Subsídios Linguístico-Computacionais para a Revisão Gramatical Automática de Redações do Ensino Médio

Ariani Di-Felippo[1,2], Dayse Simon[3], Milena França[4], Pedro F. Martins[3]

[1]Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
Rod. Washington Luís, 235, Caixa Postal 676 - CEP 13565-905, SP– Brazil

[2]Núcleo Interinstitucional de Linguística Computacional - NILC
Av. Trabalhador Sao-carlense, 400 - Centro, São Carlos, Brasil

[3]Letrus
Av. Francisco Leitão, 469 – CEP 05414-025, SP, Brazil

[4]Bacharelado em Linguística – UFSCar

{arianidf,milecardfra}@gmail.com, {pedro,dayse}@letrus.com.br

Abstract. *We evaluate the LanguageTool grammar checker in a corpus of essays written by High School students in Brazilian Portuguese. Since LanguageTool is a rule-based open-source grammar checker, the gramatical rules with lowest precision have been improved to automate the process of grammar checking of the mentioned essays.*

Resumo. *Avalia-se o corretor gramatical LanguageTool em um corpus de redações produzidas por estudantes do ensino médio nos moldes do Enem. Posto que se trata de um corretor simbólico de código aberto, tem-se buscado refinar as regras gramaticais de menor precisão com vistas à revisão gramatical automática das referidas redações.*

1. Introdução

A correção gramatical automática (em inglês, *grammar checking*) é uma das aplicações do Processamento Automático das Línguas Naturais (PLN) mais amplamente utilizadas, sobretudo acopladas a editores ou processadores de texto como o Microsof Word, LibreOffice e OpenOffice. A correção gramatical consiste na detecção de problemas gramaticais (como de concordância, regência, uso de pronomes, etc.) quanto à modalidade escrita formal da língua e, por vezes, na sugestão de correções [Soni e Thakur 2018]. Nesse cenário, destacam-se atualmente as ferramentas *open source*, ou seja, sistemas cujo código-fonte é aberto, o qual, por conseguinte, pode ser adaptado para diferentes tarefas. Para o processamento gramatical do português, citam-se o LanguageTool [Naber 2003][1], o Vero[2] [Moura 2011] e o CoGroo [Silva 2013].

Embora a revisão gramatical seja atualmente uma área bem consolidada no PLN, a revisão gramatical automática de textos como as "redações escolares", por exemplo, é

[1] O LanguageTool é, na verdade, um corretor gramatical multilíngue, que não incluía o português em sua versão original [Naber 2003]. Atualmente, esse sistema já é capaz de processar textos nas diferentes variantes do português (https://languagetool.org/pt-BR/).

[2] O Vero era originalmente um verificador ortográfico [Moura 2011]. A partir de 2009, ele passou a englobar um corretor gramatical (https://pt-br.libreoffice.org/projetos/vero/).

um desafio para o PLN, uma vez que esses textos possuem problemas variados quanto à frequência de ocorrência e complexidade de tratamento.

Neste trabalho, apresenta-se uma avaliação do LanguageTool em um *corpus* de sentenças extraídas de redações produzidas por alunos do ensino médio como treinamento para o Exame Nacional do Ensino Médio (Enem). Entre os corretores de livre acesso, o LanguageTool fora selecionado por (i) ser um sistema de PLN simbólico (isto é, a identificação dos problemas é baseada em regras manualmente descritas) e (ii) não ter sido tão amplamente avaliado. Assim, ao se identificar os tipos de problemas gramaticais que o corretor detecta com menor eficiência, este trabalho gera subsídios linguísticos para refinar as regras do sistema, contribuindo para que este possa ser utilizado, por exemplo, na revisão gramatical automática de redações do ensino médio.

2. Avaliação do LanguageTool em um *corpus* de redações

Considerando os objetivos do trabalho, utilizou-se um *corpus* constituído de um conjunto de 82.440 sentenças, as quais compõem redações de alunos do ensino médio produzidas como treinamento para o Exame Nacional do Ensino Médio (Enem). O referido *corpus* foi cedido pela Letrus[3], que é um centro de tecnologia e letramento que desenvolve ferramentas de escrita e avaliação de textos para escolas. Especificamente, as sentenças foram coletadas em formato digital da plataforma virtual da própria empresa.

Uma vez selecionadas, as sentenças foram submetidas ao LanguageTool, que identificou um conjunto amplo de problemas nesse *corpus* de sentenças com base em 36 regras (cf. Tabela 1), as quais capturam diferentes tipos de problemas classificados como (i) capitalização, (ii) confusão de palavras, (iii) gramática, (iv) miscelânea, (v) pontuação, (vi) redundância, (vii) repetição, (viii) sintaxe e (ix) tipografia.

Do total de problemas identificados pelo corretor, 3 linguistas computacionais analisaram manualmente uma amostra aleatória de 4.043 casos e classificaram os problemas em duas categorias de detecção, comumente utilizadas no PLN [Rino *et al* 2002, Silva 2013], a saber: (i) *verdadeiros positivos* (VP) (isto é, problemas corretamente detectados pelo corretor gramatical) e (ii) *falsos positivos* (FP) (isto é, problemas identificados equivocadamente pelo corretor gramatical). Na sequência, calculou-se de forma automática a tradicional medida de *precisão* (P) [Resnik e Lin 2010].

Especificamente, a medida P indica o número de problemas corretos (de acordo com os especialistas) que foram detectados pela ferramenta em relação ao total que foi detectado. Para calcular P, tem-se a fórmula: *P = (verdadeiros positivos / (verdadeiros positivos + falsos positivos))*[4]. O cálculo de P resulta em um valor entre 0 e 1, sendo que, quanto mais próximo de 1, maior é a precisão obtida pela regra.

Na Tabela 1, tem-se as 36 regras utilizadas pelo LanguageTool para a detecção dos 4.043 casos da amostra. Nessa figura, as regras estão organizadas em função das categorias a que pertencem, segundo as informações extraídas da plataforma *online* LanguageTool Community, especificamente da área que contém as regras relativas ao portugues[5]. Na última linha da tabela, tem-se a precisão média do LanguageTool (81%).

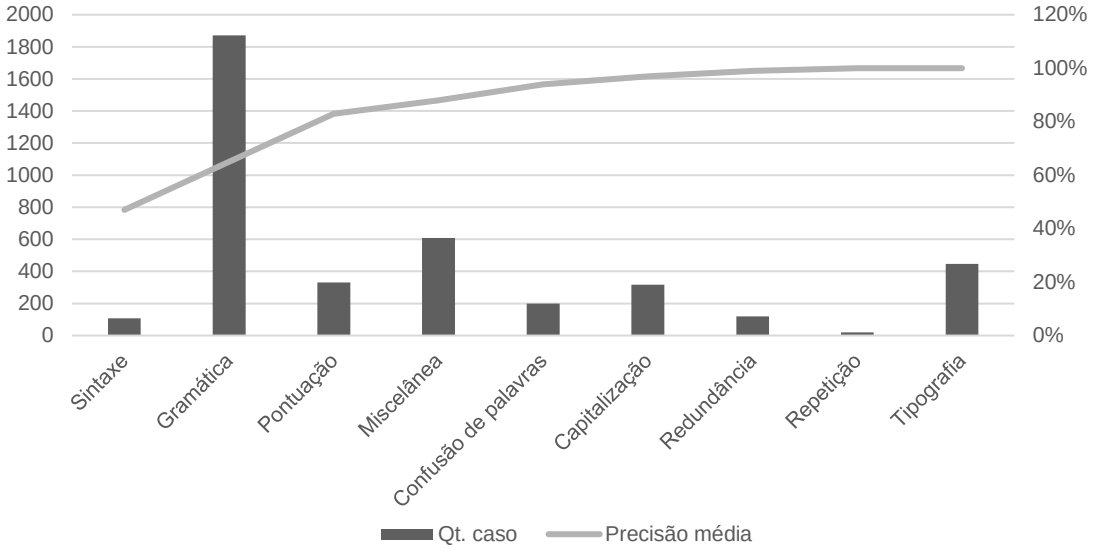
[3] https://www.letrus.com.br/.
[4] Por exemplo, a precisão P da Regra 1 da Tabela 1 foi assim calculada: P = (66 / 66 + 161) = 0.29.
[5] https://community.languagetool.org/?lang=pt-PT.

Tabela 1. Precisão das regras do LanguageTool no *corpus* de estudo.

Categoria	***Regra***	***Qt. casos***	***VP***	***FP***	***P* (%)**
Capitalização	Capitalização da frase	100	98	2	98
Confusão de palavras	Confusão entre “pratica” e “prática”	100	90	10	90
	Confusão entre “esta” e “está”	217	210	7	96
	Confusão entre “traz” e “trás”	100	99	1	99
Gramática	Concordância de gênero	227	66	161	29
	Concordância verbal	101	31	70	31
	Confusão entre verbo no passado e no futuro	67	24	43	35
	Concordância entre verbo e predicado	100	64	36	64
	Concordância de número	137	93	44	67
	Confusão entre “mau” e “mal”	142	98	44	69
	Confusão entre “a” e “há”	183	138	45	75
	Pronome oblíquo + verbo	155	117	38	75
	Concordância verbo -se + plural	100	82	18	82
	Verbo do tipo “estar” + adjetivo + “de que”	105	90	15	85
	Concordância “ser” + adjetivo	101	89	12	88
	Colocação pronominal	101	95	6	94
	Erro de crase	352	329	14	97
Miscelânea	Expressão prolixa: “mais grande”	12	7	5	58
	Expressão prolixa: “mais bom”	5	3	2	60
	Confusão entre “tem” / “têm”	109	103	6	94
	Remoção de “eu” e “nós”	101	96	5	95
	Ocorrência de “as vezes” (“às vezes”)	100	97	3	97
	Ocorrência de “afim” (“a fim”)	224	224	0	100
	Verbo “estar” + “aonde”	58	58	0	100
	Confusão entre “haver” com “a ver”	24	24	0	100
Pontuação	Locução entre vírgulas	130	86	44	66
	Ausência de pontuação final	100	83	17	83
	Abreviatura “etc”	100	100	0	100
Redundância	Conjunção redundante	100	99	1	99
	Comparativo especial: “mais melhor”	4	4	0	100
	Comparativo especial: “mais pior”	15	15	0	100
Repetição	Palavra repetida	20	20	0	100
Sintaxe	Fragmento: dois artigos seguidos	107	51	56	47
Tipografia	Espaço entre frases	105	105	0	100
	Aspas inteligentes (“ ”)	141	141	0	100
	Ocorrência de espaço antes de pontuação	200	200	0	100
	TOTAL	**4043**	**3329**	**705**	**81**

Com base na Tabela 1, pode-se dizer que, com exceção da categoria “gramática”, as demais englobam problemas bastante pontuais. A categoria “miscelânea”, por exemplo, é composta exclusivamente por regras lexicalizadas. Ademais, ao se cruzar as informações de frequência e precisão média (Figura 1), “gramática” é a categoria mais frequente (1.872 casos) e de menor precisão média na (65%).

Figura 1. Frequência e precisão média das regras por categoria.



As regras da categoria "gramática" estão organizadas na Tabela 2 em ordem crescente de precisão, posto que as regras de menor precisão ocupam o topo do ranque. Na última linha da tabela, tem-se a precisão média das regras ditas "gramaticais" do corretor (68%).

Tabela 2. Precisão das regras gramaticais do LanguageTool no *corpus* de estudo.

No.	***Regra***	***Qt. casos***	***VP***	***FP***	***P* (%)**
1ª	Concordância de gênero	227	66	161	29
2ª	Concordância verbal	101	31	70	31
3ª	Confusão entre verbo no passado e futuro	67	24	43	36
4ª	Concordância entre verbo e predicado	100	64	36	64
5ª	Concordância de número	137	93	44	67
6ª	Confusão entre "mau" e "mal"	142	98	44	69
7ª	Confusão entre "a" e "há"	183	138	45	75
8ª	Pronome oblíquo + verbo	155	117	38	75
9ª	Concordância verbo -se + plural	100	82	18	82
10ª	Verbo do tipo "estar" + adjetivo + "de que"	105	90	15	85
11ª	Concordância "ser" + adjetivo	101	89	12	88
12ª	Colocação pronominal	101	95	6	94
13ª	Erro de crase	353	439	14	97
	TOTAL	**1872**	**1426**	**546**	**68**

Partindo-se das regras mais precisas, observa-se, com base na Tabela 2, que a Regra 13, responsável por detectar os problemas de uso de crase, tem precisão de 97%, sendo a mais frequente, com 353 ocorrências na amostra de 4.043 casos. Sobre essa regra, os poucos casos de *falsos positivos* (14) se restringem à ocorrência de formas do verbo *ir* seguidas de "*até a*" e um substantivo feminino (p.ex.: "[..] *olho dentro de casa e vou até*

a porta [...]"), para as quais o corretor sugere *"até à"*[6]. A Regra 12, que identifica problemas de concordância pronominal também possui uma precisão relativamente alta de 82%. Como exemplo de *falso positivo,* cita-se o caso de "*Portanto conclui-se que* [...]". Devido à ausência de vírgula depois de "*portanto*", o LanguageTool sugere equivocadamente a anteposição do pronome ("*Portanto se conclui* [...]").

As Regras 11, 10, 9, 8, 7 e 6 são todas lexicalizadas e possuem precisão mediana, variando de 69% a 88%. Diz-se "lexicalizada" porque a aplicação destas requer a ocorrência de palavras específicas. A Regra 11, por exemplo, aplica-se somente mediante a ocorrência do verbo "*ser*". Nesse sentido, pode-se questionar a classificação da Regra 8 como lexicalizada. No entanto, essa regra, ao lidar com pronomes oblíquos, também tem seu espoco de aplicação bastante restrito.

As Regras 5, 4, 3, 2 e 1 tratam de fenômenos mais genéricos, como as concordâncias de gênero (Regra 1) e número (Regra 5). Entre elas, as Regras 3, 2 e 1, apresentam os menores índices de precisão. A Regra 3, com uma precisão sutilmente mais elevada (36%) que 2 e 1, diz respeito especificamente à confusão entre as formas verbais no passado (p.ex.: "*andaram*") e no futuro (p.ex.: "*andarão*"). Esse problema é relativamente pouco frequente, já que houve apenas 67 casos na amostra. Observa-se que a Regra 2, de concordância verbal, e a Regra 1, de concordância nominal de gênero, possuem valores de precisão muito próximos, no caso, 31% e 29%, respectivamente.

Entre os vários *falsos positivos* gerados pela Regra 2 (70) estão casos como o grifado a seguir "*A mistura cultural entre eles leva mais conhecimento* [...]", para os quais o corretor sugere que haja concordância entre o verbo e o elemento que ocorre imediatamente à sua esquerda. Para essa ocorrência em particular, o LanguageTool sugere "*A mistura cultural entre eles levam mais conhecimento* [...]". Esse tipo de equívoco parece decorrente da incapacidade do sistema em identificar dependências de mais longa distância, como é o caso da relação entre o verbo "*leva*" e o núcleo do sintagma nominal sujeito, "*mistura*", que ocorre na quarta posição à esquerda do verbo.

Entre os *falsos positivos* gerados pela Regra 1 (161) estão casos como o grifado a seguir "Com a porta dos estudos aberta, *[...]*"). Para o LanguageTool, "estudos" e "aberta" devem concordar em gênero (e também em número), sugerindo (erroneamente) a reescrita do trecho para "estudos abertos". No caso, o corretor não reconhece que o adjetivo "aberta" modifica "porta", que é o núcleo do sintagma nominal "a porta dos estudos". Isso parece ocorrer porque, embora a correção gramatical do inglês conte com um *part-of-speech tagger* derivado de [Brill 1992], que é um dos melhores da literatura, e um *chunker* (isto é, ferramenta que identifica os sintagmas), o mesmo parece não estar disponível para o processamento do português. Aparentemente, as regras do corretor são compostas por expressões regulares capazes de contemplar de forma mais ampla as concordâncias entre elementos adjacentes, como "determinante + nome" ("a porta")).

Dos dados da Tabela 2, destaca-se também que as regras, classificadas no sistema como "gramaticais", detectam no geral tipos de problemas que são cobertos pela tipologia de 17 categorias proposta por Pinheiro [Pinheiro 2008] a partir da análise manual do *corpus* CORVO, composto por 249 redações do Enem [Pinheiro 2008]. Isso indica que a referida tipologia, embora proposta em 2008, é viável para um estudo mais amplo dos tipos de problemas em redações produzidas por estudantes do ensino médio.

---

[6] A análise da aplicação dessa regra se baseou em gramáticas e dicionários que não recomendam o uso da preposição "*a*" após "*até*". Assim, entende-se que, em "*até a porta*", o "*até*" está seguido de um artigo.

3. Considerações Finais

O trabalho ora descrito revelou as regras que apresentam os maiores valores de *falsos positivos* nas redações do tipo Enem produzidas por estudantes secundaristas. Diante disso, tem-se estudado a (i) arquitetura do corretor para compreender mais amplamente como se dá a revisão gramatical e, sobretudo, (ii) as informações contidas nas regras, disponíveis *online* em formato *xml*. Na sequência, pretende-se propor refinamentos para as regras, o que pode ser feito por meio do "editor de regras" *online* do LanguageTool[7]. Acredita-se que tais refinamentos podem consistir, por exemplo, em ampliar os padrões previstos pelas expressões regulares e a indicação da necessidade de processos adicionais (p.ex.: *tagging, parsing* e correção ortográfica) para o tratamento de outras regras.

Agradecimentos. Ao CNPq, pelo suporte financeiro (Proc. Nº 401175/2018-9).

Referências

Brill, E. (1992). A simple rule-based parto f speech tagger. In Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP´92), p. 152-155.

Moura, R. (2011). Vero, the Brazilian Portuguese Spell Checker. Disponível em http://www.broffice.org/verortografico, Janeiro.

Naber, D. A. (2003). Rule-Based Style and Grammar Checker. 2003. Diplomarbeit. Technische Fakultät, Universität Bielefeld. Bielefeld.

Pinheiro, G. M. (2008). Redações do ENEM: estudo dos desvios da norma padrão sob a perspectiva de corpos. 2008. 152f. Dissertação (Mestrado em Linguística) - Faculdade de Filosofia, Letras e Ciências Humanas - FFLCH, Universidade de São Paulo.

Resnik, P.; Lin, J. (2010). Evaluation of NLP Systems. In: Clark, A; Fox, C; Lappin, S. (Ed.). The Handbook of Computational Linguistics and Natural Language Processing. Oxford: Wiley-Blackwell, p. 271-295.

Rino, L.H.M.; Di Felippo, A.; Pinheiro, G.M.; Martins, R.T.; Filié, V.M.; Hasegawa, R.; Nunes, M.G.V. (2002) Aspectos da construção de um revisor gramatical para o português do Brasil. In *Estudos Linguísticos*, v. 31. São Paulo, Brasil. ISSN 1413 0939. 1 CD-ROM.

Silva, W.D.C.M. (2013). Aprimorando o Corretor Gramatical CoGrOO. 2013. 178f. Dissertação (Mestrado em Ciência da Computação) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

Soni, M., Thakur, J.S. (2018) A Systematic Review of Automated Grammar Checking in English Language. Submitted to Computational Linguistics. Disponível em https://arxiv.org/pdf/1804.00540.pdf

[7] https://community.languagetool.org/ruleEditor2/