse nesses sistemas que forneçam *feedbacks* e sugestões de correção para os desvios detectados. Muitos dos sistemas desenvolvidos para a língua inglesa têm bom desempenho em termos de notas, em comparação com avaliadores humanos, utilizando atributos facilmente computáveis, como o tamanho das redações. Entretanto, isso não corresponde aos aspectos considerados pelos avaliadores humanos. Assim, para que se possa oferecer *feedbacks* válidos, as dimensões de avaliação devem considerar atributos que estejam relacionados às questões avaliadas por humanos.

Nesse contexto, alguns trabalhos têm mostrado que as habilidades escritas dos estudantes podem melhorar após eles receberem *feedbacks* e sugestões de correção em plataformas de avaliação automática de textos. Um exemplo é o estudo feito por Wu (2018), que investigou o impacto de uma ferramenta de avaliação automática de redações no desenvolvimento das habilidades escritas dos estudantes. O objetivo era verificar se a proficiência escrita de aprendizes chineses de inglês como L2 melhorava com a utilização da plataforma *Piagaiwang*. Para essa avaliação, os estudantes forneciam um esboço da redação, para o qual recebiam o *feedback* automático com a identificação dos desvios e as sugestões de correção. Em seguida, era submetida uma versão revisada, para a qual novamente os estudantes recebiam um *feedback*. Por fim, era submetida à plataforma a versão final da redação. Seis tipos de erros foram frequentemente mencionados pela ferramenta: ortografia, colocação, erros em verbos, erros em nomes, uso equivocado de palavras e sentença errada. Conforme a pesquisa, os desvios marcados pela ferramenta tiveram a sua ocorrência minimizada, o que demostrou que ela colaborou para o aumento da proficiência escrita desses estudantes.

Segundo Litman (2016), um dos problemas de se usar ferramentas de PLN tradicionais no contexto educacional é que elas são treinadas em textos escritos profissionalmente, como *corpora* de textos jornalísticos. Portanto, essas ferramentas em geral não têm um bom desempenho quando aplicadas a textos escritos por aprendizes. Nesse contexto, algumas pesquisas investigam o desempenho de ferramentas de PLN ao processarem textos não revisados, isto é, aqueles escritos por estudantes e que geralmente contêm desvios diversos.

Napoles *et al.* (2016) mediram a queda de desempenho do *parser* por dependências *Stanford Dependency Parser* a partir de um número controlado e de tipos específicos de desvios das sentenças. Para a pesquisa, os autores utilizaram o *corpus* de treino do NUS *Corpus of Learner English* (NUCLE), formado por redações escritas por aprendizes de inglês, cujos

desvios gramaticais estão classificados em 28 categorias, e cada desvio possui a respectiva correção. Nesse trabalho, os autores focaram nos seis tipos de desvios mais frequentes: artigo ou determinante; erros mecânicos (pontuação, maiúsculas/minúsculas, desvios ortográficos); concordância nominal; preposições; formas de palavras; desvios de tempo e forma verbal. A partir da métrica de avaliação *medida-f dos elementos etiquetados*, a pesquisa mostrou que um maior número de desvios em uma sentença diminui a acurácia do *parser*, mas há tipos de desvios que têm um maior impacto no *parsing*, sozinhos ou em combinação com outros. Por exemplo, a substituição de verbos, os desvios mecânicos (especialmente o uso de vírgulas) e preposições faltantes ou desnecessárias são mais críticos para o desempenho do *parser* avaliado. Porém, a distância maior ou menor entre desvios na sentença não impacta a acurácia.

Huang *et al.* (2018), por sua vez, avaliaram o desempenho de sete *parsers* de modo a comparar a sua acurácia e fornecer dados para pesquisadores que precisem decidir sobre qual a melhor ferramenta para textos de aprendizes. Para o estudo, eles utilizaram 1.000 sentenças extraídas do *corpus EF-Cambridge Open Language Database*, que contém textos escritos por aprendizes de inglês advindos de 188 países. Os autores também investigaram um possível viés inserido por humanos ao corrigirem a saída do *parser*, bem como o efeito dos desvios no *parsing*, e compararam o desempenho das ferramentas nos textos de aprendizes com aquele obtido em textos bem-escritos. Huang *et al.* (2018) demonstram que oferecer ao anotador apenas a saída de um *parser* aumenta o número de desvios não vistos. Já no desempenho das ferramentas, 39,2% dos erros de *parsing* continham um desvio gramatical, e 63% dos desvios dos aprendizes causaram erros de *parsing* (pontuação e desvios ortográficos foram os que mais causaram erros). Por fim, a comparação do desempenho dos *parsers* em textos revisados mostrou que, em termos percentuais, as diferenças não são tão significativas; porém, o *corpus* de aprendizes contém sentenças muito menores, em média, e o desempenho das ferramentas costuma ser melhor em sentenças mais curtas, o que justificaria o resultado. Na comparação entre *parsers*, os autores concluem que o desempenho em textos revisados pode predizer o desempenho em textos de aprendizes, isto é, a ferramenta que melhor lida com textos bem-escritos será também a melhor para textos com desvios.

Outro trabalho que estudou o desempenho de *parsers* é o de Lyashevskaya e Panteleeva (2017), que avaliaram a ferramenta escolhida pela presente pesquisa: o *UDPipe*. As autoras criaram um esquema de anotação no modelo do UD para o *corpus* de redações de aprendizes russos de inglês como L2 *Russian Error-Annotated English Learner Corpus* (REALEC)[29] e

[29] Disponível publicamente em http://realec.org.

avaliaram a ferramenta comparando-a com a anotação manual das árvores sintáticas. A escolha do *UDPipe* foi justificada em função da sua facilidade de uso e do fato de que ele fornece informações de *POS tags*, funções sintáticas, relações de dependência e estrutura que facilita tarefas de contagem. Para as 373 sentenças analisadas, o *parser* identificou corretamente os núcleos sintáticos (*heads*) em 92,9% dos casos, dos quais 91,7% tiveram suas dependências corretamente etiquetadas. Nos demais nodos, 95,8% deles foram etiquetados corretamente. Os casos em que o *parser* mais cometeu erros foram homonímia sintática, ordem de palavras agramatical e desvios ortográficos e gramaticais. As autoras também demonstraram que grande parte dos problemas de *parsing* seria resolvida com a ajuda de um corretor ortográfico, o que permitiria uma análise específica da estrutura sintática dos aprendizes.

Em comum, os trabalhos que analisam ferramentas de PLN têm os tipos de *corpora* utilizados: trata-se de textos escritos por aprendizes de inglês como língua estrangeira. Há ainda poucos trabalhos que se ocupem dos textos de aprendizes de português, tanto como língua estrangeira quanto como língua materna. Como exemplo de pesquisa desenvolvida para esse público, merece destaque o SciPo, um ambiente *web* composto por um conjunto de ferramentas integradas para auxiliar estudantes a escreverem textos acadêmicos (resumos e introduções) da área de Computação (FELTRIM, 2004). Essa ferramenta de auxílio à escrita, diferentemente das abordagens que focam em aprendizes de alguma língua estrangeira, tem como principal público universitários falantes nativos de português.

Cabe citar ainda a pesquisa de doutorado de Torres (2016), que propôs a criação de recursos para aprendizes de português que são hispanoablantes, de forma a contribuir com as suas produções textuais acadêmicas por meio de uma ferramenta de suporte à escrita acadêmica. A pesquisa teve como foco o nível lexical e tomou como ponto de partida o ferramental metodológico da Linguística de *Corpus*. Outra ferramenta criada para aprendizes de português é o Avalingua, um sistema de auxílio à escrita que busca identificar e classificar automaticamente diversos níveis de desvios em *corpora* de aprendizes de português que tenham como língua materna o galego, além de propor soluções com o objetivo de melhorar as habilidades linguísticas de seus usuários (GAMALLO *et al.*, 2015). Em termos de análise da influência de desvios em ferramentas de PLN utilizando como fonte a língua portuguesa, não foi possível encontrar ainda qualquer pesquisa que realize essa tarefa de forma sistemática.

No âmbito das pesquisas brasileiras em avaliação/correção automática de redações em português, vem crescendo o interesse do PLN pelo desenvolvimento de ferramentas que se proponham a essas tarefas. Citam-se aqui apenas alguns dos trabalhos. Bazelato e Amorim (2013) desenvolveram um classificador para a correção automática de redações que serviu

como *baseline* para pesquisas subsequentes. Para isso, os autores utilizaram um *corpus* de treino com 379 redações e de teste com 50 redações, obtidas da plataforma de redações do Portal UOL Educação. O objetivo era desenvolver, a partir de técnicas de Aprendizado de Máquina, um classificador *bayesiano* que atribuísse notas às redações de forma automática. Ainda que o desempenho do classificador tenha sido consideravelmente inferior às avaliações humanas (correlação de Pearson de 0,396, em comparação com 0,564 dos humanos), esse trabalho estabeleceu bases a partir das quais outras pesquisas puderam se desenvolver.

Santos Júnior *et al.* (2015) propuseram a automatização da análise ortográfica e gramatical de atividades escritas produzidas em ambientes de aprendizagem à distância, de forma a diminuir a sobrecarga de trabalho do tutor/professor em relação à correção dessas atividades. Tal ferramenta segue a avaliação por competências do ENEM, tendo sido projetada em uma arquitetura de dois módulos: um ortográfico e um gramatical. O módulo gramatical usa como base o corretor gramatical de código aberto CoGrOO[30]. No módulo ortográfico, a correção é feita consultando-se o dicionário JSpell e um dicionário auxiliar, no qual se podem acrescentar palavras. Para avaliar o sistema, utilizaram-se 20 redações disponíveis na *web*[31]. O módulo ortográfico foi avaliado em 10 redações, divididas em dois grupos para duas etapas. Na primeira, o algoritmo teve taxa de acerto de 60%. Após verificar que os desvios não identificados estavam em palavras ausentes no dicionário, estas foram acrescentadas. Em nova avaliação, o sistema obteve 80% de acertos. O módulo gramatical foi avaliado com o mesmo método e, no primeiro conjunto de redações, obteve uma taxa de acerto de 60%. Após avaliação e reformulação de algumas regras, o módulo teve novamente taxa de acerto de 60%. Quando os dois módulos foram aplicados simultaneamente, obteve-se uma taxa de acertos de 80%.

Almeida Júnior *et al.* (2017), por sua vez, desenvolveram um sistema de avaliação automática da Competência 1 em redações do ENEM, utilizando para isso técnicas de AM e de PLN. Os autores se basearam em 4.547 redações obtidas do Portal UOL Educação, que disponibiliza redações enviadas por alunos e as respectivas correções dos avaliadores. O objetivo desse trabalho foi apresentar um sistema que pudesse reduzir os esforços envolvidos no processo de correção/avaliação das redações, a partir da avaliação automática da Competência 1. No sistema desenvolvido, cada redação foi representada como um vetor composto pelas seguintes características: número de parágrafos, frases, palavras, caracteres e erros ortográficos identificados pelo corretor ortográfico e analisador morfológico *Hunspell*,

[30] O trabalho não traz nenhuma informação ou referência sobre o desempenho dessa ferramenta em termos de identificação dos desvios gramaticais.
[31] Os autores não informam onde buscaram as redações.

complementado por um dicionário auxiliar; 124 erros gramaticais identificados pelo corretor ortográfico CoGrOO; número de vírgulas, pontos, pontos de exclamação e interrogação, além das classes gramaticais do português. O classificador escolhido foi o SVM, e cada redação pertencia a uma classe, conforme a nota atribuída pelo avaliador humano. O resultado obtido foi de 52% de acertos considerando-se apenas a nota específica dada pelo avaliador, e 93% de acertos considerando-se notas adjacentes a 0,5 pontos de distância da nota do avaliador.

Já Amorim e Veloso (2017) propuseram uma avaliação automática multiaspectual de redações do ENEM a partir de duas questões principais: como atributos objetivos se comportam em um sistema de avaliação automática multiaspectual de redações; e quais os atributos mais relevantes para cada aspecto. O *corpus* da pesquisa consistiu em 1.840 redações extraídas do Portal UOL Educação, e os aspectos avaliados equivaliam às cinco competências do ENEM. Os atributos utilizados pelos autores se dividiram em de domínio e genéricos. O primeiro grupo considerava questões como uso de primeira pessoa do singular, uso de ênclise, número de pronomes demonstrativos. O segundo englobava os seguintes atributos:

- gramática e estilo: número de desvios gramaticais identificados pelo CoGrOO, número de desvios ortográficos, número de desvios de estilo, etc.;
- atributos sintáticos: número de sentenças com mais de 70 caracteres;
- organização e desenvolvimento: número de marcadores discursivos total e por sentença;
- complexidade lexical: índice Flesch, média de palavras por sentença, número de *tokens*, número de *types*;
- uso de vocabulário específico do *prompt*: similaridade entre vetor de frequência de palavras no *prompt* e na redação.

A classificação multiaspectual obteve valores de *kappa* de 0,4245, o que foi considerado satisfatório. Em termos de relevância dos atributos, os de vocabulário e de complexidade lexical foram os mais relevantes para a nota total. Porém, os atributos mais relevantes para cada um dos aspectos diferiram, o que, segundo os autores, indica que seria melhor treinar um classificador específico para cada competência.

Nau *et al.* (2017) desenvolveram uma ferramenta para a identificação automática de desvios de linguagem, de forma a subsidiar o processo de revisão de textos escritos por estudantes ou auxiliar na revisão de artigos e documentos. A ferramenta foi desenvolvida em duas etapas: construção do catálogo de palavras e expressões a serem detectadas, e aplicação de técnicas de PLN. Os desvios-alvo foram arcaísmos, barbarismos, cacófatos, plebeísmos, pleonasmos, chavões e clichês, e marcas de oralidade. Também foram consideradas como

desvio sentenças com trechos de mais de 45 palavras sem sinal de pontuação. Para a construção do *corpus* de teste, eles utilizaram 762 redações também extraídas do Portal UOL Educação. Entre as técnicas utilizadas, estavam remoção de *stopwords*, lematização e extração de n-gramas. Como resultado, a ferramenta identificou 3.255 desvios e foi considerada apta a identificar os desvios de linguagem pretendidos, que, segundo os autores, na maioria das vezes não são identificados pelas ferramentas embutidas nos editores de textos existentes.

O trabalho de Galhardi *et al.* (2018) descreve um analisador léxico-morfológico de redações nos moldes do ENEM. A base de dados utilizada foi novamente o banco de redações do Portal UOL Educação, do qual foram extraídas 20 redações. O sistema, cuja descrição não apresenta experimentos de avaliação, utiliza técnicas como *tokenização*, lematização e *POS tagging* para analisar questões lexicais e morfológicas das redações, permitindo que se busque uma palavra e apresentando como saída todas as formas dessa palavra que aparecem nas redações, bem como significado, frequência, classe gramatical, contexto, entre outros. A proposta era utilizar essa ferramenta como primeira etapa de análise das redações, para então seguir com os demais níveis linguísticos.

Outro trabalho que investigou um aspecto específico da avaliação automática de redações foi o de Cândido e Webber (2018), que se ocuparam da coesão textual. O seu objetivo principal foi implementar uma ferramenta de análise e avaliação automática da coesão textual a partir da *Teoria do Foco* e da *Teoria da Centragem*. Para a avaliação da ferramenta, os autores utilizaram 35 redações escritas por alunos do curso de Engenharia de uma instituição de ensino superior. Os valores obtidos na avaliação da coesão textual foram comparados com as notas atribuídas por uma banca avaliadora, composta por dois especialistas, as quais tinham valores entre 0 e 2 pontos. As notas dos dois avaliadores tiveram uma diferença média de 0,16 pontos após a remoção de redações com mais de 0,4 pontos de diferença entre ambas as notas. Na análise da ferramenta, manteve-se como base essa mesma diferença máxima de 0,4 pontos, obtendo-se 70% de compatibilidade global com as notas dadas pelos avaliadores.

Por fim, Fonseca *et al.* (2018) descrevem o desenvolvimento de um sistema de avaliação automática de redações nos moldes do ENEM. Tal sistema também considerou as cinco competências do ENEM para atribuir as notas às redações, e foi desenvolvido a partir de duas abordagens: redes neurais profundas e engenharia de *features*. Para os experimentos, os autores utilizaram 56.644 redações escritas por estudantes de Ensino Médio em uma plataforma *on-line* (a mesma na qual foram escritas as redações do *corpus* da presente pesquisa), com as respectivas notas dadas por avaliadores humanos, e divididas em conjuntos de treino, desenvolvimento e teste. Para a abordagem via engenharia de *features*, os textos foram

tokenizados e etiquetados com um *POS tagger*, e o sistema usou como referência um vocabulário composto pela lista de palavras Unitex DELAF, acrescido de algumas palavras ausentes no recurso. Os autores afirmam que, idealmente, gostariam de utilizar também um *parser*, mas não o fizeram porque *parsers* treinados em textos gramaticalmente bem-escritos em geral apresentam um desempenho ruim em textos que contêm desvios. As *features* (ou atributos) se dividiram em cinco grandes categorias:

- contagens genéricas: número de vírgulas, número de caracteres, número de parágrafos, número de sentenças, média de sentenças por parágrafo, tamanho médio de sentença (*tokens*), tamanho do vocabulário válido (número de *types* que constam no vocabulário), número de *tokens* excluindo pontuações, tamanho médio de palavra (caracteres), número de *types* fora do vocabulário, número de *tokens* fora do vocabulário, razão *type*/*token* de palavras fora do vocabulário, número de *tokens* lexicais que aparecem no *prompt*, número de *tokens* lexicais que aparecem no *prompt* de textos de apoio, frequência média e menor frequência de palavra no *corpus*;
- presença de expressões específicas: presença e número de agentes sociais; conectivos em início de parágrafo, conectivos em início e meio de sentença, número de conectivos, número de conectivos diferentes, número de marcas de oralidade, número de expressões propositivas;
- n-gramas de *tokens*: presença de n-gramas altamente correlacionados com a nota;
- n-gramas de POS *tags*: presença de n-gramas de POS *tags* altamente correlacionados com a nota da redação;
- contagem de POS *tags*: número de ocorrências de cada POS *tag* no texto.

No total, a abordagem por engenharia de *features* dos autores contava com 681 valores, mas nem todos eram relevantes para cada uma das competências do ENEM. Então, eles fizeram uma seleção específica por competência. Para essa abordagem, eles treinaram algoritmos de regressão supervisionados do tipo *gradient boosting* e de regressão linear. As métricas de avaliação utilizadas para ambas as abordagens foram RMSE (*root mean squared error*) e QWK (*quadratic weighted kappa*). Os autores identificaram que a abordagem por engenharia de *features* com o algoritmo *gradient boosting* teve melhores resultados para as Competências 1 a 4, enquanto a abordagem por redes neurais profundas se saiu melhor na Competência 5. Eles também identificaram que os modelos apresentaram melhores resultados na atribuição da nota total, em relação à atribuição das notas individuais, e que a Competência 1 se mostrou como a

mais fácil de avaliar. Todos os resultados ultrapassaram o *baseline* para a tarefa com margem significativa na avaliação das duas métricas.

Os trabalhos apresentados indicam um interesse acadêmico e comercial em questões relacionadas à análise, avaliação e correção de redações de estudantes em processo formação. Porém, os estudos nacionais ainda contam com pouca interdisciplinaridade. A pesquisa de Pinheiro (2008) destaca-se nesse contexto, tendo sido utilizada como base para diversas decisões tomadas durante esta pesquisa. Além disso, os estudos que propuseram sistemas de classificação e/ou avaliação automática de redações forneceram *insights* sobre as abordagens possíveis, especialmente sobre os atributos utilizados. Neste estudo, o objeto de análise restringe-se ao nível da sentença, e por isso os atributos que consideram questões do texto não puderam ser aplicados. No entanto, como se mostra no Capítulo 6, alguns atributos clássicos fizeram parte dos atributos extraídos e se mostraram decisivos para os resultados obtidos.

O próximo capítulo apresenta as características do *corpus* desta pesquisa, bem como as quatro etapas que a compõem: pré-processamento dos textos, anotação linguística dos desvios, extração dos atributos linguísticos e correlação com a presença de desvios sintáticos.

4 MATERIAIS E MÉTODOS

4.1 *Construção do* corpus

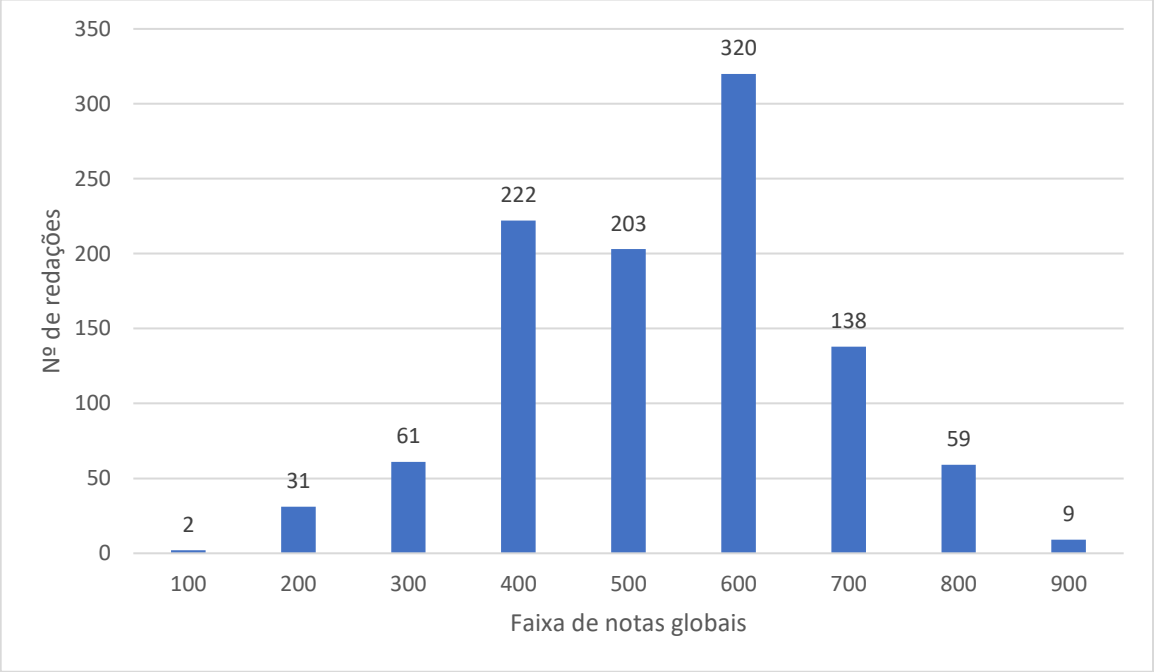
Descreve-se nesta seção o processo de construção do *corpus* utilizado na pesquisa, conforme os conceitos teóricos apresentados no Capítulo 2. Os textos foram coletados em formato digital na plataforma *on-line* da empresa Letrus[32], um centro de tecnologia e letramento que desenvolve ferramentas de escrita e avaliação de textos para escolas. A plataforma não oferece nenhum recurso de marcação ou correção de desvios de escrita de qualquer ordem. Assim, o trabalho de coleta dos textos já havia sido realizado e, portanto, faltavam ainda a sua seleção e caracterização inicial. Para utilizar o *corpus* fornecido, assinou-se um Termo de Sigilo de Dados, disponível no Anexo 1. Porém, durante a trajetória do mestrado, a empresa autorizou a disponibilização pública dos dados, contanto que fossem omitidas todas as metainformações.

O *corpus* é constituído por redações dissertativo-argumentativas escritas por estudantes do Ensino Médio nos moldes do ENEM, como simulados para esse exame. Os textos foram produzidos nos anos de 2017 e 2018 por alunos de escolas públicas de São Paulo (SP) e de Laguna (SC), sendo a grande maioria (1.014 textos) oriundos da primeira cidade, e apenas 31 textos oriundos da segunda. Uma das informações inicialmente fornecidas no *corpus* é a data de nascimento dos produtores: 1997 (5), 1998 (117), 1999 (448) e 2000 (362). Há ainda 113 textos em que as datas de nascimento não estão especificadas. Como a coleta foi realizada por terceiros, dispunha-se apenas das informações previamente fornecidas, não sendo possível preencher as informações faltantes. Os temas sobre os quais os estudantes deveriam discorrer incluíam legitimidade dos movimentos sociais, arte e cultura para transformação social, igualdade de gênero, sistema prisional brasileiro, saúde pública e importância da leitura.

As redações foram avaliadas por corretores humanos, que estabeleceram notas para as cinco competências, de acordo com a grade do ENEM, as quais foram somadas para formar a nota final. No Gráfico 1, apresentam-se as faixas de notas globais das redações nas avaliações humanas.

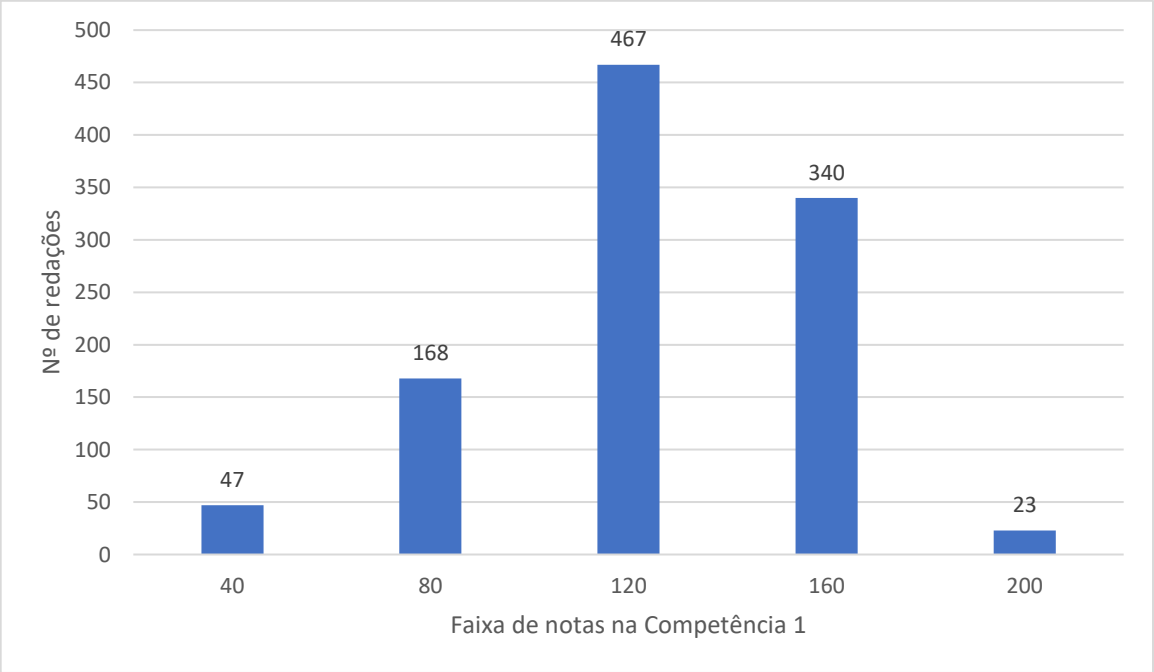
[32] Disponível em: https://www.letrus.com.br/

Gráfico 1 – Faixa de notas globais.



No Gráfico 2, exibem-se as faixas de notas obtidas pelos textos na Competência 1 do ENEM, também nas avaliações humanas:

Gráfico 2 – Faixa de notas na Competência 1.



Considerando-se as caracterizações relevantes apresentadas por Berber-Sardinha (2000) e trazidas na Seção 2.3 (p. 22), as características do *corpus* são mostradas no Quadro 2.

Quadro 2 – Características do *corpus*.

Gênero	Redação escolar nos moldes da redação do ENEM
Modalidade	Modalidade escrita formal da língua portuguesa
Modo	Escrito
Tempo	Sincrônico e contemporâneo
Seleção	De amostragem (estático)
Conteúdo	Monolíngue de gênero específico
Autoria	De aprendiz
Finalidade	De estudo/de treinamento ou teste

Originalmente, o *corpus* foi disponibilizado em planilha contendo 42 colunas, sendo que 28 delas contêm as seguintes metainformações: ID do aluno (apenas um código numérico para manter o sigilo), ID do texto, texto da introdução, texto do desenvolvimento, texto da conclusão, escola, cidade, estado, código da escola, turma, código da turma, ano letivo, número e código do teste, *scores* dos três algoritmos de avaliação automática desenvolvidos pela Letrus, data de nascimento do aluno, indicações de se a redação foi concluída, corrigida ou zerada e o motivo da anulação, seguidas das notas das cinco competências e da nota final. As 14 colunas finais trazem informações como número de palavras, número de conectivos, número de problemas ortográficos e demais caracterizações. Os dados originais foram preservados, e todas as alterações relatadas a partir daqui foram feitas em dados específicos.

Optou-se por excluir da seleção os 21 textos que foram anulados por alguma das formas de anulação previstas pelo ENEM, como "fuga ao tema" ou "não atendimento ao tipo textual" e aqueles que obtiveram nota final zero. Essa decisão foi tomada, nos textos zerados, porque a grade de correção do ENEM não prevê que textos que não foram anulados por algum dos motivos previstos recebam nota zero e, portanto, é provável que tenha havido algum equívoco no processo de avaliação humana. Além disso, como eles não possuíam notas para as cinco competências, isso poderia dificultar uma eventual correlação posterior com as notas[33]. O número final de textos que compõem o *corpus* é 1.045 redações.

A etapa de construção do *corpus* consistiu na extração dos textos e respectivas metainformações da planilha contendo o *corpus*, cujo resultado foi um arquivo em formato textual bruto por redação, em codificação UTF-8. Cada arquivo foi nomeado de acordo com o número referente à ID do texto previamente fornecida. Para a extração, a planilha foi ordenada por ordem crescente de ID do texto, que não tinha relação com as notas. Optou-se por essa identificação exatamente para evitar inserir viés na anotação, ocasionado pela ordem em que as notas (finais e referentes à Competência 1) apareciam, uma vez que se esperam menos desvios

[33] Para esta pesquisa, decidiu-se não utilizar as notas dos corretores humanos para nenhuma etapa. No entanto, como se pretende que o *corpus* possa ser disponibilizado publicamente, as notas podem servir para estudos futuros.

em textos com notas maiores. As metainformações também foram extraídas e colocadas no início do arquivo, marcadas com o símbolo "#"[34]. As três partes dos textos (introdução, desenvolvimento e conclusão) foram identificadas e aparecem na sequência do arquivo extraído, após todas as metainformações[35].

Essa etapa também englobou a geração de estatísticas básicas do *corpus* para fins de caracterização quantitativa, como número de *tokens*, número de *types*, riqueza lexical (em inglês, *type token ratio* ou TTR), número de sentenças e de parágrafos, número de desvios ortográficos, média de palavras por sentença, média de palavras por texto, média de sentenças por texto e média de parágrafos por texto. A maior parte dessas estatísticas pôde ser obtida facilmente a partir das metainformações fornecidas. Na Tabela 1, mostram-se as estatísticas iniciais.

Tabela 1 – Caracterização do *corpus*.

Descrição	**Número**
Nº palavras	325.111
Nº palavras únicas (*types*)	184.967
Riqueza lexical (TTR)	0,57
Nº sentenças	10.652
Nº parágrafos	4.572
Nº desvios ortográficos[36]	7.420
Média palavras/sentença	30,52
Média palavras/texto	311,11
Média sentenças/texto	10,19
Média parágrafos/texto	4,37

Tais estatísticas apresentam um panorama sobre o *corpus*, mas uma análise mais atenta dos dados mostra, por exemplo, que a média de palavras por sentença e a média de parágrafos por texto varia imensamente. Enquanto há sentenças compostas de uma palavra, há aquelas que chegam a ter mais de 100 palavras. Da mesma forma, enquanto há textos com apenas um parágrafo, há também exemplos de redações com 10 parágrafos. Logo, ainda que se possa ter

[34] As ocorrências desse símbolo nos textos foram substituídas pela expressão {HASHTAG}, para evitar problemas posteriores nos *scripts*, que de maneira geral deveriam ignorar linhas contendo esse símbolo.
[35] Na versão do *corpus* disponível publicamente, as metainformações não são fornecidas.
[36] Não foi fornecida pela empresa que coletou o *corpus* nenhuma informação sobre a forma de contagem de tais desvios. Logo, não é possível definir se são únicos ou não, e se estão limitados à ortografia, mas as análises parecem indicar que se trata de desvios únicos, pois, como descrito nas próximas seções, a presença de desvios ortográficos parece ser maior do que esse número aponta. Sugere-se cautela ao utilizar tal referência, mas se julgou interessante acrescentá-la à caracterização, uma vez que mostra uma presença significativa desse tipo de desvio.

uma ideia dos aspectos quantitativos, eles devem ser considerados com cautela. Quanto à representatividade do *corpus*, considera-se que ele seja representativo do fenômeno que se pretende mapear: os desvios sintáticos mais frequentes.

A próxima etapa foi a preparação dos dados para a anotação manual dos desvios, que exigiu a segmentação das sentenças, pois a anotação seria realizada nesse nível textual. Porém, na primeira tentativa de segmentação automática, percebeu-se que havia algumas características (como falta ou excesso de espaços antes ou após os sinais de pontuação, e eventuais sentenças repetidas) que causaram dificuldades na segmentação. Assim, procedeu-se à tarefa de "limpeza de ruídos", na qual os problemas de segmentação foram localizados e corrigidos manualmente.

A tarefa seguinte foi preparar os arquivos para anotação por sentenças, de acordo com a forma de segmentação utilizada pela biblioteca de Python "nltk". Para isso, criaram-se novos arquivos em formato textual (novamente um por redação) que contêm uma sentença por linha, conforme foram segmentadas automaticamente via "nltk". Para esta pesquisa, portanto, os limites de uma sentença (e o conceito de sentença em si) são definidos de acordo aquilo que o "nltk" utiliza para a segmentação automática, isto é, os sinais de pontuação de fim de sentença (ponto final, de interrogação e de exclamação), com algumas regras específicas (como desconsiderar os pontos utilizados em abreviações e números).

Como esta pesquisa enfoca os desvios sintáticos, os desvios puramente ortográficos foram corrigidos para facilitar a etapa de anotação. Essa correção foi feita para permitir que o anotador se concentrasse apenas nos desvios sintáticos. Os problemas ortográficos corrigidos foram somente aqueles identificados pelo módulo de correção ortográfica do *Microsoft® Word para Office 365 MSO (16.0.12624.20422) 32 bits* (doravante, *MS Word*)[37]. Essa etapa também definiu os limites entre o que era desvio ortográfico e o que era desvio sintático. Sabe-se que tal decisão tornou a anotação intrinsecamente vinculada a essa versão do corretor ortográfico e, portanto, datada, uma vez que, para reproduzir os experimentos realizados neste trabalho, é necessário utilizar essa mesma versão da ferramenta. Além disso, sabe-se que o *MS Word* separa desvios em ortográficos e gramaticais de maneira que os ortográficos são aqueles não encontrados no léxico que integra a ferramenta; o restante dos desvios é considerado gramatical. Logo, a abordagem escolhida para esta pesquisa torna-se limitada por uma classificação falha da ferramenta em relação aos tipos de desvios. Entretanto, por restrições de tempo, decidiu-se

[37] Preservou-se o arquivo original sem a correção dos desvios ortográficos. Assim, caso se queira futuramente proceder a uma análise dos desvios ortográficos (pelo menos aqueles identificados pelo *software* de correção automática utilizado), isso poderá facilmente ser realizado comparando-se o arquivo original com o corrigido.

não adentrar as questões teóricas envolvidas no estabelecimento das fronteiras entre desvios ortográficos e lexicais, deixando-as para trabalhos futuros que se ocupem dessa temática.

Para concluir a preparação do *corpus*, criou-se automaticamente uma planilha com todas as sentenças de todos os textos, a qual foi usada para a primeira fase da anotação manual. Tal planilha possui três colunas com as seguintes informações em sequência: (i) ID da sentença, formada pela ID do texto e pelo número que identifica a sua ordem sequencial de ocorrência no texto, (ii) o texto da sentença, (iii) a anotação de desvios, que foi totalmente preenchida pela letra N, indicando "sem desvio" (o processo de anotação é descrito na Seção 4.3, na p. 65). A fim de facilitar a visualização, inseriu-se uma linha vazia sinalizando o término de um texto e o início do outro. O processo de extração, limpeza e caracterização do *corpus* foi feito de forma automática, sempre que possível, para agilizar o trabalho[38]. Ainda que a preparação do *corpus* tenha demandado tempo considerável, um pré-processamento rigoroso foi fundamental para garantir a correção da segmentação de sentenças para a anotação, evitando retrabalhos.

Essa etapa também englobou a seleção de aproximadamente 20% das sentenças do *corpus* (1.998 sentenças), as quais foram separadas na planilha preparada para a anotação. Tais sentenças foram utilizadas como *corpus* de teste para o AM. A anotação dessas sentenças foi realizada após a última etapa da pesquisa como forma de garantir que nenhum viés de anotação fosse inserido nos algoritmos treinados, a partir da identificação de atributos que eventualmente poderiam ocorrer apenas nessa parcela do *corpus*.

4.2 *Anotação automática via* parsing

Os arquivos resultantes do *parsing* automático foram utilizados em duas etapas da pesquisa: para o carregamento de arquivos na plataforma *on-line* utilizada na segunda fase da anotação manual de desvios (descrita na Seção 4.3.3, p. 72); e na extração automática de atributos linguísticos (descrita na Seção 4.4, p. 79). Em relação à primeira etapa, a interface de anotação FLAT (*Folia Linguistic Annotation Tool*) (GOMPEL; REYNAERT, 2013) aceita como entrada o formato de arquivo CONLL-U. Para obter esse formato, executou-se o *parsing* utilizando-se o *UDPipe* (STRAKA; HAJIČ; STRAKOVÁ, 2016). Assim, as sentenças com desvio foram extraídas automaticamente da planilha utilizada na primeira fase da anotação (descrita na Seção 4.3.2, p. 71) e salvas em um arquivo textual único, o qual foi *parseado* com o *UDPipe*.

[38] Todos os *scripts* utilizados nessa etapa serão disponibilizados ao término da pesquisa.

A escolha dessa ferramenta se deu em função do seu desempenho na campanha de avaliação para a qual ela foi desenvolvida, obtendo valores de UAS (*unlabeled attachment score*) de 87,2% e de LAS (*labeled attachment score*) de 84,7% para o português. Já o *POS tagger* embutido na ferramenta obteve acurácia de 97,4% para o português. Além disso, foram relevantes na escolha a sua facilidade de uso via terminal, o fato de utilizar uma representação por dependências com anotações em variados níveis linguísticos (que facilitou a posterior extração dos atributos linguísticos) e a sua licença livre. Seguindo o que afirmam Huang *et al.* (2018) sobre a escolha da ferramenta a ser utilizada em textos de aprendizes, como consta na Seção 3.2 (p. 47), um *parser* que apresente o melhor desempenho em textos presumidamente sem desvios será o que melhor se sairá em textos de aprendizes. Em um equilíbrio entre desempenho, tempo de processamento e facilidade de uso, o *UDPipe* se mostrou a melhor escolha para esta pesquisa.

Em relação ao segundo aspecto, a extração dos atributos linguísticos foi feita a partir das análises do *parser*, bem como do *POS tagger* e analisador morfológico embutidos nele. Os atributos extraídos são descritos na Seção 4.4 (p. 79). Sabendo-se que grande parte desses atributos depende de um *parsing* consistente e que pesquisas anteriores mostraram que o desempenho do *parser* é impactado pela presença de desvios, realizou-se um experimento de correção, que tinha como objetivo verificar o impacto da correção das sentenças na extração dos atributos. Para isso, selecionou-se um conjunto de 500 sentenças, que tiveram os seus desvios corrigidos manualmente (não foi inserida nenhuma sentença com desvio de segmentação, uma vez que a correção se limitou ao nível da sentença). Esse conjunto foi *parseado* e teve parte dos seus atributos extraídos[39] e contados. Em seguida, compararam-se os resultados obtidos com as mesmas 500 sentenças, mas não corrigidas. As Tabelas 2 a 6 mostram o resultado comparativo, em termos de número de sentenças que variaram e número e percentual de sentenças que permaneceram no mesmo intervalo, iniciando-se com o número total de *tokens* da sentença.

[39] Cinco dos 17 atributos foram inseridos no *script* de extração apenas após a realização desse experimento. Não se julgou necessário refazer a comparação com os atributos restantes

Tabela 2 – Comparativo: nº de *tokens.*

Intervalo	Corrigidas	Não corrigidas	Diferença distribucional	Permaneceram no mesmo intervalo
1 – 25	87	95	–8	83 (87,36%)
26 – 50	254	248	+6	239 (96,37%)
51 – 100	139	141	–2	134 (95,03%)
101 – 150	19	15	+4	15 (100%)
> 150	1	1	0	1 (100%)

Na Tabela 2, a maior diferença foi nas sentenças entre 1 e 25 *tokens*: houve oito sentenças a menos no grupo das corrigidas. Comparando-se as sentenças que permaneceram no mesmo intervalo em ambos os conjuntos, vê-se que 87,36% das sentenças com até 25 *tokens* são as mesmas entre as corrigidas e as não corrigidas. Nos demais intervalos, mais de 95% das sentenças se manteve entre os dois conjuntos.

Tabela 3 – Comparativo: nº de *tokens* até a raiz.

Intervalo	Corrigidas	Não corrigidas	Diferença distribucional	Permaneceram no mesmo intervalo
0 – 5	242	236	+6	210 (88,98%)
6 – 10	121	114	+7	97 (85,09%)
11 – 20	97	103	–6	81 (78,64%)
21 – 30	26	26	0	15 (57,69%)
> 31	14	21	–7	8 (38,09%)

Em termos de número de *tokens* até a raiz, houve aumento de sete sentenças no conjunto de corrigidas no intervalo de 6 a 10 *tokens*, e diminuição de sete sentenças no intervalo de mais de 31 *tokens* até a raiz. No que se refere ao percentual de sentenças que permaneceram no mesmo intervalo, esse atributo foi o que mais apresentou variabilidade, principalmente nos dois últimos intervalos. Destaca-se que, no intervalo 21–30, a distribuição é a mesma, mas apenas cerca de 57% das sentenças são idênticas em ambos os conjuntos. A correção das sentenças parece ter impacto significativo na identificação da raiz pelo *parser*, o que advoga a favor da sua correção. Porém, como mostra o Capítulo 6, esse atributo não esteve entre os mais decisivos para as decisões dos algoritmos na etapa de AM. Além disso, uma análise manual de algumas sentenças verificou que, independentemente da correção ou não das sentenças, o *UDPipe* tem dificuldades de estabelecer corretamente a raiz da sentença. A questão do desempenho do *parser* na identificação da raiz merece uma verificação sistemática e aprofundada, a qual está fora do escopo da pesquisa.

Tabela 4 – Comparativo: nº de formas verbais infinitas.

Intervalo	Corrigidas	Não corrigidas	Diferença distribucional	Permaneceram no mesmo intervalo
0 – 1	319	322	–3	314 (97,51%)
2 – 5	168	165	+3	159 (96,36%)
6 – 9	12	12	0	11 (91,67%)
> 9	1	1	0	1 (100%)

Já nas formas verbais infinitas, a variação foi menor: apenas os dois intervalos menores sofreram variação de três sentenças. Em termos percentuais, todos os intervalos tiveram mais de 90% das sentenças idênticas em ambos os conjuntos. Nota-se novamente aqui a importância de verificar não apenas a distribuição, mas também o número de sentenças que são iguais em ambos os conjuntos, porque a maior variabilidade em termos percentuais (8,33%) ocorreu em um dos intervalos em que a distribuição permaneceu a mesma (o intervalo 6–9).

Tabela 5 – Comparativo: tipo de sentença.

Intervalo	Corrigidas	Não corrigidas	Diferença distribucional	Permaneceram na mesma classe
Simples	81	85	–4	75 (88,23%)
Composta	419	415	+4	409 (98,55%)

Em termos de sentenças simples ou compostas, a variação da distribuição foi de quatro sentenças entre as 500 corrigidas. Já no que se refere ao percentual de sentenças idênticas, quase a totalidade das compostas são as mesmas entre as não corrigidas e as corrigidas. Já entre as simples, pouco mais de 88% das sentenças permaneceu na mesma classe após a correção.

Tabela 6 – Comparativo: presença de atributos.

Intervalo	Corrigidas	Não corrigidas	Diferença distribucional	Permaneceram na mesma classe
Cópula	286	286	0	278 (97,20%)
Formas nominais	345	342	+3	339 (99,12%)
Passiva	150	150	0	146 (97,33%)
Pron. rel.	267	261	+6	251 (96,17%)
Pron. rel. antes de VFin	35	44	–9	33 (75%)
que conjunção	149	148	+1	141 (95,27%)
Relativa	257	268	–11	240 (89,55%)
Subjuntivo	141	140	+1	133 (95%)

Na presença de determinadas características, a maior variação na distribuição foi nas relativas, em que o conjunto das corrigidas contou com 11 sentenças a menos, o que foi a variação máxima identificada. Já no que se refere ao percentual de sentenças que permaneceu na mesma

classe, em relação ao conjunto das não corrigidas, a maior variação foi na presença de pronome relativo antes de verbo finito, em que 25% das sentenças mudaram de classe. Exceto esse atributo e o da presença de relativas, cujo percentual chegou muito perto de 90%, em todos os demais atributos, 95% ou mais das sentenças permaneceram na mesma classe.

A partir dessa comparação, decidiu-se não continuar com a correção das sentenças, uma vez que essa tarefa é manual e demandaria muito tempo para que um número suficiente de sentenças fosse corrigido, considerando-se que a extração dos atributos tinha como principal objetivo o uso de técnicas de AM. O que essa comparação sugere é que corrigir as sentenças não parece ter um impacto significativo na saída do *parser*, mas não porque os resultados sejam adequados, independentemente da presença de desvios, mas antes pelo contrário. É provável que, em função de a ferramenta ser treinada em um *corpus* de textos jornalísticos, que tem características completamente diferentes das redações, o desempenho do *parser* seja prejudicado tanto nas sentenças com desvios quanto nas corrigidas. A análise dos erros de *parsing* foge ao escopo desta pesquisa. Apesar disso, decidiu-se continuar com a extração dos atributos, cujos resultados do conjunto total de sentenças são apresentados no Capítulo 6.

4.3 *Anotação manual de desvios sintáticos*

Esta seção apresenta e explica a tipologia de desvios sintáticos utilizada como base para a anotação, descreve as duas fases da anotação manual dos desvios sintáticos, e apresenta os resultados quantitativos obtidos nessa tarefa.

4.3.1 A tipologia de desvios sintáticos

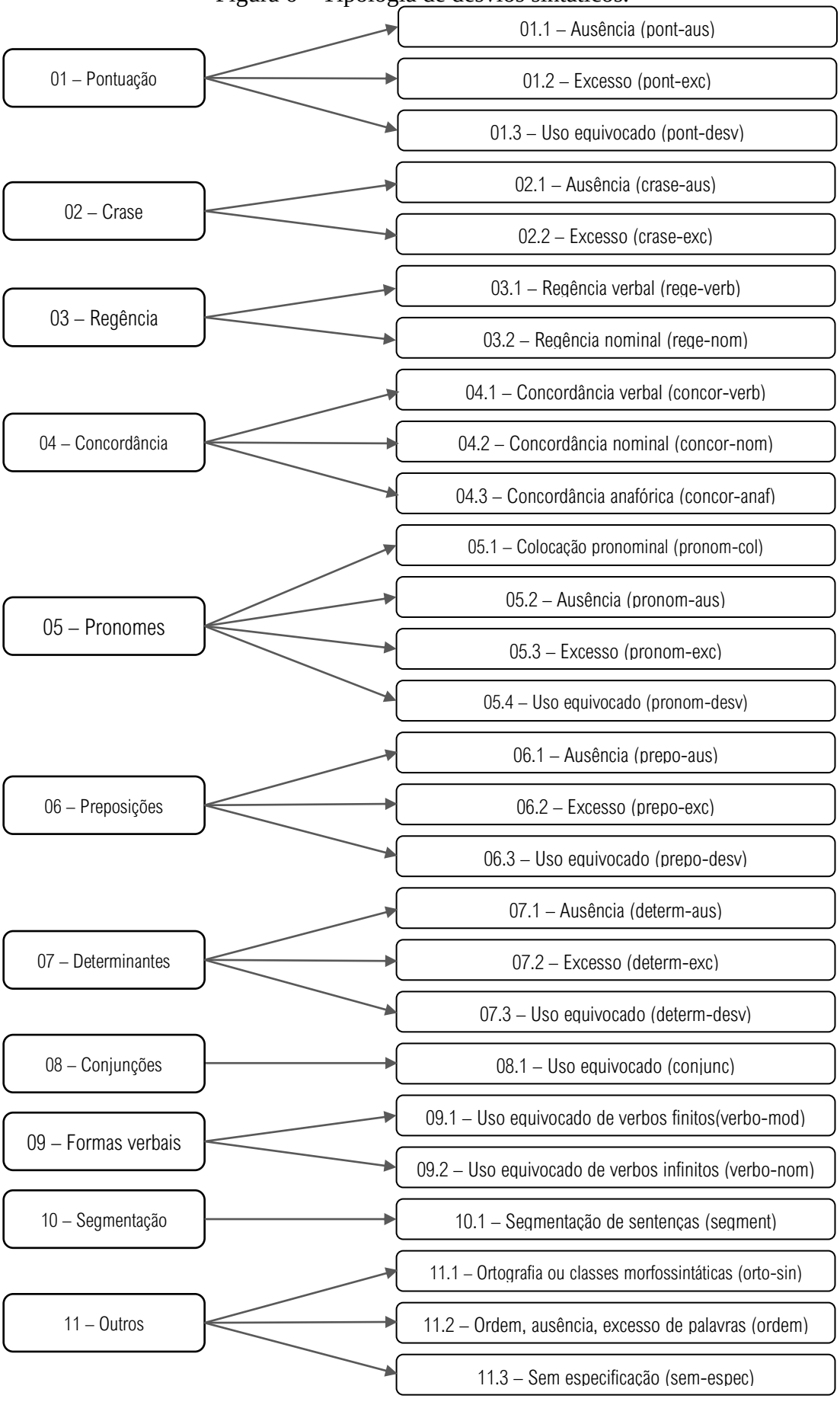
Inicialmente, a ideia era usar, sem alterações, a tipologia de Pinheiro (2008), que foi brevemente apresentada no Capítulo 3. Porém, para verificar a sua adaptabilidade aos propósitos da pesquisa, realizou-se uma anotação piloto (de caráter ilustrativo) em um conjunto de seis textos: dois textos de cada faixa de nota (alta, intermediária e baixa) na Competência 1. Essa anotação foi útil para explicitar as categorias de desvios que não eram cobertas pela tipologia inicial. Além disso, buscou-se uma maior granularidade tipológica, a fim de obter uma descrição mais aprofundada dos desvios sintáticos. Assim, a primeira tarefa dessa etapa foi a adaptação da tipologia aos propósitos desta pesquisa. Na Figura 6, mostra-se a tipologia resultante.

Na interface de anotação, cada subcategoria de desvio foi identificada por um código composto por dois números, como mostrado na Figura 6: o primeiro identificava a categoria

geral do desvio, que serviu como base na primeira fase da anotação; o segundo identificava a subcategoria, que foi usada na segunda fase. Ressalta-se que as categorias têm *caráter hierárquico*, de forma a não sobrepor desvios que se encaixem em mais de um tipo. Nesse sentido, ainda que a ausência de crase, por exemplo, decorra de um problema de regência, a sua categoria está em posição superior na ordem de categorias; logo, deve prevalecer na anotação. Isso valeu para todos os desvios que pudessem ser classificados em mais de uma categoria.

Para guiar a tarefa, criou-se uma diretriz de anotação com a descrição detalhada das categorias e subcategorias e do seu caráter hierárquico, e a explicação do escopo e do processo de ambas as fases da anotação, bem como o uso da ferramenta de anotação. A diretriz completa consta no Apêndice A, mas as definições gerais com alguns exemplos são retomadas na sequência.

Figura 6 – Tipologia de desvios sintáticos.



1) **Uso de pontuação:** ausência, excesso ou uso inadequado de sinais de pontuação, como separação de elementos por vírgulas indevidas, aglutinação de sentenças, falta de pontuação separando elementos deslocados[40]:

    a. Ela influenciou o desenvolvimento da sociedade em que vivemos< >é uma grande responsável pelas transformações ocorridas, mas o hábito de leitura diminuiu (…) [*ausência*]

    b. O leitor<,> passa então, a se apropriar dos textos, mergulhando nele, de forma profunda a estabelecer um compromisso com oque está sendo lido [*excesso*]

    c. O avanço da tecnologia está fazendo com que os jovens se limitem apenas naquilo que leem na rede virtual, nunca vão procura ir além daquilo que leram na internet<,> um exemplo disso ocorre nas escolas; quando um professor pede uma pesquisa, os alunos não recorrem aos livros e sim a internet, tendo assim os mesmos conhecimentos de uma maneira mais rápida e resumida, pois ler é um processo cansativo se comparado com as mídias digitais [*uso equivocado*]

2) **Crase:** ausência de crase em casos obrigatórios ou excesso de crase (ausência de crase em casos optativos não foi considerada desvio). O uso do acento agudo (*á*) foi considerado como crase, pois a sua utilização no lugar da crase parece ser um desvio de ortografia. Então, quando ele era usado adequadamente, não era marcado como desvio:

    a. O motivo dessas manifestações são <à> luta por moradia, reforma na Previdência Social e trabalhista, entre outros. [*excesso*]

3) **Regência:** problemas de regência nominal ou verbal. Casos com mais de uma possibilidade de regência para um mesmo sentido não foram considerados desvios:

    a. O avanço da tecnologia está fazendo com que os jovens se **limitem** apenas <naquilo> que leem na rede virtual (...) [*regência verbal equivocada*]

    b. (…) com **visitas** frequentes prefeituras e subprefeituras, os grupos de movimentos devem sim fazer protestos más em pacificação total. [*regência nominal equivocada*]

4) **Concordância:** desvios de concordância verbal, nominal ou anafórica. Esta última etiqueta foi prevista apenas para pronomes pessoais, os pronomes *esse*/*este* indicando retomada e a expressão *o mesmo*/*a mesma* com valor de retomada, e engloba desvios decorrentes da correferência entre elementos citados anteriormente, dentro da sentença, mas cujo elemento referente não concorda com o referido. A falta de acento nos verbos *ter* e *vir* na terceira pessoa do plural foi considerada desvio de concordância:

    c. <Esses dados> **mostra** que, em geral, o ato de ler está associado a uma atividade obrigatória, solitária que exige bastante paciência e atenção do leitor. [*concordância verbal*]

    d. (…) algumas vertentes não reconhecem **mulheres transexuais** como sendo do gênero feminino, mesmo quando <a mesma> identifica-se desse modo. [*concordância anafórica*]

5) **Pronomes:** desvios de colocação pronominal obrigatória, ausência, excesso ou uso equivocado de pronomes de qualquer tipo, como troca entre os pronomes oblíquos *o* e

[40] A notação utilizada para marcar elementos com desvio é < >.

lhe[41] ou uso de *onde* como pronome relativo não se referindo a lugar[42]. Não se considerou desvio a não utilização de mesóclise e nem a próclise após vírgula[43] (mas a próclise no início da sentença foi anotada como desvio):

a. (...) pois automaticamente a família <lhe> influenciou a isso, assim a criança reproduz aquilo que ver em casa para gerações futuras [*uso equivocado de* lhe *em vez de* o]

b. (…) foi possível presenciar manifestações pacificas de estudantes de todo o território brasileiro, <onde> mostraram que é desde cedo que se deve lutar (…) [*uso equivocado de* onde]

6) **Preposições:** ausência de preposições em casos obrigatórios, excesso ou uso equivocado das preposições em casos não ligados a questões de regência obrigatória. A maioria das ocorrências esteve ligada ao uso equivocado de contrações e ao excesso de preposições em expressões como *mediante a* e *muitas das vezes*:

a. Os movimentos sociais no Brasil são historicamente reconhecidos e causam reverberação desde a época< >que houve a Independência do Brasil [*ausência*]

b. <Mediante a> isso as escolas deveriam propagar através de mesas redondas aos jovens o quão importante é lutar pela democracia de forma pacífica (...) [*excesso*]

7) **Determinantes:** ausência de determinantes em casos obrigatórios, excesso (como em *cujo o* e *pela a*) ou uso equivocado de determinantes (pouco frequente):

a. O primeiro livro foi impresso no século XV, fruto da invenção da tipografia de< >alemão chamado Johannes Gutenberg. [*ausência*]

8) **Conjunções:** qualquer tipo de desvio ligado a conjunções (como o uso sequencial de *mas porém*). Essa categoria tem apenas uma subcategoria:

a. <Se caso> isso não diminua o Brasil terá que criar novas penitenciárias para receber com menos desconforto, novos detentos. [*uso sequencial de conjunções sinônimas*]

9) **Formas verbais:** uso equivocado de verbos finitos quanto a formas, tempos e modos verbais (não ligados a concordância); uso equivocado das formas verbais infinitas gerúndio, infinitivo e particípio. Casos como ausência de *-r* em infinitivos gerando forma verbal finita deveriam ser anotados na subcategoria 09.1 – *verbo-mod*:

a. Mas, no Brasil a leitura parece <está> longe de ser algo primordial na vida da grande maioria. [*uso equivocado de verbo finito*]

b. As pessoas <guardando> aquilo e assim não apoiam os movimentos. [*uso equivocado de gerúndio*]

---

[41] Nesta pesquisa, não se considerou a troca entre os pronomes *o* e *lhe* como um problema de regência, e por isso se manteve a orientação inicial da diretriz aqui. Porém, recomenda-se que essa decisão deva ser revista por estudos que pretendam utilizar essa tipologia.

[42] Sabe-se que esse uso é muito comum e tem sido cada vez mais aceito, mas como ele é penalizado pelos avaliadores do ENEM, optou-se por anotá-lo como desvio. Sugere-se que pesquisas futuras revejam essa decisão.

[43] Essa decisão se justifica porque, segundo Bechara (2009), não se inicia período por pronome átono, mas ele observa que, apesar de alguns gramáticos tradicionais não o aceitarem, isso não vale para qualquer oração dentro de um mesmo período (o que se chama aqui de sentença).

10) **Segmentação de sentenças[44]:** sentenças que indicam a continuação da anterior ou da posterior, mas que foram separadas por ponto final (como sentenças que começam com *Assim como*). Essa categoria tem apenas uma subcategoria:

    a. <Assim como> acreditam que homens não podem apoiar o movimento por não pertencerem ao grupo ao qual ele é destinado.

11) **Outros:** palavras de classes morfossintáticas diferentes das categorias anteriores, que contêm desvios que influenciam a sintaxe; desvios de ordem, ausência ou excesso de palavras lexicais; ou desvios que não se encaixem em nenhuma das categorias:

    a. Os grupos deveriam procurar cobrar das prefeituras e governos porém de uma forma mais <eficácia> [*substantivo usado no lugar de adjetivo*]

    b. Portanto, <medidas devem haver> para solucionar o impasse, segundo Immanuel Kant " o ser humano é aquilo que a educação faz dele. [*problemas de ordem de palavras*]

Cabe uma atenção especial à categoria *Outros*, uma vez que a sua utilização foi alterada durante o processo de anotação. Inicialmente, definiu-se que o seu uso deveria ser extremamente cauteloso, de forma a evitar que qualquer desvio que o anotador não soubesse categorizar fosse marcado nessa categoria. No entanto, durante o processo, percebeu-se a ausência de dois tipos de desvios importantes: aqueles ligados à ordem de palavras, que não eram esperados inicialmente em textos de falantes nativos, bem como a ausência ou o excesso de palavras lexicais; e aqueles ligados às demais classes morfossintáticas não previstas na tipologia. Nesse último tipo de desvio se incluem essencialmente os diversos problemas de ortografia que alteram a classe morfossintática de uma palavra, como a separação de elementos que deveriam ser escritos juntos, e vice-versa (p. ex., *a cerca* em vez de *acerca*). Assim, todos os desvios desses dois tipos foram inicialmente inseridos nessa categoria e, após o fim da anotação, foram marcados nas duas novas subcategorias *orto-sin* e *ordem* diretamente no arquivo de saída da plataforma de anotação. A subcategoria *sem-espec* não foi atribuída a nenhum desvio.

Sempre que possível, a anotação deveria ser feita no *token* associado ao desvio, conforme foi descrito na diretriz. Casos de ausência de elementos em geral deveriam ser marcados no *token* imediatamente anterior ao local em que tal elemento deveria ocorrer. Uma exceção foram os casos de ausência de preposição que cause problemas de regência, em que o *token* a ser anotado era o elemento regente (verbo, adjetivo, nome). As duas fases da anotação são descritas em mais detalhes nas próximas seções.

[44] No decorrer da anotação, percebeu-se que essa categoria, na verdade, se refere a problemas de pontuação, que podem ter ocorrido na sentença anterior ou na própria sentença com o desvio, quando ela deveria ter sido conectada à sentença seguinte. Sugere-se que pesquisas futuras insiram essa categoria como uma subcategoria de pontuação.

4.3.2 Primeira fase da anotação

Na primeira fase, utilizou-se uma planilha com todas as sentenças do *corpus*, identificadas de acordo com a ID do texto e ordenadas conforme o *corpus* original, cuja classificação das sentenças não tinha qualquer relação com as notas totais ou parciais. Na terceira coluna, preenchida previamente com a letra N, que marca sentenças que não possuem desvio, o anotador deveria verificar a presença de no mínimo um dos desvios sintáticos das categorias da tipologia e atribuir a letra D maiúscula à sentença (Figura 7). Sentenças em que o anotador não identificava qualquer desvio não recebiam nenhuma marcação.

Figura 7 – Arquivo de anotação: primeira fase.

ID	text	class
51305.1	Em um país democrático, onde todos tem a liberdade de se expressar, expor suas opiniões e cobrar seus direitos de cidadão.	N
51305.2	Podemos ver um estado onde a grande maioria da população tem o intuito de buscar um país melhor, criando movimentos e ações para uma vivência melhor e uma população e um governo mais corretos.	N
51305.3	O movimento social é um grupo de pessoas que tem como objetivo alcançar mudanças sociais dentro do âmbito político, em	N
51305.4	No 5º artigo no IV parágrafo da constituição brasileira fala que é livre a manifestação de pensamentos, no IX parágrafo fala que é livre o direito de expressão de comunicação.	N
51305.5	Muitas vezes nesses movimentos a grande maioria das pessoas saem ás ruas em busca de algo melhor de forma correta e prudente, porém uma minoria dessa população acabam abusando desse direito, causando danos para a cidade e muitas vezes	N
51305.6	No Brasil os movimentos ganharam mais força na década de 1960, quando surgiram os primeiros movimentos de luta contra a política vigente, ou seja, a população insatisfeita com as transformações ocorridas tanto no campo econômico e social, como	N
51305.7	As ações coletivas mais conhecidas no Brasil são o Movimento dos Trabalhadores Sem Terra (MST), o Movimento dos Trabalhadores Sem Teto (MTST) e os movimentos em defesa dos índios, negros e das mulheres.	N
51305.8	Portanto, é de muita importância a existência de cada movimento para mostrar que a população não está satisfeita com essa qualidade de vida em que levamos no país.	N
51305.9	É importante a expressão de cada indivíduo sobre sua opinião em um estado que existe tanta diversidade cultural.	N
51305.10	Para termos melhor resultado é relevante que a sociedade pense na melhor forma possível, pensando sempre em obter parcimônia em todos os movimentos obtidos ao decorrer da história brasileira, algo mais organizado e com menos vandalismo	N
51305.11		
51306.1	No Brasil, essas transformações foram se consolidando ao longo da década de 1950, e alteraram o consumo e o comportamento de parte da população que habitava os grandes centros urbanos.	N

Essa fase da anotação foi executada por uma anotadora, que foi a própria autora, anotando cerca de 600 sentenças por dia. Para testar a clareza da diretriz e, consequentemente, calcular a concordância entre anotadores, uma parcela de 150 sentenças contendo entre 3 e 25 tokens foi selecionada aleatoriamente e anotada por um segundo anotador (também mestrando em Linguística). A concordância entre anotadores foi calculada utilizando-se o *kappa de Cohen*, cujos valores variam entre 0 e 1: quanto mais perto de 1, maior a concordância entre os anotadores. O resultado da concordância da primeira fase da anotação foi de 0,543, o que, conforme Landis e Koch (1977), é considerado um nível de concordância moderada. Ainda que uma classificação binária de sentenças com ou sem desvio não pareça, inicialmente, uma tarefa complexa, estudos indicam que os anotadores divergem em relação ao que é considerado aceitável ou não em termos de desvios. Rozovskaya e Roth (2010), por exemplo, encontraram

valores de *kappa* entre 0,16 e 0,40 entre anotadores que deveriam identificar se uma sentença escrita por um falante não nativo de inglês estava correta ou não. Mesmo que a tarefa descrita pelos autores seja diferente da que se realizou aqui, os números sugerem que julgar uma sentença em termos de presença ou não de desvios pode não ser tão trivial quanto se esperava.

No entanto, cabem algumas considerações sobre a questão da concordância entre anotadores. O treinamento do segundo anotador consistiu em apenas um encontro presencial, no qual se apresentaram os principais pontos da diretriz, mas cujo foco foi a explicação sobre questões técnicas da anotação. Não houve período de treinamento nem anotação piloto para que se pudessem identificar as inconsistências, refazer o treinamento, reformular as diretrizes ou discutir casos divergentes e/ou problemáticos. Essa decisão baseou-se essencialmente em restrições de tempo, mas a ausência dessas etapas fundamentais provavelmente impactou os resultados da concordância entre anotadores. As restrições de quantidade de sentenças e número máximo de *tokens* por sentença também se colocam como limitações em relação aos caminhos definidos para a tarefa de anotação desta pesquisa. Investigações que pretendam adotar uma metodologia de anotação similar devem tomar essas questões como aprendizado, de forma a dedicar maior tempo e atenção ao treinamento dos anotadores e à resolução de divergências.

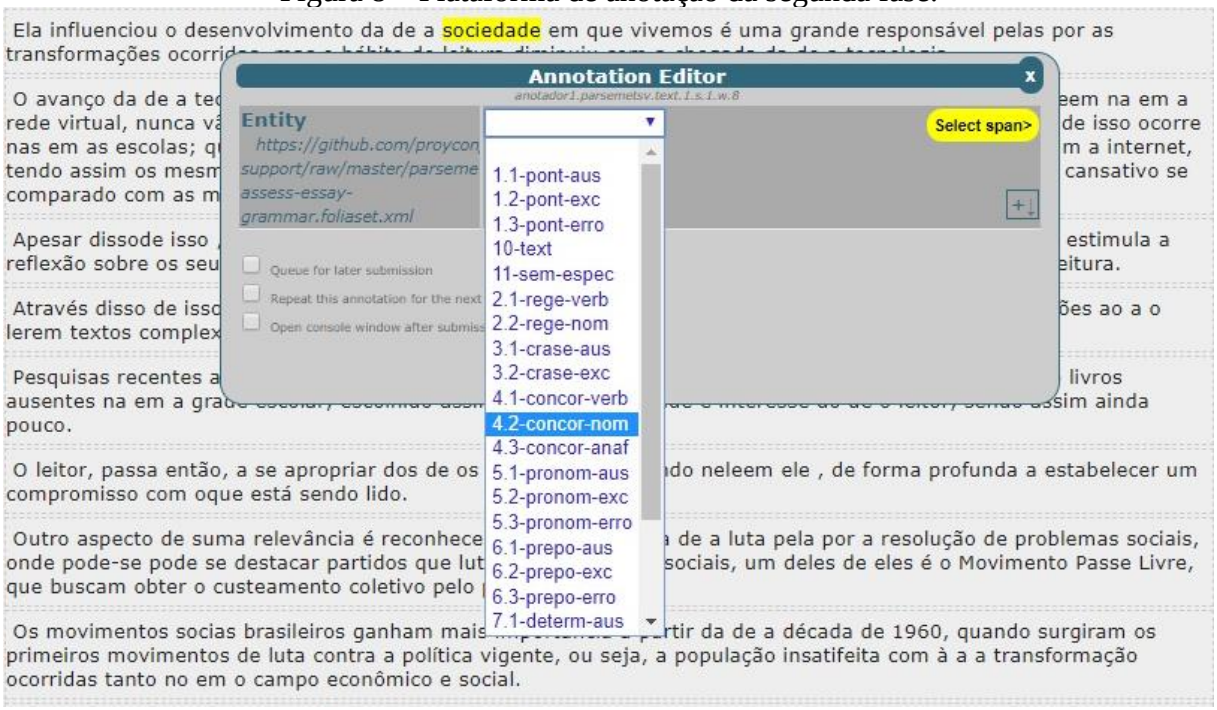
4.3.3 Segunda fase da anotação

A partir das anotações da primeira fase, uma parcela de 2.500 sentenças com desvio foi selecionada de forma que os desvios fossem tipificados. Assim, a segunda fase consistiu na etiquetação dos desvios conforme a tipologia já apresentada. Para isso, utilizou-se a interface de anotação FLAT, que é uma plataforma *on-line* para anotação linguística (GOMPEL; REYNAERT, 2013). A interface foi escolhida pela sua facilidade de uso e pelo seu *design* amigável ao usuário, e também porque já era conhecida e utilizada em outro projeto de anotação do qual a principal anotadora dos dados participa. Para isso, solicitou-se autorização de uso do servidor localizado na Universidade de Düsseldorf (Alemanha), que foi concedida.

Primeiramente, as etiquetas dos desvios foram configuradas, sendo identificadas pelo seu código numérico e pelo código abreviado da descrição, a fim de diminuir a necessidade de consultas às diretrizes. Em seguida, o arquivo em formato CONLL-U foi carregado no FLAT na pasta correspondente ao usuário previamente criado da anotadora. A plataforma de anotação carrega 60 sentenças por página, na ordem em que aparecem no arquivo de entrada. A anotação era, sempre que possível, feita por *token* (as exceções foram locuções verbais, eventuais estruturas com verbo-suporte quando o problema era de regência, e problemas de ordem de

palavras ou conjuntos de palavras), por meio de cliques com *mouse*. A anotadora anotou uma sentença por vez, classificando todos os desvios que ocorriam, alcançando em média 200 sentenças anotadas por dia. A duração da anotação foi de 15 dias consecutivos. Na Figura 8, mostram-se o *layout* da plataforma de anotação e a janela que é aberta após clicar no *token* para atribuir a etiqueta correspondente.

Figura 8 – Plataforma de anotação da segunda fase.



Após o término da anotação, os arquivos foram baixados da plataforma em formato CUPT (utilizado pelo projeto *Parseme*), que consiste nas 10 colunas com as informações correspondentes ao CONLL-U mais uma coluna adicional com as anotações de desvios. Quando o *token* não possui nenhuma anotação, a coluna é preenchida por um asterisco. Na Figura 9, vê-se o formato de saída do arquivo anotado, que pode ser aberto em qualquer editor de textos.

Figura 9 – Arquivo de saída no formato CUPT.

sent_id = 19827.1

text = Os movimentos sociais no Brasil são historicamente reconhecidos e causam reverberação desde a época que houve a Independência do Brasil.

ID	FORM	LEMMA	UPOSTAG	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC	ANNOT
1	Os	o	DET	_	Definite=Def|Gen	2	det	_	_	*
2	movimentos	movimento	NOUN	_	Gender=Masc|Nu	9	nsubj:pass	_	_	*
3	sociais	social	ADJ	_	Gender=Masc|Nu	2	amod	_	SpacesAfter=\s\s	*
4-5	no	_	_	_	_	_	_	_	_	*
4	em	em	ADP	_	_	6	case	_	_	*
5	o	o	DET	_	Definite=Def|Gen	6	det	_	_	*
6	Brasil	Brasil	PROPN	_	Gender=Masc|Nu	2	nmod	_	_	*
7	são	ser	AUX	_	Mood=Ind|Numbe	9	aux:pass	_	_	*
8	historicamente	historicamente	ADV	_	_	9	advmod	_	_	*
9	reconhecidos	reconhecer	VERB	_	Gender=Masc|Nu	0	root	_	_	*
10	e	e	CCONJ	_	_	11	cc	_	_	*
11	causam	causar	VERB	_	Mood=Ind|Numbe	9	conj	_	_	*
12	reverberação	reverberação	NOUN	_	Gender=Fem|Nun	11	obj	_	_	*
13	desde	desde	ADP	_	_	15	case	_	_	*
14	a	o	DET	_	Definite=Def|Gen	15	det	_	_	*
15	época	época	NOUN	_	Gender=Fem|Nun	11	obl	_	_	*
16	que	que	PRON	_	Gender=Fem|Nun	17	nsubj	_	_	1:06.1-prepo-aus
17	houve	haver	VERB	_	Mood=Ind|Numbe	15	acl:relcl	_	_	*
18	a	o	DET	_	Definite=Def|Gen	19	det	_	_	*
19	Independência	independência	PROPN	_	Gender=Fem|Nun	17	obj	_	_	*
20-21	do	_	_	_	_	_	_	_	_	*
20	de	de	ADP	_	_	22	case	_	_	*
21	o	o	DET	_	Definite=Def|Gen	22	det	_	_	*
22	Brasil	Brasil	PROPN	_	Gender=Masc|Nu	19	nmod	_	SpaceAfter=No	*
23	.	.	PUNCT	_	_	9	punct	_	SpacesAfter=\r\n	*

Sobre o arquivo de saída, vê-se que se manteve a informação da ID original da sentença, bem como o texto não *tokenizado*, com a anotação separada, o que caracteriza a anotação *stand-off*. Após, cada um dos *tokens* aparece numerado em uma linha, com as seguintes colunas do formato CONLL-U[45], mais a coluna ANNOT, específica do CUPT e contendo as anotações inseridas no FLAT:

1. ID: o índice do *token*, começando em 1 e reiniciando a cada nova sentença;
2. FORM: a palavra em si ou o sinal de pontuação;
3. LEMMA: a forma lematizada da palavra;
4. UPOSTAG: etiqueta de *POS*, de acordo com as categorias estabelecidas pelo UD;
5. XPOS: etiquetas de *POS* específicas de língua (é preenchido com *underscore* quando não houver informação, como no caso do português);
6. FEATS: lista de aspectos morfológicos específicos da língua ou obtidos do inventário universal de aspectos morfológicos do UD;
7. HEAD: o número que identifica o núcleo (*head*) daquela palavra, ou seja, a palavra da qual ela depende. Se o número for 0, significa que a palavra é a raiz da sentença;
8. DEPREL: relação de dependência com o núcleo ou a raiz nos moldes do UD;

[45] Disponível em: https://universaldependencies.org/format.

9. DEPS: representação de dependência aprimorada (é preenchido com *underscore* quando não houver informação, como no caso do português).
10. MISC: outras anotações (é preenchido com *underscore* quando não houver informação).
11. **ANNOT: anotação das etiquetas desta pesquisa (é preenchido com * quando não houver anotação).**

Para essa segunda etapa, escolheu-se aleatoriamente um conjunto de 100 sentenças com desvio a serem anotadas por um segundo anotador (o mesmo que anotou as sentenças da concordância entre anotadores da fase 1). Na mesma plataforma de anotação utilizada para a classificação dos desvios da segunda fase pela anotadora principal, o anotador deveria encontrar e classificar os desvios das 100 sentenças. Para essa fase, a métrica utilizada foi a medida-f, que é obtida a partir da média harmônica entre os valores de precisão (número de desvios identificados corretamente dividido pelo número de desvios identificados) e cobertura (número de desvios identificados corretamente dividido pelo número de desvios que deveriam ter sido identificados) da anotação, considerando uma das anotações como *gold standard* (a anotação "correta", com a qual se compara). Nessa avaliação, considerou-se o *gold standard* como sendo as anotações realizadas pela pesquisadora, e a medida-f obtida foi de 0,6472.

Ao analisar os casos de discordância, identificou-se que a maior parte deles se refere a anotações feitas por apenas um anotador. Houve uma grande quantidade de discordâncias relacionadas às subcategorias de pontuação e segmentação, em que um dos anotadores anotou, deixou de anotar ou classificou os desvios em subcategorias diferentes. Isso pode ser um indício de que a categoria *segmentação* pode ser incorporada na categoria de pontuação. No que se refere aos casos em que um desvio foi anotado em subcategorias diferentes por ambos os anotadores, houve alternâncias entre as subcategorias *verbo-mod* e *verbo-nom*. Em alguns casos, percebeu-se que o caráter hierárquico da tipologia de desvios não foi seguido, por exemplo, em casos de concordância verbal que receberam a etiqueta de *verbo-mod*.

Esses resultados reforçam que o treinamento do anotador foi insuficiente e que a tarefa de identificação e classificação de desvios sintáticos não é trivial. Porém, outros fatores podem ter impactado os resultados, especialmente o tamanho da diretriz de anotação e o excesso de etiquetas, uma vez que se utilizou o conjunto de etiquetas das subcategorias. Vê-se duas possibilidades de adaptação da metodologia de anotação no futuro. A primeira delas se refere à tipologia de desvios, que pode ser revista de modo que as categorias sejam reorganizadas. A questão de maior destaque é a incorporação da categoria *Segmentação* à categoria de *Pontuação*. A segunda delas diz respeito ao método de anotação. Para facilitar o processo, o ideal é dividir a classificação em três etapas: na primeira, os *tokens* em que ocorrem os desvios

são identificados como “contém desvio”; na segunda, esses *tokens* são classificados por categoria; na terceira, as categorias são classificadas novamente de acordo com as subcategorias. Entre cada uma dessas etapas, bem como entre a fase 1 e a fase 2 da anotação, é preciso calcular a concordância entre anotadores e proporcionar momentos de treinamento e discussão dos casos divergentes e/ou dos exemplos-limite. Tal adaptação à metodologia foi identificada, mas não aplicada nesta pesquisa, caracterizando-se como limitação e possibilidade de correção e aprofundamento por projetos futuros que se ocupem dessa tarefa. A próxima seção apresenta os resultados da anotação em termos quantitativos.

4.3.4 Análise quantitativa dos desvios

Esta seção apresenta os resultados das duas fases da anotação: a classificação das sentenças e a tipificação dos desvios sintáticos. Os resultados da primeira fase em termos numéricos para o *corpus* de treino podem ser vistos na Tabela 7, que mostra que mais de três quartos das sentenças contêm pelo menos um desvio sintático. Todos os textos anotados apresentaram pelo menos uma sentença com desvio, inclusive os que receberam 200 pontos na avaliação humana (pontuação máxima para a Competência 1).

Tabela 7 – Presença de desvios no total de sentenças anotadas.

Intervalo	**Nº de sentenças**
Contém desvio	6.347 (73,34%)
Não contém desvio	2.307 (26,66%)
Total	8.654

A seguir, mostram-se os resultados da segunda fase da anotação. A Tabela 8 apresenta o número de desvios por categoria, em ordem de frequência.

Tabela 8 – Tipos de desvios por categoria

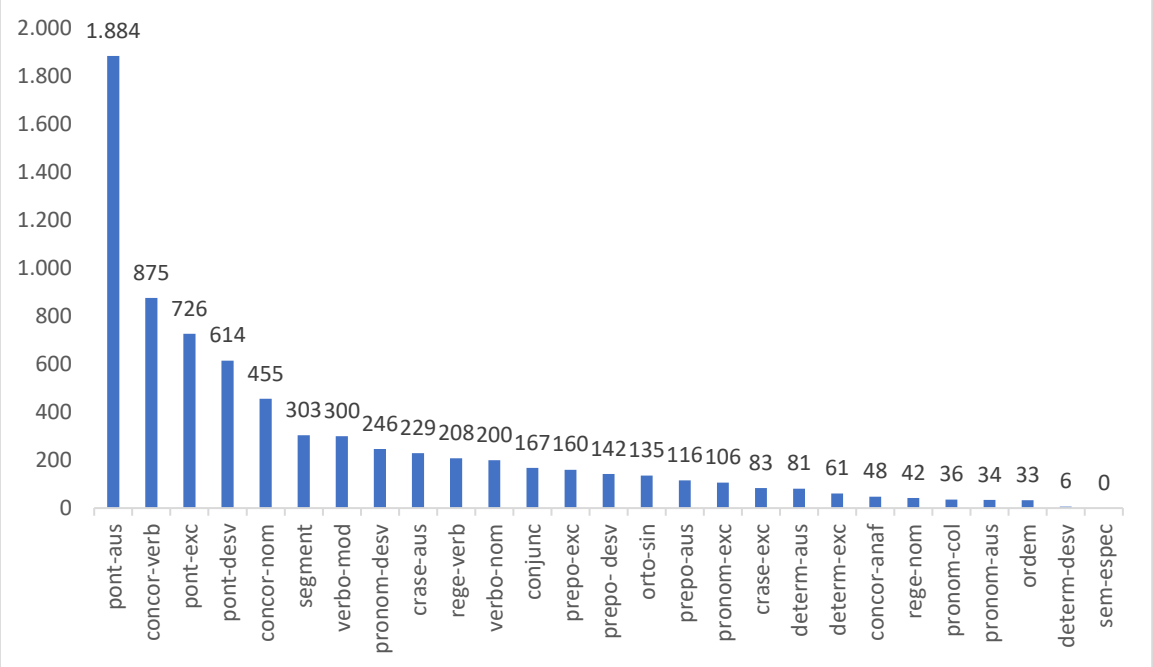
Código	**Quantidade**
01 – Pontuação	3.224
04 – Concordância	1.378
09 – Formas verbais	500
05 – Pronomes	422
06 – Preposições	418
02 – Crase	312
10 – Segmentação	303
03 – Regência	250
11 – Outros	168
08 – Conjunções	167
07 – Determinantes	148
Total	**7.290**

Em termos de distribuição do número total de desvios, em relação ao número de sentenças com desvio anotadas (2.500), tem-se uma média de 2,91 desvios por sentença. Porém, é preciso analisar com cautela essa média, pois há sentenças com apenas um desvio, enquanto outras possuem muito mais de três desvios.

Os dados por categoria mostram uma prevalência massiva de problemas de pontuação, seguidos pela concordância. Com relação à pontuação, a experiência de anotação permite afirmar, sem equívoco, que a grande maioria dos desvios está ligada ao uso da vírgula, como é demonstrado também na análise linguística do Capítulo 5. Vale destacar ainda uma particularidade: a categoria de preposições teve maior ocorrência do que a de crase ou a de regência. Mesmo considerando o caráter hierárquico da tipologia, parece que as preposições têm trazido dificuldades aos produtores de textos. Por outro lado, demais palavras gramaticais, como conjunções e determinantes, não apresentam tantos problemas. Nesse sentido, pode ser interessante um estudo comparativo com desvios de escrita realizados por aprendizes de português como L2, a fim de verificar se há diferenças nas frequências desses tipos de desvio.

Como a análise por categorias não basta para compreender adequadamente a dimensão dos fenômenos em termos quantitativos, é preciso olhar também para as subcategorias. No Gráfico 3, tem-se a distribuição numérica de desvios de cada uma das subcategorias, por ordem de frequência, facilitando a comparação dos resultados.

Gráfico 3 – Tipos de desvios por subcategoria



Identificou-se um total de 7.290 desvios, dos quais a ausência de pontuação representa a grande maioria de ocorrências (25,84% do total). A seguir, vem a concordância verbal (12% dos desvios), que pode indicar um problema dos produtores dos textos de estabelecerem as dependências corretas entre os elementos da sentença, realizando assim a concordância adequada. Vê-se ainda que o gráfico mostra uma distribuição de Zipf, indicando que um número pequeno de etiquetas corresponde à maior parte dos tipos de desvios. Nesse sentido, as cinco subcategorias de pontuação e concordância representam, juntas, 62,47% do total de desvios. Após pontuação e concordância, o próximo desvio mais frequente é o que diz respeito à segmentação sentencial. Esse é um resultado que não pode ser ignorado, uma vez que ele explicita o desafio de realizar uma análise que se limite ao nível da sentença, e não considere o texto como um todo. Assim, ao que parece, os produtores não apresentam problemas somente de construção de sentenças, mas também de segmentação e de estruturação das conexões entre as sentenças de um texto. Tais desvios provavelmente envolvem questões de pontuação, ainda que se refira à sentença anterior.

Os números referentes à subcategoria *pronom-desv* estão em grande medida relacionados ao uso de *onde* como pronome relativo sem referência a local. Assim, cabe uma reflexão sobre se esse aspecto deve continuar sendo visto como desvio. Nota-se também o fato de que os desvios de uso de crase são menos frequentes do que se esperava, e a ausência de crase é muito mais recorrente do que a sua colocação inadequada. Outro aspecto a ser considerado é que as subcategorias ligadas a questões verbais têm uma frequência significativa em relação às demais (exceto as de pontuação). Entre os 10 tipos de desvios mais frequentes, vê-se (i) concordância verbal, (ii) problemas de formas, tempos e modos verbais e (iii) regência verbal. Nota-se ainda que o desvio de concordância anafórica que se pretendia analisar se mostra pouco frequente, mas é superior aos casos de colocação pronominal, por exemplo, comumente referida por professores como uma questão problemática. É provável que isso se deva à decisão de não marcar como desvio a próclise após vírgulas.

É importante levantar aqui ainda dois pontos que merecem destaque. Em primeiro lugar, notou-se uma influência muito grande de outros níveis linguísticos na sintaxe. Sabe-se que uma análise que se propõe a isolar um nível linguístico dificilmente o fará com sucesso, uma vez que os fenômenos não ocorrem de forma independente. Porém, devido às limitações impostas pela pesquisa, foi necessário estabelecer um recorte de análise, ao que se optou pela sintaxe. Ainda assim, não podem ser ignorados os desvios sintáticos que são decorrentes de outros níveis, como as diversas questões morfológicas que têm impacto na sintaxe, no sentido de que

os produtores dos textos muitas vezes separam ou aglutinam elementos de forma inadequada (como é o caso de *em baixo*, *encima*, *a cerca*, *frequenti mente*).

Em segundo lugar, cabe ressaltar a questão da ortografia, especialmente no que se refere à acentuação. Em várias ocasiões, os desvios anotados consistiam originalmente em questões ortográficas, mas devido ao caráter hierárquico da tipologia, aparecem marcados em subcategorias diversas. Um exemplo é a ausência de acento na forma verbal *está*, que foi marcada como desvio de uso equivocado de pronome (*esta*). Da mesma forma, a não acentuação da forma verbal *dê* foi anotada como desvio de uso equivocado de preposição (*de*). Tais desvios provavelmente não seriam identificados por um corretor ortográfico, pois as palavras decorrentes deles de fato existem no léxico. Em alguns casos, percebeu-se, por exemplo, que o *UDPipe* acertou a *POS tag* mesmo que a palavra apresentasse um desvio que a colocasse em uma classe morfossintática equivocada (como em *está/esta*), mas em outras situações (como em *e/é*), a ausência do acento traz problemas à criação da árvore sintática. A falta de acentuação foi um problema recorrente, o que pode levar à reflexão sobre a influência da correção automática de editores de textos e dos *smartphones* sobre esse comportamento dos estudantes. Tal análise está fora do escopo da pesquisa, mas fica como uma possibilidade de investigação futura. O Capítulo 5 traz análises mais aprofundadas desses fenômenos.

4.4 *Extração de atributos linguísticos e correlação via Aprendizado de Máquina*

Os objetivos da etapa de identificação e extração de atributos linguísticos das sentenças foram dois: em primeiro lugar, a tarefa ajudou a compreender o fenômeno que está sendo descrito; em segundo lugar, os atributos foram utilizados como entrada para o treinamento dos algoritmos de AM descritos no Capítulo 6. Os atributos (ou *features*), para esta pesquisa, são características das sentenças que independem da presença ou ausência de desvios sintáticos, e que estão positivamente correlacionadas com a ocorrência de desvios, ou seja, que podem estar trazendo mais dificuldades aos produtores de textos.

Assim, a partir da anotação linguística, identificaram-se alguns atributos específicos desta pesquisa, que se juntaram a outros atributos clássicos na literatura, formando o conjunto de 17 atributos linguísticos extraídos automaticamente. Na literatura, os trabalhos que propõem classificadores para avaliar as redações, como descrito no Capítulo 3, muitas vezes consideram atributos que vão além do nível da sentença, uma vez que a entrada para os algoritmos é o texto todo. Aqui, foi preciso limitar-se apenas àqueles considerados relevantes e que poderiam ser extraídos de sentenças. Para complementar esse conjunto, utilizaram-se as características

observadas na anotação como possivelmente importantes para a ocorrência de desvios. Os atributos extraídos com os respectivos códigos utilizados no arquivo do AM são os seguintes:

1 – n_tokens: número de *tokens* da sentença;
2 – sent_type: tipo de sentença – simples (apenas um verbo finito) ou composta (mais de um verbo finito);
3 – n_tokens_root: número de *tokens* à esquerda da raiz da sentença;
4 – n_vfin: número de verbos finitos;
5 – n_vinf: número de verbos infinitos (infinitivo, gerúndio e particípio);
6 – n_commas: número de vírgulas;
7 – copula: presença de cópula (*ser* e *estar*);
8 – subjuntivo: presença de subjuntivo;
9 – voice_pass: presença de voz passiva;
10 – vinf: presença de formas verbais infinitas (infinitivo, gerúndio e particípio);
11 – pron_rel: presença de pronome relativo;
12 – rel_bef_vfin: presença de pronome relativo à esquerda de um verbo finito;
13 – que_conj: presença de *que* com *POS* de conjunção;
14 – relativa: presença de oração relativa;
15 – relat_in_subj: presença de relativa dentro do sujeito;
16 – tree_depth: profundidade da árvore sintática;
17 – av_dep_length: distância média entre elementos dependentes.

O arquivo a partir do qual foi feita a extração automática dos atributos foi a saída do *parser*, em formato CONLL-U. As identificações e extrações foram possíveis em função da diversidade de informações linguísticas que o *parser* e o *POS tagger* e analisador morfológico embutidos no *UDPipe* ofereciam. Assim, essa ferramenta se mostrou vantajosa em função de ter uma riqueza linguística, em termos de etiquetas, que facilitou o processo extração automática dos atributos.

Os atributos foram salvos em formato de planilha composta por 20 colunas. Nas duas primeiras constam a ID e a sentença a que se referem os atributos (uma sentença por linha), nas demais constam as informações extraídas, em termos numéricos e em termos de presença (1) ou ausência (0), e a última coluna indica a classe de cada sentença (D para sentenças com desvio, N para as sem desvio).

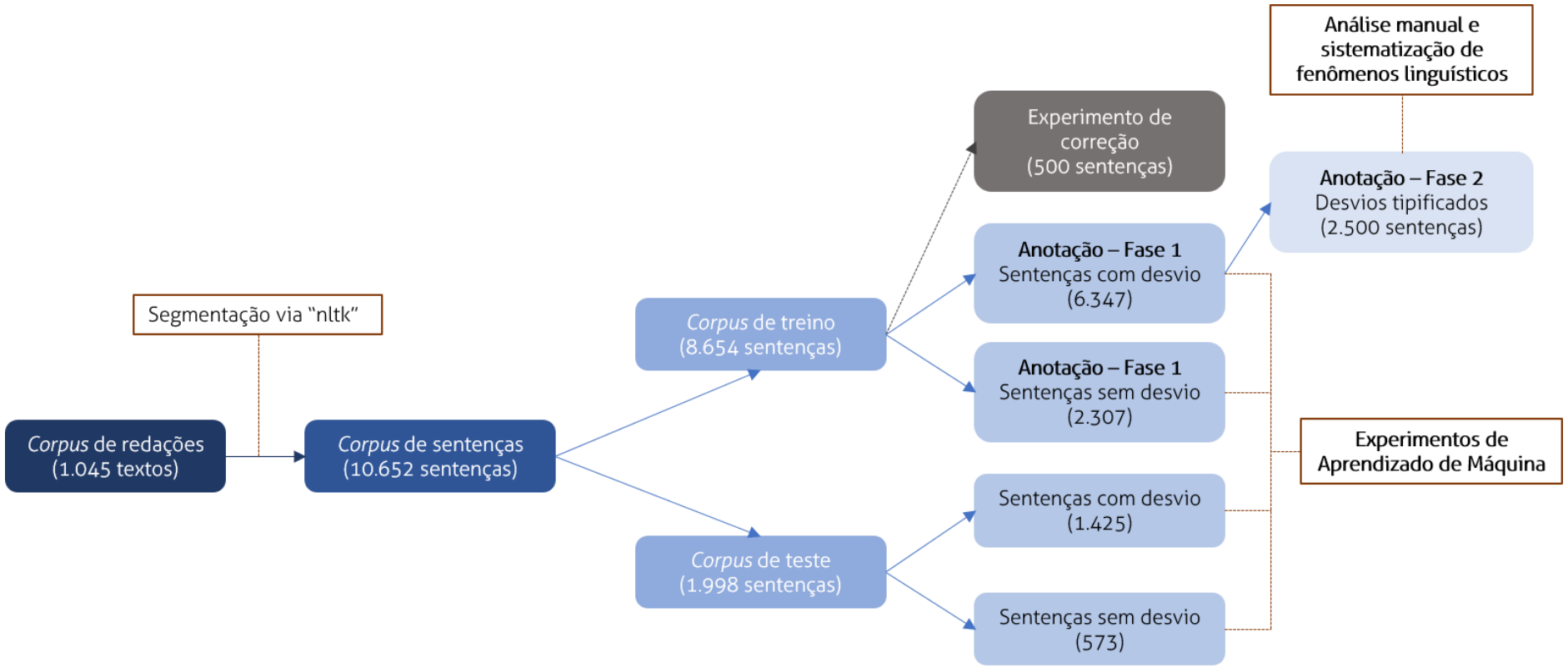
Para o AM, utilizou-se o *Weka*, citado na Seção 2.6 (p. 38), em função da sua facilidade de uso e interface amigável para pessoas sem experiência em Computação. Para preparar os arquivos para o AM, o arquivo com os atributos foi convertido automaticamente conforme as exigências do formato ARFF, que é aceito pelo *Weka*. O objetivo da etapa de AM era explicitar

correlações entre os atributos e a presença de desvios. Assim, a tarefa consistiu em uma classificação binária (com e sem desvio), em que os atributos são independentes entre si, e não se tentam predizer os tipos de desvios.

Para a classificação, buscou-se testar algoritmos clássicos de diferentes paradigmas. Os algoritmos testados no *Weka* foram *MultiLayer Perceptron* (MLP) no paradigma conexionista; *SMO*, *Logistic Regression*, e *Naïve Bayes* no paradigma matemático (probabilístico); *One-R*, *J48*, *JRip* e *Random Forrest* no paradigma simbólico. Também foram testados dois algoritmos de seleção de informações, a fim de verificar quais foram os atributos considerados mais relevantes: o *InfoGainAttributeEval* e o *CfsSubsetEval*. Essa etapa consistiu em dois experimentos de classificação e um de seleção de atributos. No primeiro experimento de classificação, o *corpus* de treino foi balanceado com o número idêntico de sentenças com e sem desvio (2.307 sentenças). No segundo experimento, utilizou-se o total de sentenças (2.307 sem desvio e 6.347 com desvio). A partir dos resultados, os três melhores classificadores foram selecionados, treinados e aplicados no *corpus* de teste. O terceiro experimento foi o da seleção de informações. Os resultados dessa etapa são apresentados no Capítulo 6.

De forma a sistematizar os conjuntos de sentenças que foram utilizados em cada uma das etapas da pesquisa, a Figura 10 mostra um esquema de divisão dos *corpora* que resume o que se descreveu neste capítulo. O próximo capítulo realiza uma análise linguística dos fenômenos a partir de uma abordagem qualitativa, buscando analisar e sistematizar os fenômenos, independentemente da sua frequência de ocorrência. Essa subdivisão e sistematização em fenômenos linguísticos nos quais os desvios ocorrem é interessante para permitir uma maior compreensão sobre os tipos de estruturas em que os desvios sintáticos ocorrem.

Figura 10 – Esquema dos conjuntos de textos utilizados na pesquisa.



5 ANÁLISES LINGUÍSTICAS QUALITATIVAS DOS DESVIOS

Para compreender melhor os fenômenos nos quais os desvios sintáticos ocorrem, foi preciso analisar os dados anotados e descrevê-los a partir de análises linguísticas. Essa descrição é relevante para o desenvolvimento de ferramentas de PLN e para a compreensão das dificuldades dos estudantes. A partir da identificação dos fenômenos nos quais os desvios ocorrem, pode-se, por exemplo, reforçar o aprendizado dessas construções no ensino de produção textual.

Este capítulo faz algumas referências à frequência dos fenômenos, mas não se propõe a apresentar sistematicamente as suas contagens. Sabe-se também que pode haver mais de um fenômeno envolvido na ocorrência de um desvio, e que alguns casos analisados cabem em mais de uma das seções propostas. Porém, os exemplos trazidos servem como ilustração dos fenômenos considerados relevantes para a ocorrência de desvios, e as sobreposições complexificam, mas não invalidam as análises individuais de cada fenômeno.

Dessa forma, o capítulo traz primeiramente os fenômenos gerais que ocorreram em mais de uma subcategoria de maneira semelhante e, em seguida, os específicos por categoria. Como algumas das subcategorias apresentavam muito mais desvios do que outras e os fenômenos iam se repetindo no decorrer da análise, optou-se por selecionar uma amostra. Para definir o tamanho dessa amostra, somou-se o número de desvios de todas as subcategorias, excluindo-se a majoritária (*pont-aus*) e a minoritária (*determ-desv*), e dividiu-se o número obtido (5.400) pelo número de subcategorias (26). Assim, chegou-se ao resultado de 208 exemplos. Apenas nove subcategorias continham mais desvios do que o tamanho definido para a amostra; logo, as demais tiveram todos os seus exemplos analisados. Nessas subcategorias com número excedente de exemplos, buscou-se selecionar a amostra de maneira aleatória.

5.1 *Possível influência de particularidades da fala na escrita*

A modalidade escrita da língua estabelece padrões para certas palavras (ou certos fenômenos) que frequentemente não correspondem à maneira como tais são produzidos na fala. Trata-se, é claro, de duas modalidades distintas que compartilham o mesmo sistema linguístico, e cada uma apresenta particularidades, não sendo a intenção ou a função de nenhuma das duas reproduzir fielmente o que acontece em uma ou na outra. Entretanto, quando se analisam redações de estudantes, encontram-se o que parecem ser influências que o modo como se fala exerce no modo como se escreve. Esta seção analisa dois fenômenos: a omissão de letras cujos fonemas não ocorrem na fala, e os desvios em palavras homônimas e parônimas.

Um dos exemplos em que são omitidas letras cujos fonemas correspondentes não são realizados refere-se ao *-r* final em infinitivos. Esse fenômeno ocorreu muitas vezes nas redações (50 ocorrências na amostra analisada), indicando que tal influência é significativa nos textos[46]:

(1) Lutar por um ideal é um direito , por isso o brasileiro está democraticamente crescendo (…) , e para que isso aconteça no Brasil os movimentos sociais devem <ocorre> .[47]

Identificou-se ainda que diversos desses desvios ocorrem em locuções verbais, como no exemplo (1), o que abre espaço para cogitar se de fato é uma influência da fala ou se na verdade os estudantes apresentam dificuldades em construções com sequências de formas verbais. Outra questão a se notar é que nem sempre a omissão do *-r* gera uma forma verbal:

(2) Deveria <te> mais conhecimento sobre estes movimentos para a população (…) .

Destacam-se também as questões relacionadas às palavras que são homônimas (com pronúncia idêntica e muitas vezes a mesma grafia, mas com significados diferentes) ou parônimas (com pronúncia e grafia parecidas, mas significados diferentes), que podem causar problemas em relação a qual das grafias se refere ao conceito pretendido. Em termos de concordância verbal, um caso de desvio desse tipo envolveu a forma da terceira pessoa do verbo *pôr*, que recebe um *-m* no plural, mas cuja pronúncia no singular e no plural é idêntica:

(3) Veja bem , movimentos sociais são reuniões de pessoas na qual se <põe> em movimento pela conquista de algo .

Já no uso das preposições, ocorreu a troca entre a forma verbal *traz* e a preposição *trás*, cujas pronúncias também são iguais. Os quatro casos analisados com desvios desse tipo (ilustrados pelo exemplo a seguir) traziam a preposição no lugar do verbo:

(4) A superlotação <trás> com ela inúmeros problemas , celas que ultrapassam o limite de detentos , ficam expostas a terem maior índice de danificação , onde pode levar a fuga dos prisioneiros .

Outra troca entre palavras com a mesma pronúncia que ocorreu uma vez na amostra foi o uso do pronome *vós* no lugar do substantivo *voz*. Nesses casos em que as classes morfossintáticas das palavras são diferentes, parece não ser de grande dificuldade para ferramentas computacionais identificar a presença de desvio. Foi possível encontrar ainda

[46] Em função de a saída do arquivo baixado da plataforma de anotação não conter a ID da sentença, optou-se por não inserir essa informação nos exemplos de todo o capítulo. Além disso, manteve-se a separação dos sinais de pontuação por espaços, uma vez que este também era o formato do arquivo de saída da plataforma.

[47] O *token* em que ocorre o desvio é marcado pelos sinais < e > em todo o capítulo. Para fins de concisão, será mantido apenas o trecho da sentença que apresenta o desvio alvo de análise de cada seção.

desvios em expressões com a mesma pronúncia, como em *nada a ver*, que foi grafada como *nada haver*, e em *ainda assim*, que foi grafada como *ainda sim*. Percebeu-se também a troca entre palavras ou terminações de palavras cujas pronúncias não são idênticas, mas são muito semelhantes. Esse foi o caso, por exemplo, dos equívocos entre as palavras *mas*, *mais* e *más*. Em termos de formas verbais, uma troca que ocorreu algumas vezes no *corpus* foi entre as formas conjugadas de futuro dos verbos *ver* e *vir* (*veem* e *vêm*).

Outra questão que merece destaque é a troca frequente entre as terminações *-ão* e *-am*, especialmente em verbos, resultando no uso do tempo verbal equivocado:

(5) (…) tem detentos que estão juntos com outros detentos que <cometerão> crimes menores com pena menores , e outros ficam presos provisório (…) .

O caso de equívoco entre as formas verbais em geral impacta mais a semântica do que a sintaxe; todavia, assim como em outros casos mencionados nesta seção, o *parser* utiliza informações morfológicas e morfossintáticas para identificar as dependências. Com isso, é possível que problemas nos tempos verbais impactem a construção da árvore sintática. Já no caso da troca entre *-am* e *-ão* que resulta em substantivo, há uma maior influência na estrutura sintática, especialmente quando o verbo em questão teria o papel de raiz da sentença:

(6) As facções criminosas <encontrão> neste ambiente local fértil para estender seus tentáculos e criar articulações onde o papel do Estado deixa verdadeiros rombos .

No exemplo (6), o corretor gramatical do *MS Word* não identifica como desvio a palavra *encontrão*, que exerce a função de raiz. O *UDPipe*, por sua vez, identifica esse elemento como verbo e o marca como raiz. É interessante notar que, em alguns casos, o *MS Word* parece conseguir identificar precisamente os desvios, e o *UDPipe* também consegue aplicar as etiquetas corretas, apesar da presença de desvios. Em outros casos, apenas uma ou a outra ferramenta consegue identificar/etiquetar corretamente as palavras que contêm desvios. Seria interessante verificar de forma sistemática o tratamento dado por ferramentas computacionais a esses desvios (tanto corretores gramaticais quanto *POS taggers* e *parsers*).

5.2 *Desvios decorrentes de problemas de acentuação, ortografia ou digitação*

Realizar uma análise isolando-se um nível linguístico é uma tarefa limitada em si, visto que os fenômenos da língua não se comportam de forma segmentada, nos níveis estabelecidos pelas necessidades de sistematização humana. Logo, limitar-se a uma análise puramente sintática implica ignorar a influência dos demais níveis linguísticos e da ortografia, que inicialmente se

havia decidido ignorar. Porém, considerou-se importante analisar neste capítulo também os problemas sintáticos decorrentes de desvios ortográficos.

Estabelecer o limite entre o que é puramente ortográfico, o que é puramente sintático e o que está na interface entre esses dois níveis mostrou-se um desafio. Inicialmente, a intenção da correção ortográfica automática era estabelecer que os desvios não identificados por essa ferramenta seriam considerados sintáticos. Entretanto, logo se percebeu que isso não seria suficiente para definir os limites entre ortografia e sintaxe, em função de que a ferramenta de correção ortográfica se baseia apenas na distinção entre palavras que não são encontradas no léxico (e, dessa forma, são marcadas como desvios ortográficos) e aquelas que pertencem ao léxico da ferramenta (estas são marcadas como desvios gramaticais). Em termos práticos, foi necessário delimitar o que seria ou não anotado; portanto, optou-se por manter a anotação dos desvios que estão mais próximos da ortografia do que da sintaxe, e analisá-los nesta seção. Porém, é necessária a problematização em relação a essa decisão, uma vez que problemas de ausência de acentuação, por exemplo, não poderiam ser considerados de fato desvios sintáticos. Pesquisas futuras que pretendem se ocupar dessa temática precisarão definir, com base em aspectos teóricos mais robustos, os limites entre o que realmente pertence à ortografia, e o que faz parte do nível da sintaxe.

Um dos mais frequentes tipos de desvios ortográficos, como já havia sido identificado por Pinheiro (2008) e por Oliveira (2013), são os problemas de acentuação[48]. Entre os desvios de concordância verbal, aqueles relacionados à acentuação ocorreram essencialmente nos verbos *ter* e *vir* na terceira pessoa (ausência de acento no plural ou excesso de acento no singular, ainda que o primeiro caso seja bem mais frequente que o segundo):

(7) Os movimentos sociais <tem> tomado grande proporção ao logo do tempo , sendo mais específico , por volta de 1950 a 1960 .

(8) Cogita-se com muita frequência que os movimentos sociais <vem> com cada vez mais força , de forma que ocorra mudanças por meio das ações políticas .

A maior parte dos casos está relacionada à ausência de acentuação na terceira pessoa do plural do verbo *ter*, com bem menos ocorrências de problemas com o verbo *vir* e raros problemas quanto ao excesso de acentuação na terceira pessoa do singular, o que ocorreu apenas no verbo *ter*. Encontraram-se ainda poucas ocorrências de troca de acento em *obter* e *conter*.

Outro tipo de desvio bastante frequente foi a ausência de acento no verbo *estar*, resultando em pronome demonstrativo *esta*, e, com menor frequência, o excesso de acento no

[48] Não se considerou como problema de acentuação qualquer desvio relacionado ao uso de crase.

pronome *esta*, resultando no verbo *estar*. Esse tipo de desvio se mostra um desafio para ferramentas de PLN, porque, ainda que os corretores gramaticais de editores de texto consigam identificar adequadamente a maioria desses casos, há alguns exemplos em que essa identificação não acontece. Em termos de *parsing*, o *UDPipe* identificou, em todas as sentenças verificadas, o verbo não acentuado como pronome, aplicando a etiqueta de *POS* incorreta e, por consequência, gerando dependências equivocadas.

(9) O voto não é apenas a única forma de participar das escolhas sociais , mas também com a opinião daquilo que <esta> ou será implantado .

No exemplo (9), possivelmente devido à estrutura de coordenação utilizada com os verbos *estar* e *ser* seguidos de particípio, o *MS Word* não foi capaz de identificar a ausência de acento. Quando o acento foi colocado equivocadamente sobre o pronome (quatro casos), o *MS Word* não sinalizou a presença de desvio. Já o *UDPipe* etiquetou três dos casos como o verbo auxiliar *estar* e um deles como um adjetivo, mostrado no exemplo (10):

(10) Deveria te mais conhecimento sobre estes movimentos para a população , (…) pois sem <estás> formas de contenções para à sociedade formas diversas para fazer estes abordagens (…) .

Um caso que ocorreu apenas uma vez no conjunto analisado, mas que também foi marcado como desvio de pronome resultante de problemas de acentuação, é o do exemplo (11):

(11) A superlotação <é> o descaso <nós> presídios pode causar revolta e revolta resultante em violência e violência que na maioria dos casos são resultantes em morte .

Nesse caso, há dois problemas de colocação inadequada de acento: um deles na conjunção *e*, resultando no verbo *é*, e outro na contração prepositiva *nos*, resultando no pronome pessoal reto *nós*. Esse exemplo contraria a tendência percebida de falta de acentos, inserindo inadequadamente o sinal gráfico onde ele não deveria existir. A acentuação da conjunção *e* ocorreu outras três vezes na amostra.

Os problemas de acentuação se manifestaram também em termos de desvios de uso de preposição. Nesse caso, os desvios estão ligados essencialmente à ausência de acentuação nos verbos *dar* e *pôr*, resultando nas preposições *de/da* e *por*:

(12) (…) por isso precisamos com urgência com que a legislação do país <de> novos planos estratégicos para agir de maneira com que ninguém se prejudique (…) .

(13) Mas precisa cada um se <por> no seu lugar , fazer sua parte que tudo entra nos eixo , mas enquanto um jogar a bola para o outro e procurando desculpas para não fazer melhor , nada , absolutamente NADA vai pra frente .

No exemplo (12), o MS Word não conseguiu identificar a ausência de acento, e o *UDPipe* etiquetou o verbo como preposição. Nos casos com o verbo *por*, que foram bem menos frequentes, todas as ocorrências foram marcadas corretamente pelo MS Word.

Analisando os desvios de conjunção, a ausência de acentuação no verbo *ser*, resultando na conjunção coordenada *e*, foi bastante frequente. Nesse caso, o corretor gramatical do MS Word não identificou nenhuma das ocorrências como desvio, o que leva a crer que a identificação de tal ausência de acentuação pode ser um desafio para as ferramentas computacionais de correção gramatical. Seguem dois exemplos:

(14) Os problemas do Brasil acarretam um ao outro , a desigualdade <e> grande , o ensino e educação é muito pobre , o que fazem os jovens infratores .

(15) Um local pouco usado são nas escolas públicas <e> onde deveria nascer a curiosidade da criança e jovem sobre o mundo as poucas aulas que são apresentadas para eles (…) .

No exemplo (14), identifica-se com facilidade que se trata do verbo *ser* que não foi acentuado. Entretanto, no exemplo (15) essa questão já não é tão evidente, visto que, em função de a estrutura sintática apresentar diversos desvios, seria possível que a intenção de fato fosse usar a conjunção, e não o verbo. A decisão por considerar este um desvio de conjunção se deu porque foi necessário estabelecer uma hipótese de correção da sentença para que se pudessem identificar os desvios, e nesse caso pareceu fazer mais sentido uma hipótese com o verbo *ser*.

Outra categoria que apresentou problemas de acentuação que influenciam na sintaxe foi a de formas verbais, no que se refere tanto ao uso equivocado de tempos, modos e formas verbais finitas quanto às formas infinitas (infinitivo, gerúndio e particípio). A primeira questão identificada foi a ausência de acento em palavras de outras classes morfossintáticas (como substantivos e adjetivos) que resulta em uma forma verbal, como em *crítica/critica*. Houve ainda omissão do acento na palavra *porém*, resultando no infinitivo flexionado *porem*.

Identificou-se também a ausência ou o excesso de acentuação que resultaram em outros tempos verbais ou em verbos inexistentes, como a ausência do acento nas formas de futuro do presente ou a aplicação equivocada de acento em verbos no presente terminados em *-i*, conforme ilustram os exemplos:

(16) Muitas pessoas tem acesso restrito a justiça e cometeram crimes sem gravidades e poderiam aguardar o julgamento fora das prisões , isso <diminuira> o excesso de contingência .

(17) O ensino de rede pública no Brasil ainda é algo que deixa a desejar , os jovens muitas vezes <concluí> os estudos , e sai deixando de aprender muitas coisas , que lhe fará falta (…) .

Tais desvios causam maior impacto na semântica do que na análise sintática em si. Porém, considerando que as ferramentas computacionais levam em conta o contexto e as

informações morfológicas e morfossintáticas das palavras para construir a árvore sintática, torna-se relevante trazer tal tipo de desvio para esta análise.

Outra subcategoria que apresentou desvios de acentuação que influenciam na sintaxe foi a de *orto-sin*, que engloba justamente os desvios em classes morfossintáticas não previstas na tipologia, mas que têm alguma influência na construção da árvore sintática. Os dois casos que merecem destaque são os do substantivo *país* e os da forma verbal *irá*. O primeiro caso, que foi o mais frequente nessa subcategoria, tendo ocorrido 37 vezes, altera a sintaxe na medida em que insere um problema de concordância, visto que a palavra *pais* é um substantivo no plural, enquanto *país* é singular. Esse fenômeno já havia sido identificado por Pinheiro (2008), como descrito na Seção 3.1 (p. 42).

Ocorreram também desvios que derivam de erros de digitação ou que se parecem com desvios ortográficos (possivelmente em função da falta de revisão do texto ou do desconhecimento de grafia), mas que não foram identificados e corrigidos durante a correção ortográfica e que, por causarem impactos na sintaxe, foram anotados na fase de tipificação dos desvios sintáticos. Tais desvios podem se referir a omissão ou acréscimo de letras ou sílabas, trocas de letras por outras ou inversões de letras dentro das palavras.

Os desvios que se relacionam à omissão de letras ou sílabas foram pouco frequentes. Nesse tipo de desvio, pode-se ponderar (ainda que não seja possível afirmar com segurança analisando os textos produzidos, e não o processo de produção) que a sua ocorrência esteja relacionada à falta de revisão do texto. Um exemplo pode ser visto na sentença a seguir:

(18) Os movimentos sociais tem tomado grande proporção ao <logo> do tempo , sendo mais específico , por volta de 1950 a 1960 .

Ocorreu ainda a omissão de sílabas, especialmente em formas verbais infinitas:

(19) Diante da crise econômica as questões estão cada vez mais <pautas> pelos cidadãos que buscam proceder através de manifestações nas ruas , na internet e redes sociais , (…) .

Outra hipótese relacionada a esse fenômeno é que, como a omissão ocorre no meio das palavras, talvez não fosse percebida mesmo com uma revisão do texto. Em função de que, em todos os casos, as palavras de fato existem na língua portuguesa, só é possível a identificação da presença de desvio quando se analisa o contexto no qual ele está inserido, tanto em termos das palavras imediatamente adjacentes quanto dos demais elementos das orações.

O acréscimo de letras mostrou-se comum em diversos exemplos. Uma das mais comuns foi o acréscimo equivocado de *-r* ao fim de palavras:

(20) A solução seria o governo implantar projetos <parar> não deixarem os jovens nas ruas (…).

(21) (…) isso acaba gerando contradições e muitas das vezes não é tão saudável e produtivo para movimento em si , na democracia ou pela <buscar> de seus direitos de igualdade .

A letra *-r* também foi inserida equivocadamente em outros pontos:

(22) Os presos normalmente voltam mais violentes das prisões <por contra de> todo o descaso (…).

Nesse caso, é mais difícil identificar se se trata de desconhecimento da expressão ou de desvio ortográfico/de digitação. Independentemente do fenômeno que deu origem ao desvio, tal equívoco influencia a sintaxe, uma vez que insere uma preposição em uma expressão fixa na qual ela não pode ocorrer sem resultar em agramaticalidade.

O fenômeno da troca de letras por outras foi um dos mais frequentes dentro desse tipo de desvios relacionados à ortografia ou à digitação, ocorrendo em palavras diversas:

(23) (…) , agora mais que nunca os jovens <então> na rua , e onde a faixa etária se encontra .

(24) (…) pois ele é a <porte> de entrada pra arte e cultura que pode muda a socialização (…).

Outras trocas comuns ocorreram nas palavras *deste/desde*, *tendo/tento/tanto*, *portanto/portando*, entre outras. O exemplo (24) merece destaque em função do tipo de desvio decorrente da troca de letras. A hipótese de escrita que parece fazer mais sentido pelo contexto é a palavra *porta*, na expressão *porta de entrada*. No entanto, a troca do *a* pelo *e* gerou um substantivo existente, o que dificulta a identificação da presença de desvio. Caso a hipótese de escrita esteja correta, a troca das letras gera um desvio de concordância entre o artigo e o substantivo ao qual ele se refere. Esse exemplo é desafiador para ferramentas de PLN, pois é mais simples marcar esse fenômeno como um problema de concordância, checando as informações morfológicas e as diferenças de gênero entre os dois elementos, do que identificar a troca de letras. Porém, se não houvesse a troca da letra, também não haveria qualquer problema de concordância nominal.

Por fim, tem-se ainda a troca de lugar das letras dentro da palavra:

(25) Um bom exemplo , são os black blocks que , de um jeito agressivo , tentam passar seus <ideias> destruindo patrimônios públicos e privados .

No exemplo (25), tem-se um problema parecido com o que ocorre no exemplo (24), já que tanto a palavra pretendida quanto a utilizada existem, mas possuem gêneros diferentes. Neste exemplo, porém, o que acontece é a inversão das letras *i* e *a* na palavra, o que é suficiente para resultar em um problema de concordância.

A grande quantidade de desvios de ortografia que influenciam a sintaxe e a sua ocorrência em diversas categorias demonstram que não é possível isolar completamente os

níveis linguísticos ao se propor uma análise sintática. Todavia, também se percebeu que corretores gramaticais (ou pelo menos o do *MS Word*, usado nesta análise) já são capazes de detectar diversos desses problemas. Vale lembrar que os desvios analisados aqui são aqueles que a ferramenta de correção ortográfica não identificou. Como uma possibilidade futura, pode-se investigar mais a fundo os desvios ortográficos presentes em redações, buscando formas de melhorar o desempenho de ferramentas de correção ortográfica existentes.

5.3 *Influência de questões ligadas à morfologia na sintaxe*

Assim como a ortografia influenciou os desvios sintáticos, também as questões morfológicas tiveram impacto na ocorrência de alguns tipos de desvios. Nesta seção, analisam-se dois fenômenos da morfologia que se mostraram relevantes: o primeiro é a aglutinação ou separação equivocada de palavras; o segundo é a flexão de palavras que são invariáveis.

Os desvios de separação ou aglutinação equivocada podem ou não gerar palavras existentes, da mesma classe morfossintática pretendida ou de outras classes. Um exemplo desse último caso é mostrado a seguir:

(26) Temos o direito de participar de movimentos sócias <com tanto> que isso não venha prejudicar as pessoas ao nosso redor e que seja algo produtivo de maneira segura .

Nesse caso, ambos os termos existem, o que pode justificar a sua separação equivocada. A análise mostrou que tal separação é mais comum em palavras com função de ligação, como conjunções. Outros exemplos frequentes de separações equivocadas ocorreram nas palavras *todavia* (como *toda via*), *contudo* (como *com tudo*), *portanto* (como *por tanto*), *acima* (como *a cima*), entre outras. Destaca-se o desvio contido no exemplo abaixo:

(27) (…) no dia 28 de abril de 2017 teve uma grande manifestação contra a reforma da previdência parou o Brasil todo não tinha nada em <circula mento> como , transporte , escolas e etc..

Notam-se no exemplo (27) dois fenômenos distintos relacionados ao termo anotado. O primeiro deles se refere à separação equivocada de elementos, gerando palavras existentes. Uma busca no Vocabulário Ortográfico da Língua Portuguesa (VOLP)[49] mostra que a palavra *mento* existe e, segundo o dicionário Michaelis On-Line[50], restringe-se aos campos da anatomia e da zoologia[51]. No entanto, é muito provável que essa não era a palavra que se pretendia empregar.

[49] Disponível em http://www.academia.org.br/nossa-lingua/busca-no-vocabulario
[50] Disponível em https://michaelis.uol.com.br/
[51] Conforme a definição, trata-se da parte saliente que fica abaixo do lábio inferior em animais e humanos.

Assim, chega-se ao segundo desvio: a seleção equivocada do sufixo nominalizador. É provável, pelo contexto, que a palavra pretendida fosse *circulação*. Esse tipo de desvio seria mais esperado em textos de aprendizes não nativos de português, mas ocorreu algumas vezes no *corpus* de redações de nativos dessa língua. Logo, parece válido aprofundar, como trabalho futuro, os estudos sobre os equívocos de formação de palavras em redações, a fim de compreender melhor os aspectos envolvidos nesse fenômeno.

Outra questão que talvez seja menos desafiadora às ferramentas é quando a separação de palavras gera elementos que não existem na língua portuguesa:

(28) Porém , se existem leis a serem cumpridas e artigos que defendem os direitos humanos qual seria o porque de tantos movimentos sociais que ocorreram <na quele> período , e que ocorrem até hoje ?

No exemplo (28), é possível que a separação tenha resultado do fato de que o elemento *na* existe como contração prepositiva, levando à separação equivocada. Analisando as aglutinações indevidas, identificou-se que tais desvios também podem resultar em palavras existentes ou não. No caso das não existentes, verificou-se a ocorrência de algumas contrações indevidas, especialmente entre *desde* e *a/as*:

(29) Só podemos inverter essa situação ensinando os futuros cidadãos e futuras lideranças <desdas> primeiras palavras a serem cidadãos de educação em tempo absoluto .

Já no caso daquelas que existem na língua, mas têm significado diverso, uma ocorrência frequente foi a aglutinação de elementos na expressão *a fim de*. A palavra *afim* geralmente tem *POS* de adjetivo, e o *MS Word* é capaz de marcar tal item como contendo desvio gramatical. O *UDPipe*, por outro lado, etiquetou todas as ocorrências dessa aglutinação como um advérbio de modo. Nesse sentido, ferramentas computacionais que se proponham a lidar com textos que contenham desvios sintáticos precisam ser projetadas de forma a considerar esse tipo de fenômeno, levando em conta as aglutinações, mas também as separações, cuja análise deverá ir além dos limites do *token* para a sua identificação.

A segunda questão ligada à morfologia foi a flexão de palavras invariáveis. Na língua portuguesa, há palavras que permitem flexão de gênero, de número, de pessoa, etc., e outras que são chamadas de invariáveis, cuja forma não muda. Porém, percebeu-se nas análises que os produtores dos textos nem sempre reconhecem as palavras que não permitem flexão. Em alguns casos, identificou-se que o desvio de flexão ocorreu em palavras que permitem flexão quando assumem determinados *POS* na sentença, mas não o permitem quando o *POS* é diferente. Esse é o caso do exemplo abaixo:

(30) Relativo à legitimidade dos movimentos sociais no Brasil , pode-se destacar <tantos> aspectos positivos quanto negativos .

Aqui a expressão *tanto… quanto* é invariável, mas a palavra *tanto* permite flexão quando exerce o papel de modificador do substantivo (*Ela tem tantas tarefas atrasadas que nem sai mais de casa*). Outro exemplo semelhante é o caso de flexão do particípio:

(31) Entretanto , em meados de 1950 outras ONGs já haviam se <movimentados> para defender seus direitos .

Quando tem *POS* de adjetivo, a palavra *movimentado* concorda com os elementos a que se refere, ou seja, pode ser flexionada (como em *ruas movimentadas*). No entanto, no caso da construção do exemplo (31), a forma verbal complexa exige um particípio que não é variável, isto é, não permite a flexão de plural realizada. Também ocorreu com frequência a flexão de preposições ou de elementos que compõem locuções prepositivas:

(32) (…) se juntam em uma ação para reivindicar algo na qual possa se beneficiar , <por meios de> protestos e impondo sua opinião ; pode acontecer por meio de paralisações e ocupação (…) .

Nesse exemplo, vê-se a flexão da primeira ocorrência da expressão *por meio de*, mas não da segunda. Novamente, a palavra *meio* pode ser flexionada quando assume outros papéis na sentença, mas não dentro dessa expressão. Outra expressão em que ocorreu fenômeno semelhante foi *por conta de*, em que o termo *conta* foi flexionado. Também se identificou flexão equivocada da preposição *contra*, entre outras ocorrências.

5.4 *Desvios de colocação, regência e concordância de palavras*

Após abordar alguns dos fenômenos que têm influência na sintaxe, mas que fazem parte de outros níveis linguísticos, chega-se àqueles fenômenos que são o foco desta pesquisa. A sintaxe se ocupa das relações estabelecidas entre os elementos de uma sentença, especialmente no que se refere a três aspectos: ordem (ou sintaxe de colocação), questões ligadas à regência (ou sintaxe de regência) e concordância entre palavras (ou sintaxe de concordância). Assim, esta seção se subdivide nos fenômenos que dizem respeito a esses três aspectos.

5.4.1 Inversões da ordem canônica SVO

Segundo o *World Atlas of Language Structures Online* (WALS Online)[52], uma grande base de dados com diversas propriedades estruturais das línguas, a ordem predominante das orações do português (chamada de *ordem canônica*) é aquela em que o sujeito, o verbo e o objeto (SVO) aparecem nessa sequência. Assim, diz-se que o português é uma língua com ordem SVO, em que o sujeito e/ou o objeto podem ser omitidos, dependendo das características do verbo. Porém, a inversão dessa ordem não é incomum, sendo considerada um recurso de estilística.

Nas redações analisadas nesta pesquisa, notou-se a ocorrência de inversões diversas entre esses três elementos principais. Nesse sentido, também se identificou que tais inversões parecem suscitar a ocorrência de desvios sintáticos. Esta seção analisa os desvios que ocorrem em sentenças em que a ordem canônica SVO foi invertida, como os fenômenos relacionados à colocação do sujeito posposta ao verbo.

A primeira ocorrência a ser analisada é o emprego equivocado de vírgulas em estruturas com inversões. Tal fenômeno foi identificado apenas três vezes dentro da amostra de desvios da subcategoria de excesso de pontuação, como ilustram os exemplos:

(33) Torna-se evidente <,> a importância da inclusão de respeito em movimentos envolvendo um de conjuntos de pessoas .

(34) Contudo , seria a legitimidade destas organizações coletivas <,> dada apenas posteriormente o " boom " da Revolução Industrial ?

No exemplo (33), inverteu-se a ordem do sujeito, mantendo-se a sequência entre verbo e predicativo de sujeito no início da sentença. Nesse caso, vê-se que o elemento trazido para o início da oração é mais curto do que o sujeito. Pode-se argumentar que o motivo da inversão seja realçar os elementos que foram deslocados para o início da oração, mas também é válido notar a questão do tamanho de cada elemento. Uma possibilidade de investigação seria verificar se há correlação entre esse tipo de inversão e o número de *tokens* dos elementos invertidos.

No exemplo (34), a estrutura é interrogativa; segundo alguns gramáticos, esse tipo de inversão é mais comum em interrogativas. Porém, nota-se que a inversão separou o verbo *ser* do particípio *dada* na estrutura passiva, o que parece ter complexificado a sintaxe da sentença, e isso poderia justificar o uso equivocado do sinal de pontuação. Não é possível definir categoricamente o tipo de fenômeno envolvido na colocação equivocada das vírgulas nesses exemplos, mas é válido questionar se a inversão da ordem influencia de forma significativa nesses desvios.

[52] Disponível em https://wals.info/

Outro caso frequente de inversões é em orações subordinadas, em que o objeto muitas vezes aparece anteposto às estruturas de sujeito e verbo. Um dos fenômenos que se destaca pela sua frequência (47 ocorrências) são os desvios de regência verbal em que o elemento regente (o verbo) aparece posposto ao elemento regido (a preposição), o que ocorre essencialmente em orações subordinadas restritivas:

(35) Compreendendo melhor o que é os movimentos sociais (…), e estão reivindicando o que é prometido pelas pessoas que eles <votam> , acreditando que irá fazer aquilo por eles .

(36) Atualmente , com essa crise em que o país está <passando> , o maior foco da maioria dos políticos e brasileiros é em ganhar dinheiro (…) .

No exemplo (35), o desvio foi a ausência da preposição que é requerida pelo verbo *votar*. Nesse caso, a inversão está ligada à subordinada restritiva, que exige que a preposição seja colocada antes da conjunção que a introduz, entre os dois elementos sublinhados. O exemplo (36) também traz um caso de subordinada restritiva, mas nessa sentença foi utilizada a preposição equivocada em relação ao que exige o verbo *passar*. Na regência nominal, tal inversão só ocorreu uma vez na amostra:

(37) Na contemporaneidade , é comum serem realizados movimentos sociais com o intuito de alcançar transformações de questões que grande parte da população não está satisfeita .

No exemplo (37), utilizou-se a preposição equivocada, de acordo com a regência do adjetivo *satisfeita* (termo regente). Vê-se que, nesse caso, a inversão também deriva da construção de uma subordinada restritiva, em que a preposição precisa ser colocada antes da conjunção *que*, a qual introduz a estrutura subordinada. Nas análises, percebeu-se que a ausência ou o uso equivocado da preposição ocorreu com mais frequência do que o uso excessivo de preposição no fenômeno de inversão da ordem canônica dos elementos. Questiona-se se há alguma relação entre a inversão e os desvios de regência, mas também se há alguma influência da distância existente entre o elemento regente e o elemento regido.

Em termos de concordância verbal, identificaram-se 28 casos de desvio de concordância quando houve a inversão entre o sujeito e o verbo (sujeito posposto). Nesses casos, nem sempre a inversão esteve relacionada às orações subordinadas:

(38) Ao analisar o atual cenário brasileiro , <é> notório os diversos grupos que integram o quadro social do país e suas diferentes necessidades .

(39) O movimento já obteve grandes conquistas , e continua seguindo em busca da quebra de paradigmas , <é> realizada diversas manifestações e os temas abordados são a violência , respeito , (…) .

Nos exemplos (38) e (39), o verbo com o qual a concordância é feita de maneira inadequada é o *ser*. No primeiro caso, tem-se uma estrutura com predicativo de sujeito, com um sujeito posposto dentro do qual há uma oração subordinada restritiva. Novamente surge a hipótese de que o número de *tokens* pode ter alguma relação com a opção pela inversão da ordem canônica. Já no segundo caso, tem-se um problema de concordância envolvendo uma estrutura passiva, o que também poderia dificultar a construção adequada da concordância. Nesse segundo exemplo, é interessante notar que há também um problema de concordância no particípio, que concorda em gênero, mas não em número com o elemento ao qual se refere. Diferentemente do exemplo (38), no caso do exemplo (39) os elementos invertidos têm exatamente o mesmo número de *tokens*. Seguem mais exemplos:

(40) A intervenção seria a organização dos líderes desses movimentos , para que não <aconteça> atitudes de forma contrária ao propósito principal do movimento .

(41) Através dela pode ser passado situações em que <se encontram> o mundo neste momento .

O exemplo (40) ilustra a maioria dos problemas de concordância com o sujeito posposto, em que o verbo aparece do singular, e o sujeito com o qual ele deveria concordar está no plural. Nesse caso, a sentença na qual ocorre o desvio é uma subordinada, o que também se mostrou frequente nesse tipo de fenômeno. No exemplo (41), por sua vez, ocorre o contrário: o verbo é colocado no plural, mas o sujeito está no singular. Todavia, o substantivo que antecede o verbo está no plural, o que talvez possa justificar a ocorrência do desvio. Novamente, o fenômeno ocorre dentro de uma oração subordinada, o que reforça a hipótese de que pode haver uma correlação da ocorrência de desvios com inversões da ordem canônica e orações subordinadas.

Os desvios de concordância com inversão de elementos também ocorreram na concordância nominal. A maioria dos casos está ligada à concordância com o particípio anteposto em passivas ou com construções do tipo *é* + adjetivo:

(42) De fato , nesses movimentos sociais é <mostrado> a revolta do povo que vem aumentando que , por muitos anos acomodada , agora abriu os olhos para enxergar a injustiça e tomar as devidas atitudes.

(43) É <constante> as noticias referentes às reivindicações no Brasil recorrente aos inúmeros grupos insatisfeitos com as propostas do atual governo no país .

Ainda no que se refere a estruturas inversas, ocorreram alguns poucos casos em que as inversões de ordem causaram desvios de estrutura sintática:

(44) Informações essas <que> não são questionadas e que se disseminam para as muitas regiões do país com acesso limitado à educação formal e crítica .

No exemplo (44), a sentença não apresentaria qualquer desvio se fosse utilizada, por exemplo, como uma aposição. Porém, isoladamente, ela não tem verbo principal, visto que o pronome relativo *que* introduz uma subordinada, e antes dele não há verbo. A inversão se dá entre substantivo e demonstrativo, o que exige que a sequência seja uma subordinada, mas tal fenômeno resulta em desvio porque se constrói uma sentença sem verbo principal.

Uma questão que é analisada nesta seção, mesmo que tal decisão seja questionável, é a colocação pronominal equivocada em início de sentenças (próclise). Segundo Bechara (2009), já não se considera mais este um problema puramente sintático, e sim uma questão fonético-sintática. No entanto, uma vez que a colocação do pronome átono na frente do verbo em geral leva a uma inversão entre verbo e objeto, decidiu-se por analisar esse fenômeno nesta seção sobre inversões. Considerando que tal colocação pronominal é típica da fala, mesmo nos níveis cultos, esperava-se uma ocorrência frequente desse desvio. Entretanto, tal posição de pronomes só ocorreu em oito sentenças. Destas, três sentenças se referem ao verbo *sentir-se*, outras quatro envolvem verbos reflexivos ou recíprocos (*expressar-se*, *aliar-se*, *movimentar-se*, *ver-se*) e apenas uma indica índice de indeterminação do sujeito:

(45) Nos sentimos injustiçados e oprimidos por motivos individuais e coletivos .

(46) Se deve deve lutar , pelos motivos certos que trás benefícios ao todo e não ao individual .

No exemplo (45), o verbo está no plural, o que teria exigido a omissão do *-s*, em caso de ênclise (*sentimo-nos*). Como tal fenômeno só ocorre na escrita e, ainda assim, é pouco usual, mas a não omissão do *-s* causa estranhamento em termos de sonoridade (**sentimos-nos*), é possível que resida aí a justificativa para a anteposição do pronome. Já o exemplo (46), que é aquele em que se poderia argumentar mais fortemente que o critério da inversão da ordem canônica não se aplica, traz duas particularidades: a primeira é que essa foi a única ocorrência de próclise equivocada com índice de indeterminação do sujeito; a segunda é a repetição da forma verbal *deve*. Ademais, o verbo *trazer* apresenta um desvio de grafia e um problema de concordância, e há mais de um desvio de pontuação. Assim, os problemas de construção da estrutura sintática vão além da colocação pronominal equivocada.

Como último exemplo desta seção, destaca-se uma das poucas sentenças que apresentaram desvio de ordem de palavras:

(47) Portanto , <medidas devem haver> para solucionar o impasse , segundo Immanuel Kant " o ser humano é aquilo que a educação faz dele .

Nesse exemplo, a inversão da ordem com a anteposição do objeto gerou uma estrutura que, se não é totalmente agramatical, no mínimo causa estranheza na leitura. Ainda, a flexão da

locução verbal com o verbo *haver* também foi feita de maneira equivocada, o que levanta a hipótese de que a construção com esse verbo seja a justificativa para a inversão.

5.4.2 Presença de elementos deslocados ou intercalados

Além das inversões entre sujeito, verbo e objeto, ocorrem com muita frequência as anteposições de estruturas adverbiais e de orações subordinadas de finalidade, entre outros deslocamentos comuns. Também ocorre nas estruturas de sentenças a inserção de elementos intercalados, como estruturas apositivas explicativas, entre outros. Considera-se que, assim como as inversões da ordem canônica, a presença de elementos deslocados torna as estruturas mais complexas, uma vez que frequentemente aumenta a distância entre os elementos dependentes.

Esta seção analisa os diversos fenômenos relacionados a elementos deslocados e intercalados que possam estar associados aos desvios sintáticos. O primeiro caso analisado é o das aposições explicativas, inicialmente no que se refere a problemas de uso de pontuação:

(48) O <Brasil> como uma sociedade rica em <cultura> não esta aproveitando-a como deveria .

No exemplo (48), a expressão sublinhada é uma aposição que traz uma explicação sobre o país. Dessa forma, precisa ser sinalizada com vírgulas em seu início e final, como mostram as marcas de anotação[53]. Tal aposição foi inserida entre o sujeito e o verbo; logo, a presença de vírgulas marca um elemento intercalado que rompe a estrutura central da sentença, aumentando a distância entre o núcleo do sujeito e a estrutura predicativa da qual ele depende. No exemplo a seguir, também há um problema de pontuação em uma aposição explicativa intercalada:

(49) (…) no Brasil o movimento se intensificou no século XX , quando a <população> insatisfeita com transformações feitas no campo econômico e social , agiu contra a polícia vigente (…) .

No exemplo (49), a estrutura apositiva sublinhada foi marcada com vírgulas apenas em seu término, e não no início, mas ambas as vírgulas são obrigatórias. Além disso, trata-se de uma estrutura complexa de vários *tokens*, dentro da qual há ainda uma construção passiva. Nesse caso, a inserção da aposição também aumentou significativamente a distância entre o núcleo do sujeito (*população*) e o verbo do qual ele depende (*agiu*).

[53] Na ausência de pontuações, a anotação foi realizada no token imediatamente anterior ao local onde a pontuação deveria ter ocorrido.

Outra questão frequente envolveu desvios de pontuação na anteposição de orações subordinadas ou de construções com funções adverbiais. A ausência de pontuação nos casos com mais de três *tokens* foi frequente[54]:

(50) Desde a década de <60> os movimentos sociais são ações recorrentes na sociedade brasileira .

Esse caso típico de adjuntos adverbiais de tempo ou de local deslocados para o início da sentença teve apenas algumas ocorrências de ausência de vírgulas. Identificou-se ainda a ocorrência de desvios em que a estrutura adverbial continha uma oração subordinada restritiva:

(51) No país em que <vivemos> a arte é fundamental na sociedade , ela tem a função de mostrar caminhos ou trazer conhecimentos e de mostra uma visão apurada sobre o nosso cotidiano , ela tem um papel fundamental dentro das comunidades .

Relativamente recorrente, porém, foi a ausência de vírgulas em adjuntos deslocados inseridos no meio da sentença, tanto intercalados quanto antepostos a uma oração:

(52) A criminalidade penal tem crescido a cada ano , jovens de 15 anos já começam a praticar furtos e trafico de drogas , nas prisões do <Brasil> estão crescendo e ficando cada vez mas pior , de fato que as fundações casa também estão cheias , (…)

Nesse caso, a estrutura adverbial de local se encontra em uma sentença com sequências de orações coordenadas, inserida entre o sujeito (*tráfico de drogas*) e os gerúndios coordenados a que ele se refere (*estão crescendo e ficando*). Nota-se que foi gerado um problema de concordância verbal, possivelmente em função da presença de um elemento intercalado na dependência, o que aumenta a distância entre os dependentes. Porém, não foi só a vírgula que encerra a estrutura adverbial que foi omitida; também ocorreu a omissão da vírgula inicial:

(53) (…) tendo em vista a idealização de sociedade perfeita que foi tirada dos pobres durante a revolução <industrial> no século XIX , é um dos principais papeis de movimentos sociais , ONGs e sindicatos.

No exemplo (53), é obrigatória a colocação de vírgulas intercalando o trecho sublinhado por dois motivos: o primeiro é que se trata de uma aposição explicativa sobre o período de tempo em que ocorreu o evento anterior (*a revolução industrial*); o segundo é que há uma sequência de duas estruturas com função adverbial, que precisam ser separadas por vírgulas.

Outra questão identificada em termos de estruturas com função adverbial foi a ausência de vírgulas marcando orações com a presença de formas verbais finitas e função adverbial:

(54) Quando se fala em movimentos <sociais> é comumente dizer que são grupos de pessoas que se organizam para lutar contra alguma ideia imposta por algum agente público superior (…) .

[54] Conforme se havia definido na diretriz, somente seria considerada ausência de vírgulas quando tais estruturas tivessem mais de três *tokens*, tanto no início quanto no meio da sentença.

No exemplo (54), a oração sublinhada tem função adverbial temporal e foi deslocada para o início da sentença. Nesse caso, é obrigatória a colocação de vírgulas intercalando as duas orações. Tal caso também ocorreu no meio da sentença:

(55) O governo cobra muito de todos os trabalhadores , pessoas que batalham (…) e <quando> alguns cidadãos resolvem lutar pelos seus <direitos> em boa parte das vezes é isso o que acontece .

Vê-se no exemplo (55) que a oração adverbial sublinhada não foi marcada com vírgulas nem em seu início, nem no final. Porém, como foi inserida em posição anteposta entre duas orações coordenadas, a colocação das vírgulas é obrigatória. Outra ocorrência frequente foi a anteposição de orações subordinadas introduzidas pela locução conjuntiva *para que* ou as reduzidas de infinitivo introduzidas pela preposição *para*:

(56) No entanto , para que todas essas ações sejam <tangíveis> é necessário que a Receita Federal invista uma maior parcela dos impostos em arte e cultura .

(57) Para sair dessa crise que passa em nosso sistema <penitenciário> é imprescindível a privatização , que é a principal .

No exemplo (56), o trecho sublinhado corresponde a uma subordinada adverbial que foi anteposta e que, por isso, precisa ser intercalada por vírgulas. Esse tipo de inversão é comum no conjunto de dados analisados, e frequentemente está associada à ausência de pontuação. Nota-se ainda o uso do subjuntivo como tempo verbal das duas orações subordinadas dessa sentença (*sejam* e *invista*), o que poderia influenciar a ocorrência do desvio. No exemplo (57), a estrutura anteposta (sublinhada) é uma subordinada reduzida de infinitivo, dentro da qual também ocorre uma adjetiva restritiva. Dessa forma, para marcar o deslocamento da estrutura, é necessária a colocação de vírgulas ao seu término. O exemplo a seguir merece destaque:

(58) Entretanto , analisando as últimas manifestações <ocorridas> é possível notar que elas não estão respeitando as leis , desta forma , acabam causando destruição , prejuízo e desordem para as cidades.

O trecho sublinhado também se refere a uma oração subordinada que foi anteposta à oração principal e que, portanto, deveria ter sido marcada por vírgulas. Todavia, ressalta-se nesse caso o fato de que se trata de uma estrutura com gerúndio (*analisando*) dentro da qual há ainda uma estrutura com particípio (*ocorridas*). É possível que a presença dessas duas formas verbais infinitas também esteja associada à ocorrência do desvio.

A última questão ligada a ausência de pontuação em elementos deslocados se refere às orações condicionais intercaladas:

(59) Sabe-se que <se> o Brasil quiser investir na <educação> ele pode .

No exemplo (59), a subordinada condicional intercala a conjunção subordinativa *que* e a subordinada subjetiva que ele introduz (*ele pode*); por isso, precisaria ser intercalada por vírgulas no seu início e final. Nessa sentença, há uma inversão de elementos: a ordem direta seria "*Sabe-se que o Brasil pode investir na educação se ele quiser*". Destaca-se que foram transpostos para a subordinada condicional o sujeito (*o Brasil*) e o complemento do verbo *poder* (*investir na educação*), que constitui núcleo da subordinada subjetiva. Cogita-se que a construção da árvore sintática automática de sentenças com tais inversões possa trazer desafios às ferramentas computacionais, especialmente diante da ocorrência de desvio sintático.

Relativamente à colocação ou à ordem dos elementos dentro das sentenças, diz-se que, em português, as estruturas com funções adverbiais têm posição flexível, podendo ocorrer em locais diversos, seja como anteposições ou posposições às estruturas que elas modificam. Todavia, tal variabilidade não é totalmente livre e, portanto, determinadas posições de advérbios podem gerar estruturas agramaticais. Uma vez que se trata de redações escritas por falantes nativos de português, não se esperava que houvesse desvios de ordem de palavras. De fato tais desvios foram muito pouco frequentes, mas destacam-se dois exemplos marcantes:

(60) A reforma da previdência está sendo uma das intervenções mais impactantes deste ano , porque fere a cultura dos direitos trabalhistas construídos em nosso país , mesmo sendo necessária para melhorar nossa economia é uma mudança que deixa <extremamente o movimento do trabalho irritado> .

(61) Pois sempre haverá um fato a ser relatado <ao longo que a história da humanidade se desenvolve durante vários anos que existe e ainda vário existirão .>

No exemplo (60), nota-se que o produtor do texto tem conhecimento das orientações e regras da modalidade escrita formal do português, uma vez que a sentença de maneira geral apresenta poucos desvios. Nesse sentido, destaca-se a colocação do advérbio em posição irregular. Uma hipótese para esse caso é que a intenção era inserir o advérbio antes de *irritado* e que, por desatenção ou falta de revisão do texto, gerou-se o desvio de ordem de palavras.

No exemplo (61), por outro lado, identificam-se problemas significativos de construção sintática da sentença, ao ponto de a compreensão da mensagem ficar prejudicada. Nesse caso, é difícil estabelecer uma hipótese de reescrita que não altere significativamente não apenas a ordem das palavras, mas também os itens lexicais e a estrutura da sentença como um todo. Logo, tal caso explicita uma questão que foi evidenciada durante as análises: dentro de um mesmo fenômeno, há desvios que comprometem mais e aqueles que comprometem menos a compreensão da mensagem pretendida.

Nesse sentido, o desenvolvimento de ferramentas de correção/avaliação automática de redações deve considerar que desvios semelhantes podem não ter o mesmo peso em termos de

comprometimento da estrutura sintática e que, por isso, talvez não devam ser contabilizados ou penalizados da mesma forma. Ao mesmo tempo em que o nível de comprometimento da compreensão da mensagem é de difícil mensuração em termos objetivos, negligenciar esse fato na arquitetura de tais ferramentas computacionais pode gerar inconsistências em termos de avaliação de redações, em comparação com a avaliação humana. Conforme as experiências em processos de avaliação de redações dos quais a pesquisadora fez parte, estes últimos tendem a avaliar de maneira diferente textos com desvios sintáticos semelhantes, mas que são mais ou menos compreensíveis em função das suas particularidades ou da sua quantidade na redação.

5.4.3 Formação de construções clivadas

Orações clivadas são compostas por uma estrutura iniciada pelo verbo *ser*, seguida pelo termo clivado (essencialmente substantivos), colocando-se *que*/*quem* na sequência, e seguindo-se com o restante da sentença. Um exemplo de estrutura clivada seria "*Foi a buchada que me fez mal*". Segundo Perini (1995, p. 215), "apesar da evidente complexidade da relação, as orações clivadas são de uso muito corrente tanto na fala quanto na escrita". Nas análises, verificou-se a presença de desvios nas tentativas dos produtores de construírem tais sentenças clivadas, os quais se deram essencialmente em duas situações: acrescentou-se apenas o verbo *ser*, sem completar a estrutura com o *que* (tal construção ocorreu apenas uma vez); inseriu-se apenas o *que*, sem iniciar a construção com o verbo *ser* (esse caso se mostrou relativamente frequente):

(62) Entretanto quando a voz do povo já não é mais ouvida , geralmente tendem a (…), e é a partir <disso> surgem os gritos que já não são tão compreensíveis , pois já querem calar a voz do outro .

(63) A partir de tais lutas <que> essas classes sociais conseguiram grande parte dos direitos possuídos , mas , como tudo na vida , o extremo desses movimentos é extremamente prejudicial (…) .

Houve, no entanto, um caso particular:

(64) No cenário brasileiro não é diferente , pois , <são> através das falhas do sistema carcerário que os detentos estabelecem a divisão de poder entre si , resultando no crescimento de facções criminosas e em violências internas .

No exemplo (64), a estrutura clivada contém os dois elementos necessários, mas a conjugação do verbo foi feita mantendo-se a concordância com o substantivo inserido no elemento clivado (*falhas*). Todavia, em função de a estrutura clivada ser iniciada pela locução prepositiva *através de*, a conjugação adequada deveria ser o singular. Considerando, conforme Perini (1995), que estruturas clivadas são complexas em si, e que tal caso é acrescido de

complexidade devido à estrutura escolhida para ser clivada, pode-se pensar que o desvio se dá justamente em função de tal complexidade estrutural.

5.4.4 Estruturas de coordenação

Sentenças compostas (isto é, aquelas em que há mais de uma forma verbal finita) foram alvo de investigação da pesquisa, uma vez que se pretendia verificar se elas poderiam estar mais fortemente correlacionadas com a presença de desvio sintático do que as sentenças simples (com apenas um verbo finito). Assim, um dos fenômenos gerais analisados foi a coordenação.

O primeiro tipo de desvio ligado à coordenação é a ausência de pontuação separando orações coordenadas introduzidas por conjunção coordenada. Bechara (2009) afirma que um dos usos da vírgula é separar orações coordenadas proferidas com pausa. Já que a noção de "proferidas com pausa" é vaga, utilizou-se como regra a anotação de ausência de pontuação quando o sinal não ocorria em sentenças com orações estabelecendo relação de coordenação e presença de conjunção coordenada entre ambas:

(65) Isso mostra que há corrupção esta presente por todas as <partes> pois o dinheiro que deveria ser investido na segurança , educação e saúde são desviados e para outros fins (…).

A ausência de pontuação é a subcategoria com maior número de desvios, e somente uma amostra de 208 sentenças foi utilizada para a análise dos fenômenos linguísticos. Além disso, o limite entre as regras de uso de pontuação e as questões que são essencialmente estilísticas não é claro, mesmo após consultas a mais de uma gramática. Nesse sentido, como afirma Oliveira (2013), "as funções da pontuação são tão diversificadas que não haverá jamais manual que possa abarcar todas as suas potencialidades" (p. 43).

Assim, as questões de pontuação foram desafiadoras para a anotação, com decisões frequentemente baseadas na experiência da autora. Dessa forma, em função da complexidade do uso de sinais de pontuação e da diversidade de fenômenos nos quais ocorrem os desvios, abre-se espaço para uma investigação mais aprofundada em relação às orientações das gramáticas e do próprio ENEM sobre o que é desvio de pontuação. No caso do ENEM, seria interessante verificar se as redações são penalizadas pelos corretores de acordo com percepções individuais, e se há possibilidade de uniformizar as avaliações por meio de ferramentas computacionais ou de orientações mais rigorosas.

Identificou-se que a concordância também parece ser afetada por estruturas de coordenação. Quando há coordenação entre os elementos que compõem o sujeito, nem sempre a concordância do verbo é feita adequadamente:

(66) Porém , a repressão e a violência sofrida pelos manifestantes <interfere> no direito do cidadão fazendo com que nós voltamos ao tempo da ditadura .

No exemplo (66), cogita-se que a concordância tenha sido realizada com o elemento mais próximo, levantando questionamentos sobre a influência da distância entre os dependentes nos desvios de concordância verbal. Outra questão possível é que é justamente a coordenação entre elementos que complexifica a estrutura, dificultando a identificação de que os núcleos da construção são dois elementos coordenados.

(67) A corrupção e a desigualdade no Brasil <está> desencadeando movimentos sociais de forma contínua e <está> gerando polêmica ao ponto da população desacreditar nos atos (…) .

No exemplo (67), destaca-se a presença de duas estruturas de coordenação: uma entre os elementos do sujeito e a outra entre os dois tempos verbais compostos (verbo *ser* + gerúndio). Ambos os verbos auxiliares têm problemas de concordância com o sujeito, mas concordam entre si. Novamente a coordenação parece ser uma estrutura que traz dificuldades aos produtores de redações no que se refere à concordância.

Também se identificaram problemas de concordância nominal em coordenações:

(68) As condições de vida dentro de uma cadeia é <precária> , <desumana> e <cruel> (…) .

(69) (…) esses movimentos iriam ser passivos e <produtivo> para sociedade .

(70) (…) todos os seus valores e opiniões acerca dessas ações serão <recebidas> de forma negativa (…).

No exemplo (68), a sequência de adjetivos coordenados mantém o mesmo problema de concordância em relação ao núcleo ao qual se referem; ao mesmo tempo, todos eles concordam com o substantivo mais próximo (*cadeia*). Já no exemplo (69), destaca-se o fato de que o primeiro elemento da coordenação mantém a concordância adequada, e só o segundo incorre em desvio. O exemplo (70) traz uma estrutura passiva em que o particípio não concorda em gênero porque os dois elementos coordenados aos quais ele se refere têm gêneros diferentes e, por isso, seria necessário utilizar o masculino. Todavia, o particípio concorda em gênero com o elemento mais próximo da coordenação.

Por fim, identificaram-se ainda alguns casos de diversas coordenações intercaladas por conjunções *e*. Também houve casos de repetição sequencial de *e* e algumas ocasiões em que ele foi inserido em locais nas sentenças em que não desempenhava função nenhuma. Assim,

vê-se que as variedade dos desvios ligados à coordenação parecem reforçar a afirmação de que se trata de uma estrutura de execução complexa por parte dos produtores dos textos.

5.4.5 Estruturas de subordinação

Outra hipótese que esta pesquisa investigou é a de que subordinações estão fortemente correlacionadas à presença de desvio sintático, pois constituem estruturas complexas e, por isso, podem ser de difícil execução por parte dos produtores de textos em formação. Assim, durante as análises linguísticas, também se buscou investigar os desvios que ocorreram nesse tipo de estrutura. Tais desvios foram bastante variados e ocorreram em diversas categorias. Entretanto, de forma a manter a generalização, esta seção analisará os diferentes fenômenos ligados à subordinação, ainda que eles apresentem características e funcionamentos diversos.

A primeira categoria de desvios ligados às subordinadas é a de pontuação. Ainda que a colocação de pontuações possa ser uma questão estilística, como dito anteriormente, o uso da vírgula para separar orações intercaladas e subordinadas adjetivas explicativas ou adverbiais é comumente recomendado. Nas análises, foram identificadas 20 ocorrências de ausência de vírgulas em subordinadas adjetivas em que não havia dúvida de que eram explicativas (nas quais, para diferenciar daquelas de valor restritivo, a presença de vírgula é obrigatória):

(71) É bem provável que os presos acabem se tornando pessoas piores (…), mas não é tão simples mudar essa situação no <Brasil> que é um país que prende tantas pessoas (…) .

Ainda no que se refere ao uso de pontuação em subordinadas adjetivas explicativas ou restritivas, identificou-se também o fenômeno inverso: o excesso de vírgula em restritivas. Tal fenômeno foi muito menos frequente do que a ausência de vírgulas nas explicativas:

(72) Todavia esse grupo de pessoas <,> que compõem o movimento , já foram tratadas com violência pelo governo , que se recusa a aprovar e apoiar a razão do movimento .

Ainda sobre pontuação, identificou-se a ocorrência frequente de separação de sujeito e verbo com vírgula quando, dentro do sujeito, havia uma estrutura relativa:

(73) Outro fator <u>que colabora com o problema</u> <,> é a falta de comprometimento de pena , onde os presos não cumprem o tempo imposto pela justiça , tendo como consequência a desordem (…) .

Bechara (2009) aceita que, quando a oração adjetiva restritiva tiver “certa extensão”, tal pontuação pode ser empregada, especialmente quando os verbos de duas orações diferentes se juntam. Porém, como a noção de extensão não é retomada e é de difícil definição, decidiu-se

anotar como desvios os casos de colocação de vírgulas como o apresentado no exemplo (73). Tal situação ocorreu 23 vezes na amostra analisada, mas nem sempre na oração principal:

(74) Sabe-se que uma sociedade que clama por mudança <,> tem que primeiramente se ouvir para assim o " poder " os ouvir .

No exemplo (74), tem-se uma estrutura de subordinação ligada ao verbo principal *saber*, e uma relativa restritiva ligada ao núcleo do sujeito *sociedade*, que é dependente do verbo *ter*. Trata-se de uma relativa de apenas quatro *tokens*, o que coloca em dúvida novamente o que se caracteriza como "de certa extensão". Todavia, nota-se que a construção sintática da sentença como um todo é complexa, com variadas formas verbais (finitas ou não) e mais de uma relação de subordinação. Uma estrutura complexa também pode ser vista no exemplo a seguir:

(75) Para poder reduzir essa taxa de aumento séria necessário investimentos na área da educação , para aqueles **que tem menos condições** <,> não levarem o crime como uma opção , já que muitas crianças e adolescentes das periferias abandonam os estudos para entrarem no crime .

A sentença do exemplo (75) possui três estruturas de subordinadas reduzidas de infinitivo com valor de finalidade (sublinhadas), estando uma delas anteposta à oração principal. Além disso, há ainda uma subordinada introduzida pela locução conjuntiva *já que*. Dentro de uma das reduzidas há ainda uma relativa restritiva, marcada em negrito, que separa o núcleo *aqueles* do infinitivo flexionado (*levarem*) ao qual ele está ligado, e é justamente nessa estrutura que ocorre o desvio de pontuação. Há ainda outros desvios na sentença, como ausência de pontuação, excesso de acentuação e falta de concordância verbal, mas considerando a complexidade das relações estabelecidas e as distâncias entre os elementos dependentes, tal construção sintática é compreensível e relativamente bem-executada.

Também ocorreram problemas de concordância verbal nas construções em que o verbo está dentro de uma subordinada restritiva ou explicativa, e o seu sujeito é o elemento sendo explicado ou restrito:

(76) E isso só nos mostra cada vez mais o quão importante é o voto consciente pois estamos elegendo políticos que nos <represente> e faça o melhor pelo povo .

Um caso que merece destaque é o do exemplo abaixo:

(77) (...) , o erro , diversas vezes vem de corrupções , desvios de dinheiro que <deveriam ser> implantados em algo maior , em intolerâncias , entre outros .

No exemplo (77), a concordância foi feita com o núcleo nominal *desvios*, mas a subordinada restritiva na verdade está restringindo o complemento *dinheiro*. Nesse caso, cogita-se que o estabelecimento das dependências entre os elementos, especialmente em estruturas

oracionais complexas (como aquelas que contêm subordinação), pode ser um desafio para os produtores de textos. As análises linguísticas destacaram ainda outros desvios relacionados ao estabelecimento equivocado das dependências entre elementos, como problemas de concordância entre sujeito e verbo diante da presença de relativas dentro do sujeito:

(78) O embate que alguma dessas organizações traçam com governos e outros tipos de poderes legislativos <colocam> em duvida os direitos dos seres humanos e a liberdade de se expressarem através de protestos .

No exemplo (78), há ainda a questão de que as dependências se caracterizam como sendo de longa distância, ou seja, há vários elementos e dependências internas que se estabelecem entre o sujeito *embate* e o verbo *colocam*. Vê-se que há vários substantivos no plural que estão mais próximos na sentença, o que pode ter levado à ocorrência do desvio. Nesse sentido, seria interessante investigar como são estabelecidas as dependências sintáticas pelos produtores de textos em formação, a fim de verificar se há dificuldades no ensino ou na aprendizagem desse tipo de relação, ou se tais desvios estão ligados a uma distância maior entre os dependentes, comparando-se com aqueles que são adjacentes entre si.

Outra questão ligada às estruturas de subordinação são os desvios em termos de uso dos pronomes relativos. Há dois fenômenos comuns que dizem respeito a esse grupo de palavras. O primeiro deles é a alternância entre os diversos pronomes relativos, como *no/na qual* em vez de *cujo/cuja*. O segundo, ainda considerado desvio pelos mais conservadores, mas que já se impõe há algum tempo na modalidade escrita, nos mais diversos gêneros textuais, é o uso de *onde* sem referência a lugar, com papel de pronome relativo universal.

Entre o primeiro conjunto de desvios, as trocas entre os pronomes relativos foram as mais diversas, mas destaca-se o uso de pronomes variados no papel de *cujo/cuja*:

(79) O conceito social se volta á ação coletiva de um determinado grupo , <que> os objetivos são alcançar mudanças sociais .

Também foi frequente o uso de *no/na qual* no lugar de *que*:

(80) (…) movimentos sociais são reuniões de pessoas <na qual> se põe em movimento (…).

Já o uso de *onde* sem referência a lugar ocorreu 137 vezes no conjunto analisado, impondo-se como o mais frequente entre as ocorrências de um mesmo fenômeno. Em função da sua recorrência e do seu uso também em outros gêneros textuais, sugere-se que uma próxima tarefa de anotação não considere esse fenômeno como desvio, também de forma a não inserir nas anotações uma grande quantidade de ocorrências que provavelmente derivam de evoluções

da língua e que talvez nem possam mais ser consideradas desvio. Segue um exemplo para fins de ilustração do fenômeno, que é bastante sistemático na sua manifestação:

(81) Isso ocorreu no ano de 2013 , <onde> as manifestações cresceram através das mídias sociais (…).

Ainda no que se refere a pronomes relativos, identificou-se a ausência sistemática de preposição diante do relativo *que*, especialmente em construções com função adverbial:

(82) Os movimentos sociais no Brasil são historicamente reconhecidos e causam reverberação desde a época <que> houve a Independência do Brasil .

A inversão das sentenças por meio da construção de uma oração relativa pode justificar a ausência da preposição do exemplo (82). Porém, vê-se que o uso de preposições diante de relativos causa incertezas nos estudantes quando se identifica que o contrário também ocorreu:

(83) O mais conhecido no Brasil , seria o (MST) (…) , forma um coletivo e selecionam ideias <nas> quais são colocadas em formas de protestos , manifestações e greves .

Na sentença do exemplo (83), inseriu-se equivocadamente a preposição *em* diante do relativo em uma subordinada relativa na qual a forma verbal está na voz passiva. Nesse sentido, não apenas a inversão pode justificar o excesso de preposição, mas também a presença de passiva. Identificou-se, na amostra analisada, que é frequente a inserção equivocada da preposição *em* diante do pronome relativo *o qual/a qual*.

A ausência de preposição também ocorre na presença de *que* com papel de conjunção:

(84) Muitos deles acabam perdendo a sua essência pelo fato <que> as pessoas que participam dos mesmos vão até as ruas com um conceito totalmente adverso ao qual eles querem passar .

Caso a sentença do exemplo (84) fosse construída sem a conjunção, o substantivo *fato* estabeleceria relação de ligação com a oração seguinte por meio da preposição: "(…) *pelo fato de as pessoas participarem dos mesmos* (…)". Porém, não se poderia considerar este como um problema de regência nominal, uma vez que tal substantivo não exige esse complemento.

Outro aspecto ligado a pronomes diz respeito à sua colocação. Um dos critérios de colocação de pronomes pessoais átonos e do demonstrativo *o* se refere à orientação gramatical de que "não se pospõe, em geral, pronome átono a verbo flexionado em oração subordinada" (BECHARA, 2009, p. 588). Tal posição equivocada ocorreu 12 vezes, das quais nove estão relacionadas ao uso de pronome relativo, e apenas três ocorrem em casos de conjunção:

(85) A arte é um caminho para a conscientização , onde <pode-se> quebrar paradigmas (…).

(86) Todos esses movimentos são legais de acordo com o artigo 5º da Constituição Federal de 1998 , dito que <pode-se> reunir todos pacificamente , no entanto , sem armas e em locais abertos (…).

No exemplo (85), a colocação equivocada ocorre com o uso do *onde* sem referência a lugar, conforme o fenômeno descrito anteriormente nesta seção. Assim, vê-se que os desvios envolvendo subordinações são de diversos tipos e ocorrem com frequência, permitindo reforçar a hipótese de que se trata de um fenômeno complexo do ponto de vista da produção textual.

5.4.6 Voz passiva

A construção da voz passiva poderia ser considerada como uma inversão da ordem canônica, em termos semânticos, já que o objeto ou paciente da voz ativa se torna o sujeito na passiva, e o sujeito se torna objeto com função de agente da passiva. Todavia, em termos puramente sintáticos, a passiva em geral segue a ordem canônica dos elementos, com sujeito; forma verbal composta pelos verbos *ser*, *estar* ou *ficar* seguida de particípio; e eventualmente o agente da passiva (mas esse elemento não é obrigatório e em geral não é explicitado na sentença).

Durante as análises, identificou-se que a construção da voz passiva pode estar relacionada à ocorrência de tipos variados de desvios. O primeiro a ser analisado se refere à ausência de acento indicativo de crase em estruturas passivas:

(87) A arte é mostrada <a> população de diversas formas , porém ela tem uma grande repercussão quando ela desperta senso crítico e sensibiliza aquele que está vendo e (…) .

A estrutura argumental do verbo *mostrar* exige um complemento direto, indicando *o que é mostrado*, e um complemento indireto, introduzido por preposição, indicando *a quem se mostra*. Dessa forma, na transformação para a voz passiva, por questões de regência do verbo, exige-se a colocação de crase no complemento ligado ao particípio. Esse tipo de desvio não foi muito frequente, ocorrendo apenas cinco vezes na amostra analisada.

Um desvio um pouco mais frequente relativo à voz passiva foi a falta de concordância do verbo *ser*, que em geral acarretou problemas de concordância também no particípio:

(88) Os movimentos sociais no Brasil <é marcado> por oposições e revoltas , exercido contra os governos opressores e lutando pela democracia e liberdade .

No caso do exemplo (88), a concordância estaria adequada se verbo e particípio estivessem em suas formas plurais. Porém, o gênero do particípio concorda com o gênero do sujeito da passiva. Os casos em que os elementos concordam em gênero, mas não em número, e em que as construções verbais estão no singular, mas o sujeito está no plural, foram os mais frequentes dentro da amostra analisada, ocorrendo também com sujeitos femininos:

(89) (…) seria a melhor forma de organizar e protestar de forma pacífica onde as ideias <seria exposta> de forma clara e prudente , causando que ambas das partes <possa ser ouvida e entendida> .

No exemplo (89), vê-se que há dois equívocos de concordância em estruturas passivas no feminino que concordam em gênero, mas não em número. Além disso, na segunda construção, em que também há a coordenação entre dois particípios, destaca-se a presença do verbo *poder* como modalizador da estrutura passiva, que apresenta o mesmo problema de concordância. Merece destaque o caso ilustrado pelo exemplo a seguir:

(90) (…) e por isso pode se tornar um opressor , partindo do conceito de que <foram ameaçados> .

O exemplo (90) mostra dois aspectos particulares: a distância entre o sujeito (*opressor*) e a estrutura passiva; e o fato de que, semanticamente, tal sujeito carrega a ideia de grupo, referindo-se aos membros do movimento social que podem se tornar opressores enquanto grupo. A distância entre a palavra que faz o papel de sujeito da passiva e a estrutura verbal em si, somada ao fato de que ambos os elementos se encontram em orações diferentes, poderia justificar por si só a dificuldade de realização da concordância adequada. No entanto, como se discutirá mais adiante neste capítulo, a concordância semântica entre conceitos que expressam sentido de grupo se fez presente em outros tipos de desvios, reforçando a hipótese de que há forte influência da semântica nos aspectos relacionados à concordância. Isso evidencia mais uma vez as limitações impostas pela tentativa de analisar um nível linguístico isolado.

As análises identificaram 27 ocorrências em que o desvio envolveu apenas o particípio. Destas, sete foram problemas de concordância somente de número, 16 não concordaram em gênero, e quatro apresentaram falta de concordância de gênero e de número:

(91) Algumas soluções e decisões para a precariedade do sistema carcerário brasileiro poderiam ser <feita> , como por exemplos .

(92) Finalizando , o movimento social é <representada> por todos aqueles que buscam mudança (…) .

(93) (…) uma lei que decretasse que , a ação dos movimentos radicais , fossem <multados> e (…) .

No exemplo (91), ainda que haja uma dependência com uma distância significativa entre o sujeito da passiva (sublinhado, em que há também uma estrutura de coordenação), vê-se que o verbo modal *poder* está no plural, mas o mesmo não ocorre com o particípio da passiva. Além disso, a sentença termina abruptamente, caracterizando-se como incompleta. O exemplo (92) mostra a ausência de concordância de gênero. Por fim, o exemplo (93) ilustra a ausência de concordância de gênero e número entre sujeito e estrutura verbal de passiva, mas ressalta-se o fato de que o particípio concorda com o núcleo nominal que está mais próximo (*movimentos*

radicais), o que reforça a ideia de que a distância entre os elementos que estabelecem relações de dependência tem influência na ocorrência de desvios.

Ainda na concordância de passivas, identificaram-se problemas em subordinadas reduzidas de particípio que, quando desenvolvidas, referem-se a estruturas passivas:

(94) (…) os temas abordados são a violência , respeito , estupros , diferença salarial entre os gêneros , poucos cargos políticos <preenchido> por mulheres e o preconceito contra a mulher .

No exemplo (94), a respectiva versão desenvolvida da construção, mantendo-se a concordância adequada, seria *cargos públicos que são preenchidos por mulheres*. Assim, exige-se a concordância de número entre o sujeito da passiva (*cargos*) e o particípio. Na sua versão reduzida, ainda que o auxiliar de passiva não seja realizado, mantém-se a mesma regra: o particípio deve concordar com o elemento ao qual se refere. Os problemas de concordância envolveram tanto gênero quanto número, e se percebeu que frequentemente o particípio concorda com o termo posterior, e não com aquele com o qual estabelece dependência. No exemplo (95), o particípio concorda com o núcleo nominal seguinte (a palavra *impostos*), e não com o elemento dependente *dinheiro*:

(95) (…) esta gerando o dinheiro <pagos> em impostos para conseguir manter o sistema carcerário , ou até mesmo para próprios fins , como já foi ocorrido .

Por fim, houve três casos ligados à escolha da forma de particípio (alguns verbos permitem duas formas, como *pago* e *pagado*). Em uma das sentenças, ocorreu a passiva *é dizido*, que não permite tal forma de particípio. Nos outros dois casos, foi utilizado o particípio *tragos* (*vários temas estão sendo debatidos e <tragos> à luz*), que também não existe.

5.4.7 Segmentação de sentenças e uso de pontuação

Esta seção descreve os fenômenos identificados na ocorrência de desvios de segmentação intra ou suprassentencial. Analisam-se inicialmente os casos em que uma sentença teve os seus elementos internos segmentados por sinais de pontuação equivocados, como aglutinação de várias orações separadas por vírgulas em uma mesma sentença, mas também os casos que envolvem, por exemplo, a segmentação por vírgulas da continuidade entre sujeito e verbo sem elementos intercalados. Na sequência, descrevem-se aqueles casos em que uma sentença que deveria estar conectada com a anterior ou a posterior foi segmentada com ponto final.

Inicia-se a análise descrevendo os problemas relacionados à segmentação que ocorre internamente em uma sentença. O primeiro fenômeno, que se mostrou frequente, mas cujo

estabelecimento de regras claras de anotação que não envolvessem questões de estilística se mostrou desafiador, foi o uso de vírgulas onde deveriam ocorrer pontos finais. Nesse caso, estabeleceu-se que tal desvio só deveria ser anotado quando houvesse ruptura clara do foco temático entre as sentenças, e em que todas as sentenças resultantes fossem completas (isto é, não deveriam ser segmentadas sentenças que tivessem apenas subordinadas, por exemplo). Dentro do conjunto de sentenças anotadas como desvio de uso de pontuação, tal desvio ocorreu 205 vezes, caracterizando-se como o mais frequente entre todos os fenômenos analisados. O exemplo a seguir mostra um caso típico de sentença com um número muito elevado de *tokens* na qual se podem identificar diversas sentenças completas aglutinadas:

(96) $_{1}$[Hoje em dia quem batalha mais pelos direitos do povo são os movimentos sindicais que estão mais presentes nos protestos contra o governo e contra as leis que eles querem implantar] <,> $_{2}$[a mídia não dá muita atenção a esses movimentos , só vai quem acompanha pelas redes sociais ou que já estão no movimento a muito tempo , pois não tem muita visibilidade na TV ou em outras mídias] <,> $_{3}$[mesmo não tendo a atenção da mídia os sindicalistas estão sempre batendo de frente com o governo , como nos últimos dias que fizeram greves gerais e passeatas contra a reforma da previdência] <,> $_{4}$[Grupos populares se juntam em redes sociais e marcam dias mais acessíveis para que muitos possam comparecer no protesto e mostrar para o governo o que o povo quer] <,> $_{5}$[alguns meses atrás podemos ver o que os movimentos populares podem fazer , eles pararam o Brasil por estarem insatisfeitos com a ex-presidente Dilma , e conseguiram então o impeachment da presidenta].

A sentença do exemplo (96) é composta por 169 palavras, e a aglutinação de sentenças torna difícil a sua compreensão. A redação na qual essa sentença se insere é composta por apenas três sentenças: uma corresponde ao parágrafo de introdução; uma ao parágrafo de desenvolvimento; e uma ao parágrafo de conclusão. Isso pode evidenciar um desconhecimento da função dos sinais de pontuação para a estruturação das sentenças em um texto. Ao analisar o exemplo mais atentamente, vê-se que os desvios sintáticos não se limitam ao uso de vírgulas onde deveriam ocorrer pontos finais: há ausência de outras vírgulas dentro das sentenças, há desvios de concordância, entre outros. Em termos de sentenças aglutinadas, identificou-se a divisão em cinco sentenças internas que foram consideradas mais evidentes (tal divisão pode ser feita de diferentes formas, sendo esta uma proposta possível), marcadas entre colchetes e numeradas para facilitar a identificação. Destaca-se que a sentença 4 inicia-se por maiúscula após a vírgula, o que poderia ser mais um indício de que ali se pretendia de fato iniciar uma nova sentença. Essa mesma questão do uso de maiúsculas foi percebida em outras sentenças com desvios de segmentação.

Todavia, há casos em que a segmentação de sentenças é ainda menos clara, uma vez que é difícil estabelecer o limite entre elas:

(97) $_{1}$[Movimentos Sociais são constituídos por movimentos populares , sindicais , e ONGs] <,> $_{2}$[as ONGs são movimentos sociais , cada uma com o seu objetivo assim como o Criança Esperança ajuda crianças sem suporte familiar ou foram deixadas pelas famílias nas ruas a AACD também é uma ONG que ajuda crianças com doenças físicas ou mentais com ajuda de doações de pessoas , essas são ONGs que a muito tempo vem ajudando pessoas com suas dificuldades do dia a dia] <,> $_{3}$[mas os movimentos não só rosas , grupos como o KKK , Talibã e o Nazismo foram originalizados como movimentos sociais mas ai no final todo mundo sabe aonde essa história terminou , a começar pelo Nazismo que matou milhões de judeus entre outros que eles consideravam diferentes] <,> $_{4}$[o KKK é a junção de muitos Movimentos com ideologias quase iguais , eles também são considerados muito violentos e assassinos] .

A sentença do exemplo (97) tem 153 palavras; porém, na sua estrutura, é menos clara a segmentação interna em sentenças completas do que no exemplo anterior. Aqui também as maiúsculas são usadas aparentemente sem critério específico, e por isso não podem ser utilizadas como evidência para eventuais tentativas de segmentação. Novamente, identifica-se que os desvios sintáticos vão além da segmentação das sentenças, inclusive dificultando em alguns pontos a compreensão da mensagem. A hipótese de segmentação estabelecida divide a sentença em quatro, mas se assume que tal segmentação é questionável. Vale lembrar que se consideram possibilidades de segmentação desse tipo apenas quando há a presença da vírgula, mesmo que haja outros pontos de segmentação nos quais não ocorre sinal de pontuação.

Ainda no que se refere à segmentação equivocada dentro da sentença, outro desvio frequente foi o uso de pontuação segmentando elementos contínuos. O fenômeno mais comum foi a segmentação de sujeito e verbo com vírgula dentro da oração principal. Como uma das hipóteses era a correlação entre dependências de longa distância e a presença de desvios, as análises subdividiram esse fenômeno em dois: separação de sujeito e verbo com vírgula quando a estrutura do sujeito tinha menos de cinco *tokens*; e separação quando o sujeito tinha mais de cinco *tokens*. O primeiro caso ocorreu 33 vezes; o segundo, 20 vezes. Logo, o número maior ou menor de elementos dentro do sujeito não parece ter influência na ocorrência de vírgulas. Casos típicos de segmentação de sujeitos com menos de cinco *tokens* são mostrados a seguir:

(98) Essa influência <,> se torna perigosa , quando a vítima se transforme em um opressor .

(99) Em suma , os movimentos sociais <,> são uma grande arma contra o governo .

(100) No Brasil , na década de 1950 , o aumento da globalização e da população urbana <,> ocasionou no crescimento da visibilidade de movimentos e organizações sociais , principalmente (…) .

O tamanho da sentença também não parece estar relacionado à colocação da vírgula entre sujeito e verbo, visto que os exemplos (98) e (99) consistem em sentenças curtas, enquanto o exemplo (100) engloba um número maior de *tokens*. No primeiro exemplo, não há elementos antes do sujeito; no exemplo (99), a sentença vem precedida pela expressão *em suma*; e o exemplo (100) vem precedido por duas estruturas deslocadas com função adverbial. Assim,

mesmo analisando sistematicamente diversas outras ocorrências desse fenômeno, não foi possível estabelecer correlações entre a sua ocorrência e uma característica linguística específica. Cogita-se, portanto, que os estudantes têm dificuldades na compreensão das funções da pontuação, em especial da vírgula, independentemente das estruturas sintáticas pretendidas.

Outro caso menos frequente de segmentação foi o uso de vírgula para separar o verbo de seus objetos ou objeto direto e indireto:

(101) Tornar tais grupos legítimos , portanto , é dar às comunidades <,> a chance de mostrar a sua voz e a sua opinião sobre determinado assunto .

(102) A história do Brasil é marcada <,> por diversos movimentos sociais que causaram impacto nos dias atuais , dando ênfase no movimento " Diretas Já " , que ocorreu na época da ditadura militar (…) .

(103) Com tudo , o governo poderia ter flexibilidade (…) ou então tentar explicar o porque ele não pode <,> atender as exigências pedidas , naquele momento pelos movimentos sociais .

No exemplo (101), ocorre a inversão entre o objeto direto e o indireto, provavelmente o primeiro é muito mais longo que o segundo, e ambos foram segmentados por vírgula. Já no exemplo (102), a forma verbal utilizada aparece na voz passiva, que, como visto na Seção 5.4.6 (p. 109), esteve relacionada à presença de outros tipos de desvios. Por fim, o exemplo (103) ilustra a segmentação com vírgula da locução verbal *poder atender*. Desvios em locuções verbais se fizeram presentes em outros fenômenos, como se verá na Seção 5.5.2 (p. 130).

Houve ocorrências ligadas ainda a vírgulas após expressões como *visto que*, *de modo/forma/maneira que*, *desde que*, entre outras. Por fim, citam-se ainda poucas ocorrências de vírgula inadequada após pronome relativo e após preposição. Novamente, ressalta-se que a quantidade e as características diversas dos desvios de pontuação tornam a análise de correlações entre fenômenos um desafio. Cogita-se, mais uma vez, que os desvios desse tipo estejam associados ao fato de que a pontuação tem uso restrito à escrita e de que os estudantes só aprendem as suas funções durante a trajetória escolar. Ao que parece, é justamente a compreensão da função da pontuação que influencia a ocorrência de desvios, mais do que qualquer característica linguística particular.

Quando os problemas de segmentação vão além do nível da sentença, em geral não é possível identificar com certeza a qual das sentenças aquela que é foco de análise deveria estar conectada, pois a análise se restringe às partes segmentadas com ponto final. Assim, por exemplo, se uma sentença segmentada equivocadamente devia estar conectada à anterior, isso significa que o desvio de pontuação ocorreu naquela que a antecede. Logo, tais sentenças foram anotadas como um todo, e não em um *token* específico. Também é interessante notar que a tipologia contava com uma categoria específica para problemas de segmentação desse segundo

tipo, mas, na sua essência, trata-se de uma ocorrência particular de desvio de pontuação. Cabe avaliar então, para tarefas de anotação futuras que pretendam usar a mesma tipologia, a possiblidade de mover essa categoria como uma subcategoria de pontuação.

Nesse fenômeno, identificou-se a ocorrência frequente de sentenças em que a oração principal ou a sentença toda continha apenas formas verbais infinitas. Como se verá na Seção 5.5.1 (p. 126), essas formas verbais (também chamadas de formas nominais, porque podem assumir papel de nomes) estiveram envolvidas em diversas categorias de desvios. Quando havia um verbo conjugado em orações subordinadas, mas não na oração principal, a análise mostrou a prevalência de sentenças com gerúndio (50 ocorrências) e infinitivo (9 ocorrências) ocupando a função de verbo principal:

(104) <Mostrando> que a vida com o crime é dura , ruim , não é livre e leva muitas vezes a morte dos criminosos ou de familiares queridos .

(105) Assim , <reconhecer> que para ser um artista , basta amar a arte e respeitar a cultura que ela carrega.

Em ambos os exemplos, tem-se uma forma verbal infinita iniciando a sentença, seguida de uma subordinada com verbos finitos. No exemplo (104), a sentença parece apresentar uma justificativa ou explicação de algum elemento que estava na sentença anterior, indicando que esta foi segmentada inadequadamente. Já no exemplo (105), a sentença parece fazer parte de uma lista de propostas de intervenção para o problema proposto, conforme é solicitado pelo ENEM. Na maioria massiva dos desvios desse tipo envolvendo gerúndio, a forma verbal infinita foi o primeiro elemento a ocorrer na sentença. Em termos de infinitivo, as ocorrências foram bem menos numerosas, mas em três delas havia o sujeito precedido da forma verbal infinita. No exemplo (106), ocorreu ainda uma estrutura apositiva não marcada por vírgulas dentro do sujeito, que está sublinhado na sentença para fins de clareza:

(106) O Ministério da Educação e da Cultura junto as mídias e por meio de propagandas <conscientizar> os indivíduos sobre a importância organização na luta por seus direitos , e não haja conflitos e nenhum movimento fique manchado .

Analisando as sentenças que continham apenas uma das três formas verbais infinitas (uma ou mais do mesmo tipo na sentença), novamente o gerúndio foi o mais frequente, com 28 ocorrências. O número de ocorrências com o infinitivo foi bem inferior (sete sentenças), e houve apenas dois exemplos só com particípios. Seguem exemplos ilustrativos:

(107) Mudando a penalidade no Brasil .

(108) Ter ações comunitárias dentro das penitenciarias , investir na educação , ter incentivo dentro das escolas , propagandas nas televisões .

(109) Realidade relatada diariamente em jornais e revistas .

Enquanto a maior parte das sentenças ligadas ao uso de gerúndios é consideravelmente curta, as sentenças com infinitivo se caracterizam como listas de ações com duas ou mais dessas formas verbais infinitas, sem a presença de sujeitos dos infinitivos. As duas sentenças com uso de particípio (uma delas mostrada no exemplo (109)) tinham estrutura e tamanho similares. Nota-se também que sentenças como as dos exemplos (107) e (108) parecem compor a parte final do texto e com frequência parecem fazer parte das propostas de intervenção.

Outro fenômeno identificado foram as sentenças com ausência total de formas verbais:

(110) Um dos problemas mais visíveis e sem solução .

(111) Outros , porém , com a expansão do processo de globalização .

Essas ocorrências têm diversas formas e estruturas, mas em geral são sentenças curtas, e algumas delas se parecem com apostos explicativos referentes a elementos da sentença anterior ou à sentença como um todo, como o caso do exemplo (109). O exemplo (111), por sua vez, parece uma sequência da sentença anterior.

Foi possível identificar 61 sentenças iniciando por conjunções ou locuções conjuntivas de coordenação que, segundo a gramática padrão, não devem assumir posição inicial. Um exemplo muito comum foram as sentenças iniciadas pela conjunção *e*. Esses casos foram anotados como desvios, mas tais sentenças ocorrem em outros gêneros textuais, muitas vezes utilizadas como recursos estilísticos. Também ocorreram sentenças começando por *pois, ou, bem como, assim como, como também, além de* (seguido de infinitivo).

Entre as conjunções e locuções conjuntivas de subordinação iniciando sentenças (sem oração principal posposta), as ocorrências foram variadas. Os exemplos mais comuns envolvem locuções conjuntivas formadas pela conjunção *que*, como *visto que, já que, sendo que, para que, de forma/modo que*, o próprio *que*. Ocorreram ainda sentenças iniciando por *devido a, porque, onde*, entre outras. Outro grupo de expressões identificado no início de sentenças e que foi marcado como desvio de segmentação foram as explicativas como *isto é, ou seja* e (*como*) *por exemplo*. O último grupo de palavras identificado no início de sentenças com problemas de segmentação foram as preposições. Esse fenômeno ocorreu 16 vezes e, em todas elas, as sentenças geradas não possuíam oração principal e eram claramente parte da sentença anterior ou posterior. Entre as preposições mais comuns, estão *com*, *para* e *através de*.

Por fim, o último fenômeno analisado em termos de segmentação suprassentencial foram as sentenças em que se identificaram claramente partes faltantes, tendo como justificativa provável a desatenção ou falta de revisão do texto. Também não é impossível que tais sentenças

resultem de problemas na plataforma na qual os textos foram escritos ou de eventuais manipulações dos textos antes do fornecimento do *corpus* para esta pesquisa. Durante a etapa de anotação, tais sentenças foram buscadas no *corpus* original, a fim de verificar se houve erros nas manipulações de arquivos feitas durante o pré-processamento do *corpus*, mas os problemas já constavam nos textos originais. Dois exemplos desse desvio podem ser vistos a seguir:

(112) Com o objetivo que ocorra uma transformação social , a cultura e a arte são as principais ferramentas para o país evoluir como sociedade e <assegurarmos> .

(113) É muito comum ver mensagens que a mídias podem passar sobre a realidade em que convivemos , (…) que por sinal , pode causar um impacto no senso crítico <do> .

Os diversos tipos de fenômenos relacionados à segmentação de sentenças sugerem que essa questão seja problemática aos produtores de textos. Da mesma forma, acredita-se que ela seja desafiadora também para ferramentas que se proponham a lidar com redações de estudantes, visto que muitos fenômenos são difíceis de serem mapeados por meio de regras explícitas. Uma possibilidade para aqueles desvios que vão além do nível da sentença seria inserir nos dados uma segunda camada de anotação, em nível textual, que desse conta de marcar a relação estabelecida entre duas sentenças que foram segmentadas. Também poderia ser útil, em ferramentas de auxílio à escrita e corretores gramaticais, identificar automaticamente a ausência de oração principal ou mesmo de formas verbais (finitas ou infinitas), de modo a alertar o produtor do texto quanto a um possível problema de segmentação.

5.4.8 Sintaxe de regência: demais fenômenos relevantes

Regência é um conceito que se refere à relação sintático-semântica que se estabelece entre os termos ditos regentes (o verbo na regência verbal; o substantivo, adjetivo ou advérbio na regência nominal) e aqueles que estabelecem uma relação de dependência com eles, isto é, os termos regidos. Alguns dos casos de regência são abordados em outras seções, já que estão ligados a fenômenos variados de ocorrência de desvios. Esta seção aborda aqueles fenômenos específicos da sintaxe de regência que não foram inseridos em outras seções. Inicia-se analisando os casos de regência nominal e, na sequência, descrevem-se os de regência verbal.

A primeira questão de regência nominal foi a ausência da preposição exigida por substantivos como termo regente. Entre esses desvios, 26 exemplos foram de ausência de crase[55]. Desconsiderando os casos de crase antes de substantivos em construções com verbo-

[55] Ainda que não seja possível saber, em ausência de crase, se o elemento faltante é a preposição ou o artigo, decidiu-se analisar o fenômeno como um problema de regência, para facilitar a sistematização desse tipo de desvio.

suporte, que são analisados na Seção 5.6.1 (p. 133), a maior parte dos substantivos regentes ocorreu apenas uma ou duas vezes, com exceção de *(re)integração*, com quatro ocorrências:

(114) Sabe-se que a finalidade de nossos presídios é a reintegração do preso <a> sociedade (…)

Nos fenômenos ligados à regência nominal de substantivos que não envolviam crase, foi identificado apenas um caso de ausência da preposição, que tem uma estrutura particular:

(115) Pensando nessas em essas propostas ainda há <esperança> que esse número diminua .

No exemplo (115), está ausente a preposição *em* ou *de*, exigida pelo substantivo *esperança*; porém, a estrutura que segue é uma subordinada introduzida pela conjunção *que*. Como já se viu anteriormente, a construção das subordinadas está associada ocorrências diversas de desvios; logo, pode-se justificar a ausência da preposição aqui também em função da subordinada. Isso leva a crer que, enquanto os estudantes parecem apresentar incertezas diversas quanto ao uso da crase (como também é demonstrado nos casos de excesso de crase, analisados na Seção 5.8.2, p. 147), os problemas de ausência de preposição em casos de regência que não a envolvem podem estar ligados a outros tipos de fenômenos, como a presença de subordinação.

Ainda no que se refere à regência nominal em substantivos, ocorreram alguns casos em que se utilizou a preposição equivocada para construir a relação de regência:

(116) O que estes movimentos buscavam é a conquista <por> direitos e acabar com a repressão (…) .

Em relação à ausência de crase em regência nominal em que o termo regente não era um substantivo, houve 24 ocorrências de falta de crase em estruturas regidas por adjetivos, e nenhum desvio ligado a advérbios. No entanto, destaca-se uma particularidade entre aqueles termos identificados como adjetivos: cinco ocorrências estavam ligadas a formas que, segundo consulta no Dicionário On-Line Priberam da Língua Portuguesa[56], funcionam apenas como particípios (duas de *direcionado* e três de *relacionado*). Das demais ocorrências, 11 casos se referiam a termos que podem ter função tanto de adjetivo quanto de particípio, conforme consulta no mesmo dicionário citado acima, entre as quais sete se referem ao termo *ligado*. As outras ocorrências envolvem os termos *chegado*, *preso* e *servido*. Entre as sentenças em que os termos regentes eram adjetivos que funcionam como particípio, duas envolveram o termo *contrário*, e as outras foram de *atento*, *beneficente*, *benéfico*, *irrelevante*, *prejudicial*, *vantajoso*.

[56] Disponível em https://dicionario.priberam.org/

As questões de regência verbal envolveram fenômenos muito mais variados do que as de regência nominal. Em relação a esse tipo de regência, a primeira questão abordada se refere à ausência de preposição obrigatória para fins de regência verbal, iniciando-se pelos desvios de crase. Nesse aspecto, identificou-se que diversas das ocorrências continham o verbo *ir*:

(117) E outros tipos de movimentos sociais , as mulheres vão <as> ruas pelos direitos iguais , (…).

A recorrência desse desvio provavelmente se justifica em função de um dos temas mais frequentes do conjunto de redações que compõem o *corpus* ser "a legitimidade dos movimentos sociais". Assim, há vários desvios de ausência de crase ligados à expressão *ir às ruas*, como ilustrado pelo exemplo (117): das 28 ocorrências de ausência de crase com esse verbo, 27 se referem a essa expressão e uma envolve *ir à procura*. Logo, é importante considerar que a temática da redação também pode ser um fator que influencia o tipo de construção mais comum do *corpus*, bem como os desvios mais recorrentes.

Houve ainda várias ausências de crase para fins de regência ligadas a outros verbos, sendo o mais comum deles o verbo *levar*:

(118) No Brasil o Conselho Nacional de Justiça (CNJ) aponta com um dos motivos da superlotação , as prisões em flagrante que levam <a> prisão provisória .

No exemplo (119), o verbo que exige preposição se refere a dois objetos (um direto e outro indireto), e o objeto direto está intercalado ao verbo regente e ao termo regido:

(119) Portanto a atitude de reverter esse quadro tem que partir do governo , tendo o interesse de querer ressocializar essas pessoas <a> sociedade .

Os casos que envolveram a ausência de outra preposição também foram frequentes, tendo ocorrido 30 vezes. A maior parte dos verbos ocorreu apenas uma vez, mas alguns se repetiram, como *assistir*, *conscientizar*, *prezar* e *usufruir*. Algumas ocorrências são particularmente interessantes:

(120) O Brasil é considerado um país da diversidade de raças , e isso <vem> da colonização dos portugueses e a ordem e importância das imigrações que ancoraram suas embarcações no país .

No exemplo (120), há uma estrutura de coordenação em que, no primeiro elemento, foi utilizada a preposição exigida pelo termo regente, mas no segundo não. Por questões de paralelismo sintático e para evitar ambiguidades, a preposição é obrigatória nesse caso. Cogita-se que a distância entre termo regente e termo regido, além da presença da coordenação, possa estar associada à ocorrência do desvio. A sentença a seguir também traz um caso particular:

(121) (…) temos que nos mobilizar e <falar> com nossos amigos e familiares quais são mesmo os legítimos movimentos sociais na nossa cidade ou estado , (…) .

No exemplo (121), o verbo exige dois complementos preposicionados (*falar sobre algo com alguém*), que aparecem na ordem inversa na oração. Além disso, o segundo complemento é uma subordinada, o que pode influenciar a ausência da preposição, além da distância entre termo regido e termo regente. O último exemplo referente a esse fenômeno é visto a seguir:

(122) Certamente os grupos surgem quando a sociedade civil <se dá conta> que não pode aceitar uma situação de forma passiva .

Esse caso merece destaque por duas particularidades: a primeira é a forma do termo regente em si, a segunda é a presença de oração subordinada como complemento do termo regente. A expressão que exerce a função de regente é uma expressão multipalavra idiomática, já que o sentido de cada parte não corresponde ao sentido do todo e que ambos os elementos ocorrem como uma expressão fixa, não permitindo inversões como **a conta que a sociedade se dá*. Logo, o termo regente é a expressão inteira *dar-se conta*, que exige a preposição *de*. Ademais, o complemento da expressão multipalavra é uma subordinada introduzida pela conjunção *que*, o que também pode ter influência na ocorrência do desvio.

Em termos de uso excessivo de preposição junto a verbos que são transitivos diretos, os casos de excesso de crase foram bem menos frequentes do que a ausência desse sinal:

(123) É preciso sempre buscar a melhor forma de manifestar e expor <ás> ideias .

Entre os casos de ausência de preposição que não envolvem crase, percebeu-se a recorrência de usos excessivos da mesma preposição junto a verbos transitivos diretos que são semanticamente próximos. O primeiro exemplo envolve os verbos *afetar* e *prejudicar*:

(124) (…) a sociedade é feita a base da educação e isso afeta diretamente <no> sistema carcerário .

(125) (…) que , prejudica <no> crescimento intelectual e <no> desenvolvimento do pensamento (…).

No exemplo (124), há um advérbio intercalado entre o verbo e a preposição, mas houve sentenças em que ambos os elementos eram contínuos. Já no exemplo (125), há dois complementos coordenados do verbo *prejudicar* que utilizam a preposição *em*, estando o primeiro contínuo ao verbo. Um caso semelhante ocorreu com os verbos *ocasionar* e *acarretar*, com os quais também foi utilizada a preposição *em*:

(126) (…) que ocasiona distorção do real propósito da ação .

(127) Isso acarreta causas que muitas vezes fariam ia diferença para a sociedade (…) .

Outro verbo em que se identificou o excesso de preposição foi *utilizar*:

(128) (…) os ancestrais da época utilizavam <de> substâncias com algumas tonalidades de cor (…) .

Nesse caso, a provável justificativa para o uso da preposição é a existência da variação *utilizar-se*, que exige a preposição *de*. No entanto, sem o reflexivo, o verbo *utilizar* se torna transitivo direto. Com os demais verbos, também foi frequente o uso excessivo de preposição: houve 40 ocorrências desse fenômeno na amostra analisada. Novamente, a maior parte dos verbos ocorreu apenas uma vez, com preposições variadas, exceto *alienar*, *esquecer*, *evitar* e *tornar*, que ocorreram de duas a três vezes cada.

A última questão relacionada à regência verbal diz respeito ao uso da preposição equivocada. Entre os verbos com maior número de ocorrências, destaca-se o *chegar*:

(129) (…) de forma passiva e educacional , pois assim é mais fácil chegar <no> objetivo .

O uso da preposição *em* com o verbo *chegar* é bastante comum, assim como ocorre com o verbo *ir*. Nesse sentido, é possível que, em breve, tal uso já não deva mais ser considerado desvio de regência, impondo-se também nos usos mais formais da língua escrita, como prova da evolução linguística. A maior parte dos demais verbos também ocorreu apenas uma vez, exceto *contribuir*, *refletir*, *reintegrar*, *resultar*, *limitar-se*, *resumir-se* (duas ocorrências cada), *visar* (três ocorrências), *sair* (quatro ocorrências). Cabe ressaltar ainda que, entre as sentenças analisadas, nove delas envolvem verbos reflexivos ou recíprocos:

(130) (…), e isso não se deve apenas <pelo> aumento da criminalidade , mas sim a fatores que (…).

(131) O povo vem revelando seu poder a cada dia , se opondo <contra> o governo tão intolerante (…).

Um exemplo a ser destacado é visto a seguir:

(132) Algumas das escolhas que foram tomadas no passado ainda remetem <nos> dias de hoje (…) .

No exemplo (132), há um problema ligado à regência do verbo *remeter*. Porém, ao analisar a sentença do ponto de vista semântico, levanta-se a hipótese de que o verbo pretendido é *refletir*, para o qual a preposição estaria correta. Apesar das percepções e intuições, pode-se apenas criar hipóteses sobre possíveis equívocos de escolha lexical.

5.4.9 Sintaxe de concordância: demais fenômenos relevantes

Entre os desvios mais frequentes, aqueles relacionados à concordância verbal e nominal permitiram análises interessantes no que se refere aos fenômenos linguísticos envolvidos. Esta

seção sistematiza as questões de concordância que não foram descritas nas demais seções. A análise inicia com os desvios ligados à concordância nominal e, em seguida, aborda a concordância verbal.

Uma das hipóteses relacionadas aos desvios de concordância é que poderia haver influência entre dependências de longa distância e presença de desvios de concordância. Ainda que em algumas circunstâncias essa hipótese tenha sido parcialmente validada, como se verá ao longo desta seção, os desvios de concordância são muito frequentes também entre elementos adjacentes. O fenômeno da ausência de concordância nominal entre elementos adjacentes foi bastante recorrente no conjunto de sentenças analisadas: 44 ocorrências envolvendo nomes e modificadores adjacentes, e 47 envolvendo determinantes e nomes adjacentes. A maior parte desses casos envolveu problemas de concordância apenas em número, mantendo-se o gênero adequado:

(133) (…) as formas de protestos e a busca de nossos direitos que chamamos de movimento <sociais> .

(134) (…) foi gerado uma série de necessidades da população , as atividades produtivas <na> cidades grandes obtiveram um aumento de urbanos desordenado .

Para fins de padronização da anotação, estabeleceu-se que o nome carregaria a informação pretendida, e os elementos à sua volta seriam anotados como contendo desvio de concordância. Para tarefas de anotação futuras que utilizem as mesmas diretrizes, pode-se pensar em aprofundar essa questão.

(135) (…) sendo assim , as manifestações feitas de forma <passivas> devem ser legitimadas (…) .

No exemplo (135), vê-se um caso em que todos os demais elementos da sentença estão no plural, exceto aquele que estabelece a relação de dependência com o modificador (*de forma passiva*). Pode haver dificuldade dos estudantes na identificação e no estabelecimento do núcleo da dependência, de forma que seja possível realizar a concordância adequada. Logo, caberia avaliar se o uso de ferramentas como *parsers* por dependência poderia auxiliar nesse aspecto durante o processo de formação de produtores de textos.

Na maior parte dos casos de ausência de concordância de número entre nome e determinante ou modificador, foi o nome que apareceu no plural, e o determinante/modificador ocorreu no singular. Isso vai ao encontro de uma tendência da modalidade falada de não realizar a marcação redundante de plural em todos os elementos dependentes, realizando-a apenas no determinante ou no nome.

Em alguns dos casos analisados, o modificador estabelecia a relação de dependência por meio de preposição:

(136) O movimento social (…) são grupos de <pessoa> que se manifestam para adquirir seus direitos e deveres , buscando reconhecimento , igualdade , justiça , honestidade , etc.

Ocorreram alguns casos, também, em que o modificador foi antecedido ao nome:

(137) No entanto , esse movimento conseguiu <importantes> conquista de cidadania ao longo da historia brasileira , como os direitos de terra , reconhecimento na sociedade .

As sentenças em que houve ausência de concordância apenas de gênero entre nome e determinante ou modificador foram bem menos frequentes. Apenas cinco casos de ausência de concordância entre nome e modificador foram identificados, e em duas delas o desvio parece ter sido ocasionado por um desvio ortográfico de troca entre os termos *sociais* e *sócias*.

Entre os casos envolvendo determinantes e nomes, houve 14 ocorrências de ausência de concordância de gênero, envolvendo diversos tipos de determinantes, como *todo(s)*, *este(s)/esse(s)*, *um/uma*, *o/a*, *outro/outra*:

(138) Com isso , podemos concluir que a manifestação deve sim acontecer , pois é <um> forma do direito que está em luta ser visto pelo governo .

(139) (…) pois sem estás formas de contenções para à sociedade formas diversas para fazer <estes> abordagens em geral de todos os movimentos .

Em apenas um caso, ocorreu a ausência de concordância tanto de gênero quanto de número entre nome e modificador (no trecho *precariedades nos complexos <carcerária>*), não havendo nenhuma sentença em que isso ocorreu entre determinante e nome. A quantidade e os tipos de desvios relacionados à ausência de concordância entre nome e modificador podem indicar a dificuldade dos produtores de textos de compreenderem as relações de dependência, mas sobretudo parece dar indícios da ausência de revisão dos textos escritos.

Outro fenômeno bastante frequente em que houve problemas de concordância nominal foram as sentenças envolvendo verbos considerados como tendo função de ligação (verbos copulativos ou verbos relacionais) e os chamados complementos predicativos de sujeito.:

(140) (…) que a população julguem a quem foi preso mesmo que algumas pessoas sejam <inocente> (…).

Nesse exemplo, ainda que a construção esteja no modo subjuntivo, o sujeito e o elemento predicativo são intercalados apenas pelo verbo. No entanto, esse caso foi muito pouco frequente na amostra analisada. Na maioria das sentenças, havia dois ou mais elementos entre o núcleo do sujeito e o adjetivo predicativo estabelecendo relação de dependência:

(141) Contudo , manifestações em prol de direitos são <legítimos> desde que esses não agridam (…) .

No exemplo (141), a ausência de concordância se dá em termos de gênero. Entretanto, há um elemento mais próximo do adjetivo predicativo, que não é o núcleo do sujeito ao qual ele se refere, mas que pode ter influenciado a marcação de concordância (o termo *direitos*). Nesse caso, há quatro *tokens* intercalando o núcleo do sujeito e o verbo que faz a função de ligação com o complemento predicativo.

(142) Devido a tantas informações que são " jogadas " pelas mídias , os leitores recebem informações e não procuram ir além , defendendo ou indo contra sem ao menos saber se <é> <verídico> .

O exemplo (142) representa bem a complexidade da construção: há uma estrutura de coordenação na sentença que contém o elemento sublinhado que atua como sujeito do complemento predicativo, há uma estrutura condicional na qual o predicativo ocorre, e há uma elipse do sujeito que faz papel de objeto da sentença na qual ocorre. Nota-se também a distância de vários *tokens* entre o elemento que tem função de sujeito e o complemento predicativo. Na maioria dos representantes desse fenômeno, a distância entre os elementos dependentes e a complexidade das estruturas nas quais ele ocorria (presença de coordenações, subordinações, estruturas condicionais, entre outras) parecem ter sido relevantes para a ocorrência do desvio.

Por fim, houve ainda casos de problemas de concordância em estruturas específicas, por exemplo, com as expressões *entre outros* e *muitas vezes*. Outro caso recorrente de desvio de concordância envolveu expressões como *a si mesmo* ou *a si próprio*, *um ao outro*, *a maioria de* e estruturas comparativas como *um dos mais/menos*, entre outras.

Relativamente aos fenômenos de concordância verbal, vários deles já foram analisados em seções anteriores, visto que envolveram outros fenômenos de interesse desta pesquisa, como problemas de concordância em que havia uma relativa entre o núcleo do sujeito e o verbo. Assim, o fenômeno mais relevante analisado nesta seção é o de ausência de concordância entre um sujeito simples (sem estruturas de coordenação, relativas ou elementos intercalados, exceto modificadores) e o verbo do qual ele depende:

(143) As " saidinhas " dos detentos <causa> preocupações na sociedade , pois aqueles que entram saem mais violentos e até mesmo piores .

Outro caso identificado foi a presença de locuções verbais nesse tipo de fenômeno:

(144) Esses movimentos <pode ocorrer> por alguns motivos , possui um leque amplo (…) .

A presença de pelo menos um modificador do nome intercalando o núcleo do sujeito e o verbo com desvio de concordância também foi identificada como recorrente:

(145) (…) as pessoas estavam insatisfeitas com o modo em que o governo militar <delimitavam> as terras do país , tendo como objetivo implantar a reforma agrária no país .

(146) (…) conflito de interesses é onde os movimentos sociais <torna> uma ferramenta de intervenção .

Nota-se que os exemplos (145) e (146) apresentam o problema inverso em termos de concordância de número. No primeiro, identifica-se que *governo militar* tem sentido de grupo de pessoas, o que poderia justificar a opção pelo plural, de acordo com uma possível preferência pela concordância semântica, em detrimento da sintática. Já o caso do exemplo (146) parece seguir a tendência da fala de não marcar o plural redundante no verbo, mantendo as marcas morfológicas apenas no nome e no modificador. Outra possibilidade é que a concordância tenha seguido o núcleo do elemento seguinte (*ferramenta*).

Identificou-se também que muitas vezes a concordância foi feita com o núcleo nominal mais próximo, como ilustrado a seguir:

(147) Os direitos dos movimentos sociais no Brasil <teve> mais valor no início de 1960 , pois foram criados os primeiros movimentos legítimos no Brasil .

Vê-se no exemplo (147) que todos os elementos que constituem o sujeito mantêm a marca de plural. No entanto, a estrutura adverbial intercalada (*no Brasil*) é a mais próxima do verbo, que manteve a mesma marca de número. Tal fenômeno ocorreu 16 vezes na amostra analisada, reforçando a hipótese de que há uma tendência, em termos de concordância, de estabelecer a dependência com o núcleo nominal mais próximo.

Outros dois fenômenos identificados como relacionados aos desvios de concordância foram a presença de números e porcentagens (como em "*44% da população não <lê>*"), e os problemas de concordância ligados ao verbo *existir* (como em "*<Existe> vários movimentos*"). Nota-se, portanto, que as questões de concordância verbal e nominal, em função da pluralidade dos fenômenos envolvidos, ainda parecem bastante complexas e de difícil execução por parte dos produtores de textos em formação. Da mesma forma, as experiências com corretores gramaticais sugerem que tal fenômeno é desafiador também para ferramentas de PLN. Portanto, abre-se aí uma possibilidade de aprofundamento do estudo desses fenômenos com vistas ao aprimoramento dessas ferramentas.

5.5 *Questões relacionadas a formas verbais finitas e infinitas*

Segundo Bechara (2009), "Entende-se por verbo a unidade de significado categorial que se caracteriza por ser um molde pelo qual o falar organiza seu significado lexical". Tradicionalmente, as estruturas verbais são, na maioria das vezes, os elementos predicadores

das orações. Assim, são centrais para que se defina a estrutura argumental da sentença e, quando finitos não auxiliares, correspondem à raiz da árvore sintática de dependência. Em função da sua importância, os fenômenos ligados às formas verbais finitas e infinitas são analisados separadamente nesta seção, que descreve tanto os fenômenos nos quais os verbos parecem ser protagonistas na ocorrência de desvios, quanto aqueles nos quais têm papel coadjuvante. Inicia-se analisando os desvios das formas infinitas e, na sequência, descrevem-se aqueles ligados a questões de tempo, modo e demais usos das formas finitas.

5.5.1 Uso ou presença de formas verbais infinitas

As formas verbais chamadas de infinitas são o infinitivo, o gerúndio e o particípio, e recebem essa designação porque também podem desempenhar função de substantivo (BECHARA, 2009) — por isso, são chamadas ainda de formas nominais. Exceto o infinitivo, que permite flexão em algumas circunstâncias, as formas verbais infinitas são invariáveis. Em alguns dos fenômenos analisados, os desvios se dão diretamente em termos de uso dessas formas verbais; em outros, a sua presença parece influenciar a ocorrência de desvios de diversas ordens.

O primeiro fenômeno a ser analisado é a presença de vírgulas separando elementos com papel de sujeito quando, dentro desse sujeito, ocorrem formas verbais infinitas[57]. A separação de sujeito e verbo por vírgula não se restringe a esse fenômeno, sendo muito frequente em diversos tipos de sentenças. Todavia, identificou-se um número significativo de casos em que elas se fizeram presentes, o que reforça a hipótese de que esse fenômeno esteja positivamente correlacionado com tais desvios de uso de pontuação. Um exemplo ilustrativo consta a seguir:

(148) Essa forma de pensar e agir <,> contribuiria para uma melhor qualidade de vida e faria com que a estimativa de vida se elevasse .

Estruturas semelhantes ocorrem com a presença de particípios, ainda que sejam menos frequentes do que aquelas com infinitivos:

(149) Portanto , decisões tomadas pela sociedade <,> podem ser consideradas insignificantes por não terem uma abrangente consciência de que determinado assunto é valoroso para outro indivíduo (…).

Não se identificou nenhuma ocorrência desse fenômeno na presença de gerúndio. Destaca-se uma sentença em que a forma verbal presente no sujeito é um infinitivo flexionado:

[57] Nesta seção, as formas verbais infinitas a que se está referindo nas análises dos exemplos aparecem sublinhadas, de forma a especificá-las claramente, visto que outras dessas formas podem ocorrer nas sentenças analisadas.

(150) Portanto , o fato desses movimentos estarem interligados <,> facilita em uma resolução , por exemplo um bom planejamento para que não houvesse o desemprego , violências , (…) .

Cogita-se que, nesse caso, a colocação da vírgula se justificaria em função de uma possível interpretação das estruturas como orações intercaladas ou deslocadas, que precisariam ser marcadas por vírgulas. Todavia, não é impossível que tal desvio nada tenha a ver com as formas verbais infinitas, justificando-se pelo número de *tokens* do sujeito, pela presença de coordenação ou pela distância entre o núcleo do sujeito e o elemento predicador com o qual ele estabelece relação de dependência. A quantidade e as características dos exemplos encontrados no conjunto de dados analisados não permitem tecer conclusões sobre esse aspecto, abrindo espaço para investigações mais aprofundadas e em maior quantidade de dados.

A segunda questão envolvendo formas verbais infinitas restringe-se ao infinitivo: trata-se do uso de crase antes de verbos que têm essa forma. Tal caso teve apenas oito ocorrências e está entre os diversos problemas relacionados ao uso de crase.

(151) Pode-se dizer que a manifestação contra o aumento da tarifa de ônibus em 2013 acarretou grande parte da população <à> lutar pelo que se via necessário , sendo que , através dela outros Estados brasileiros começaram <à> protestar por questões importantes como a corrupção , política (…) .

No exemplo (151), há uma locução verbal que exige o uso da preposição *a* em posição intermediária, na qual a crase foi aplicada de maneira equivocada. No contexto das locuções verbais compostas por verbos infinitos que exigem o uso de preposição, identificaram-se 11 ocorrências em que tal preposição foi omitida: duas na locução *passar a*, três em *vir a*, três em *levar a* e três em *começar a*. Do total de ocorrências, sete dessas estruturas são contínuas, e as demais possuem elementos intercalados. Também no âmbito das locuções verbais, outro fenômeno que ocorreu apenas uma vez na amostra analisada diz respeito ao uso de ênclise após particípio (*tem agido-se*).

Voltando às análises do infinitivo, outra questão bastante recorrente, que foi marcada como desvio durante a anotação, foi o uso de contração prepositiva antes de orações reduzidas de infinitivo como a que segue:

(152) O movimento nada mais é do que o modo <da> sociedade acordar e lutar pelos seus direitos .

Segundo Bechara (2009), essa construção não é aceita por alguns gramáticos porque o sujeito não pode ser regido por preposição. Porém, segundo o autor, tal utilização não trata "*de regência preposicional de sujeito, mas do contato de duas palavras que (...) costumam ser incorporadas na pronúncia*. Se tais combinações parecem contrariar a lógica da gramática, cumpre observar que não repugnam a tradição do idioma" (p. 536 – grifo do autor). Nesse

sentido, talvez a anotação de tal fenômeno como desvio precise ser repensada, já que ele é bastante frequente não só nas redações, mas também em textos jornalísticos ou de gêneros que prezam pela utilização da modalidade escrita formal. A defesa por se manter a anotação traz outra questão apontada por Bechara (2009, p. 568): "A não combinação da preposição com o sujeito garante o valor expressivo da preposição e a ênfase posta no sujeito".

Nos desvios ligados diretamente ao uso de formas verbais, identificaram-se diversas formas finitas onde deveria ter ocorrido um infinitivo, flexionado ou não. Algumas dessas ocorrências parecem estar mais associadas a desvios de ortografia ou de digitação do que propriamente de desconhecimento quanto ao uso do infinitivo:

(153) A leitura é extremamente importante , não apenas para a formação do nosso intelectual , mas também para <temos> novos conhecimentos , questionamentos , para ter a formação de um novo vocabulário.

A maior parte das ocorrências desse tipo de desvio se deu quando a forma que se pretendia usar era um infinitivo flexionado. No entanto, ressalta-se o exemplo do verbo *vir*:

(154) Mas com toda essa nossa privatização , tiramos nossos próprios direito de ir e <vim> (…)

Mesmo que se pudesse argumentar um possível erro de digitação, a coordenação entre infinitivos indica um provável desconhecimento da forma infinitiva do verbo *vir*. Desvios de uso do infinitivo com esse verbo ocorreram mais de uma vez no conjunto dos dados analisados, o que reforça a hipótese de que ele esteja trazendo dificuldades de realização da sua forma infinitiva.

O infinitivo é a única forma verbal infinita que permite variação de pessoa; a esse fenômeno costuma-se chamar *infinitivo flexionado*. Segundo Bechara (2009), não se flexiona um infinitivo pertencente a uma locução verbal. Fora da locução, a utilização da forma de infinitivo flexionado é uma questão de escolha entre reforçar ou não o sujeito do verbo. Todavia, tais orientações parecem não ser totalmente compreendidas pelos produtores dos textos, já que houve diversas ocorrências de flexão do infinitivo em locuções verbais:

(155) (…) passando o resto da vida na prisão , aguardando que algum dia possam <serem> julgados .

(156) No ano de 1960 esses movimentos começaram a <serem> importantes , pois a população (…).

No exemplo (155), o infinitivo foi flexionado em uma estrutura passiva com presença de verbo com função de modalização (*possam*). Já no exemplo (156), a locução verbal contém uma preposição intercalada, o que poderia justificar o uso do infinitivo flexionado. Entretanto, tal desvio foi muito frequente e ocorreu em diversas estruturas diferentes, o que reforça a hipótese de desconhecimento das orientações de uso desse infinitivo.

Identificou-se ainda a troca frequente entre as três formas, isto é, o uso de um tipo onde deveria ocorrer outro. Uma dessas alternâncias envolve o uso do gerúndio no lugar de infinitivo ou de particípio:

(157) (…) assaltando pessoas , conseguindo <vendendo> o pertence da vitima ou trocando pela droga .

O contrário também ocorreu em algumas sentenças: utilizou-se particípio ou infinitivo onde deveria haver um gerúndio, especialmente em locuções verbais:

(158) Grandes movimentos vem <ganhado> espaço na sociedade nos últimos anos (…).

No exemplo (158) fica a dúvida se se pretendia usar a locução *vem ganhando*, em que o desvio ocorreu na forma verbal infinita, ou a locução *tem ganhado*, o que implica desvio no verbo auxiliar. As análises anteriores e a intuição sobre esse tipo de desvio levam a crer que se trata do primeiro caso, e que o desvio da forma verbal é, na verdade, um desvio de grafia.

Nas análises, dois tipos de desvios em formas verbais infinitas suscitaram a dúvida sobre a melhor forma de descrição. Trata-se da ocorrência dessas formas onde deveria ter ocorrido um verbo no modo indicativo ou no subjuntivo. Decidiu-se, de maneira arbitrária, inserir tais casos na seção que descreve os desvios em formas verbais infinitas, considerando-os como representando a interface entre ambos os tipos de verbos. Ademais, os *tokens* marcados como contendo o desvio são, de fato, formas verbais infinitas, o que justifica a sua análise nesta seção.

Assim, o primeiro dos casos se refere ao uso de formas verbais infinitas onde necessariamente deveria ocorrer uma forma no indicativo:

(159) Esse método beneficia muito a sociedade em si , mas também <gerando> capital para o país .

(160) (…) por conta desses movimentos os jovens <começando> a trilhar caminhos e adquirir conhecimentos através da manifestação da arte .

(161) (…) um exemplo radical sobre isso é o grupo Estado Islâmico (…) ele mata pessoas e rompem as leis de países para chamar atenção para o que eles <exigindo> .

No exemplo (159), o problema de uso da forma infinita se dá por questões de paralelismo sintático, uma vez que ambas as sentenças estabelecem relação de coordenação e, por isso, precisam utilizar as mesmas formas verbais. O exemplo (160) traz uma locução verbal em que a forma do verbo auxiliar foi substituída equivocadamente por gerúndio. Nota-se também que o verbo funciona como auxiliar para dois infinitivos em coordenação (*trilhar* e *adquirir*). No exemplo (161), a estrutura escolhida exige uma forma verbal finita; no entanto, há duas hipóteses possíveis: utilizar o verbo *exigem* ou inserir o auxiliar *estão*.

O segundo caso se refere ao uso de formas verbais infinitas onde necessariamente deveria ocorrer um verbo no subjuntivo. Esse caso frequentemente envolveu a utilização da estrutura subordinativa *para que*, a qual exige uma forma verbal subjuntiva:

(162) Assim para que todos os movimentos <serem> legítimos , agrupa-los os quais tem os interesses (…).

(163) Se o governo deixasse de tomar atitudes equivocadas e <começar> a pensar em ressocializar (…).

O primeiro exemplo desse conjunto ilustra as estruturas mencionadas no parágrafo anterior. Já o exemplo (163) traz uma estrutura condicional em que ocorre a coordenação de dois verbos (*deixar* e *começar*), intercalados por cinco *tokens*, entre eles um infinitivo que compõe com o primeiro verbo no subjuntivo uma locução verbal (*deixar de tomar*) e uma construção com verbo-suporte[58] (*tomar atitudes*). O desvio ocorre na segunda forma verbal da coordenação, que é realizada como um infinitivo. Essa sentença ilustra a questão levantada no início deste capítulo sobre a influência de mais de um fenômeno na ocorrência de um desvio.

5.5.2 Uso ou presença de formas verbais finitas: questões de tempo e modo

Os verbos finitos são os principais responsáveis pelas estruturas de predicação que constituem as orações. Assim, os desvios sintáticos associados a eles também adquirem as mais diversas formas. Diversos desses desvios encontram-se dispersos entre os demais fenômenos linguísticos descritos neste capítulo, mas há alguns que dizem respeito especificamente à sua ocorrência nas sentenças, seja em termos de modo, tempo ou de formas verbais finitas.

O primeiro desvio descrito se refere a problemas na construção do modo subjuntivo. Uma das ocorrências frequentes é o uso de formas indicativas no papel de subjuntivos, possivelmente por desconhecimento em relação à forma correta do modo subjuntivo:

(164) (…) exigir ao governo e a sociedade que respeitem , aceitem e principalmente <cumprem> as leis e se preciso criem novas para que todos tenham seus direitos atendidos (…) .

No exemplo (164), nota-se que outros verbos na mesma sentença foram utilizados na sua forma subjuntiva adequada, o que indica conhecimento em relação ao modo a ser utilizado na construção. É possível que o desvio se justifique porque algumas formas de subjuntivo substituem o *a* do indicativo por *e*, como é o caso de *aceitar*, que também ocorre na sentença. Logo, é possível que, por semelhança, se tenha usado a mesma lógica com o verbo *cumprir*, cuja forma de subjuntivo faz exatamente o contrário: troca o *e* do indicativo por *a*.

[58] Tais construções são conceituadas e descritas em seção específica (Seção 5.6.1, na página 131).

Outro desvio comum é o uso do tempo equivocado do modo subjuntivo (por exemplo, uso do pretérito perfeito do subjuntivo no lugar do presente, também do subjuntivo). Tal desvio ocorreu frequentemente em estruturas condicionais, como mostra o exemplo a seguir:

(165) Se os telespectadores não procuram em outras fontes , (…) não <saberiam> que os manifestantes estavam , na verdade , se defendendo de um primeiro ataque a polícia .

No exemplo (165), é o verbo contido na estrutura condicional que define o tempo verbal do subjuntivo: ambos devem estar no presente. Um caso particular pode ser visto a seguir:

(166) Se a política <ajuda-se> nos colégios para poder ter palestras e passeios culturais para os (…)

O problema do exemplo (166) não está relacionado ao modo subjuntivo em si, já que o tempo e o modo estão adequados à estrutura em que o verbo se insere. Nesse caso, o problema está na forma do verbo: utilizou-se hífen e a partícula *se* quando se pretendia utilizar a forma subjuntiva *ajudasse*. Na fala, ambas as formas têm pronúncia muito similar, o que pode ter motivado a ocorrência desse desvio.

Descritas as questões de desvios de modo, identificaram-se também aquelas ligadas aos tempos verbais. Com frequência, ocorreu a troca entre as formas de presente e pretéritos:

(167) (…) todos <lutam> e conquistaram de uma forma pacifica o direito de pagar o valor do preço (…).

(168) Os movimentos sociais brasileiros <ganham> mais importância a partir da década de 1960 , quando surgiram os primeiros movimentos de luta contra a política vigente (…) .

No caso mostrado no exemplo (167), ressalta-se a estrutura de coordenação na qual um dos verbos aparece no presente e o outro no pretérito perfeito do indicativo. O mesmo ocorre no exemplo (168), mas nesse caso as formas verbais estão em duas orações diferentes. Um caso a ser destacado é o da sentença abaixo:

(169) Por exemplo : A lei <permiti> que quando uma pessoa estiver idosa (uma idade determinada) , é possível ela herdar uma quantia em dinheiro até falecer .

A forma verbal marcada contém um desvio de tempo e de pessoa, mas uma análise atenta identifica que, na verdade, trata-se de uma influência da fala na escrita. Caso se queira identificar esse tipo de desvio por meio de ferramentas computacionais, será necessário projetá-las de forma a considerar fatores internos e externos ao texto.

Em alguns casos, o desvio da forma verbal esteve ligado apenas à pessoa. Destaca-se um tipo de desvio que ocorreu algumas vezes:

(170) A população sempre <estás> antenada sobre os movimentos , pois eles defende os direitos (…).

(171) (…) , pois eles <estarias> dando informações em prol a eles mesmo é a sociedade , (…) .

No exemplo (170), houve a troca da terceira pela segunda pessoa do singular do presente, a partir do acréscimo da letra *-s* no final do verbo. Já no exemplo (171), o *-m* marcador de terceira pessoa do plural do futuro do pretérito foi trocado por *-s*, que marca a segunda pessoa do singular desse mesmo tempo. Porém, destaca-se que o *-s* em geral é a marca morfológica de plural em substantivos, o que talvez tenha alguma influência na ocorrência do desvio. Por fim, houve ainda desvios variados em termos da forma dos verbos, como nos casos a seguir:

(172) (…) para essa sociedade que <habitavem> no Brasil quando os portugueses chegaram .

(173) (…) quando os europeus <vinheram> para o Brasil , para poderem passar para (…) .

Nos exemplos (172) e (173), identifica-se o uso de formas verbais existentes em alguns contextos de fala informal, mas que em geral são vistas como menos prestigiadas. Nessas ocorrências, o corretor ortográfico do *MS Word* identificou os desvios, mas apenas no verbo do exemplo (172) a sugestão de substituição estava adequada. Logo, evidencia-se a complexidade envolvida nos desvios e a necessidade de tratá-los com soluções que vão além dos limites do texto, analisando também desvios comuns advindos da fala.

5.6 *Aspectos semânticos, lexicais ou ligados a expressões compostas/multipalavras*

Durante toda a pesquisa e, principalmente, nas definições da tarefa de anotação, buscou-se restringir ao máximo o foco de interesse no nível da sintaxe. Assim, por exemplo, sentenças bem-construídas sintaticamente, mas que não faziam sentido foram consideradas como “sem desvio”. No entanto, as análises linguísticas evidenciaram a limitação envolvida em tal restrição. Alguns dos fenômenos em que os desvios sintáticos ocorreram parecem ter forte relação com aspectos semânticos ou com formas e expressões lexicais específicas. Esta seção analisa os fenômenos que são sintático-semânticos, lexicais ou relacionados a expressões compostas/multipalavras, e que de alguma forma puderam ser correlacionados com a presença de desvio sintático.

5.6.1 Problemas em construções com verbo-suporte

Verbos plenos são caracterizados por possuírem conteúdo semântico e estrutura argumental canônica (p. ex. *fazer um bolo*). Já os verbos-suporte[59] são vazios ou quase vazios de significado — mas carregam informações de tempo, modo e aspecto (p. ex. *fazer uma denúncia*). Construções com verbos-suporte (como ***fazer*** *um* ***passeio*** e ***dar*** *uma* ***olhadinha***) em geral são consideradas complexas, em oposição à sua forma simples, isto é, o verbo pleno (*passear* ou *olhar*). Essas construções com verbo-suporte costumam ser inseridas na interface entre os aspectos sintáticos e semânticos.

Durante as análises dos fenômenos linguísticos, identificou-se que tais construções, chamadas aqui de LVCs (em inglês, *light-verb constructions*), parecem influenciar a ocorrência de desvios de regência. Nesses casos, é difícil definir se se trata de um fenômeno de regência verbal ou nominal, uma vez que é a construção inteira (verbo-suporte + nome predicativo) que rege o uso da preposição. Um exemplo desse fenômeno pode ser visto a seguir:

(174) (…) mantendo assim a ordem e fazendo jus <a> pluralidade de ideais existente no Brasil .

No exemplo (174), não se pode dizer que é o verbo *fazer* nem que é o nome *jus* quem exige o uso da preposição *a* e, por consequência, do acento indicativo de crase[60], mas sim a construção *fazer jus*. LVCs podem ter os seus elementos contínuos ou não:

(175) (…) mas seu líder acaba atingindo outras pessoas com suas razões as vezes pessoais e errôneas sobre o que realmente se deve lutar , causando grandes impactos <a> humanidade .

O exemplo (175) traz um caso particular de LVC, uma vez que o verbo tem sentido causativo, isto é, o sujeito do verbo é a causa do evento expresso pelo nome predicativo. Tais construções se comportam como um tipo diferente de LVC, mas o seu funcionamento em termos de regência para o fenômeno descrito aqui é muito similar. Nesse caso, nota-se ainda que os elementos da LVC não são contínuos, pois há um modificador anteposto ao nome intercalando a construção. Todavia, mais uma vez é a construção *causar impacto* que exige a preposição *a* (ou *em*), exigindo o uso de crase. Outro exemplo particular pode ser visto a seguir:

(176) O governo deveria dar mais ouvidos <a> população e o cidadão tem o dever de pesquisar e acompanhar seus representantes governamental .

---

[59] Alguns autores fazem distinção entre *verbo-suporte* e *verbo leve*. Para esta análise, cujo tema dos verbos-suporte/verbos leves não é central, consideram-se esses dois termos como sinônimos que definem o mesmo fenômeno, ao que se optou aqui pela expressão *construções com verbo-suporte* e pelo acrônimo LVCs.

[60] Inclusive, tal substantivo praticamente só ocorre junto ao verbo *fazer* na linguagem corrente, como mostra uma busca na plataforma *Corpus do Português*, utilizando-se o concordanciador.

No caso do exemplo (176), é muito difícil argumentar que é o substantivo que esteja regendo o uso de preposição. No entanto, trata-se de uma LVC que na verdade apresenta características de expressão cristalizada ou idiomática, uma vez que não é possível inferir o sentido de *dar ouvidos* a partir do sentido individual de cada um dos elementos. Logo, identifica-se um problema recorrente quando se pretende analisar elementos isolados da língua: é muito comum que eles não funcionem isoladamente. Esse é o caso das LVCs, das expressões cristalizadas e idiomáticas e de outras expressões multipalavras, bem como das chamadas locuções (conjuntivas, prepositivas, verbais, etc.), cujo funcionamento em termos sintáticos tem particularidades que ainda carecem de investigações amplas e sistemáticas.

Várias das LVCs que ocorreram na amostra analisada permitem que se estabeleça o substantivo como termo regente. Esse é o caso, por exemplo, de construções como *ter acesso*, *ter direito*, *dar apoio* e *dar valor*. Porém, não é o que ocorre nos casos descritos anteriormente ou em LVCs com *dar início*. Nesse sentido, percebe-se que é sistemática a ocorrência de desvios de regência com essas construções (30 sentenças na amostra analisada), o que reforça o fato de que elas merecem ser analisadas separadamente. Para além de todas as possibilidades que se abrem a partir desta investigação, aquelas que mais motivam reflexões se referem justamente à influência de questões semânticas na ocorrência de desvios sintáticos. As LVCs são apenas uma das questões passíveis de uma investigação mais aprofundada, mas esta poderia se expandir para escolhas lexicais que levam a desvios sintáticos, uso equivocado das locuções em termos semânticos, entre outras possibilidades que se inserem na interface entre a sintaxe e a semântica, e dentro dos diversos aspectos lexicais e semânticos.

5.6.2 Problemas com o uso do verbo *haver*

Um dos verbos que apresentaram desvios em diversas subcategorias é o verbo *haver*; portanto, julgou-se interessante analisá-lo separadamente. Enquanto na categoria de concordância, tais desvios estavam ligados à flexão do verbo, nas demais os problemas parecem estar muito mais ligados à grafia. Na subcategoria *crase-exc*, houve sete ocorrências de uso de *à/á* onde se pretendia utilizar o verbo *haver*:

(177) (…) resultado contrário do esperado e até mesmo em morte , quando <á> confronto com a polícia .

Na subcategoria *prepo-desv*, identificaram-se 23 casos do verbo *haver* grafado como *a*:

(178) Não <a> fórmulas mágicas para nos tornarmos leitores assíduos (…) .

Já na subcategoria *determ-desv*, que contou com apenas seis ocorrências, três dos exemplos estavam relacionados ao verbo *haver* grafado como *a*, mas que, nesses casos, não poderia ser considerado como um desvio de preposição:

(179) (…) em média uma pessoa por dia é assassinada nos presídios do país , <a> falta de condição sanitária , lotação nas celas , os presidiários estão se misturando (…) , estão se matando entre si .

No que se refere aos problemas identificados como de concordância, conforme já era esperado, estes estão relacionados à flexão irregular do verbo *haver*, que deve se manter impessoal quando não exerce a função de auxiliar:

(180) É necessário que por parte dos movimentos <hajam> palestras que conscientizem as pessoas (…).

Ocorreu ainda o uso de *há* onde deveria ocorrer um artigo ou uma preposição:

(181) Com <há> chegada da internet ficou cada vez mais fácil e prático ter acesso a tudo (…).

(182) (…) pessoas se manifestando em oposição <há> uma causa é extremamente motivador (…).

Por fim, identificou-se ainda um caso em que se pretendia usar o verbo *haver*, mas se omitiu a letra inicial, resultando no verbo *ouvir* (exemplo (183)). Isso poderia ser compreendido como uma influência da pronúncia na escrita, pois a letra *h* é muda em início de palavras:

(183) Que foi quando <ouve> uma grande movimento social pela insatisfação da população (…).

Em termos de ferramentas computacionais, alguns desses desvios podem ser (e são) facilmente identificados pelos corretores ortográficos e gramaticais, como a flexão do verbo. Em outros casos, porém, torna-se mais difícil estabelecer regras que identifiquem exatamente quando tal verbo foi escrito de maneira equivocada.

5.6.3 Desvios em expressões fixas preposicionais, adverbiais ou conjuntivas

De acordo com Ranchhod (2003), expressões fixas se referem aos termos "usados para designar as categorias constituídas por sequências coesas de elementos lexicais". Nesta seção, consideram-se apenas as expressões fixas (ou locuções) preposicionais, adverbiais ou conjuntivas que contenham desvios na sua formação.

Na categoria de crase, identificou-se, com uma grande frequência de ocorrências, a ausência desse sinal gráfico em diversas expressões, como *à medida que* e *às vezes*, cuja presença de crase é obrigatória em qualquer circunstância:

(184) <A medida que> a população carcerária cresce outros problemas são acarretados como , a superlotação dos presídios , as péssimas condições em que os presos são submetidos (…) .

(185) Porque já começa se justificando , (…) infelizmente <as vezes> muitos já se vitimizam se estereotipando e não fazendo nada pra ser diferente e mudar tudo isso .

Além disso, foram identificadas expressões em que o sinal de crase não faz parte da expressão, sendo utilizado apenas em circunstâncias que permitam o seu uso, como é o caso de *junto a*, *quanto a*, *devido a* e *em relação a*. Outro fenômeno frequente em que se notou a ausência de crase foi nas expressões fixas do tipo *à tona*, *à base*, *à espera*, *às custas*, entre outras. Inseriu-se ainda na análise das expressões fixas a ausência de crase em indicativo de horas, que contou com apenas uma ocorrência. Também se identificaram algumas expressões fixas nas quais a crase foi aplicada de forma irregular, como em *pouco a pouco* e em *a favor*.

Em termos de regência nominal, identificou-se a ausência de preposição ou o uso da preposição equivocada nas expressões *em/a favor de*, *em relação a*, *a respeito de*, entre outras:

(186) O governo deixou que acontece-se essas revoltas e fossem as ruas para lutarem <em favor aos> seu direitos .

Na categoria referente às preposições, foi frequente na análise desse fenômeno a ausência de um dos elementos em locuções prepositivas, como na expressão *devido a*:

(187) <Devido> pessoas que acabam indo para roubar e cometer vandalismo .

Outro caso que chamou atenção foi a ausência de preposição na expressão *de vez em quando*, como no exemplo (188) abaixo. Cogita-se que esse desvio está mais relacionado à falta de revisão do texto produzido após a sua conclusão do que ao desconhecimento da expressão:

(188) Os movimentos sociais não se limitam só a manifestações públicas que acontecem <de vez quando> , ou seja , nunca são manifestações que acontecem frequentemente , (…) .

Em termos de excesso de preposição, identificou-se o uso de preposições formando locuções prepositivas que na verdade (ainda) não existem, como *mediante a* e *perante a*. Entretanto, uma busca rápida no jornal Folha de São Paulo[61] indicou que tal uso parece começar a se fazer presente também em textos escritos por profissionais e que passam por processos de revisão. Encontraram-se ocorrências de ambas as expressões em reportagens do referido

[61] É evidente que apenas esta fonte de consulta não é suficiente para afirmar que tais construções devem ser aceitas. Porém, como se trata de um jornal que possui um manual de redação publicado e é reconhecido como um meio de comunicação que faz uso da modalidade formal da língua, a referência teve como objetivo apenas ilustrar a ocorrência de tais locuções prepositivas também em outros gêneros textuais.

jornal[62], abrindo espaço para o questionamento sobre o quanto tais construções devem ser consideradas como desvios ou como variação linguística.

Dois exemplos interessantes de excesso de preposição em expressões são o uso *grosso modo* com uso irregular da preposição *a*, e a construção equivocada *tendo em vista de*:

(189) Ou o governo que ta muito falho com suas leis , (…) temos que enxergar <a grosso modo> de fato o que está errado no sistema carcerário brasileiro .

(190) O nosso sistema carcerário vai se tornando pior e mais perigoso (…) mediante ao tempo vivido na cadeia , e tendo em vista <do> que se passa dentro de um presídio , mudaram (…) .

No exemplo (189), a expressão fixa do latim não permite o uso de preposição, ainda que essa inadequação seja caso recorrente de dúvidas, sendo inclusive alvo de *blogs* e *sites* que se dedicam a resolver dúvidas sobre a língua portuguesa[63]. Já no exemplo (190), pode-se cogitar se o uso da preposição estaria relacionado a outra expressão fixa: *em vista de*.

Uma expressão bastante frequente que apresentou excesso de preposição foi a construção *muitas vezes*, que ocorreu tanto como *muita das vezes* ou *muitos das vezes* (apenas uma ocorrência) quanto como *muitas das vezes*. Nos dois primeiros casos, há desvios de concordância na expressão; no terceiro, porém, cabe o questionamento se essa construção de fato deve ser considerada desvio ou se, mais uma vez, trata-se de evolução linguística.

No que se refere ao uso de determinantes em expressões fixas, identificou-se a ausência do determinante em expressões como *a maioria* e *a maior parte,* e na construção *o que* retomando a sentença anterior:

(191) <Maioria> das ações sociais são iniciadas quando os cidadãos se sentem violados .

(192) (…) existem muitas propostas com o intuito de impulsionar <maior parte> do público (…) .

(193) (…) diversos presos relataram que ficaram encarcerados por meses até ver o juiz , <que> acaba causando um revolta , por não lutar pela sua liberdade em seu tempo estimulado .

Na categoria que engloba os desvios de uso de conjunções, a maior parte dos casos de expressões fixas contínuas (sem nenhum elemento intercalado) ocorre apenas uma vez e está ligada à ausência ou ao excesso de conjunções em expressões como *enquanto que* (uso irregular do *que*) ou *sendo que* e *por exemplo* (ausência de um dos elementos). Em termos de expressões fixas descontínuas, os exemplos estão associados essencialmente ao uso de duas expressões:

---

[62] *Links* para dois dos textos que apresentaram tais construções: https://www1.folha.uol.com.br/mercado/2019/11/congresso-quer-barrar-pontos-da-mp-do-emprego-entre-eles-taxacao-de-desempregado.shtml e https://acervofolha.blogfolha.uol.com.br/2016/07/13/folha-diz-como-thatcher-virou-a-dama-de-ferro/.

[63] Vide https://duvidas.dicio.com.br/grosso-modo-ou-a-grosso-modo/.

tanto… quanto e *mais… (do) que*. Em ambos os casos, apenas uma parte da expressão foi empregada, enquanto a outra foi omitida ou substituída por outro termo:

(194) Ou seja , a sociedade inconformada com as mudanças <tanto> no campo econômico <e> social , mas só a partir de 1950 , que começaram a ser reconhecidos .

(195) Prisões provisórias tem sido usadas como regras <do que> exceção - e que ela se tornou uma forma de antecipar a execução de pena delas .

Cabe notar que, na primeira expressão descontínua, a primeira parte é que ocorreu, e a segunda foi omitida ou substituída; já na segunda expressão, foi a segunda parte que apareceu nas sentenças, sendo a primeira frequentemente omitida. Houve ainda duas ocorrências da expressão *nada mais é que*, e em ambas o elemento omitido foi o *que*:

(196) Movimento social <nada mais é> a organização de grupos , de uma determinada população com intuito de fazer mudanças através de oposições políticas .

No que tange às possibilidades de identificação e/ou correção automática desse tipo de desvios, pode-se avaliar possibilidades de inserir, em ferramentas computacionais, um léxico (ou uma consulta a tal recurso) com as expressões fixas mais comuns, como as mostradas aqui.

5.6.4 Ausência ou excesso de palavras gramaticais ou lexicais

Uma das questões genéricas que se identificou em mais de uma categoria de desvios foi a ausência ou o excesso de palavras sem relação com nenhum dos fenômenos descritos até agora. Algumas dessas ocorrências parecem estar ligadas à falta de revisão ou à desatenção, mais do que à falta de compreensão dos fenômenos sintáticos e/ou semânticos envolvidos. Um exemplo foi a repetição de artigos definidos e indefinidos, iguais ou diferentes entre si:

(197) Dessa forma <os os> conflitos diminuem e a população evolui .

(198) Nosso país vive <uma o> maior déficit carcerário da historia , com mais de 600 mil presos , com apenas 371 mil vagas , com 250 mil presos sendo provisórios , e com isso , (…) .

Também ocorreram poucos casos de repetição ou uso excessivo de preposições:

(199) (…) , mas quando passam <a> de certa maneira <a> serem os agressores .

Outro exemplo é o uso excessivo de conjunções, como mostrado a seguir:

(200) (…) a inclusão delas é essencial para que <que> saibam o que irão buscar em suas vidas .

(201) <Se caso> isso não diminua o Brasil terá que criar novas penitenciárias para receber (…) .

No exemplo (201), talvez o desvio se dê em função do desconhecimento de que ambas as conjunções têm o mesmo sentido e que, por isso, são redundantes. Em termos de elementos faltantes que possivelmente são resultado de falta de revisão do texto, identificou-se a ausência de preposições que estabelecem relações de subordinação entre um nome e o seu modificador:

(202) Os movimentos sociais que ocorreram na <história Brasil> tem grande reflexos nos dias (…).

Outros casos pouco frequentes foram a ausência de determinantes diante de nomes que causou construções estranhas, a ausência de determinantes após a preposição *para*, e a ausência da conjunção *que*:

(203) Em síntese com o que <psicólogo> e filósofo William James abordou (…) .

(204) (…) esses movimentos iriam ser passivos e produtivo para <sociedade> .

(205) Além disso , <percebe-se> o povo brasileiro vem dando seu jeito para as mudanças serem feitas (…)

Identificaram-se também algumas situações em que os desvios pareciam relacionados ao desconhecimento do funcionamento de alguma estrutura particular. Um dos exemplos mais comuns foi a repetição do artigo após contrações, especialmente em *pelo/pela* e *cujo/cuja*. Um caso que merece destaque é o que ocorre na sentença a seguir:

(206) Aos longos anos da <a> história brasileira , é marcada com diversas lutas e revoltas contra governos e contra circunstâncias impostas pela sociedade , ganhando maior destaque e expandindo-se a partir da década de 70 , contradizendo o regime militar implantado nesta .

O exemplo (206) representa a única ocorrência de repetição de artigo após alguma contração prepositiva diferente das citadas anteriormente. Nota-se, porém, que tanto a estrutura sintática quanto as escolhas semânticas apresentam diversos problemas em sua construção, apesar da aparente tentativa de escrita de uma sentença complexa e de acordo com a modalidade escrita formal. Nesse contexto, é importante refletir sobre a questão do gênero textual *redação escolar* exposta no capítulo de fundamentação teórica desta dissertação. Ainda que tal discussão esteja fora do escopo da pesquisa, não se pode negligenciar a influência, na ocorrência de desvios dos mais diversos tipos, da tentativa de escrita de sentenças “que impressionem”.

Os casos anotados na subcategoria *ordem* também estiveram associados à ausência ou ao excesso de palavras gramaticais ou lexicais. Nesses casos, uma revisão do texto provavelmente ressaltaria a presença do desvio, já que as ocorrências se parecem muito mais com esquecimentos ou repetições de palavras por equívoco ou desatenção do que desconhecimentos da estrutura sintática ou da semântica dos elementos. A maior parte das ocorrências foi a ausência de palavras, em especial no que se refere a formas verbais ou ao *que*:

(207) No ano de 2014 ocorreu o maior movimento social da história Brasileira , (…) <2 depois> das manifestações ela sofreu impeachment e passou a presidência para seu vice Michel Temer .

(208) (…) a penalidade <poderia reduzida> de 30 anos para 25 anos , apenas 5 anos a menos , mas os prisioneiros terão punições , e terão que trabalhar muito .

(209) São pequenas atitudes que devem <ser primeiramente pelas pessoas> , ter empatia pelo próximo para que as nossas crianças , o futuro do pais , aprendam a dar valor a vida (…) .

(210) O que o governo é a sociedade <tem pensar> é o que fazer quando os presos tiverem nos presídios como desenvolver algo lá dentro entre eles , e não pensar em acabar com a vida deles ao pouco (…).

No exemplo (207), ilustra-se um dos poucos casos de provável ausência de substantivos, uma vez que parece faltar o substantivo indicando o tempo (*meses*, *semanas*, *anos*…). Já nos exemplos (208) e (209), faltam elementos que compõem as formas verbais da voz passiva: no primeiro, falta o auxiliar *ser*; no segundo, falta o particípio. Vê-se que em (208) a mensagem ainda é compreendida, apesar da ausência do verbo auxiliar, mas o mesmo não ocorre em (209), porque falta a ação que se pretendia expressar pela estrutura passiva. Por fim, no exemplo (210) falta o *que* na estrutura *ter que pensar*. Logo, os três últimos exemplos do conjunto reforçam a frequência de desvios em formas verbais complexas, como passivas ou locuções verbais.

Nos poucos casos de excesso de palavras, também não foi possível ter certeza de que havia de fato termos excessivos, ou se faltavam elementos de conexão entre eles. A única ocorrência em que não há dúvida do excesso de palavras envolve a repetição do verbo, que é colocado em posição irregular:

(211) (…) na escola não aprende a matéria necessária , então por que não seria <ir> mais fácil <u>ir</u> roubar ?

Esse exemplo ilustra a situação de falta de revisão dos textos, uma vez que é muito pouco provável que, após uma segunda leitura, se mantivesse a repetição do verbo na posição em que ele está. Nas experiências com revisão de textos da pesquisadora, uma ferramenta que auxilia muito na detecção desse tipo de desvio é a leitura em voz alta pelo leitor automático embutido no próprio editor de textos. Assim, vale considerar esse tipo de ferramenta como uma das etapas de auxílio nas plataformas e ferramentas que oferecem suporte à escrita. Citam-se ainda alguns desvios relacionados à ausência de preposições ou determinantes para fins de paralelismo. No entanto, como esse é um fenômeno que facilmente se confunde com questões estilísticas, não será abordado aqui.

5.6.5 Trocas entre classes morfossintáticas

Entre os desvios anotados na subcategoria *orto-sin*, identificou-se que houve o uso equivocado de determinadas classes de palavras gramaticais quando se pretendia utilizar outra classe. Tais desvios não foram muito frequentes, mas nota-se que quase todas as ocorrências estão ligadas ao uso de adjetivos quando se pretendia outra classe. Entre as trocas ocorridas, identificou-se o uso de adjetivos no lugar de advérbios como a mais frequente:

(212) Pode-se dizer <essencial> que devido ao alto índice , (…) a capacidade desses presídios estão extremamente excedidas .

Vê-se que a intenção, no exemplo (212), era utilizar o advérbio *essencialmente*. Nesse fenômeno, a maior parte dos casos envolveu o uso de adjetivo onde deveria ter ocorrido um advérbio terminado em *-mente*. O contrário ocorreu apenas uma vez: utilizou-se o advérbio *comumente* onde deveria ser utilizado o adjetivo *comum*. Adjetivos também ocorreram em lugar de substantivos em quatro ocasiões, das quais três parecem derivar de desvios ortográficos (acréscimo de *-l* em substantivos, formando adjetivos). O contrário, porém, foi mais frequente: substantivos foram utilizados em lugar de adjetivos em sete situações. Todavia, a maior parte dos casos também parece derivar de desvios ortográficos, como a troca entre *essências* e *essenciais*, entre *sócias* e *sociais*, e entre *polícias* e *policiais*. Um exemplo menos evidente nesse sentido pode ser visto a seguir:

(213) Os grupos deveriam procurar cobrar das prefeituras e governos porém de uma forma mais <eficácia> , como baixos assinados e com visitas frequentes em prefeituras e subprefeituras (…) .

Ainda que sejam pouco frequentes, tais desvios se mostram como desafios, uma vez que geram palavras existentes na língua.

5.7 *Questões de anáfora e correferência*

O processamento computacional de anáforas e correferências ainda é um desafio às ferramentas de PLN. Conforme Vieira *et al.* (2008), esses fenômenos envolvem alta complexidade cognitiva. Nesse sentido, uma das questões que se pretendia investigar foi a correlação entre os desvios sintáticos e as questões de retomada de elementos apresentados anteriormente na sentença. Entende-se como uma expressão anafórica aquela que retoma um elemento anterior, que é chamado de antecedente; a relação entre a expressão anafórica e o seu antecedente é

chamada de correferência. Consideraram-se aqui apenas os desvios de correferência por meio de pronomes.

Em alguns desvios anotados na categoria de concordância, identificou-se que, a fim de não repetir um elemento, utilizou-se um determinante e omitiu-se o termo a que ele se referia (elipse de substantivos). Porém, o determinante nem sempre concordava com o elemento que ele pretendia retomar. Tal fenômeno ocorreu apenas três vezes e é ilustrado na sentença a seguir:

(214) (…) eles vão as ruas para reivindicar seus direitos e também <o> da população que acabam apenas assistindo o governo mudar suas vidas sem falar nada .

Percebeu-se que a retomada por pronome relativo se deu de forma equivocada em diversas ocasiões. Grande parte dos casos ocorreu na expressão *o qual/a qual*, seguida ou não por preposição:

(215) Portanto usar o modelo da democracia , como uma alternativa possível em uma sociedade complexa , <as quais> vivem atualmente , tornou-se um instrumento incapaz de responder (…) .

(216) Em consequência disso , vê-se , a todo instante movimentos <na qual> possuem finais trágicos , como destruições de órgãos públicos e privados , brigas , e até mortes .

Todos os casos de retomada de pronome relativo são contínuos, exceto o mostrado no exemplo (215), que é intercalado por uma vírgula (que está posta de maneira equivocada). Nesse exemplo, a expressão anafórica concorda com o antecedente apenas em gênero. Além disso, o substantivo que é o núcleo do antecedente possui um modificador como seu dependente. Já o exemplo (216) traz um caso em que o pronome relativo é precedido equivocadamente pela preposição *em* e não concorda nem em gênero, nem em número com o seu antecedente. Como já foi descrito anteriormente, as estruturas de subordinação de maneira geral parecem estar mais fortemente correlacionadas à presença de desvio sintático do que estruturas consideradas menos complexas, uma vez que aquelas estiveram envolvidas na ocorrência de desvios de vários tipos e em várias categorias.

Foi possível identificar ainda casos de falta de concordância entre pronome anafórico e antecedente envolvendo os pronomes *dele*, *nele*, *lhe* e *o/a*[64], em que a distância entre as expressões anafóricas e os seus antecedentes foram maiores do que no caso dos relativos:

(217) Estamos lidando com locais que deveriam abrigar 6 presos e existem 20 dentro <delas> .

[64] Todas as referências a palavras no texto que não se referem a exemplos específicos aparecem na sua forma lematizada, para manter a concisão e evitar a repetição das palavras em todas as suas formas: femininas e masculinas no singular e no plural.

(218) (…) que também ocasiona em vitimismo por dizer que sofrem na sociedade brasileira com o sistema do governo , não pensando em que eles mesmos <os> elegeram .

No exemplo (217), a distância entre expressão anafórica e antecedente é de vários *tokens*. Além disso, pelo contexto, infere-se que o termo *locais* se refere a *celas*, o que justificaria a expressão anafórica no feminino. No exemplo (218) ocorre um caso interessante de concordância semântica, mas não sintática. Inicialmente, vê-se que também há certa distância entre retomador e antecedente, mas o que se destaca é que o termo *governo* se refere a um conjunto de pessoas: *os políticos*. Nesse sentido, o pronome oblíquo *os* poderia ter seu uso justificado como uma retomada semântica das pessoas que compõem o *governo*, que é um substantivo no singular, mas se refere a um coletivo.

Ao analisar os desvios de retomada de elementos por pronomes pessoais ou por *o mesmo/a mesma*, os quais foram anotados na subcategoria *concordância anafórica*, nota-se que esse mesmo fenômeno de concordância semântica ocorreu em diversas sentenças. Nos 10 casos em que o elemento retomador (a expressão anafórica) aparece no plural, mas retomando um antecedente no singular, apenas um envolve o uso de *o mesmo*.

(219) Além disso , um dos movimentos mais impactantes na sociedade atual é o movimento negro , que luta principalmente por questões étnicas visando a inserção <dos mesmos> em sociedade , (…) .

No exemplo (219), vê-se que a relação de correferência ocorre com vários *tokens* intercalando expressão anafórica e antecedente. Além disso, o antecedente *movimento negro* se refere ao coletivo *negros*. Assim, a retomada se refere à *inserção dos negros*, o que justificaria o uso da expressão anafórica no plural.

Os casos em que o pronome retomador (a expressão anafórica) aparece no singular, mas o seu antecedente está no plural foram bem mais frequentes que o fenômeno anterior, ocorrendo 21 vezes na amostra investigada. Destas, cinco se referem ao uso de *o mesmo*, uma se refere à retomada com a palavra *tal*, e todos os outros envolvem a retomada por pronome pessoal reto.

(220) (…) os jovens muitas vezes concluí os estudos , e sai deixando de aprender muitas coisas , que <lhe> fará falta quando <ele> for ingressar no mercado de trabalho , (…) .

No exemplo (220), há duas retomadas equivocadas do antecedente: uma pelo *lhe* e outra pelo *ele*. Porém, elas concordam entre si em número. Aqui é possível novamente argumentar em relação à concordância semântica, uma vez que o antecedente se refere ao grupo de pessoas *jovens*, e as expressões anafóricas se referem a uma situação que pode acontecer com *algum jovem específico*. Porém, nesse caso os verbos que estabelecem dependência com o termo

jovens também aparecem no singular, estando de acordo com os demais retomadores. Em relação a esse fenômeno, um caso merece destaque:

(221) Porém essas mudanças , precisa de tempo , dinheiro , mão de obra , dinheiro , pensamento estratégico , dinheiro , mas essas obrigação não parte de nós , <cidadão> , sim daqueles que estão no poder .

Trata-se do único exemplo de catáfora encontrado na amostra em que o elemento retomador (a expressão catafórica) *nós* aparece antes do elemento que ele retoma (*cidadão*). Aqui o pronome retomador plural está muito próximo em distância do retomado, que está no singular, e é mais difícil considerar a concordância semântica. Porém, como esse não é o único desvio de concordância da sentença, pode-se cogitar que tal tema traz dificuldades ao produtor para além de questões de anáfora ou catáfora.

Em oito ocorrências de ausência de concordância entre retomador e retomado, os elementos concordaram em número, mas não em gênero. Os pronomes envolvidos nessas sentenças foram os seguintes: *eles* (duas ocorrências), *elas* e *ela* (uma ocorrência cada), *o mesmo* (duas ocorrências), *os mesmos* e *os próprios* (uma ocorrência cada). A prevalência de retomadores no masculino levanta a hipótese de certa preferência por esse gênero, talvez porque o gênero considerado neutro no português em geral receba a marca morfológica de masculino.

A hipótese de certa preferência do masculino como gênero de retomada "universal" se reforça ao analisar os casos em que retomador e retomado não concordaram nem em gênero, nem em número, ainda que o número de ocorrências tenha sido pequeno para embasar conclusões nesse sentido. Em todos os sete casos em que não houve nenhuma das duas marcas de concordância, o gênero do pronome anafórico foi o masculino. Em praticamente todos os casos, também, pode-se considerar a concordância semântica como influência para os desvios, uma vez que os antecedentes foram *sociedade*, *população* e *pessoas*.

Por fim, o último caso analisado tem forte influência de um fenômeno da fala: a repetição do sujeito por pronome pessoal reto antes da ocorrência do verbo do qual depende. Tais casos só tiveram sete ocorrências, o que mostra que, ainda que tal fenômeno seja muito comum na fala, não tem sido transportado para a modalidade escrita da língua:

(222) A arte <ela> é designada para estimular a criatividade de um ou mais indivíduos com objetivo de criar um senso crítico , ou seja , com expressões corporais e imagens registradas (…) .

(223) Por conta do vasto contexto histórico , grupos que antes não tinha tantos ouvidos , após a construção dos movimentos , <eles> ganharam reconhecimento e tem seus direitos estabelecidos por lei .

No exemplo (222), o substantivo que faz papel de sujeito e o pronome pessoal encontram-se adjacentes um ao outro. Já no exemplo (223), há uma oração relativa e uma

estrutura adverbial intercaladas entre o núcleo do sujeito e o pronome pessoal que o retoma. Nesse fenômeno, a distância entre dependentes não parece fazer diferença na ocorrência de desvios, já que o número de casos de elementos adjacentes e o número daqueles em que havia elementos intercalados foi semelhante (três do primeiro caso, quatro do segundo). Para identificar demais fenômenos envolvidos nesses desvios, é necessário analisar um conjunto maior de exemplos.

5.8 *Fenômenos específicos de algumas categorias*

Alguns dos desvios mapeados foram específicos das categorias previstas na tipologia e, por isso, não puderam ser sistematizados em fenômenos gerais. Assim, esta seção, dividida em subseções de acordo com as categorias nas quais isso ocorreu, traz fenômenos diversos. Por questões de restrição de tempo e de espaço, apresentam-se análises menos aprofundadas desses desvios, apenas elencando os fenômenos e trazendo exemplos pontuais de cada um.

5.8.1 Desvios de uso de pontuação

Na categoria de pontuação, identificaram-se tipos diversos de desvios, relacionados também a sinais de pontuação variados. Anotaram-se quatro casos de ausência de ponto final segmentando duas sentenças completas, com a segunda iniciando por letra maiúscula (cogita-se que se trata de um desvio ligado à desatenção ou à falta de revisão no texto):

(224) Uma verdadeira democracia se faz a partir da soberania popular e do respeito aos direitos <fundamentais> Diante dessas informações , traduzir o que seria legítimo para movimentos sociais poderia ser resumido em não fazer o uso de violência (…) .

Um problema de pontuação mais frequente que o anterior foi a ausência de pontuação marcando expressões de explicação, correção, continuação, conclusão ou concessão, e conjunções e advérbios adversativos[65], principalmente quando pospostos:

(225) Entretanto , apenas em 1985 (…), com o intuito de promover , gerenciar , financiar <e> portanto , possibilitar que todos tenham acesso a diversas formas de expressão de cultura e arte .

(226) Os Ministérios , da educação , trabalhista , entre <outros> devem ver as leis que mais favorecem a população que esta lutando pelo direito , e que são as que mais sofrem sem eles , (…).

[65] Bechara (2009) recomenda o uso de pontuação para intercalar elementos como *por exemplo*, *entre outros*, *não obstante*, *aliás*, *porém*, etc.

Em termos de excesso de pontuação, tem-se o uso excessivo de parênteses, de ponto-e-vírgula e de ponto final, que ocorreram somente em algumas sentenças, e de dois-pontos, que foi um pouco mais frequente que os anteriores, como ilustram os exemplos a seguir:

(227) Já os que cometeram crimes como <(> Assaltos <)> a penalidade poderia ser reduzida de 5 anos para 3 anos , e também terão punições , e terão que trabalhar muito .

(228) A arte pode ser de grande ajuda para transformar vidas , pois através dela pessoas melhoram de vida , e , acabam lidando com alguns problemas como <;> ansiedade e nervosismo .

(229) As prisões do país são um tanto injustas , por que adolescentes / crianças podem ser roubar e sair imunes por não serem de maior ? <.>

(230) Falta de utensílios , superlotação em celas , falta de manutenção , entre outros problemas estão levando presos a um caminho de <:> " briga por seu espaço " , fazendo rebeliões , causando (…) .

Por fim, o uso de um sinal de pontuação no lugar de outro englobou pontuações diversas, mas com poucas ocorrências em cada caso. Houve dois casos de uso de ponto de interrogação em sentenças afirmativas, e também dois casos de uso de ponto final em perguntas:

(231) (…) e na mesma via , cresce o número de presos provisórios , muitos passam meses e até anos esperando o seu julgamento , e a prisão provisória se tornou um refúgio para a crise carcerária <?>

(232) Existem juízes corruptos , como policiais corruptos também , como queremos deter a lei , por um pais melhor com uma lei como essa que temos hoje em dia <.>

No exemplo (231), é evidente que a sentença não se caracteriza como uma interrogação. Já no exemplo (232), em função da aglutinação de sentenças, a hipótese que parece fazer mais sentido é a de que o trecho sublinhado é uma sentença interrogativa. Porém, a estrutura sintática problemática não permite afirmar isso inequivocamente. Outra troca que aconteceu algumas vezes foi entre o ponto-e-vírgula e o ponto final. Tais ocorrências foram anotadas como desvio quando ambas as sentenças segmentadas com o ponto-e-vírgula eram plenas e quando a segunda iniciava por maiúscula:

(233) (…) agressão verbal se tornaram as formas mais viáveis de se resolver um problema em meio a crise do nosso país <;> Devemos compreender os motivos pelos quais alguns movimentos sociais agem .

Ainda no que se refere ao uso de ponto-e-vírgula, também foram identificados três casos em que esse sinal de pontuação foi utilizado no lugar de dois-pontos:

(234) No Brasil os movimentos mais conhecidos se dividem em três <;> MST (…) , MSTS (…) e os movimentos em favor do índios , negros e mulheres .

Já o sinal de dois-pontos foi utilizado duas vezes onde deveria ter ocorrido uma vírgula:

(235) Diante dos fatos expostos acima conclui-se que <:> pelo fato do país não apoiar a maioria da sua população que é pobre , muitas pessoas que estão inseridas nesse contexto social encontram-se (…).

O contrário também ocorreu em três sentenças, como ilustrado a seguir:

(236) Em virtude disso os movimentos sociais vistos hoje contra , (…) traz uma minoria que faz desses movimentos um ato violento e criminoso que desvia o foco principal <,> alcançar a mudança .

Também foi identificado um caso de vírgula onde deveria haver uma interrogação:

(237) E como podemos não ter um governo tão injusto <,> sabendo escolher na hora de votar .

Em função de os desvios de pontuação se apresentarem em fenômenos tão diversos, mas com poucas ocorrências em cada um deles, torna-se difícil tentar encontrar sistematizações de estruturas que poderiam justificar tais desvios.

5.8.2 Casos particulares de uso de crase

As regras de uso do acento indicativo de crase são, de forma geral, bem estabelecidas pelas gramáticas, especialmente no que se refere aos casos em que o acento não deve ocorrer. Aquelas ocorrências ligadas a questões de regência já foram analisadas na Seção 5.4.8 (p. 117), mas há alguns casos de excesso de crase que são particulares dessa categoria e que, por isso, não se inseriram em nenhum dos fenômenos descritos anteriormente. Tais casos se referem em grande medida a grupos de palavras ou a características das palavras que impedem a aplicação da crase, e serão foco de análise desta seção.

O caso mais típico em que a contração da preposição *a* e do artigo *a*, resultando na aplicação do acento indicativo de crase, não pode ocorrer é diante de palavras masculinas. Essa é a questão considerada mais evidente de desvio de crase, uma vez que não é possível usar o artigo *a*, que antecede substantivos femininos, diante de palavras de gênero masculino sem incorrer em problemas de concordância. Porém, tal desvio ocorreu 16 vezes na amostra, o que pode ser um indício de que os produtores de textos não têm certeza quanto a função da crase:

(238) Muitos fatores implicam quando o assunto trata-se do aumento de criminalidade através dos presídios (…) , onde o presidiário não tem acesso <á> saneamento básico (…) .

(239) É necessário também , que o Poder Judicial crie cadastros <à> estes grupos , para que estes , ganhem visibilidade perante a lei e possam ter uma maior fiscalização de sua legitimidade , (…) .

No exemplo (238), o acento utilizado foi o agudo[66], e esse foi o único caso de uso de crase diante de palavra masculina no singular sem elementos intercalados. Destaca-se ainda o fato de o desvio se dar numa estrutura em que o termo regente é, na verdade, uma construção

[66] Conforme as orientações da diretriz de anotação, anotaram-se os casos de *á* como sendo craseados.

com verbo-suporte (*ter acesso a*). No exemplo (239), o substantivo masculino que segue está no plural. Nesse caso, diferentemente daqueles de ausência de crase descritos na Seção 5.4.8 (p. 117), o número do substantivo parece ter relação com o desvio, porque a maioria dos casos envolveu substantivos no plural. Ressalta-se o uso do pronome demonstrativo entre o *a* craseado e o substantivo masculino no plural. Esse tipo de desvio em que havia um elemento intercalando o termo regente e o regido ocorreu várias vezes, e os elementos intercalados foram pronomes demonstrativos, possessivos, indefinidos, entre outros.

Também se identificaram quatro ocorrências de uso de crase diante de palavras femininas no plural, mas em que o *a* craseado permaneceu no singular. Destes, um ocorreu na construção *defronte a*, um foi na construção *quanto a*, um esteve relacionado a uma construção com verbo-suporte (*dar atenção a*) e um ocorreu na construção passiva *estar relacionado a*. Nove ocorrências de excesso de crase diante de palavras femininas envolveram artigos indefinidos ou pronomes pessoais, demonstrativos e indefinidos:

(240) Adquiri-se assim , a capacidade de tomar decisões , tendo a objetividade de fazer uma reflexão sobre alguma causa social , ou até mesmo se juntar <à> ela .

(241) E que por conseguinte , ocasiona uma autoestima afetada , a desvalorização de seu povo , assim como de suas capacidades e competências enquanto indivíduo pertencente <à> uma sociedade (…).

Houve ainda algumas ocorrências de crase após uma preposição diferente de *a*, como *para* e *com*. Por fim, ocorreram 11 casos de uso de crase em estruturas em que não poderia ocorrer uma preposição, mas sim apenas o artigo definido, e duas aplicações de crase sobre o pronome oblíquo *a*. A diversidade de desvios de crase reforça a hipótese de que as suas funções não são totalmente compreendidas ou causam incerteza para os produtores de textos em formação. Isso poderia se justificar, assim como no caso da acentuação gráfica e da pontuação, pelo fato de ser um fenômeno restrito à modalidade escrita da língua, que só começa a ser aprendida durante as etapas da educação formal.

5.8.3 Casos específicos envolvendo pronomes

Pronomes foram um grupo de palavras associado a diversos tipos de desvios sintáticos. Entretanto, há alguns casos específicos dessa categoria que não puderam ser sistematizados em nenhum dos fenômenos descritos anteriormente. Um desses casos se refere ao uso de pronomes no papel de objetos. Em português, somente os pessoais retos podem exercer a função de sujeito ou predicativo, e somente os oblíquos podem assumir o papel de complemento ou objeto. Em 11 ocorrências, porém, os primeiros foram usados onde só poderiam aparecer os segundos:

(242) (…) o governo como um todo desvalorizam essas pessoas , deixando <elas> largadas como se fosse ninguém .

Também ocorreu, em poucas sentenças, a troca do pronome em termos da pessoa equivocada, especialmente em verbos reflexivos ou recíprocos:

(243) (…) , ou seja , está praticamente definido que no mínimo vamos <se> aposentar aos 65 anos .

Em termos de uso da partícula ou do pronome *se*, os desvios ainda englobaram a sua ausência em casos de ocorrência obrigatória, tanto como marcador de reflexividade ou reciprocidade quanto como indeterminador de sujeito:

(244) Atualmente , a sociedade tem <manifestado> sobre os direitos da Previdência Social (…) .

(245) Claro , é importante agir com cautela , pois também não adianta sair para as ruas em busca de mudanças , sendo que no dia a dia <não põe> em prática a cidadania e a humanidade .

O contrário também contou com algumas ocorrências. Nesses casos, o *se* foi usado como pronome reflexivo com verbos sem sentido reflexivo na sentença, ou como índice de indeterminação de sujeito quando o sujeito já havia sido explicitado ou em casos de verbos que não permitem sujeito algum, como o verbo *haver*:

(246) (…) , com isso estão "<se> resistindo" e expondo sua opinião para tentar uma melhora .

(247) (…) , pois , protestos pacíficos quase não <se> há atenção do governo , entretanto , (…) .

(248) No entanto <deve-se> ocorrer protestos no qual possam propor as suas próprias soluções (…) .

A última questão ligada ao uso de pronomes se deve à colocação pronominal no caso mais clássico de palavra atrativa: a negação. Sabe-se que a questão da colocação de pronomes se deve a fenômenos que vão além da sintaxe pura, como comentado anteriormente. No entanto, o caso mais típico de atratividade de pronomes são as negações, que tornam obrigatório o uso de próclise. Casos de ênclise após palavra negativa ocorreram apenas quatro vezes no conjunto analisado (p. e., em "*Não <deve-se> impor uma ideia* (…)"), o que indica que os produtores de textos provavelmente estão cientes dessa regra de colocação de pronomes.

5.8.4 Preposições e suas particularidades

Entre os desvios ligados ao uso de preposições que não se encaixam em fenômenos anteriores, destacam-se as ausências das preposições *para* e *em*:

(249) <Início> de século XXI houve perdas de referências artísticas , frente à tecnologia , poucas pessoas possuem o interesse pela arte e a cultura consequente .

(250) <Já> os que cometeram crimes como (Assaltos) a penalidade poderia ser reduzida de 5 anos para 3 anos , e também terão punições , e terão que trabalhar muito .

Chamam atenção os demais desvios das estruturas sintáticas das sentenças, indicando que o problema não se refere apenas a desatenção ou esquecimento da preposição.

Outro desvio de preposição foi o uso de locução inexistente:

(251) O brasileiro está sendo visto dessa forma para que não volte essa escravidão <contra ao> pessoal trabalhador , e vão as ruas para poder deixar uma forma de que eles possam ter voz sim (…).

(252) <Desde da> idade da pedra a arte é usada para marca os principais momentos da sociedade , um ótimo exemplo seria o surgimento do fogo .

Por fim, o último desvio ligado especificamente ao uso de preposição diz respeito à sua colocação entre elementos que são obrigatoriamente contínuos, por exemplo, entre os verbos que compõem uma locução verbal:

(253) Isso ocorria para que a população não desenvolvesse o senso crítico e não pudessem ter ideias e atitude que poderiam <para> modificar as situações difíceis em que eram obrigados a viver .

Também se identificaram alguns casos de repetição da mesma sílaba inicial da palavra seguinte, dando indícios de desvio decorrente de erro de digitação:

(254) As paralisações e protestos são para defender os direitos de seres humanos , e sendo assim não deixam as atitudes de quem estar no poder sempre a se <de> desejar , (…)

As análises linguísticas propostas neste capítulo pretendem servir como uma fonte inicial de descrição dos fenômenos envolvidos na ocorrência de desvios sintáticos. Assim, buscou-se sistematizar os fenômenos de acordo com aquilo que mais pareceu relevante ou interessante em termos linguísticos. Outras abordagens, é claro, são possíveis, e outros caminhos poderiam ter sido escolhidos para este capítulo. Espera-se que a contribuição desta pesquisa seja sobretudo fomentar novas possibilidades, especialmente em termos de descrição e análise linguística no contexto do Processamento Automático das Línguas Naturais, para que se possa oferecer subsídios linguísticos ao desenvolvimento de ferramentas para além dos métodos puramente computacionais e pouco interessados na língua que vêm ganhando cada vez mais espaço no cenário nacional e internacional.

6 EXTRAÇÃO DE ATRIBUTOS E CORRELAÇÃO COM OS DESVIOS VIA AM

Como já foi dito no capítulo metodológico desta dissertação, a identificação e a extração de atributos linguísticos proporcionaram uma maior compreensão dos fenômenos em que os desvios ocorrem, permitindo comparar em termos numéricos como se comportam esses atributos em sentenças com e sem desvio. Além disso, os atributos foram utilizados como entrada dos algoritmos para o Aprendizado de Máquina Supervisionado, de forma a explicitar possíveis correlações entre atributos e presença de desvios sintáticos. Este capítulo apresenta os resultados obtidos em ambas as tarefas: a extração dos atributos, na primeira seção, e o AM, na segunda seção do capítulo.

6.1 *Extração automática de atributos e as comparações possíveis*

A seguir, apresentam-se os resultados da extração dos atributos a partir da comparação de ocorrências em sentenças com e sem desvio. Para essa análise, foi utilizado o *corpus* de treino, com 2.307 sentenças sem desvio e 6.347 sentenças com desvio, apresentando-se os dados tanto em termos numéricos quanto em percentuais. Iniciando pelo atributo referente ao número de *tokens*, mostra-se na Tabela 9 como se distribuem os desvios em cinco intervalos:

Tabela 9 – Número de *tokens* (n_tokens).

Intervalo	Desvio	Sem desvio
1 – 25	1.549 (24,41%)	1.241 (53,79%)
26 – 50	3.039 (47,88%)	957 (41,48%)
51 – 100	1.603 (25,26%)	109 (4,72%)
101 – 150	135 (2,13%)	0 (0%)
> 150	21 (0,33%)	0 (0%)
Total	**6.347**	**2.307**

Vê-se que mais da metade das sentenças sem desvio possuem até 25 *tokens*. Além disso, a partir de 51 *tokens*, a prevalência de sentenças com desvio é nítida, sendo que não há nenhuma sentença sem desvio nos dois intervalos finais. Esses dados indicam que sentenças mais longas têm mais probabilidade de apresentar desvios do que as que são mais curtas. A próxima seção mostrará também que esse atributo é significativo para os resultados da etapa de AM.

O atributo seguinte se refere ao número de *tokens* que ocorrem antes da raiz (*root*) da sentença. Na Tabela 10, tem-se a distribuição desse atributo em termos numéricos e percentuais nas sentenças com e sem desvios, divididos em seis intervalos.

Tabela 10 – Número de *tokens* até a raiz (n_tokens_root).

Intervalo	Desvio	Sem desvio
0	555 (8,74%)	137 (5,94%)
1 – 5	2.599 (40,95%)	1.111 (48,16%)
6 – 10	1.446 (22,78%)	597 (25,88%)
11 – 20	1.126 (17,74%)	357 (15,47%)
21 – 30	354 (5,58%)	86 (3,73%)
> 31	266 (4,19%)	19 (0,82%)

Nota-se uma distribuição mais regular desse atributo em relação ao anterior, com diferenças menos significativas entre sentenças com e sem desvio em cada um dos intervalos. Porém, mais da metade da classe das sentenças sem desvio possui até cinco *tokens* antes da raiz, enquanto esse é o caso para um pouco menos da metade das sentenças com desvio, como mostram os percentuais. A diferença mais significativa ocorre no último intervalo, em que a maioria das sentenças com mais de 31 *tokens* antes da raiz apresenta desvios.

Na Tabela 11, descrevem-se os valores e percentuais referentes ao número de formas verbais finitas, divididos em cinco intervalos:

Tabela 11 – Número de verbos finitos (n_vfin).

Intervalo	Desvio	Sem desvio
0	228 (3,59%)	40 (1,73%)
1	1.054 (16,61%)	849 (36,80%)
2 – 5	4.261 (67,13%)	1.389 (60,21%)
6 – 9	693 (10,92%)	29 (1,26%)
> 9	111 (1,75%)	0 (0%)

Aqui, nota-se uma distribuição similar no terceiro intervalo, mas as sentenças sem desvio apresentam um menor número de verbos finitos do que as com desvio. Nota-se que 40 sentenças foram classificadas como não contendo desvio, mesmo não tendo nenhum verbo conjugado. É provável que sejam sentenças constituídas de apenas uma palavra, que alguns teóricos chamam de frase. Os resultados do AM na próxima seção mostram que o número de verbos finitos é decisivo para a presença de desvios, segundo os algoritmos testados.

Na Tabela 12, descrevem-se os valores e percentuais referentes ao número de formas verbais infinitas (infinitivo, gerúndio e particípio):

Tabela 12 – Número de formas verbais infinitas (n_vinf).

Intervalo	Desvio	Sem desvio
0	905 (14,26%)	631 (27,35%)
1	1.367 (21,54%)	665 (28,83%)
2 – 5	3.379 (53,24%)	974 (42,22%)
6 – 9	610 (9,61%)	36 (1,56%)
> 9	86 (1,35%)	1 (0,04%)

Na tabela, pode-se perceber que, em sentenças com mais de cinco formas verbais infinitas, prevalecem aquelas que possuem desvios. É provável que, nesses casos, sejam sentenças longas e com várias orações aglutinadas. Nota-se também que sentenças sem desvio têm um percentual maior de ocorrência de no máximo uma forma verbal infinita. Também se vê que é maior o percentual de sentenças sem desvio que não têm nenhuma forma verbal infinita.

Na Tabela 13, veem-se os resultados numéricos e percentuais do número de vírgulas das sentenças, divididos em seis intervalos:

Tabela 13 – Número de vírgulas (n_commas).

Intervalo	Desvio	Sem desvio
0	854 (13,46%)	617 (26,74%)
1	1.465 (23,08%)	790 (34,24%)
2 – 3	2.418 (38,10%)	730 (31,64%)
4 – 5	1.053 (16,59%)	147 (6,37%)
6 – 9	490 (7,72%)	23 (1%)
> 9	67 (1,06%)	0 (0%)

Entre as sentenças sem nenhuma vírgula, o percentual das que não contêm desvio é maior do que as que contêm desvio. É provável que se trate de sentenças curtas, que sabidamente costumam apresentar menor probabilidade de conter desvio sintático. Conforme aumenta o número de vírgulas, o percentual de sentenças com desvio é maior do que o das sem desvio, sendo que no último intervalo há apenas sentenças com desvio.

Na Tabela 14, têm-se os resultados em função do tipo de sentença: simples (com apenas um verbo finito) e composta (com mais de um verbo finito). A última coluna traz o total de sentenças simples e compostas do *corpus* de treino, bem como o respectivo percentual:

Tabela 14 – Tipo de sentença (sent_type).

Intervalo	Desvio	Sem desvio	Total
Simples	1.282 (20,20%)	889 (38,53%)	2.171 (25,08%)
Composta	5.065 (79.80%)	1.418 (61,47%)	6.483 (74,91%)
Total	**6.347**	**2.307**	**8.654**

Vê-se que quase três quartos das sentenças do *corpus* têm mais de uma forma verbal finita, contrariando alguns manuais e orientações disponíveis na internet, que sugerem a construção de sentenças simples e curtas. Dessa forma, pode-se inferir que os produtores dos textos escrevem sentenças com certa complexidade, tentando encadear e conectar mais de uma frase. Porém, o percentual das simples que contêm desvio é menor do que o das que não contêm. Da mesma forma, há mais sentenças compostas com desvio, em termos percentuais, do que as sem desvio. Isso sugere que as tentativas dos estudantes de construir sentenças maiores e possivelmente mais complexas têm falhado com bastante frequência, sugerindo um problema na aprendizagem desse tipo de estrutura na escrita.

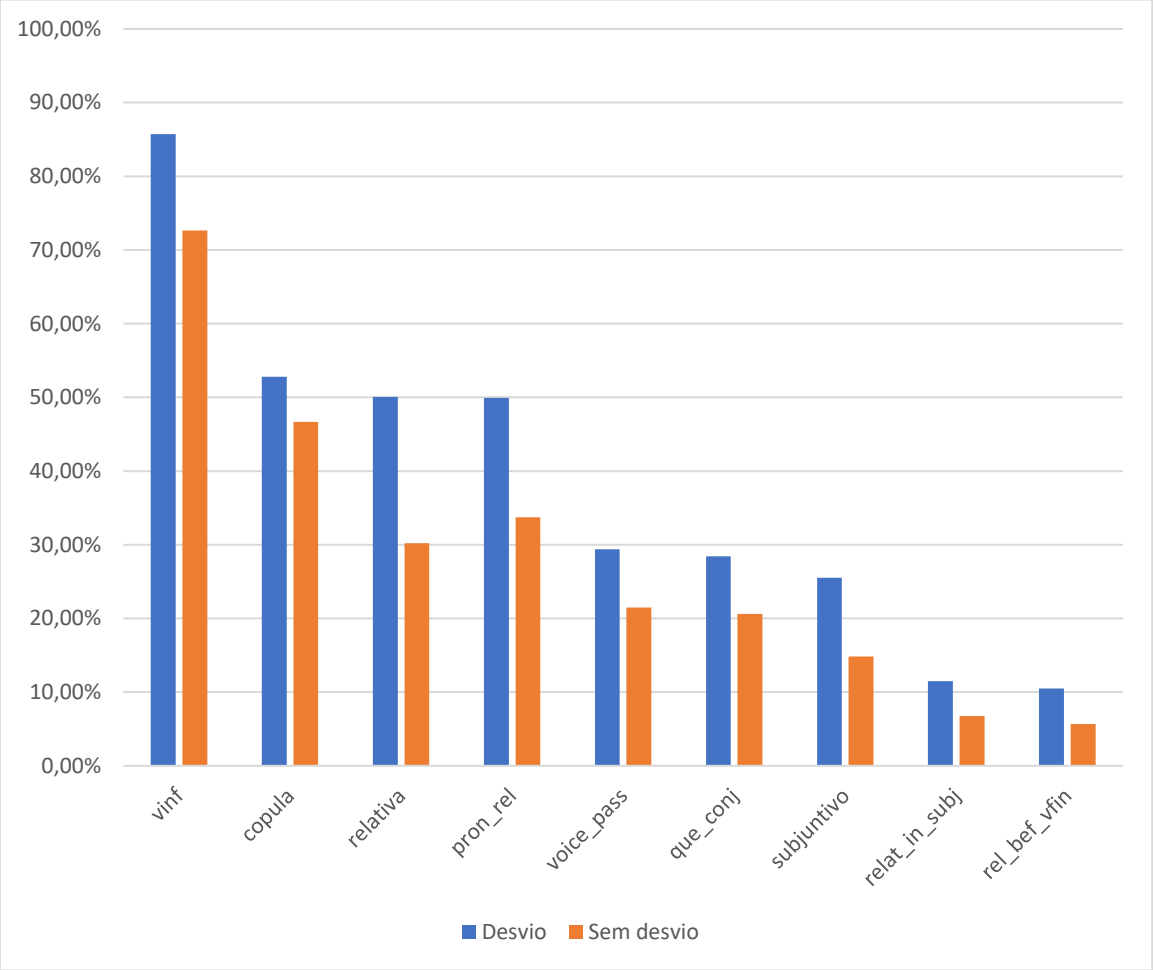
Na Tabela 15, apresentam-se os resultados numéricos e percentuais da presença dos atributos que foram extraídos a partir da sua ocorrência ou não nas sentenças:

Tabela 15 – Presença de determinados atributos.

Intervalo	Desvio	Sem desvio
Cópula (copula)	3.350 (52,78%)	1.077 (46,68%)
Subjuntivo (subjuntivo)	1.620 (25,52%)	342 (14,82%)
Passiva (voice_pass)	1.866 (29,40%)	496 (21,50%)
Formas verbais infinitas (vinf)	5.442 (85,74%)	1.676 (72,65%)
Pronome relativo (pron_rel)	3.168 (49,91%)	778 (33,72%)
Pron. rel. antes VFin (rel_bef_vfin)	666 (10,49%)	131 (5,68%)
Conjunção *que* (que_conj)	1.806 (28,45%)	476 (20,63%)
Oração relativa (relativa)	3.176 (50,04%)	697 (30,21%)
Oração relativa no sujeito (relat_in_subj)	728 (11,47%)	156 (6,76%)

No Gráfico 4, é possível visualizar o comparativo dos percentuais da ocorrência de tais características sintáticas nas sentenças com e sem desvio, ordenadas por frequência:

Gráfico 4 – Presença de determinadas características: percentual.



O gráfico evidencia que todos os atributos estão mais presentes proporcionalmente em sentenças com desvio do que naquelas que não apresentaram desvio sintático. Vale lembrar, conforme os resultados da Seção 4.3.4 (p. 76), que a anotação já demonstrou uma ocorrência significativa de desvios relacionados às formas verbais (terceira categoria de desvios mais frequente), e que a subcategoria específica das formas verbais infinitas foi a 11ª mais frequente. Ressalta-se ainda que a subcategoria *Segment* também está muitas vezes relacionada à presença de formas verbais infinitas, já que identifica, por exemplo, sentenças que contenham exclusivamente essas formas. Essa subcategoria é a sexta mais frequente.

No atributo relacionado à presença de cópula, as diferenças em sentenças com e sem desvio são menores. O mesmo vale para as relativas dentro do sujeito. Já a grande frequência de orações relativas com desvio pode servir como um indicativo de que os produtores de textos têm dificuldade de produzir sentenças complexas. Da mesma forma, como esse tipo de sentença aumenta a complexidade em termos de leitura, conforme afirmam Scarton e Aluísio (2010), esses resultados parecem dar indícios de que se trata de um fenômeno mais complexo também

da perspectiva da produção textual. O mesmo vale para os atributos da voz passiva, do *que* como conjunção e da presença de subjuntivo.

A Tabela 16 apresenta os resultados referentes à profundidade da árvore sintática, isto é, ao número de núcleos de níveis diferentes que contêm dependentes, somando-se a raiz. Em princípio, quanto mais profunda a árvore, mais complexas as relações sintáticas estabelecidas:

Tabela 16 – Profundidade da árvore sintática (tree_depth).

Intervalo	Desvio	Sem desvio
0 – 4	401 (6,32%)	458 (19,85%)
5 – 7	2.733 (43,06%)	1.338 (58%)
8 – 10	1.969 (31,02%)	425 (18,42%)
11 – 13	817 (12,87%)	74 (3,21%)
14 – 17	330 (5,20%)	11 (0,48%)
> 17	97 (1,53%)	1 (0,04%)

Vê-se que, no primeiro intervalo, predominam as sentenças sem desvio. Já a partir do terceiro intervalo, o percentual de sentenças com desvio é bem maior. Isso mais uma vez pode indicar que as sentenças com maior profundidade também são as mais longas. Esses resultados sugerem também que a maior complexidade da estrutura sintática está positivamente correlacionada com a presença de desvio sintático.

Por fim, a Tabela 17 traz os resultados referentes à distância média entre elementos dependentes. Isso é calculado somando-se a distância entre cada elemento dependente e dividindo-se esse resultado pelo número total de dependências. Assim, por exemplo, elementos adjacentes têm distância 1, os dependentes intercalados por um elemento têm distância 2, e assim por diante:

Tabela 17 – Distância média entre dependentes (av_dep_lenght).

Intervalo	Desvio	Sem desvio
0 – 1,99	637 (10,04%)	538 (23,32%)
2,00 – 2,50	2.540 (40,02%)	1.087 (47,12%)
2,51 – 3,00	2.052 (32,33%)	515 (22,32%)
3,01 – 4,00	999 (15,74%)	157 (6,81%)
> 4,00	119 (1,87%)	10 (0,43%)

Nesse atributo, vê-se que há um percentual bem maior de sentenças sem desvio no primeiro intervalo do que aquelas com desvio. O segundo intervalo não difere tanto em termos de percentual, mas os três últimos indicam que as sentenças com desvio têm distância média entre

dependentes maior. Isso sugere que a hipótese de que uma distância maior entre elementos que estabelecem relações de dependência pode ser corroborada. A partir das análises linguísticas do Capítulo 5, em que as dependências de longa distância se fizeram presentes nos mais diversos tipos de desvios, e dos resultados da Tabela 17, é possível afirmar que de fato uma maior distância entre dependentes complexifica a estrutura e está positivamente correlacionada com a presença de desvio sintático.

Salienta-se novamente que a extração de tais atributos foi realizada com base nas análises de ferramentas automáticas que não são projetadas para lidar com sentenças com uma presença massiva de desvios sintáticos. Logo, é preciso encarar tais resultados com ressalvas. Como perspectiva futura, seria interessante investigar o impacto dos desvios nos resultados das ferramentas computacionais utilizadas (p. ex. *POS tagger* e *parser*).

6.2 *Resultados do Aprendizado de Máquina*

Para a etapa de AM, foram realizados dois experimentos de classificação e um de seleção de atributos. Esta seção descreve os resultados obtidos com os algoritmos testados, iniciando pelos experimentos de classificação.

Experimento 1 – *corpus* balanceado

Nesse experimento, utilizou-se um *corpus* balanceado com o mesmo número de sentenças com e sem desvio (2.307 de cada tipo). O objetivo era testar inicialmente como se comportavam os algoritmos, de modo a ter uma base de comparação e verificar, posteriormente, se o fornecimento de mais instâncias poderia fazer diferença nos resultados. Para isso, a partir do *upload* do arquivo ARFF com os 18 atributos e a classe de cada instância no *software Weka*, testaram-se oito algoritmos de classificação clássicos, com validação cruzada de 10 conjuntos. Isso significa que o *corpus* é dividido em 10 conjuntos de instâncias, e que o algoritmo é treinado em nove deles e testado no décimo, alternando-se 10 vezes entre conjuntos de treino e de teste. Em todos os algoritmos testados, mantiveram-se os parâmetros padrão sugeridos pelo *Weka*. A Tabela 18 abaixo traz a acurácia de cada um dos algoritmos testados:

Tabela 18 – Resultados da classificação no *corpus* balanceado.

Algoritmo	Acurácia
Logistic	**68,68**
Random Forrest	68,51
JRip	68,09
SMO	68,05
Naïve Bayes	67,77
One-R	67,75
MLP	66,66
J48	65,23

Nota-se na Tabela 18 que os três algoritmos com os melhores resultados se inserem em paradigmas diferentes: o *Logistic Regression* pertence ao paradigma matemático, enquanto o *Random Forrest* e o *JRip* pertencem ao simbólico. Para verificar se a inserção de mais dados poderia melhorar os resultados, o Experimento 2, descrito a seguir, utilizou todas as sentenças do *corpus* de treino.

Experimento 2 – *corpus* de treino total

Para o segundo experimento, testaram-se os mesmos oito algoritmos. Também foi utilizada a validação cruzada em 10 conjuntos, e mantiveram-se os parâmetros padrão do *Weka*. A diferença foi que o número de instâncias era maior, com 6.347 sentenças com desvio e 2.307 sem desvio, totalizando 8.654 instâncias. A Tabela 19 apresenta os resultados obtidos na classificação:

Tabela 19 – Resultados da classificação no *corpus* de treino.

Algoritmo	Acurácia
Random Forrest	**74,87**
Logistic	74,54
JRip	74,31
MLP	74,1
One-R	73,77
J48	73,74
SMO	73,34
Naïve Bayes	63,03

Nota-se na Tabela 19 que os três algoritmos com os melhores resultados são os mesmos do experimento anterior, invertendo-se apenas a ordem entre os dois primeiros. Já entre os demais algoritmos, há uma mudança considerável na ordem de desempenho em ambos os experimentos. O SMO e o *Naïve Bayes*, por exemplo, eram o quarto e o quinto com melhor

desempenho no *corpus* balanceado, mas caíram para as últimas posições no *corpus* de treino. Porém, o mais interessante a se notar é que o único algoritmo que teve desempenho pior no *corpus* do experimento 2, comparado com o experimento 1, foi o *Naïve Bayes*. Houve um ganho significativo em acurácia ao se utilizar mais dados anotados, o que reforça o fato de que *corpora* anotados manualmente são essenciais para os trabalhos que utilizam o Aprendizado de Máquina Supervisionado.

Após esse treinamento, os três melhores algoritmos foram aplicados ao *corpus* de testes. Os resultados obtidos foram similares, mas houve uma pequena melhora no desempenho de dois dos algoritmos. O algoritmo *Logistic Regression* obteve o melhor desempenho, com acurácia de 75,62%, seguido pelo *Random Forrest*, com acurácia de 75,07%. Já o algoritmo *JRip* apresentou queda no desempenho, com um resultado de 73,17%.

Para esta tarefa, não se tinha um *baseline* com o qual os resultados poderiam ser comparados, uma vez que não se encontraram trabalhos que propusessem uma classificação binária semelhante à que foi feita aqui. Considera-se que tais resultados foram satisfatórios à pesquisa, mas, comparando-se com os classificadores descritos na revisão da literatura (Seção 3.2, p. 47), que se propõem a tarefas muito mais complexas, servem apenas para ilustrar que a tarefa não é trivial. Uma vez que a proposta da pesquisa não é desenvolver um classificador, não se aprofundará muito esse experimento no que se refere à análise de erros, matriz de confusão, etc. O que se pode cogitar é que os atributos linguísticos sejam insuficientes e/ou inadequados para a proposta, o que abre espaço para trabalhos que estendam ou especializem a lista de atributos linguísticos desta pesquisa.

Entre as possibilidades de atributos elencadas durante a pesquisa, mas que, por restrições de tempo e de ferramental, não puderam ser postas em prática, citam-se (i) a presença de sujeito posposto (que não pôde ser extraída porque o *parser* etiqueta esses casos como objetos), (ii) inversões da ordem canônica, (iii) presença de aposições, (iv) número de *tokens* do sujeito e dos objetos, (v) número de modificadores ligados a cada núcleo, (vi) número de complementos preposicionais. Sugere-se ainda uma pesquisa mais aprofundada na literatura, abrangendo os trabalhos que investigam questões de complexidade textual, a fim de identificar atributos que possam se mostrar relevantes também a estudos como o que foi proposto aqui.

Experimento 3 – Seleção de atributos

A principal hipótese que se pretendia corroborar ou refutar com esta pesquisa era a correlação entre determinados atributos linguísticos e a presença de desvios sintáticos. Isso foi feito, em partes, por meio da análise linguística manual descrita no Capítulo 5, e também pela análise

manual da extração dos atributos realizada na Seção 6.1 (p. 151). A intenção de executar algoritmos de AM para evidenciar a existência ou não de tais correlações é ir além das possibilidades das análises manuais, possivelmente enviesadas por todo o percurso da pesquisa.

Assim, para verificar se há determinados atributos linguísticos que são mais relevantes para que os algoritmos decidam entre classificar uma sentença como tendo ou não tendo desvio sintático, utilizaram-se dois algoritmos de seleção de atributos. A etapa de identificação e seleção de atributos relevantes também é chamada de *engenharia de features*, e demanda muito tempo e conhecimento especializado. Uma seleção inadequada de atributos pode impactar significativamente os resultados, tanto quando eles são insuficientes para descrever o problema proposto, quanto quando se sobrepõem ou são excessivos. Um exemplo de problema que pode ser gerado é o *overfitting*, quando os atributos são adaptados demais ao conjunto de dados de treino, não podendo ser utilizados (ou gerando resultados muito inferiores) em conjuntos de dados semelhantes. Sabe-se que os atributos escolhidos para essa tarefa provavelmente precisariam ser analisados com mais cautela, mas esse experimento busca, antes de verificar a adequação dos atributos à etapa de classificação, investigar como eles se comportam em termos de relevância.

O primeiro algoritmo para o ranqueamento dos atributos foi o *InfoGainAttributeEval*, que mede o ganho de informação de cada um dos atributos. Ele foi aplicado no *corpus* de treino, no qual se obtiveram os resultados mostrados na Tabela 20:

Tabela 20 – Resultados da seleção de atributos via *InfoGainAttributeEval*.

Atributo	Ganho de informação
1 n_tokens	0,10312
16 tree_depth	0,07572
4 n_vfin	0,07473
6 n_commas	0,05475
5 n_vinf	0,04447
17 av_dep_length	0,0415
2 sent_type	0,02389
14 relativa	0,023
10 vinf	0,01547
11 pron_rel	0,01516
3 n_tokens_root	0,01322
8 subjuntivo	0,00983
13 que_conj	0,0046
9 voice_pass	0,00458
12 rel_bef_vfin	0,00428
15 relat_in_subj	0,00368
7 copula	0,0021

Destaca-se, nessa tabela, que os atributos considerados mais relevantes por esse algoritmo, que ranqueia todos os atributos conforme o ganho de informação, são os que envolvem contagens genéricas semelhantes às realizadas por Fonseca *et al.* (2018). Os atributos linguísticos que se esperava estarem mais correlacionados com a presença de desvios estão entre os últimos, como presença de oração relativa dentro do sujeito, presença de voz passiva e atributos ligados à presença de orações subordinadas, como o *que* com *POS* de conjunção. Um resultado semelhante a esse pôde ser obtido a partir da análise manual dos atributos, como mostrado na Seção 6.1 (p. 151).

Esses resultados mostram que, pelo menos na análise desse algoritmo, a hipótese inicial é parcialmente refutada. Porém, ao identificar a profundidade da árvore sintática como o segundo atributo com maior ganho de informação, o resultado do algoritmo parece sugerir que sentenças mais complexas, isto é, com mais núcleos e subnúcleos abaixo da raiz, estão mais fortemente correlacionadas com a presença de desvios. Isso leva ao questionamento de que pode ser necessário investigar a questão da complexidade para além dos fenômenos considerados complexos, mas da sua interação com outros fenômenos na sentença. Além disso, reforça-se a questão levantada anteriormente de que seria interessante utilizar índices de complexidade e outros atributos relacionados para a correlação pretendida aqui.

O segundo algoritmo testado foi o *CfsSubsetEval*, que seleciona um conjunto de atributos relevantes entre todos os disponíveis. Os resultados com validação cruzada são mostrados na Tabela 21:

Tabela 21 – Resultados da seleção de atributos via *CfsSubsetEval*.

Número de conjuntos (%)	Atributo
10 (100%)	**1 n_tokens**
0 (0%)	2 sent_type
0 (0%)	3 n_tokens_root
10 (100%)	**4 n_vfin**
8 (80%)	**5 n_vinf**
10 (100%)	**6 n_commas**
0 (0%)	7 copula
0 (0%)	8 subjuntivo
0 (0%)	9 voice_pass
2 (20%)	10 vinf
0 (0%)	11 pron_rel
0 (0%)	12 rel_bef_vfin
0 (0%)	13 que_conj
10 (100%)	**14 relativa**
0 (0%)	15 relat_in_subj
10 (100%)	**16 tree_depth**
10 (100%)	**17 av_dep_length**

A seleção no *corpus* de treino foi de sete atributos, indicados em negrito na Tabela 21: (i) número de *tokens*, (ii) número de verbos finitos, (iii) número de verbos infinitos, (iv) número de vírgulas, (v) presença de orações relativas, (vi) profundidade da árvore sintática e (vii) distância média entre dependentes. Nesse caso, vê-se que um atributo está associado à hipótese inicial de que orações relativas estão mais fortemente correlacionadas à presença de desvio sintático.

Também se analisou qual foi o atributo escolhido pelo algoritmo *One-R*, que define apenas uma regra para as decisões tomadas, com base no atributo mais relevante. O atributo escolhido foi a distância média entre os elementos dependentes. Então, excluiu-se esse atributo e rodou-se o algoritmo novamente, obtendo uma acurácia de 73,43%, sendo o atributo escolhido o número de *tokens*. Removendo-se também esse atributo, obteve-se o melhor resultado desse classificador (73,95% de acurácia), com o atributo da profundidade da árvore sintática como o escolhido para a classificação.

Vê-se que, de maneira geral, os algoritmos de seleção de atributos apresentam resultados similares em termos dos atributos que parecem ser os mais relevantes. Não é surpresa que o

número de *tokens* esteja fortemente correlacionado com a presença de desvios sintáticos, uma vez que, quanto maior a sentença, mais complexas se tornam as relações. Também é influenciado por esse aspecto o fato de que tais sentenças costumam consistir em sequências de orações concatenadas, separadas por vírgulas ou coordenações, como já mostrou a análise linguística do Capítulo 5.

Ressalta-se também a importância das formas verbais na presença de desvios, sejam elas finitas ou infinitas. Ao que parece, para os algoritmos, quanto mais formas verbais uma sentença tem (o que também se relaciona com o seu comprimento), maior a chance de haver um desvio sintático.

Por fim, destaca-se ainda a importância dos dois últimos atributos acrescentados à extração automática dos desvios: a profundidade da árvore sintática e a distância média entre dependentes. A partir desse resultado, pode-se inferir que uma sentença mais longa, com uma árvore mais profunda e com mais dependências de longa distância é de difícil construção, mostrando-se um desafio aos produtores de textos e provavelmente também às ferramentas de PLN.

7 CONCLUSÃO

O estudo apresentado aqui investigou os desvios sintáticos presentes em redações nos moldes do ENEM produzidas por estudantes que estão cursando o Ensino Médio. Para isso, expandiu-se a tipologia de desvios proposta por Pinheiro (2008) e propôs-se uma metodologia de anotação de desvios no nível da sentença em duas fases. Na primeira, classificaram-se as sentenças entre contendo e não contendo desvio. A seguir, uma parcela de 2.500 sentenças teve os seus desvios tipificados conforme a tipologia proposta. Dessa forma, um dos recursos gerados por esta pesquisa foi um *corpus* com 10.652 sentenças classificadas em com desvio e sem desvio, e 7.290 desvios sintáticos anotados em 27 subcategorias. O *corpus* também foi *parseado* automaticamente com o *parser UDPipe*, e extraíram-se 17 atributos linguísticos para posterior correlação com a ocorrência de desvios via AM. Também se realizou a análise e sistematização de diversos fenômenos linguísticos nos quais esses desvios ocorrem, o que pode ser utilizado tanto para o desenvolvimento e o aprimoramento de ferramentas de PLN quanto para o aperfeiçoamento das habilidades de escrita e fonte de consulta para estudantes e professores.

Entre os principais resultados, identificou-se que cerca de três quartos das sentenças do *corpus* contêm desvio sintático, e que os mais frequentes são os de pontuação e os de concordância. Também se verificou que alguns dos atributos observados parecem estar fortemente correlacionados com a presença de desvios, enquanto outros não se mostraram tão relevantes.

A seguir, elencam-se as hipóteses iniciais da pesquisa e as respectivas validações:

- Existe correlação entre determinadas construções sintáticas e a presença de desvios sintáticos em redações escritas por estudantes do Ensino Médio: a hipótese inicial foi validada, visto que foi possível identificar diversos fenômenos sintáticos correlacionados à ocorrência de desvios sintáticos nas sentenças estudadas.
- Estruturas de coordenação e subordinação estão positivamente associadas à ocorrência de desvios: tanto a análise manual dos fenômenos quanto o AM corroboraram a hipótese, indicando que tais estruturas estão positivamente correlacionadas à presença de desvios.
- A voz passiva está positivamente associada à presença de desvios: tal hipótese foi parcialmente validada nas análises manuais, mas as passivas não foram um atributo relevante para os algoritmos de AM.

- Sentenças com maior número de tokens têm maior probabilidade de conterem desvios sintáticos do que aquelas que têm menos tokens: o tamanho da sentença foi o atributo mais relevante para o AM e se mostrou fortemente correlacionado à presença de desvios também nas análises manuais.
- Sentenças cujas dependências sintáticas entre elementos são mais distantes (isto é, em que há vários tokens intercalados entre dois elementos dependentes diretamente um do outro) apresentam maior probabilidade de conterem desvios sintáticos do que aquelas em que os dependentes diretos são adjacentes entre si: tal hipótese foi reforçada nas análises manuais dos fenômenos, e corroborada na etapa de AM, uma vez que esse atributo foi relevante para mais de um algoritmo.

Considera-se que o objetivo geral e os objetivos específicos da pesquisa, retomados a seguir, foram cumpridos:

- Investigar a recorrência de desvios sintáticos presentes em redações produzidas por estudantes do Ensino Médio, nos moldes exigidos pelo ENEM, bem como caracterizá-los: a pesquisa forneceu um panorama sobre o número de sentenças com desvio em um conjunto de redações de estudantes, bem como uma descrição abrangente dos fenômenos linguísticos nos quais eles ocorrem.
- Identificar e caracterizar os desvios sintáticos mais frequentes em textos de estudantes: os desvios mais frequentes são os de pontuação (especialmente a sua ausência) e de concordância, com destaque para a concordância verbal.
- expandir a tipologia de desvios gramaticais proposta por Pinheiro (2008), que também analisou desvios em redações de estudantes, e propor uma tipologia de desvios sintáticos abrangente, que possa ser útil para outras tarefas de PLN: a tipologia proposta e utilizada pela pesquisa é composta por 11 categorias e 27 subcategorias de desvios, e se mostrou adequada para a tarefa de anotação manual de desvios sintáticos.
- disponibilizar um *corpus* de redações nos moldes da redação do ENEM, anotado tanto manual quanto automaticamente (via *parsing*) no nível da sentença: construiu-se um *corpus* de redações e anotou-se automaticamente com o *parser* UDPipe, bem como manualmente em duas fases de anotação. O *corpus* pode ser solicitado e será disponibilizado publicamente em repositório adequado após a publicação da dissertação.
- analisar os fenômenos nos quais tipicamente ocorrem os desvios sintáticos encontrados durante a anotação: as análises manuais dos fenômenos cumpriram este objetivo.

- identificar e extrair atributos linguísticos das sentenças do *corpus*: foram extraídos automaticamente 17 atributos linguísticos das sentenças.
- analisar as correlações entre tais atributos linguísticos e a presença de desvios sintáticos, utilizando Aprendizado de Máquina (AM) e algoritmos de seleção de informações: identificou-se a correlação de determinados atributos, especialmente o número de *tokens* e de formas verbais das sentenças.

Na sequência, apresentam-se as principais contribuições da dissertação.

7.1 *Contribuições da pesquisa*

Entre as contribuições da pesquisa, destaca-se como principal a construção e anotação de um *corpus* de redações de estudantes do Ensino Médio com um número considerável de amostras. Esse conjunto de textos se divide em *subcorpora* que foram utilizados para as diversas etapas da pesquisa, conforme mostrou o esquema de divisão da Figura 10, que consta ao final do capítulo metodológico desta dissertação (Seção 2.6, página 82). Destacam-se: i) um *corpus* com 1.045 redações; ii) um *subcorpus* com 10.652 sentenças classificadas em contendo ou não desvio sintático; iii) um segundo *subcorpus* de 2.500 sentenças cujos desvios foram tipificados de acordo com a tipologia proposta. Esses conjuntos secundários também podem ser utilizados por demais pesquisas, conforme as tarefas pretendidas.

Outra das contribuições foi a escolha e a avaliação de diversas ferramentas utilizadas nas etapas que compõem a pesquisa. Inicialmente, escolheu-se o *parser* com base em aspectos como licença de uso livre, bom desempenho reportado, facilidade de uso. Na sequência, fez-se um experimento que explicitou que, em textos com muitos desvios e de um gênero completamente diferente daquele no qual a ferramenta foi treinada, o seu desempenho é questionável, independentemente da correção dos desvios. Também se analisou a melhor plataforma de anotação para a tipificação dos desvios. A decisão novamente teve por base a facilidade de uso e a possibilidade de uso livre, bem como a adaptação das etiquetas à tarefa pretendida.

Além disso, a pesquisa propôs uma metodologia de anotação de desvios em *corpora* de aprendizes. Tal metodologia pode ser adaptada e utilizada para diversos estudos que se ocupem dessa mesma temática. Nesse sentido, a tipologia construída é um recurso que pode ser reaproveitado para outros fins, e a diretriz de anotação utilizada aqui pode ser adaptada para diversos contextos, como anotação de erros de tradução, anotação de desvios ortográficos, anotação de marcas de oralidade em comentários da internet, entre outros. Também, como se

verá na próxima seção, a pesquisa oferece sugestões de alterações da metodologia e da tipologia, de forma que os aprendizados obtidos neste percurso possam ser incorporados a trabalhos futuros.

Entre as contribuições específicas para a área da Linguística, a sistematização e descrição dos fenômenos nos quais ocorrem os desvios sintáticos no *corpus* de redações analisado é útil para compreendê-los de forma mais aprofundada. Ademais, pode gerar materiais pedagógicos ou *feedbacks* para professoras e professores de redação, de forma que possam propor novas maneiras de ensino ou atividades de reforço especificamente onde se identificou que os estudantes apresentam mais dificuldade. Para o PLN, a identificação dos atributos mais relevantes para a predição da presença de desvios via AM é uma contribuição no sentido de que pode oferecer *insights* para o desenvolvimento de modelos e ferramentas capazes de lidar com textos que tenham essas características.

7.2 *Limitações e lições aprendidas*

O percurso da pesquisa foi constituído de muitas decisões que limitaram o escopo do trabalho e diversos aprendizados a partir dos caminhos percorridos. Uma das limitações impostas pelas decisões tomadas foi justamente fixar o nível de análise em fenômenos sintáticos, principalmente porque a língua não se comporta em "caixinhas" bem-delimitadas, mas antes em combinação e inter-relação entre os diversos níveis estabelecidos pelas necessidades de sistematização características dos seres humanos. O primeiro desafio imposto foi limitar o que era desvio sintático e o que era ortográfico. A decisão tomou um rumo mais prático do que teórico. A partir das definições instrumentais, tentou-se manter o foco apenas na sintaxe, mas isso não foi totalmente possível, como mostraram as análises linguísticas do Capítulo 5.

Outra das limitações que se constituiu como um aprendizado importante foi a metodologia de anotação proposta. A tarefa se mostrou mais complexa do que inicialmente se supunha. A comparação com a experiência em revisão de textos e avaliação de redações foi parcial, porque nessa atividade, os desvios são apenas identificados e quantificados, e não tipificados. Assim, a etapa de classificação ou tipificação dos desvios precisa ser revista para disponibilizar uma anotação confiável, uma vez que anotar 27 categorias diferentes ao mesmo tempo, no mesmo nível de anotação, é assumir o risco de que muitos desvios poderão ser ignorados. Propõe-se, portanto, uma reavaliação do método de anotação, dividindo a tarefa em etapas menores: pode-se primeiro anotar os *tokens* em que ocorrem os desvios com uma etiqueta genérica; depois classificá-los por categoria e, com base nisso, em outra etapa, classificá-los

por subcategoria. Outra opção seria fazer um recorte das categorias de maior interesse e anotá-las primeiro, para só depois anotar as demais. Cabem ainda algumas redefinições no que se refere às orientações de anotar determinados fenômenos que são muito comuns (p. ex. *onde* sem sentido de lugar e uso de contração prepositiva antes de infinitivo).

No que se refere à tipologia de desvios, considera-se que ela foi específica o suficiente para mapear os fenômenos-alvo com clareza, e cumpriu a função para a qual foi desenvolvida. No entanto, identificou-se pelo menos um problema que exige a sua revisão, relacionado à categoria de segmentação de sentenças. O processo de anotação mostrou que os casos inseridos nessa categoria na verdade se referem a problemas de pontuação, tanto na própria sentença quanto na sentença anterior. Assim, sugere-se que ela seja reformulada como uma subcategoria da categoria de pontuação.

Ainda no quesito anotação, uma das limitações da metodologia escolhida por esta pesquisa foi o cálculo da concordância entre anotadores. É necessário que, entre cada uma das etapas das fases 1 e 2 de anotação, calcule-se a concordância de uma parcela maior de sentenças sem limite de número de *tokens*, e que se discutam e resolvam as discordâncias antes de dar sequência à tarefa. Esse ajuste na metodologia provavelmente garantirá resultados mais satisfatórios em termos de concordância entre anotadores e trará à luz casos problemáticos e demais inconsistências da tipologia e das diretrizes, permitindo o aprimoramento da tarefa.

7.3 *Trabalhos futuros*

A presente pesquisa abriu diversas possibilidades de investigações sobre aspectos que, por limites de tempo, de espaço, ou de escopo, não puderam ser realizadas. Algumas delas foram elencadas ao longo do texto, mas cabe retomar as mais relevantes aqui. Uma possibilidade interessante para a Linguística seria um estudo comparativo dos tipos de desvios identificados aqui com aqueles que ocorrem em produções textuais escritas por aprendizes de português como língua estrangeira. Essa investigação poderia explicitar fenômenos semelhantes e diferentes, a fim de compreender as dificuldades de cada público.

Para o PLN, uma possibilidade de investigação é a análise sistemática da influência de desvios de vários tipos no desempenho de ferramentas computacionais como *POS taggers* e *parsers*. O Capítulo 3 desta dissertação trouxe alguns exemplos de estudos que fazem isso com textos de aprendizes de inglês como língua estrangeira. No entanto, não se identificou qualquer pesquisa que realize essa tarefa para o português, nem com redações de falantes nativos, nem com textos de aprendizes de português como L2. Tal análise poderia fornecer *insights* para o

aprimoramento de ferramentas ou para o desenvolvimento de plataformas e aplicações capazes de lidar com textos que contenham muitos desvios. Isso seria interessante não apenas para tratar redações de aprendizes, mas também comentários e textos escritos na *web*, que em geral contêm diversos desvios e marcas de oralidade.

Outro caminho interessante que se abre a partir dessa investigação é a ampliação do campo de análise para demais níveis linguísticos, especialmente em termos de ortografia e acentuação. A interface e as influências entre os diversos níveis, bem como questões semânticas e de escolha lexical em redações de estudantes, podem esclarecer algumas das lacunas que não puderam ser preenchidas pela presente pesquisa. Também seria interessante conduzir essa análise com uma definição mais precisa dos tipos, da quantidade e das características dos desvios que levam os avaliadores humanos a atribuírem determinadas notas às redações.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA JÚNIOR, C. R. C. de. **Proposta de um sistema automático de avaliação de redações do ENEM, foco na Competência 1:** demonstrar domínio da modalidade escrita formal da língua portuguesa. 2017. 84 f. Dissertação (Mestrado em Informática) – Programa de Pós-Graduação em Informática. Vitória: Universidade Federal do Espírito Santo, 2017.

ALUÍSIO, S. M.; ALMEIDA, G. M. de B. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. **Calidoscópio**, v. 4, n. 3, p. 156–178, 2006. Disponível em: http://www.revistas.unisinos.br/index.php/calidoscopio/article/view/6002. Acesso em: 16 jun. 2019.

AMORIM, E.; VELOSO, A. A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese. *In:* Student Research Workshop at Conference of the European Chapter of the Association for Computational Linguistics, 15., 2017, Valência. **Anais**... Valência: Association for Computational Linguistics, 2017.

ATTALI, Y.; BURSTEIN, J. Automated Essay Scoring With e-rater® V.2. **The Journal of Technology, Learning and Assessment**, v. 4, n. 3, 2006. Disponível em: https://ejournals.bc.edu/index.php/jtla/article/view/1650. Acesso em: 7 mar. 2020.

BAZELATO, B.; AMORIM, E. A Bayesian Classifier to Automatic Correction of Portuguese Essays. *In:* Nuevas Ideas en Informática Educativa – TISE, 2013, Porto Alegre. **Anais**... Santiago: Universidad de Chile, 2013.

BECHARA, E. **Moderna Gramática Portuguesa**. Rio de Janeiro: Nova Fronteira, 2009.

BERBER-SARDINHA, T. Lingüística de Corpus: histórico e problemática. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, v. 16, n. 2, p. 323–367, 2000.

BERBER-SARDINHA, T. **Lingüística de corpus**. Barueri: Manole, 2004.

BICK, E. Automatic Parsing of Portuguese. *In:* Workshop on Computational Processing of Written Portuguese, 2., Curitiba. **Anais**... Curitiba: 1996.

BIDERMAN, M. **Teoria lingüística:** lingüística quantitativa e computacional. Rio de Janeiro: LTC, 1978.

BRASIL. **Base Nacional Comum Curricular:** Ensino Médio. Brasília: MEC/CONSED/UNDIME, 2017.

BRASIL. **Redação no ENEM 2018:** Cartilha do Participante. Brasília: INEP/MEC, 2018.

BURSTEIN, J.; CHODOROW, M. **Progress and New Directions in Technology for Automated Essay Evaluation**. Oxford: Oxford University Press, 2010.

BURSTEIN, J.; CHODOROW, M.; LEACOCK, C. Criterion SM: Online Essay Evaluation: An Application for Automated Evaluation of Student Essays. *In:* Annual Conference on Innovative Applications of Artificial Intelligence, 15., Acapulco, México. **Anais**… Acapulco: Association for the Advancement of Artificial Intelligence, p. 3–10, 2003.

CÂNDIDO, T. G. de; WEBBER, C. G. Avaliação da coesão textual: desafios para automatizar a correção de redações. **RENOTE**, v. 16, n. 1, 21 ago. 2018. Disponível em: https://seer.ufrgs.br/renote/article/view/86013. Acesso em: 25 jun. 2019.

CASTALDO, M. M. **Redação no vestibular:** a língua cindida. 2009. 277 f. Tese (Doutorado em Educação) – Programa de Pós-Graduação em Educação. São Paulo: Universidade de São Paulo, 2009.

DALE, R.; KILGARRIFF, A. Helping Our Own: The HOO 2011 Pilot Shared Task. *In*: European Workshop on Natural Language Generation (ENLG), 13., Nancy, França. **Anais**... Nancy: Association for Computational Linguistics, 2011. Disponível em: https://aclweb.org/anthology/papers/W/W11/W11-2838/. Acesso em: 25 jun. 2019.

DI-FELIPPO, A.; DIAS-DA-SILVA, B. C. O processamento automático de línguas naturais enquanto engenharia do conhecimento linguístico. **Calidoscópio**, v. 7, n. 3, p. 183–191, 2009.

DIAS-DA-SILVA, B. C. O estudo lingüístico-computacional da linguagem. **Letras de Hoje**, v. 41, n. 2, p. 103–138, 2006. Disponível em: https://core.ac.uk/download/pdf/25532027.pdf. Acesso em: 15 jan. 2020.

DROLIA, S. et al. Automated Essay Rater using Natural Language Processing. **International Journal of Computer Applications**, v. 163, n. 10, p. 44–46, 2017.

EVERS, A. **A redação engaiolada:** padrões lexicais e ensino de redação em cursos pré-vestibulares populares. 2018. 229 f. Tese (Doutorado em Letras) – Programa de Pós-Graduação em Letras. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2018.

FELTRIM, V. D. **Uma abordagem baseada em corpus e em sistemas de crítica para a construção de ambientes Web de auxílio à escrita acadêmica em português**. 2004. Tese (Doutorado em Computação) – Instituto de Ciências Matemáticas e de Computação. São Carlos: Universidade de São Paulo, 2004.

FERREIRA, M.; LOPES, M. Linguística Computacional. *In:* FIORIN, J. L. (Ed.). **Novos caminhos da Linguística**. São Paulo: Contexto, 2017. p. 195–2014.

FONSECA, E. et al. Automatically Grading Brazilian Student Essays. *In:* International Conference on Computational Processing of the Portuguese Language, 13., Canela. **Anais**... Berlim: Springer, 2018.

FRANCESCON, P. K.; FERNANDES, R. B. A textualidade nas redações de cotistas e não-cotistas no vestibular da Unioeste 2009. In: Seminário Nacional em Estudos da Linguagem: Diversidade, Ensino e Linguagem, Cascavel. **Anais**... Cascavel: Unioeste, 2010.

GALHARDI, L. B. et al. Analisador Léxico-Morfológico de Redações de Estudantes no Estilo do ENEM. *In:* Nuevas Ideas en Informática Educativa – TISE, 2018, Santiago. **Anais**... Santiago: Universidad de Chile, 2013.

GAMALLO, P. et al. Avalingua: Natural language processing for automatic error detection. In: CALLIES, M.; GÖTZ, S. (Ed.). **Learner corpora in language testing and assessment**. Studies in Corpus Linguistics. Amsterdã: John Benjamins, 2015. p. 35–58.

GOMPEL, M. van; REYNAERT, M. FoLiA: A practical XML format for linguistic annotation – a descriptive and comparative study. **Computational Linguistics in the Netherlands Journal**, v. 3, p. 63–81, 1 dez. 2013. Disponível em: https://clinjournal.org/clinj/article/view/26. Acesso em: 26 jun. 2019.

GRAMA, D. F. WordSmith Tools: para uma análise da coesão sequencial em redações dissertativas argumentativas. **Estudos Linguísticos**, v. 45, n. 2, p. 540–554, 2016. Disponível em: https://revistas.gel.org.br/estudos-linguisticos/article/view/598. Acesso em: 25 jun. 2019.

HEILMAN, M. et al. Predicting Grammaticality on an Ordinal Scale. *In:* Annual Meeting of the Association for Computational Linguistics, 52., 2014, Stroudsburg, Estados Unidos. **Anais**... Stroudsburg: Association for Computational Linguistics, 2014. Disponível em: http://aclweb.org/anthology/P14-2029. Acesso em: 25 jun. 2019.

HERREIRA, A. da S. **Produção textual no ensino fundamental e médio:** da motivação à avaliação. 2000. 125 f. Dissertação (Mestrado em Linguística Aplicada) – Programa de Pós-Graduação em Linguística Aplicada. Maringá: Universidade Estadual de Maringá, 2000.

HOLMES, G.; DONKIN, A.; WITTEN, I. H. **WEKA:** a machine learning workbench. (Working paper 94/09). Hamilton, Nova Zelândia: Waikato University, Depatment of Computer Science, 1994. Disponível em: https://researchcommons.waikato.ac.nz/handle/10289/1138. Acesso em: 26 jun. 2019.

HOVY, E.; LAVID, J. Towards a "science" of corpus annotation: A new methodological challenge for corpus linguistics. **International Journal of Translation Studies**, v. 22, p. 13–36, 2010.

HUANG, Y. et al. Dependency parsing of learner English. **International Journal of Corpus Linguistics**, v. 23, n. 1, p. 28–54, 2018.

JOSÉ, J.; PAIVA, R.; BITTENCOURT, I. I. Avaliação automática de atividades escritas baseada em algoritmo genético e processamento de linguagem natural: Avaliador Ortográfico-Gramatical. *In:* Workshops do Congresso Brasileiro de Informática na Educação, 4., Maceió. **Anais**…, Maceió: Sociedade Brasileira de Computação, p. 95–104, 2015.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing:** an introduction to natural language processing, computational linguistics, and speech recognition. 3 (draft) ed. London: Pearson Prentice Hall, 2017.

KÖHN, C.; KÖHN, A. An annotated corpus of picture stories retold by language learners. In: Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), Santa Fe, Estados Unidos. **Anais**... Santa Fe: Association for Computational Linguistics, 2018.

LEACOCK, C. et al. Automated Grammatical Error Detection for Language Learners. **Synthesis Lectures on Human Language Technologies**, v. 3, n. 1, p. 1–134, 11 jan. 2010. Disponível em: http://www.morganclaypool.com/doi/abs/10.2200/S00275ED1V01Y201006HLT009. Acesso em: 26 jun. 2019.

LEACOCK, C. et al. **Automated Grammatical Error Detection for Language Learners**. 2. ed. Toronto: Morgan & Claypool, 2014.

LIMA, A.; MASAGÃO, V. R.; CATELLI JR., R. **Indicador de Alfabetismo Funcional - INAF: Estudo especial sobre alfabetismo e mundo do trabalho**. São Paulo: Instituto Paulo Montenegro/Ação Educativa, 2016.

LITMAN, D. Natural Language Processing for Enhancing Teaching and Learning. *In:* Conference on Artificial Intelligence (AAAI 2016), 30., Phoenix, Estados Unidos. **Anais…** Phoenix: AAAI, p. 4170–4176, 2016.

LUNA, E. Á. D. A. **Avaliação da produção escrita no ENEM:** como se faz e o que pensam os avaliadores. 2009. 157 f. Dissertação (Mestrado em Linguística) – Programa de Pós-Graduação em Letras. Recife: Universidade Federal de Pernambuco, 2009.

LYASHEVKAYA, O.; PANTELEEVA, I. **Automatic Dependency Parsing of a Learner English Corpus Realec.** (Working Papers). Serié Linguistics. Moscou, Rússia: National Research University Higher School of Economics, Department of Linguistics, 2017.

MACARTHUR, C. A.; GRAHAM, S.; FITZGERALD, J. **Handbook of writing research**. 2. ed. Nova York: Guilford Press Brand, 2016.

MANNING, C. D.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Massachusetts: MIT Press, 1999.

MARCUSCHI, B. Redação Escolar: características de um objeto de ensino - Introdução. **Revista da Faced**, v. 9, p. 139–155, 2005.

MARCUSCHI, L. A. Gêneros textuais: definição e funcionalidade. *In:* DIONISIO, A. P.; MACHADO, A. R.; BEZERRA, M. A. (Ed.). **Gêneros textuais e ensino**. 5. ed. Rio de Janeiro: Lucerna, 2007. p. 19–36.

MARSLAND, S. **Machine learning:** an algorithmic perspective. Boca Raton: CRC Press, 2009.

MARTINS, A. F. T. et al. Turbo Parsers: Dependency Parsing by Approximate Variational Inference. *In:* Conference on Empirical Methods in Natural Language Processing, 11, Massachusetts, Estados Unidos. **Anais**... Massachusetts: Association for Computational Linguistics, 2010.

MARTINS, R. T. et al. Linguistic issues in the development of ReGra: A grammar checker for Brazilian Portuguese. **Natural Language Engineering**, v. 4, n. 4, p. 287–307, 1 dez. 1998.

MITCHELL, T. M. **Machine Learning**. Nova York: MIT Press/McGraw-Hill, 1997.

NAPOLES, C.; CAHILL, A.; MADNANI, N. The Effect of Multiple Grammatical Errors on Processing Non-Native Writing. *In:* Workshop on Innovative Use of NLP for Building Educational Applications, 11., San Diego. **Anais**... San Diego: Association for Computational Linguistics, 2016.

NASCIMENTO, R. I. do; ISQUERDO, A. N. Frequência de palavras: um diagnóstico do vocabulário de redações de vestibular. **ALFA: Revista de Linguística**, v. 47, n. 1, p. 71–84, 2003.

NAU, J. et al. Uma Ferramenta para Identificar Desvios de Linguagem na Língua Portuguesa. *In:* Symposium in Information and Human Language Technology, Uberlândia. **Anais**... Uberlândia: Sociedade Brasileira de Computação, 2017. Disponível em: https://aclweb.org/anthology/papers/W/W17/W17-6601/. Acesso em: 25 jun. 2019.

NG, H. T. et al. The CoNLL-2014 Shared Task on Grammatical Error Correction. *In:* Conference on Computational Natural Language Learning: Shared Task, 18., Baltimore, Estados Unidos. **Anais**... Baltimore: Association for Computational Linguistics, 2014. Disponível em: https://www.aclweb.org/anthology/W14-1701. Acesso em: 25 jun. 2019.

NG, H. T. et al. The CoNLL-2013 Shared Task on Grammatical Error Correction. *In:* Conference on Computational Natural Language Learning: Shared Task, 17., Sofia, Bulgária. **Anais**... Sofia: Association for Computational Linguistics, 2015. Disponível em: https://aclweb.org/anthology/papers/W/W13/W13-3601/. Acesso em: 25 jun. 2019.

NIVRE, J. et al. MaltParser: A language-independent system for data-driven dependency parsing. **Natural Language Engineering**, v. 13, n. 2, p. 95–135, 12 jun. 2007.

NIVRE, J. et al. Universal Dependencies v1: A Multilingual Treebank Collection. *In:* International Conference on Language Resources and Evaluation (LREC 2016), 10., Portorož, Eslovênia. **Anais**... Portorož: LREC, 2016. Disponível em: https://www.aclweb.org/anthology/papers/L/L16/L16-1262/. Acesso em: 25 jun. 2019.

NUNES, M. das G. V. et al. Desafios na construção de recursos linguísticos para o processamento do português do Brasil. *In:* BERBER-SARDINHA, T. (Ed.). **A língua portuguesa no computador**. Campinas: Mercado de Letras, 2005. p. 33–70.

OLIVEIRA, F. S. de. **Erros linguísticos em textos formais de professores e de alunos:** um estudo de caso. 2013. 137 f. Dissertação (Mestrado em Estudos Linguísticos e Culturais) – Departamento de Estudos Linguísticos e Culturais. Madeira: Universidade da Madeira, 2013.

OLIVEIRA, K. M. S. T.; LOURA, M. do S. D. Semântica e produção de texto no Ensino Médio. **Labirinto**, v. 25, p. 457–479, 2017.

PERINI, M. A. **Gramática descritiva do português**. São Paulo: Ática, 1995.

PERINI, M. A. **Gramática descritiva do português brasileiro**. Petrópolis: Vozes, 2017.

PINHEIRO, G. M. **Redações do ENEM:** estudo dos desvios da norma padrão sob a perspectiva de corpos. 2008. 151 f. Dissertação (Mestrado em Estudos Lingüísticos e Literários em Inglês) – Faculdade de Filosofia, Letras e Ciências Humanas. São Paulo: Universidade de São Paulo, 2008.

PUSTEJOVSKY, J.; STUBBS, A. **Natural Language Annotation for Machine Learning**. Sebastopol: O'Reilly, 2013.

RANCHHOD, E. M. O lugar das expressões 'fixas' na gramática do Português. *In:* CASTRO, I.; DUARTE, I. (Eds.). **Razão e Emoção:** miscelânea de estudos em homenagem a Maria Helena Mira Mateus. V. 2. Lisboa: Imprensa Nacional-Casa da Moeda, p. 239–254, 2003.

ROZOVSKAYA, A.; ROTH, D. Annotating ESL errors: Challenges and rewards. *In:* Workshop on innovative use of NLP for building educational applications., 5., Los Angeles, Estados Unidos. **Anais**... Los Angeles: Association for Computational Linguistics, 2010.

SANDOVAL, A. N.; ZANDOMÊNICO, S. C. M. de R. Concordância verbal em redações do Exame Nacional de Ensino Médio produzidas por alunos da Educação de Jovens e Adultos no Brasil. **Agália: Revista de Estudos na Cultura**, n. 114, p. 117–132, 2016. Disponível em: https://agalia.net/Agalia/114.pdf#page=118. Acesso em: 25 jun. 2019.

SANTOS, D. et al. Floresta Sintá(c)tica: um "treebank" para o português. *In:* Encontro Nacional da Associação Portuguesa de Linguística, 17., Lisboa, Portugal. **Anais**... Lisboa: APL, 2001.

SANTOS, D. Corporizando algumas questões. *In:* TAGNIN, S. E. O.; VALE, O. A. (Ed.). **Avanços da Linguística de Corpus no Brasil**. São Paulo: Humanitas, 2008. p. 41–66.

SANTOS, P. P. dos; MOTTA, V. R. A. Leitura e produção textual no ensino médio: uma proposta a partir da linguística textual. **UniLetras**, v. 37, n. 2, p. 177–186, mar. 2017. Disponível em: http://www.revistas2.uepg.br/index.php/uniletras/article/view/8141. Acesso em: 25 jun. 2019.

SCARTON, C. E.; ALUÍSIO, S. M. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. **Linguamática**, v. 2, n. 1, p. 45–61, 2010. Disponível em: http://www.linguamatica.com/index.php/linguamatica/article/view/44. Acesso em: 25 jun. 2019.

SCOTT, M. **WordSmith tools manual**. Oxford: Oxford University Press, 1998.

SILVA, J. et al. Out-of-the-box robust parsing of Portuguese. *In:* International Conference on Computational Processing of the Portuguese Language, Berlim. **Anais**... Berlim: Springer, 2010.

SINCLAIR, J. **Corpus, concordance, collocation**. Oxford: Oxford University Press, 1991.

SOUSA, V. B. R. **A representação de atores sociais em corpus de redações estilo ENEM:** uma análise sob a ótica da semântica de papéis. 2016. 112 f. Dissertação (Mestrado em Linguística) – Programa de Pós-Graduação em Estudos Linguísticos. Uberlândia: Universidade Federal de Uberlândia, 2016.

SOUZA, J. W. da C.; DI-FELIPPO, A. Caracterização Da Complementaridade Temporal: Subsídios Para Sumarização Automática Multidocumento. **ALFA: Revista de Linguística**, v. 62, n. 1, p. 125–150, 2018.

STRAKA, M.; HAJIČ, J.; STRAKOVÁ, J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *In:* International Conference on Language Resources and Evaluation (LREC 2016), 10., Portorož, Eslovênia. **Anais**... Portorož: LREC, 2016.

TORRES, L. S. **Escrita científica em português por hispano falantes:** recursos linguísticos-computacionais baseados em métodos de alinhamento de textos paralelos. 2016. Tese (Doutorado em Computação) – Instituto de Ciências Matemáticas e de Computação. São Carlos: Universidade de São Paulo, 2016.

VIEIRA, R. et al. Processamento computacional de anáfora e correferência. **Revista de Estudos da Linguagem**, v. 16, n. 1, p. 263–284, 2008. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/2490. Acesso em: 7 mar. 2020.

WU, Z.-Y. Can an Automatic Essay Scoring System Be Used to Improve Students' Writing Skills? **International Journal of English Research**, v. 4, n. 2, p. 21–24, 2018.

ANEXO 1 – Termo de Compromisso de Manutenção de Sigilo

TERMO DE COMPROMISSO DE MANUTENÇÃO DE SIGILO – TCMS

RENATA RAMISCH, BRASILEIRA, inscrita no Cadastro de Pessoas Físicas do Ministério da Fazenda sob o NÚMERO 018.363.920-01, portador do Registro Geral 1089474652 expedido por ÓRGÃO EXPEDIDOR SJS-RS em 12 de JUNHO de 2007, perante o Centro de Autoria e Cultura Ltda, inscrito no CNPJ 21.590.974/0001-42, declara ter ciência inequívoca sobre tratamento de informação classificada cuja divulgação possa causar risco ou dano à segurança da Instituição Centro de Autoria e Cultura Ltda e da sociedade, e comprometendo-se a guardar o sigilo necessário, nos seguintes termos:

a) Tratar as informações classificadas em qualquer grau de sigilo ou os materiais de acesso restrito que me forem fornecidos e preservar o seu sigilo, de acordo com a legislação vigente;
b) Preservar o conteúdo das informações classificadas em qualquer grau de sigilo, ou dos materiais de acesso restrito, sem divulgá-los a terceiros;
c) Não praticar quaisquer atos que possam afetar o sigilo ou a integridade das informações classificadas em qualquer grau de sigilo, ou dos materiais de acesso restrito; e
d) Não copiar ou reproduzir, por qualquer meio ou modo: (i) informações classificadas em qualquer grau de sigilo; (ii) informações relativas aos materiais de acesso restrito do Centro de Autoria e Cultura Ltda, salvo autorização da autoridade competente.

Declara por estar de acordo com o presente Termo, assina na presença das testemunhas abaixo identificadas.

São Carlos, 22 de agosto de 2018.

RENATA RAMISCH
Mestranda

ARIANI DI FELIPPO
Orientadora

APÊNDICE A – Diretriz de anotação

INTRODUÇÃO

Esta tarefa de anotação tem como objetivo identificar os desvios sintáticos que ocorrem em textos escritos por estudantes do Ensino Médio. Pretende-se utilizar a anotação para descrever os desvios mais frequentes e identificar atributos linguísticos correlacionados à presença de desvios. Espera-se que o *corpus* anotado possa servir de base para o desenvolvimento de ferramentas computacionais capazes de lidar com textos que contenham muitos desvios de escrita, como *parsers*, corretores ortográficos e gramaticais, e ferramentas de auxílio à escrita.

Ainda que você não concorde com todas as orientações presentes nesta diretriz, é importante que você as siga, de forma que haja uma padronização na anotação e uma concordância adequada entre os anotadores. Caso você encontre incoerências, erros conceituais ou demais questões que considere relevantes de serem alteradas, por favor entre em contato com a organizadora da tarefa *antes* de iniciar a anotação.

O *CORPUS*

O *corpus* utilizado nesta tarefa de anotação é constituído por sentenças retiradas de 1.045 redações dissertativo-argumentativas nos moldes do ENEM, coletadas em formato digital na plataforma virtual da empresa Letrus (fornecedora deste *corpus*).

A TAREFA DE ANOTAÇÃO

Esta tarefa de anotação é constituída de duas fases. Na primeira, são classificadas as sentenças que contêm ou não contêm desvios sintáticos. Na segunda, apenas as sentenças com desvio são utilizadas, nas quais os desvios são classificados de acordo com a tipologia apresentada a seguir.

Primeira fase

Nesta fase, o anotador receberá um arquivo em formato de planilha com três coluna: ID da sentença, constituída pela ID do texto e pelo número sequencial da sentença; texto da sentença a ser classificada; coluna de anotação, que foi previamente preenchida com a letra N, que deverá identificar somente as sentenças que **não possuem** desvios. Para marcar sentenças com desvio, o anotador deverá atribuir a **letra D maiúscula** na terceira coluna. Ao final de cada texto, há uma linha preenchida na cor azul, identificando o início de um novo texto. Nesta linha vazia, não deverá ser feita nenhuma anotação, pois ela tem a função exclusiva de facilitar a visualização. Veja um exemplo do arquivo:

ID	text	errors
11201.1	A leitura é extremamente importante, não apenas para a formação do nosso intelectual, mas também para temos novos conhecimentos, questionamentos, para ter a formação de um novo vocabulário	N
11201.2	Ela influenciou o desenvolvimento da sociedade em que vivemos é uma grande responsável pelas transformações ocorridas, mas o hábito de leitura diminuiu com a chegada da tecnologia	N
11201.3	No Brasil "44% da população não lê e 30% nunca comprou um livro"	N
11201.4	Os filhos desde a infância espelham-se em seus pais, veem que a leitura não é um hábito em casa, eles não encontram uma referência para a leitura, pois automaticamente a família lhe influenciou a isso, assim a criança reproduz aquilo que ver em casa para gerações futuras	N
11201.5	O avanço da tecnologia está fazendo com que os jovens se limitem apenas naquilo que leem na rede virtual, nunca vão procura ir além daquilo que leram na internet, um exemplo disso ocorre nas escolas; quando um professor pede uma pesquisa, os alunos não recorrem aos livros e sim a internet, tendo assim os mesmos conhecimentos de uma maneira mais rápida e resumida, pois ler é um processo cansativo se comparado com as mídias digitais	N
11201.6	Apesar disso, ainda há pessoas que tem o habito de leitura que procuram novos conhecimentos, estimula a reflexão sobre os seus princípios, valores, pensamento crítico, atitudes, entre outros benefícios da leitura	N
11201.7	Através disso se tornam pessoas mais sabias com um rico vocabulário e tendo fácies interpretações ao lerem textos complexos	N
11201.8	Diante dos fatos apresentados posso concluir que a leitura tem um poder de transformação na vida do ser humano, mas que por conta da tecnologia e da falta de incentivo o ser humano cada vez tem menos o hábito de leitura, para combater isso devemos desperta a curiosidade para que possam descobrir coisas novas através da leitura, ter um incentivo dentro da própria casa	N
11201.9	Para que a leitura vire um hábito devemos prática, pois a prática leva a perfeição	N
11201.10		
11207.1	O primeiro livro foi impresso no século XV, fruto da invenção da tipografia de alemão chamado Johannes Gutenberg	N
11207.2	Após 1500, com o aperfeiçoamento da imprensa, o livro foi se modificando, desde o tipo de papel até os detalhes formais ligados a disposição das páginas hoje podemos afirmar que a presença do livro em nossa cultura foi de extrema importância, pois foi o que norteou a abrir a porta da História, a	N

Categorias de desvios
Deverão ser marcadas com D sentenças que apresentem no mínimo um dos seguintes desvios sintáticos:

1) **Uso de pontuação:** ausência, excesso ou uso inadequado de sinais de pontuação, como separação de elementos por vírgulas indevidas, aglutinação de sentenças por vírgulas em lugar de pontos finais, falta de pontuação separando elementos deslocados. Exemplos:
    a. *Ela influenciou o desenvolvimento da sociedade em que vivemos é uma grande responsável pelas transformações ocorridas, mas o hábito de leitura diminuiu com a chegada da tecnologia* (ausência de pontuação)
    b. *O avanço da tecnologia está fazendo com que os jovens se limitem apenas naquilo que leem na rede virtual, nunca vão procura ir além daquilo que leram na internet, um exemplo disso ocorre nas escolas; quando um professor pede uma pesquisa, os alunos não recorrem aos livros e sim a internet, tendo assim os mesmos conhecimentos de uma maneira mais rápida e resumida, pois ler é um processo cansativo se comparado com as mídias digitais* (uso equivocado de pontuação – aglutinação de sentenças)
    c. *O leitor, passa então, a se apropriar dos textos, mergulhando nele, de forma profunda a estabelecer um compromisso com oque está sendo lido* (excesso de pontuação)
2) **Crase:** ausência de crase em casos obrigatórios ou excesso de crase (ausência de crase em casos optativos não será considerada desvio). Exemplos:
    a. *O motivo dessas manifestações são à luta por moradia, reforma na Previdência Social e trabalhista, entre outros* (excesso de crase)
3) **Regência:** problemas de regência nominal ou verbal obrigatórios (atenção para casos como a regência do verbo *visar*, que não deverão ser marcados como desvio). Exemplos:
    a. *O avanço da tecnologia está fazendo com que os jovens se limitem apenas naquilo que leem na rede virtual (...)* (regência verbal equivocada)
    b. *(...) no Brasil, por sua vez, a população vem lutando acerca de seus direitos de forma vigente e consensual desde o século IX com movimentos na zona rural dos sem terras, por exemplo* (regência verbal equivocada)
4) **Concordância:** desvios de concordância verbal, nominal ou anafórica (quando o elemento que retoma não concorda em gênero ou número com o elemento retomado). Será considerado um desvio de concordância a falta de acento circunflexo nos verbos *ter* e *vir* na terceira pessoa do plural. Exemplos:
    a. *Esses dados **mostra** que, em geral, o ato de ler está associado a uma atividade obrigatória, solitária que exige bastante paciência e atenção do leitor* (falta de concordância verbal)
    b. *No movimento feminista radical, algumas vertentes não reconhecem **mulheres transexuais** como sendo do gênero feminino, mesmo quando **a mesma** identifica-se desse modo* (falta de concordância anafórica entre os termos em negrito)
5) **Pronomes:** desvios de colocação pronominal obrigatória, ausência, excesso ou uso equivocado de pronomes de qualquer tipo (como troca entre os pronomes oblíquos *o* e *lhe* ou uso de *onde* como pronome relativo não se referindo a lugar). Exemplos:
    a. *(...) pois automaticamente a família **lhe** influenciou a isso, assim a criança reproduz aquilo que ver em casa para gerações futuras* (uso equivocado de *lhe* em vez de *o*)
    b. *Ademais, recentemente foi possível presenciar manifestações pacificas de estudantes de todo o território brasileiro, **onde** mostraram que é desde cedo que se deve lutar pelos seus direitos, e mostraram um verdadeiro "espetáculo" de como agir **onde** demonstraram que é se unindo pacificamente que irá se obter retorno* (uso equivocado de *onde* como pronome relativo)
6) **Preposições:** ausência de preposições em casos obrigatórios, excesso ou uso equivocado das preposições. Exemplos:

a. *Os movimentos sociais no Brasil são historicamente reconhecidos e causam reverberação desde a época que houve a Independência do Brasil* (ausência de preposição)

b. ***Mediante a** isso as escolas deveriam propagar através de mesas redondas aos jovens o quão importante é lutar pela democracia de forma pacífica (...)* (excesso de preposição)

7) **Determinantes:** ausência de determinantes em casos obrigatórios, excesso (como em *cujo o* e *pela a*) ou uso equivocado de determinantes (pouco frequente). Exemplos:

a. *O primeiro livro foi impresso no século XV, fruto da invenção da tipografia de alemão chamado Johannes Gutenberg* (ausência de determinante)

8) **Conjunções:** ausência, excesso ou uso equivocado de conjunções (como o uso sequencial de *mas porém*). Exemplos:

a. *Pesquisas recentes afirmam que, os brasileiros leem em média 4,7 livros por ano, **sendo** apenas 1,3 são livros ausentes na grade escolar, escolhido assim pela vontade e interesse do leitor, sendo assim ainda pouco* (uso equivocado da locução conjuntiva *sendo que*) – esse caso não poderia ser classificado como ausência de pronome, uma vez que fica evidente a tentativa de uso da locução conjuntiva

9) **Formas verbais:** uso equivocado de verbos finitos quanto a formas, tempos e modos verbais; uso equivocado das formas verbais infinitas gerúndio, infinitivo e particípio. Exemplos:

a. *Mas, no Brasil a leitura parece **está** longe de ser algo primordial na vida da grande maioria* (uso equivocado de verbo finito)

b. *As pessoas guardando aquilo e assim não apoiam os movimentos* (uso equivocado de gerúndio)

10) **Segmentação de sentenças:** sentenças sem verbo finito; sentenças que indicam a continuação da anterior, mas que foram separadas por ponto final (como sentenças que começam com *Assim como*). Exemplos:

a. *Assim como acreditam que homens não podem apoiar o movimento por não pertencerem ao grupo ao qual ele é destinado*

b. *Assim evidenciando que o Brasil é sinônimo de luta e resistência*

Nesta etapa, o tipo de desvio ocorrido não tem relevância. Basta que ocorra no mínimo uma entre qualquer das 10 categorias acima para que a sentença seja classificada como tendo desvios. Em caso de dúvidas, sugere-se a consulta à *Moderna Gramática Portuguesa*, de Evanildo Bechara, que é a gramática base desta pesquisa. Caso não haja a possibilidade de consulta à gramática ou a dúvida permaneça após a consulta, o anotador deverá marcar a sentença duvidosa e enviar um e-mail à organizadora da tarefa (renata.ramisch@gmail.com), que será a responsável por julgar e decidir sobre casos duvidosos ou não consensuais.

Segunda fase

Esta fase consiste na classificação dos desvios sintáticos de acordo com a tipologia proposta. Logo, somente serão anotadas as sentenças previamente classificadas como **contendo desvios**. A anotação será realizada na plataforma de anotação FLAT, que será descrita mais à frente. A tabela a seguir traz a tipologia de desvios, identificando o código da categoria e subcategoria do desvio, seguida da descrição do código:

Nº	Código	Descrição
01.1	pont-aus	Ausência de pontuação
01.2	pont-exc	Excesso de pontuação
01.3	pont-desv	Uso equivocado de pontuação
02.1	crase-aus	Ausência de crase
02.2	crase-exc	Excesso de crase
03.1	rege-verb	Regência verbal
03.2	rege-nom	Regência nominal
04.1	concor-verb	Concordância verbal
04.2	concor-nom	Concordância nominal
04.3	concor-anaf	Concordância anafórica
05.1	pronom-col	Colocação pronominal
05.2	pronom-aus	Ausência de pronome
05.3	pronom-exc	Excesso de pronome
05.4	pronom-desv	Uso equivocado de pronome
06.1	prepo-aus	Ausência de preposição
06.2	prepo-exc	Excesso de preposição
06.3	prepo-desv	Uso equivocado de preposição
07.1	determ-aus	Ausência de artigo/determinante
07.2	determ-exc	Excesso de artigo/determinante
07.3	determ-desv	Uso equivocado de artigo/determinante
08.1	conjunc	Uso equivocado de conjunção
09.1	verbo-mod	Uso equivocado de tempos e modos e formas verbais finitas
09.2	verbo-nom	Uso equivocado de formas verbais infinitas
10.1	segment	Outros desvios de ordem textual
11.1	sem-espec	Sem especificação

A tipologia é dividida em 25 subcategorias de desvios, separadas nas mesmas 10 categorias utilizadas na Etapa 1. A subcategoria 11.1 é restrita a desvios que não caibam em nenhuma das demais, sendo o seu uso **evitado ao máximo**. As subcategorias da tabela têm **caráter hierárquico**, de forma a não sobrepor desvios que se encaixem em mais de um tipo. Por exemplo, ainda que a ausência de crase pudesse ser entendida como um problema de regência, sua subcategoria está em posição superior na tabela; logo, deve prevalecer.

Esta é uma anotação sintática. Ainda que se tenha consciência da dificuldade de separar os níveis linguísticos, não deverão ser considerados desvio problemas de ordem exclusivamente semântica. Assim, se a sentença estiver construída adequadamente em termos de estrutura sintática, será considerada como livre de desvios, mesmo que não faça nenhum sentido.

Pontuação
Sabe-se que diversas questões de pontuação estão relacionadas à estilística ou não são consensuais. Porém, as anotações deverão ser feitas com base no que consta aqui. Casos duvidosos deverão ser encaminhados para avaliação.

01.1 pont-aus: Ausência de pontuação
Deverão ser inseridos nesta subcategoria os casos em que há ausência de pontuação onde ela seria obrigatória. Esses casos estarão, na maioria das ocorrências, ligados à falta de vírgulas. Na plataforma de anotação, a etiqueta deverá ser inserida no **token imediatamente anterior** ao ponto em que a pontuação deveria ter sido inserida. Seguem algumas orientações:

- A regra geral para uso de vírgulas é que ela deve ocorrer quando uma sentença não estiver na ordem direta (SVO), ou seja, deverá marcar elementos deslocados.
- As vírgulas após adjuntos adverbiais deslocados são **obrigatórias** sempre que tais adjuntos tiverem mais de três palavras.
- Há ausência de vírgulas em apostos quando elas estiverem apenas no início ou final, ou quando não há vírgulas (isso vale para vocativo, mas espera-se ocorrência rara deste).
- Ausências de vírgulas antes de conjunções conclusivas ou adversativas (*mas, pois, porém, embora, portanto, logo,* etc.) deverá ser marcada, mas o uso de vírgulas *após* essas conjunções não é obrigatório.
- As expressões *isto é, ou seja, a propósito, quer dizer, além disso* sempre devem ser intercaladas por vírgulas. A ausência destas deve ser anotada como desvio.
- Em orações adjetivas explicativas ou restritivas, só deverá ser marcada a ausência de vírgulas quando a intenção de construir uma oração explicativa estiver evidente.
- Vírgula antes de *etc.* é facultativa, mas vírgula antes de *entre outros* é **obrigatória**.
- Não deverá ser anotada como desvio a ausência de pontuação em abreviaturas.

01.2 pont-exc: Excesso de pontuação
Nesta subcategoria estão os casos em que há pontuação onde ela **não pode** ocorrer. Esses casos estarão, na maioria das ocorrências, ligados ao uso de vírgulas. Na plataforma de anotação, a etiqueta deverá ser inserida **no sinal de pontuação excessivo**. Seguem orientações:

- Vírgula após a conjunção *mas* deverá ser marcada como desvio.

01.3 pont-desv: Uso equivocado de pontuação
Nesta subcategoria estão os casos em que os sinais de pontuação são utilizados de maneira equivocada, os quais estarão, na maioria das ocorrências, ligados ao uso de vírgulas. Na plataforma de anotação, a etiqueta deverá ser inserida **diretamente no sinal de pontuação equivocado**. Seguem algumas orientações gerais:

- Sentenças aglutinadas por vírgulas em que fica clara a mudança do foco temático, indicando a função do ponto final, deverão ser marcadas como uso equivocado de pontuação.
- Questões de estilística como uso de ponto-e-vírgula, opção por travessões e parênteses em apostos, uso excessivo de aspas **não** deverão ser considerados desvios.

Crase
Deverão ser inseridas nessa categoria as ocorrências relacionadas ao uso da crase.

02.1 crase-aus: Ausência de crase
O acento indicativo de crase marca a contração entre o artigo *a* e a preposição *a*. Logo, é obrigatória a crase diante de palavras femininas quando o verbo regente exige a preposição *a*. Seguem algumas orientações:

- Crase é obrigatória em contração da preposição *a* e dos demonstrativos iniciados em *a*. Exemplo:
    - *Refiro-me **àquele** homem que estava ao seu lado.*
- É **obrigatório** o uso da crase em locuções adverbiais constituídas de substantivo feminino (*às vezes, às claras, às três da manhã*, etc.).
- Em casos em que a presença da crase é obrigatória para evitar ambiguidades, a sua ausência deverá ser marcada como desvio (p. ex. *à tarde, à noite, à vista*).
- A crase diante de pronomes possessivos só é obrigatória se o substantivo estiver oculto (*Foi à casa do irmão, e não à sua*). Os demais casos **não** deverão ser anotados.

02.2 crase-exc: Excesso de crase
O uso de crase quando ela **não** é permitida deverá ser anotado com esta etiqueta. Seguem mais algumas orientações em relação ao uso da crase:

- O acento agudo sobre o *a* isolado (*Vou á praia*) será considerado crase. Se for usado quando há crase, **não** deve ser anotado; caso contrário, deverá ser marcado como excesso.
- O uso de crase após a preposição *até* é facultativo e **não** deverá ser anotado como desvio.

Regência
Desvios de regência consistem em **ausência** ou **uso inadequado de preposição**. Consideram-se dois tipos de regência aqui, conforme a categoria gramatical da palavra regente: regência verbal e regência nominal.

03.1 rege-verb: Regência verbal
Nesta subcategoria deverão entrar problemas de regência cujo termo regente é o **verbo**, seguido de objeto ou adjunto. Seguem algumas considerações:

- Em caso de **ausência** de preposição, o token a ser anotado é o **termo regente**, isto é, o verbo que exige a preposição ausente, independentemente da sua posição na sentença.
- No **uso equivocado** de preposição, o token a ser anotado é **a própria preposição**. Se ela estiver na forma contraída (*naquele, pelo,* etc.), seleciona-se a forma contraída.
- Somente deverão ser anotados os casos unânimes de problemas de regência (casos como a regência do verbo *visar* não são consensuais e por isso não devem ser anotados).
- O verbo *chegar* seguido da preposição *em*, quando se referir a um local de chegada, deverá ser anotado como desvio sintático mesmo que o seu uso tenha se mostrado muito

usual, uma vez que tal ocorrência é penalizada na Competência 1 de avaliação do ENEM. Isso vale para o verbo *assistir* não seguido da preposição *a*.

- Desvios em que o termo regente é um *particípio* devem ser anotados como problemas de **regência verbal** (*O PC está **conectado** à rede*) quando não for evidente que é um adjetivo.
- Questões de paralelismo entre termos de regências diferentes **não** deverão ser anotadas se a preposição estiver correta para pelo menos um deles (*Entrei* [em] *e saí de casa.*)

03.2 rege-nom: Regência nominal
São desvios de regência nominal todos aqueles em que o termo regente não é um verbo, podendo ser um **nome**, **adjetivo** ou **advérbio**. Sugere-se consultar listas de regência na gramática ou na web em casos de dúvida. Seguem algumas observações:

- Em caso de **ausência** de preposição, o termo a ser anotado é o **termo regente**, isto é, o *nome* que exige a preposição ausente, independentemente da sua posição na sentença.
- Em caso de **uso equivocado** de preposição, o termo a ser anotado é **a própria preposição**.
- Somente deverão ser anotados os casos unânimes de problemas de regência.

Concordância
Devem ser anotados aqui desvios em que determinados termos não concordam em gênero, número ou pessoa com os seus dependentes. Há três subcategorias de concordância: verbal, nominal e anafórica. Casos com mais de uma opção de concordância não devem ser anotados.

04.1 concor-verb: Concordância verbal
Serão anotados com essa etiqueta casos em que o sujeito não concorda em número e pessoa com o verbo da sentença. Nesse caso, a etiquetagem sempre deverá ser feita no **verbo** ou na **locução verbal** na plataforma de anotação (em locuções verbais, ambos os verbos deverão ser selecionados e etiquetados em conjunto). Seguem algumas orientações:

- Quando houver mais opções de concordância, como em porcentagens ou expressões do tipo *a maioria, grande parte, a maior parte*, estas não devem ser anotadas.
- Caso haja na mesma sentença problemas de concordância nominal no sujeito, o termo de referência deverá ser o substantivo ou pronome (termo determinado). Exemplo:
    - *Os <u>dado</u> **mostram** um crescimento na violência doméstica.* (Deve-se marcar o verbo como desvio de concordância verbal.)
- A falta de acento circunflexo nos verbos *ter* e *vir* na terceira pessoa do plural deverá ser anotada como desvio de concordância verbal.

04.2 concor-nom: Concordância nominal
Serão anotados nesta subcategoria casos em que adjetivos, artigos, numerais ou particípios (determinantes) não concordam em gênero e número com o substantivo ou pronome (determinado) a que se referem. Aqui a etiquetagem deverá ser feita nos **termos determinantes** na plataforma de anotação. **TODOS** os termos que não concordam deverão ser anotados

separadamente (exceto em locuções, que devem ser anotadas em conjunto). Seguem algumas orientações:

- Nos casos de termos determinados masculinos e femininos, os determinantes devem seguir o padrão da língua portuguesa, isto é, devem concordar no masculino plural.
- Casos em que há mais de uma opção de concordância não deverão ser marcados, especialmente quanto a concordâncias com termos determinados de mais de um gênero.
- O termo *menas* sempre deverá ser anotado como desvio.
- As expressões *é necessário, é preciso, é bom* sucedidas de palavras no feminino e/ou no plural **não** deverão ser anotadas como desvios.

04.3 concor-anaf: Concordância anafórica

Esta subcategoria mostrou-se uma possibilidade de categorizar desvios decorrentes da retomada de elementos citados anteriormente, mas cujo elemento retomador não concorda com o retomado. Deverão ser anotados com essa etiqueta pronomes pessoais, pronomes *esse/este* indicando retomada e a expressão *o mesmo/a mesma* com valor de retomada.

- O termo a ser anotado é o elemento retomador, isto é, o pronome ou a expressão que indica a retomada.
- O nível de anotação é o da **sentença**, então só devem ser anotados desvios em que o elemento retomado está na mesma sentença do retomador. Exemplo:
    - *Para dar mais visibilidade aos projetos, é importante que o Ministério das Comunicações se junte a essa aliança e divulgue **o mesmo** nas mais diversas mídias sociais.* (desvio de concordância anafórica entre *projetos* e *o mesmo*).
- As expressões *o mesmo* e *a mesma* deverão ser anotadas em conjunto, isto é, tanto o artigo quanto o termo *mesmo/mesma* deverão ser selecionados e etiquetados.

Pronomes

Nesta categoria deverão constar desvios de uso e colocação de pronomes. Há quatro subcategorias, sendo a primeira relacionada apenas à posição dos pronomes pessoais. As demais envolvem todos os tipos de pronomes em relação a presença, ausência ou uso equivocado.

05.1 pronom-col: Colocação pronominal

Nesta subcategoria, apenas devem se anotados pronomes pessoais colocados em posição irregular. Valem algumas considerações em relação à colocação pronominal:

- A mesóclise, quando utilizada corretamente, não será considerada desvio. Exemplo:
    - *Dever-se-ia pensar em soluções para o problema.*
- Será considerado desvio o uso de próclise no início da sentença, mas **não deverá ser anotada como desvio** a próclise após vírgulas. Exemplos:
    - ***Se pode** criar leis que legitimem os movimentos sociais. (há desvio)*
    - *O surgimento de movimentos sociais, os quais podem ou não envolver questões políticas, **se dá** diante de alguma injustiça social. (**não há** desvio)*
- O uso facultativo de próclise ou ênclise no meio da sentença **não** deverá ser anotado. Exemplo:

- *Meu pai **ajudou-me** com o dever de casa. / Meu pai **me ajudou** com o dever de casa.*

- Palavras negativas (*não*, *nunca*, *ninguém*), conjunções subordinativas (*embora*, *se*, *conforme*, etc.), pronomes (exceto pessoais), frases interrogativas e expressões formadas por preposição e gerúndio (*em se tratando*) atraem pronomes pessoais, exigindo próclise.
- Não é desvio o uso de pronomes não ligados a nenhum verbo em locuções verbais. Exemplo:
  - *Sua mãe **queria lhe dar** um presente.*

05.2 pronom-aus: Ausência de pronome
Esta subcategoria engloba casos de ausência de quaisquer pronomes quando seu uso é obrigatório. Também deverá ser anotada aqui a omissão de pronomes relativos.

- A anotação da ausência de pronomes deverá ser feita no token imediatamente anterior ao ponto da sentença onde o pronome deveria ter sido empregado.

05.3 pronom-exc: Excesso de pronome
Esta subcategoria destina-se à anotação de pronomes que são usados em excesso, especialmente na retomada do sujeito por meio de pronome pessoal.

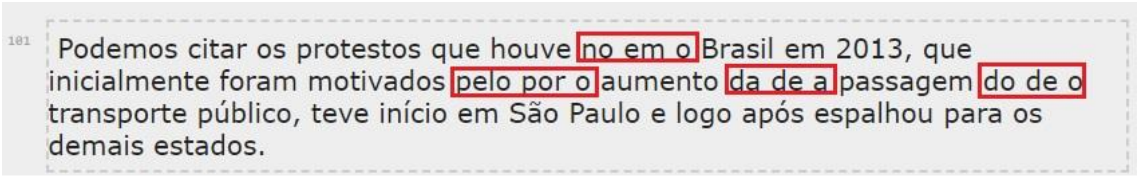
- *A criança, desde cedo, **ela** se espelha nas atitudes dos pais. (excesso de pronome)*

05.4 pronom-desv: Uso equivocado de pronome
Nesta subcategoria devem ser anotados os usos equivocados de pronomes. Por exemplo:

- A troca entre *o* e *lhe;*
- O uso de *onde* como pronome relativo não se referindo a lugar;
- O uso equivocado de *cujo/cuja*;
- ATENÇÃO! A troca entre *isso* e *isto* **não deverá** ser anotada como desvio.

Preposições
A categoria de preposições restringe-se aos desvios de uso de preposições **não relacionado à regência**, como locuções prepositivas nas quais falta uma das preposições, ausência de preposição em adjuntos adverbiais, ausência de preposição antes de pronomes relativos. Verifique se o desvio realmente não envolve questões de crase ou de regência.

O texto foi previamente parseado com o *parser* UDPipe, que separa automaticamente as contrações. Assim, aparecem no texto as formas contraídas, seguidas de cada uma das palavras que as compõem. Para esta anotação, deverão ser consideradas APENAS as formas contraídas. No exemplo a seguir, devem ser considerados para a anotação os tokens *no*, *pelo*, *da* e *do*.



06.1 prepo-aus: Ausência de preposição

Utiliza-se esta etiqueta quando faltar uma preposição obrigatória na sentença:

- O uso equivocado de *onde* quando a forma correta seria *aonde* (**não** ligado a casos de regência) deve ser anotado diretamente no token *onde*.
- Se faltar uma preposição antes de um pronome relativo, o token anotado deve ser o pronome relativo ou, na ausência deste, o token imediatamente anterior ao lugar em que a preposição deveria ter sido inserida.
- Se faltar uma das preposições em uma locução prepositiva, o token a ser etiquetado é a preposição que de fato ocorre no texto.
- Falta de preposições para fins de paralelismo **não será** considerada desvio.

06.2 prepo-exc: Excesso de preposição

Utiliza-se esta etiqueta para marcar o excesso de preposições quando elas não são necessárias ou quando são repetidas.

- O uso equivocado de *aonde* quando a forma correta seria *onde* (**não** ligado a casos de regência) deve ser anotado diretamente no token *aonde*.
- Na expressão *mediante a*, o *a* deverá ser anotado, pois esta não é uma locução prepositiva.
- Na expressão *muitas das vezes*, o *das* deverá ser etiquetado como excesso de preposição.
- A locução prepositiva *para com* **não** deverá ser considerada desvio.

06.3 prepo-desv: Uso equivocado de preposição

Esta etiqueta deve ser aplicada quando houver uso equivocado de preposições. Vale lembrar que não devem ser anotados casos de uso equivocado de preposição por questões de regência.

- Contrações utilizadas equivocadamente quando o correto seria preposição e artigo separados (ou contrações que não existem) deverão ser anotadas como desvio. Exemplo:
    - *A possibilidade **da** criança se tornar violenta se crescer num ambiente violento é grande.* (Não se deve usar contrações quando após o nome houver um infinitivo)
- A utilização de *a* no lugar de *há* e de *de* no lugar de *dê* deverá ser anotada aqui.

Determinantes

Nesta categoria deverão ser inseridos desvios ligados ao uso de artigos e demais determinantes. Essa categoria parece ser menos frequente no *corpus*.

07.1 determ-aus: Ausência de determinante

A ausência de determinantes ocorrerá, por exemplo, em paralelismo obrigatório ou omissão de determinantes que cause estranheza ou ambiguidade. Sugere-se cautela ao anotar essas questões.

07.2 determ-exc: Excesso de determinante
Esta subcategoria refere-se principalmente à repetição de determinantes após contrações nas quais eles já se fazem presentes, como o uso de *cujo o* ou *pela a*. Nesses casos, deve-se anotar **apenas** o determinante excessivo, e não a contração.

- ATENÇÃO! Não deverá ser anotada como excesso de determinantes a separação automática de contrações feita pelo *parser*, conforme exemplificado pela Figura acima.

07.3 determ-desv: Uso equivocado de determinante
Esta subcategoria está reservada às ocorrências de uso equivocado de determinantes, mas espera-se uma utilização pouco frequente desta etiqueta, uma vez que não foi possível encontrar na anotação piloto nenhum exemplo desse tipo de desvio.

Conjunções
A categoria das conjunções não possui subcategorias. Portanto, todo desvio relacionado ao uso de conjunções (ausência, excesso ou uso equivocado) deverá ser etiquetado da mesma forma.

08.1 conjunc: Uso equivocado de conjunção
O uso equivocado de conjunções pode estar ligado aos seguintes casos:

- o uso equivocado das formas *porque, por que, por quê, porquê*.
- repetição de conjunções da mesma natureza, como uso sequencial de *mas porém*;
- ausências de conjunções – estas deverão ser marcadas no token imediatamente anterior ao local em que a conjunção deveria ter ocorrido;
- falta de uma das conjunções nas locuções conjuntivas (como o *que* em *sendo que*).

Não deverão ser marcadas como desvio as seguintes sequências de conjunções:

- como por exemplo;
- tal como/tais como.

Formas verbais
A categoria das formas verbais engloba qualquer desvio de uso de verbos que não esteja em nenhuma das demais categorias. Aqui o token anotado deverá ser o verbo que contém o desvio.

09.1 verbo-mod: Uso equivocado de formas verbais, tempos e modos
Nesta subcategoria, deverão ser anotadas ocorrências de desvios ligadas a formas, tempos e modos verbais. Nesse caso, deve-se considerar o verbo como ele ocorre no texto, ainda que o problema decorra de um desvio ortográfico.

- A ausência do “r” no final de verbos que deveriam estar no infinitivo deve ser anotada como desvio dentro desta subcategoria.
- Quando a falta de acentuação do substantivo originar um verbo e não houver ambiguidade de interpretação (como quando há um determinante diante do substantivo não acentuado), por exemplo, em *hábito -> habito*, essa ocorrência **não deverá ser anotada como desvio**.

09.2 verbo-nom: Uso equivocado de formas verbais infinitas
Esta subcategoria se refere ao uso equivocado de infinitivo, gerúndio e particípio (inclusive as duas formas de particípio possíveis). Vale lembrar que se deve considerar a forma verbal como ela ocorre no texto, ainda que esta seja decorrente de um desvio de ortografia.

Segmentação de sentenças
Trata-se dos desvios de segmentação inadequada de sentenças. Esta categoria não está dividida em subcategorias.

10.1 segment: Desvios de segmentação de sentenças
Deverão ser anotados os seguintes casos:

- Sentenças que deveriam estar ligadas à anterior, mas que foram separadas por ponto final. Nesses casos, a anotação deverá ser feita **no primeiro token**.
- Sentenças que comecem com conjunções aditivas (como *assim como, pois, e* etc.). Nesses casos, a anotação deverá ser feita **na conjunção/locução conjuntiva**.
- Sentenças sem verbo algum ou apenas com formas verbais infinitas (como sequências de infinitivos ou substantivos). Nesses casos, a anotação deverá ser feita **no último token**, que geralmente é o ponto final (na ausência deste, anotar o último token).
- Demais questões de ordem textual que gerem desvios e que extrapolam o nível da sentença — nesses casos, a anotação deverá ser feita no **último token** da sentença.

A PLATAFORMA DE ANOTAÇÃO

A anotação será realizada na plataforma de anotação FLAT (*FoLiA Linguistic Annotation Tool*). Na maioria das ocasiões, o token a ser anotado é único e, por isso, a plataforma se mostrou intuitiva e de fácil utilização. A anotação será feita com o mouse, por meio da etiquetação específica, conforme a tipologia e as orientações descritas acima.

Acessando o FLAT
Para entrar no FLAT, acesse o seguinte endereço web no navegador de sua preferência:

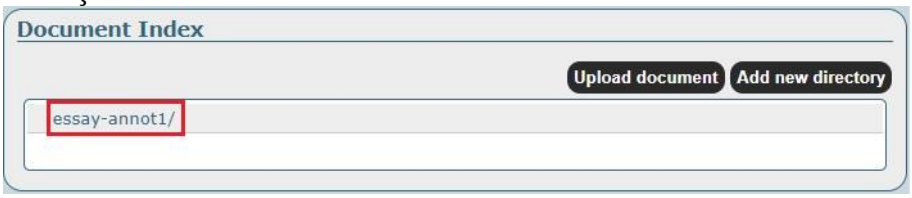
http://mwe.phil.hhu.de/

A página exige um login para acesso. Siga os seguintes passos:

1. Em ***Username***, digite: **essay-annot[X] –** substitua os colchetes pelo seu respectivo número de anotador (1, 2 ou 3) enviado por e-mail
2. Em ***Password***, digite: **XXXXXXXXXXXXX** – digite a senha que você recebeu por e-mail
3. Em ***Configuration***, selecione **ESSAY ASSESSMENT**
4. Clique em **Log In**

Abrindo a pasta do arquivo de anotação

Na página seguinte, clique na pasta ***essay-annot[X]***. É nessa pasta que estará o documento destinado à anotação.



Abrindo o arquivo

Foi designado um número para cada anotador participante da tarefa, que deverá ser usado para o login, a pasta de acesso ao arquivo de anotação e o arquivo de anotação em si. Caso o seu número não corresponda a qualquer uma dessas etapas, entre em contato com a organizadora. Por exemplo, o anotador a quem foi designado o número 1 deverá ter na sua pasta **apenas** o arquivo nomeado como **anotador1.parsemetsv**; isso vale para os demais anotadores.

Para iniciar a anotação, clique no nome do arquivo que está na pasta.

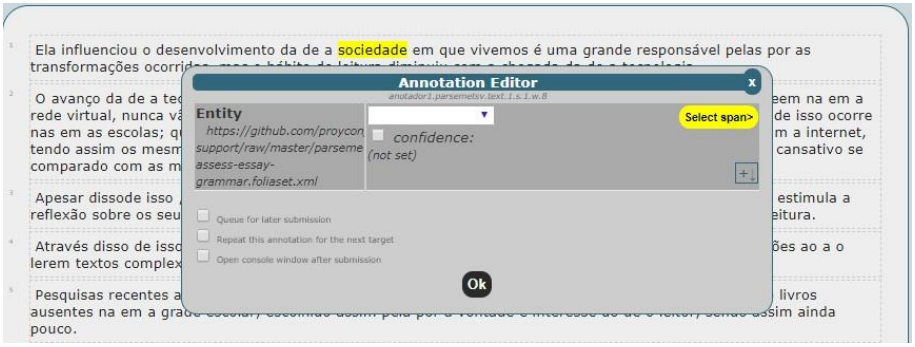
O servidor pode demorar um pouco até carregar o arquivo. Cada arquivo de anotação contém 60 sentenças, que aparecerão em uma única página na plataforma. As sentenças estão previamente separadas e numeradas de 1 a 100. Se você precisar interromper a anotação, **anote** o número da sentença na qual você parou para que possa recomeçar deste ponto.

Fazendo a anotação

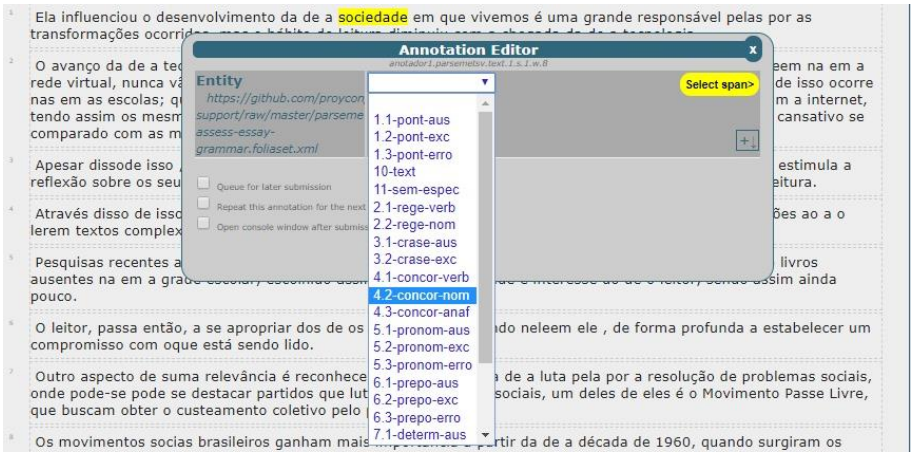
A anotação será feita clicando nos tokens que contêm os desvios descritos e atribuindo as etiquetas correspondentes, nomeadas de acordo com os códigos de cada subcategoria. Os códigos foram nomeados de forma que sejam tão intuitivos quanto possível, evitando a necessidade de consulta constante à diretriz. No entanto, recomenda-se que você tenha este documento em mãos durante a anotação, para que possa consultá-lo sempre que necessário.

Atribuindo etiquetas

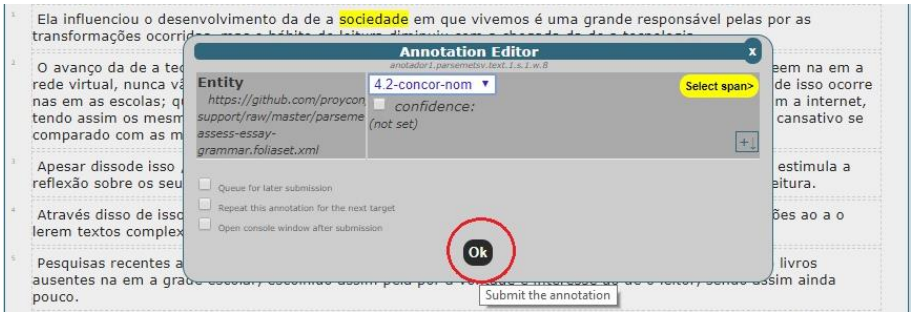
Para atribuir as etiquetas, clique sobre o token que contém o desvio (ele deverá ficar em destaque **amarelo**). A plataforma abrirá uma caixa na qual você deverá inserir a etiqueta.



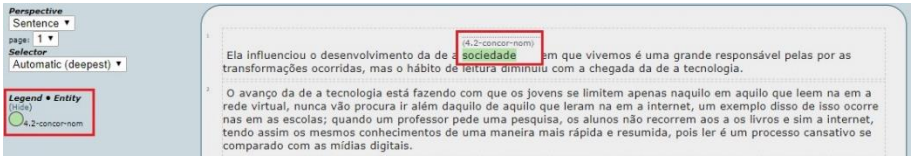
Na caixa de seleção, clique na seta de forma que apareçam as opções de etiqueta.



A seguir, clique no botão **OK.**



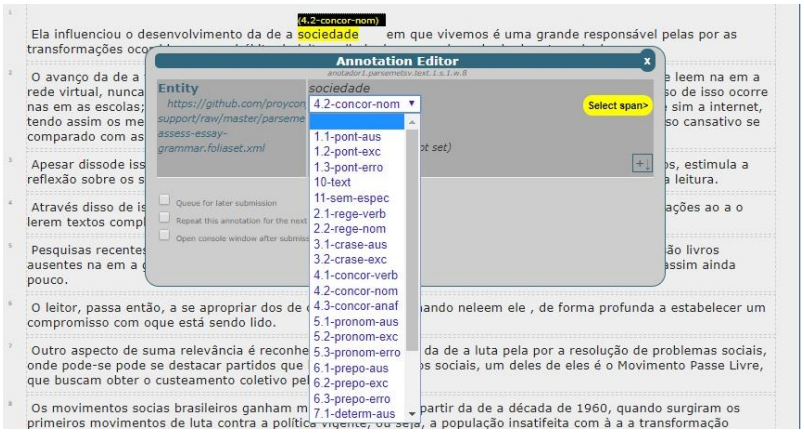
A anotação será salva automaticamente no servidor e a etiqueta atribuída aparecerá sobre o token etiquetado e no canto esquerdo da plataforma.



Você deverá anotar **todos os desvios** de cada sentença e **todas as sentenças** do seu arquivo. Sugere-se que cada anotador revise a anotação ao final, para verificar se não esqueceu nenhum desvio ou sentença, e se não há anotações equivocadas. Após concluir a anotação, envie um e-mail à organizadora da tarefa comunicando o seu término.

Deletando uma anotação

Para deletar uma anotação já submetida, basta clicar no token anotado e, na caixa de seleção, selecionar a etiqueta vazia, clicando em seguida em OK.

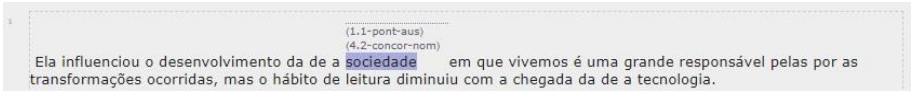


Não se preocupe se continuar aparecendo uma linha pontilhada sobre o token do qual a anotação foi removida. Eles não comprometem o status da anotação.

Anotando desvios sobrepostos

É possível que um token contenha mais de um tipo de desvio (por exemplo, um mesmo verbo apresenta um desvio de modo verbal e um desvio de concordância verbal). Para inserir mais de uma etiqueta sobre o mesmo token, faça a primeira anotação normalmente. A seguir, clique novamente no token e, logo abaixo da caixa de seleção, clique no N maiúsculo. A seguir atribua a segunda etiqueta — faça isso quantas vezes for necessário.

Ambas as anotações deverão aparecer sobre o token anotado.



Ao clicar novamente sobre o token com as duas anotações, você verá duas caixas de seleção contendo as duas anotações. Se você quiser alterar ou deletar alguma delas, siga os mesmos procedimentos descritos para uma anotação simples em cada uma das caixas de seleção.

APÊNDICE B – Forma de extração dos atributos linguísticos

A identificação automática de atributos foi realizada da seguinte forma:

1 – nº de *tokens* da sentença: contagem simples do nº de *tokens* da sentença;

2 – tipo de sentença: nº de ocorrências de “VerbForm=Fin” na coluna FEATS (se o nº de VerbForm=Fin for 1, a sentença é simples; se for maior que 1, é composta);

3 – nº de *tokens* até a raiz: contagem do nº de elementos antes de 0 na coluna HEAD;

4 – nº de verbos finitos: nº de ocorrências de “VerbForm=Fin” na coluna FEATS;

5 – nº de verbos infinitos (infinitivo, gerúndio e particípio): nº de ocorrências de “VerbForm=Inf”, “VerbForm=Ger” e “VerbForm=Part” na coluna FEATS;

6 – nº de vírgulas: contagem simples do nº de vírgulas na sentença;

7 – presença de cópula: ocorrência de “cop” na coluna DEPREL, que marca os verbos auxiliares *ser* e *estar* no português;

8 – presença de subjuntivo: ocorrência de “Mood=Sub” na coluna FEATS;

9 – presença de voz passiva: ocorrência de “Voice=Pass” na coluna FEATS;

10 – presença de formas verbais infinitas (infinitivo, gerúndio e particípio): ocorrência de “VerbForm=Inf”, “VerbForm=Ger” ou “VerbForm=Part” na coluna FEATS;

11 – presença de pronome relativo: ocorrência de “PronType=Rel” na coluna FEATS;

12 – presença de pronome relativo antes de um verbo finito: ocorrência de “PronType=Rel” na coluna FEATS antes da ocorrência de “VerbForm=Fin” na mesma coluna;

13 – presença de *que* com *POS* de conjunção: presença de um *token que* marcado como SCONJ na coluna UPOS;

14 – presença de oração relativa: ocorrência da etiqueta “acl:relcl” na coluna DEPREL;

15 – presença de relativa dentro do sujeito: ocorrência da etiqueta “acl:relcl” na coluna DEPREL somente dentro da parte da árvore identificada como NSUBJ;

16 – profundidade da árvore sintática: medida da profundidade contando a raiz e o nº de núcleos da sentença;

17 – distância média entre elementos dependentes: calcula a distância entre cada um dos elementos dependentes e divide pelo nº de dependências da sentença.