Proceedings of Symposium in Information and Human Language Technology. Uberlândia, MG, Brazil, October 2–5, 2017.

Uma Ferramenta para Identificar Desvios de Linguagem na Língua Portuguesa

Jonathan Nau[1], Aluizio Haendchen Filho[1], Guilherme Passero[1,2], Vinicius Cavaco[1]

[1]Núcleo de Inteligência Artificial e Sistemas Inteligentes (NIASI)
Centro Universitário de Brusque (UNIFEBE) – Brusque, SC – Brasil

[2]Laboratório de Inteligência Aplicada (LIA)
Universidade do Vale do Itajaí (UNIVALI) – Itajaí, SC – Brasil

{jonathan.naau,aluizio.h.filho,guilherme.passer,vinicavaco3}@gmail.com

Abstract. *The revision of formal texts is a complex task and occurs in several areas. The objective of this work is to create a tool to support the revision of texts and promote studies in automatic correction of descriptive texts. We propose a reviewer for automatic identification of language deviations in formal descriptive texts using natural language processing techniques. A case study was carried out to evaluate the proposed approach in a public set of essays. The tool identified 3,255 deviations in a universe of 762 essays.*

1. Introdução

Os desvios de linguagem são palavras ou construções que ferem as normas gramaticais e costumam ocorrer por descuido ou desconhecimento das regras da língua [Leal 2012, Pliger 2009]. De acordo com Rino et al. (2002), o público-alvo dos revisores gramaticais tem se revelado insatisfeito com a restrição de intervenção aos problemas ortográfico-sintáticos, exigindo a consideração de problemas relacionados ao estilo, muito mais vinculados à eficácia comunicativa do que a simples adequação à norma gramatical.

Nesse contexto, este trabalho apresenta uma ferramenta para identificação de desvios de linguagem em textos descritivos formais. Uma ferramenta com tal finalidade pode ter várias aplicações, tanto no âmbito acadêmico quanto empresarial, desde sistemas de correção automática de redações até revisão de artigos e documentos. Em pesquisas na literatura, identificamos a ferramenta LanguageTool [Miłkowski 2010] que detecta apenas alguns tipos de desvios considerados nesta pesquisa.

2. Desvios de Linguagem

Dentre os problemas relacionados ao desvio de linguagem, podem ser citados [Pinheiro 2007]: o uso de clichês/chavões, o emprego de marcas de oralidade, a repetição exagerada de palavras, sentença longa, palavras inadequadas e os vícios de linguagem. Os vícios de linguagem citados na literatura [Leal 2012, Pliger 2009] abrangem os seguintes tipos: (i) ambiguidade: uso de palavras com duplo sentido; (ii) arcaísmo: abrange expressões que caíram em desuso; (iii) barbarismo: emprego desnecessário de

palavras estrangeiras; (iv) cacófato: abrange junção de palavras que resultam em som desagradável ou obsceno; (v) colisão: aproximação de sons consonantais idênticos ou semelhantes; (vi) eco: repetição desagradável de terminações iguais; (vii) hiato: aproximação de vogais idênticas; (viii) plebeísmo: abrange qualquer desvio que caracteriza a falta de instrução (p. ex. gíria); (ix) pleonasmo: repetição desnecessária da palavra ou da ideia contida nela; (x) preciosismo: uso excessivo de palavras para exprimir ideias simples; e (xi) solecismo: desvio em relação à sintaxe.

3. Proposta de Solução

O Quadro 1 mostra o escopo da proposta de solução e as técnicas empregadas para cada tipo de desvio. Conforme o quadro, desvios como ambiguidade e preciosismo não foram considerados, visto que exigem técnicas de natureza semântico-pragmática, relacionados a níveis maiores do que a oração [Rino 2002]. O solecismo é um tipo de desvio já resolvido por alguns corretores gramaticais, p. ex. o ReGra [Rino 2002]. Os procedimentos adotados para a coleta, análise e desenvolvimento da solução estão divididos em duas partes: (i) construção dos catálogos e (ii) aplicação das técnicas.

Quadro 1. Proposta para Identificação de Desvios de Linguagem

Tipo		Proposta de solução	Catálogo	Técnica		
				Stopword	Lematização	N-gramas
Vícios de linguagem	Ambiguidade	Não	-	-	-	-
	Arcaísmos	Sim	x	x	x	x
	Barbarismo	Sim	x	-	-	x
	Cacófato	Sim	x	-	-	x
	Colisão	Sim	-	-	-	x
	Eco	Sim	-	-	-	x
	Hiato	Sim	-	-	-	x
	Plebeísmo	Sim	x	-	-	x
	Pleonasmo	Sim	x	-	x	x
	Preciosismo	Não	-	-	-	-
	Solecismo	Não	-	-	-	-
Clichês/chavões		Sim	x	-	x	x
Marcas de oralidade		Sim	x	-	-	x
Repetição exagerada de palavras		Sim	-	x	-	x
Sentença longa		Sim	-	-	-	x
Palavras inadequadas		Sim	x	-	-	x

Para a construção do corpus para teste, obteve-se 762 redações do Banco de Redações UOL (https://educacao.uol.com.br/).

3.1. Criação dos catálogos

Os catálogos são listas de palavras e expressões para detecção dos tipos de desvios relacionados, que foram elaborados com buscas de palavras e expressões em livros que tratam do assunto e complementados com pesquisas na *internet*, posteriormente foram revisados por um especialista da língua portuguesa.

O catálogo de arcaísmos foi construído utilizando expressões que não são mais utilizadas em textos formais, como “quiçá”, “jórnea”, “fatexa” e outras. Composto por 570 palavras, foi embasado na obra de Viterbo (1993), obtido na literatura portuguesa. Na construção do catálogo de barbarismo, foram utilizadas palavras estrangeiras que já possuem expressões em português, como por exemplo: “*show”, “ok”, “stop”* etc. Foi compilado com base em Gobbes & Medeiros (2009), sendo composto por 1428 verbetes.

Para construir o catálogo de cacófatos foi elaborada uma lista de palavras que juntas causam a ocorrência de uma nova palavra, alguns exemplos são, "culpa nela", "vez passada" e outras, como base para a construção do catálogo foi usado a obra de Tatiana Belinky (2010) e obtido 100 expressões. O catálogo de plebeísmo foi criado utilizando expressões que caracterizam a falta de instrução, como por exemplo, "saco cheio", "cacete", "nas quebradas" e outras formas. Composto por 265 verbetes, foi obtido com pesquisas na *Internet e* complementada por Gobbes & Medeiros (2009).

O catálogo de pleonasmo é uma lista de expressões redundantes, como "Cego dos olhos", "Regra geral", "Fato verídico", entre outras. Possui 340 expressões com base na obra de Krivochein (2015), complementadas por pesquisa na Internet. No catálogo de clichês e chavões foram adicionadas expressões bastante conhecidas, por exemplo, "via de regra", "caixinha de surpresas" e outras formas. Como base para a construção, foi usado a obra de Valente et al. (2004) e obtido 629 expressões.

Para construir as marcas de oralidade foram identificadas expressões comuns da fala e também os regionalismos, como por exemplo, "né", "aí", "tchê", "pa tu", "mermão" e outras expressões. Composto por 1121 palavras, foi embasado na obra de Negreiros (2009), intitulado Marcas de oralidade na poesia de Manuel Bandeira, e complementado por pesquisas na *Internet*. Por fim, na construção do catálogo de palavras inadequadas, foi utilizado o livro de Souto Maior (2010) intitulado Dicionário do Palavrão e Termos Afins. Neste dicionário, estão incluídos mais de 3 mil termos de uso inadequado em textos formais, que são utilizadas nas diversas regiões do Brasil, dos quais foram selecionados 654 termos para compor o catálogo.

3.2. Técnicas Utilizadas

As técnicas são aplicadas em duas fases: no pré-processamento e no processamento. No pré-processamento, o algoritmo transforma nos corpora todas as palavras para minúscula e realiza a remoção dos acentos. Na etapa de processamento são aplicadas técnicas de remoção de *stopwords* e lematização para a análise de alguns tipos de desvios. Também são extraídos os n-gramas presentes no texto e é aplicado um conjunto de procedimentos específicos para cada tipo de desvio.

A técnica de remoção das *stopwords* é utilizada para remover palavras que não contribuem para o processamento da linguagem natural. Esta técnica foi utilizada para extrair palavras não significativas no arcaísmo e repetição exagerada de palavras.

Lematização é o processo de deflexionar uma palavra para determinar o seu lema. Para Chrupala et al. (2008), a lematização é particularmente crítica para linguagens morfologicamente ricas, como o português, sendo útil para lidar com a escassez de formas não-modificadas. A lematização foi utilizada para auxiliar na identificação de arcaísmos, clichês e chavões.

De acordo com Broder et al. (1997) nos campos da linguística computacional e da probabilidade, um n-grama é uma sequência contínua de "n" itens de uma dada sequência de texto ou fala. Os itens podem ser fonemas, sílabas, letras, palavras ou pares de bases de acordo com a aplicação. A técnica de n-grama foi utilizada em todos

os tipos de desvios e vícios de linguagem para obter uma sequências de palavras no texto.

Foram gerados trigramas para identificar vícios de colisão, eco e hiato, que caracterizam repetição de sons consonantais idênticos em sequência, repetição desagradável de terminações iguais, e aproximação de vogais idênticas no início de palavras, respectivamente. Para detectar sentença longa sem ocorrência de um sinal pontuação, considerado um desvio, foi utilizado o limite de 45 palavras. Para identificar repetição exagerada, foram gerados unigramas que verificam a ocorrência mais de duas vezes de palavras (excluídos pronomes, artigos e preposições) na mesma sentença ou em sentenças próximas.

4. Resultados e Discussões

Numa amostragem de 762 redações, foram identificados pela ferramenta 3.255 desvios. O Quadro 2 apresenta a quantidade de desvios por tipo detectados nas redações. Além disso, para ilustrar, mostra alguns exemplos de desvios encontrados em cada tipo.

Quadro 2. Quantidade de desvios encontrados por tipo

Desvios	Quantidade de erros encontrados	Exemplos encontrados	
Arcaísmo	48	Libertinagem	Quiçá
Barbarismo	235	Buffet	Free
Cacofato	12	Desde então	Por tal
Colisão	273	Pequena parte possui	Nas novas necessidades
Eco	928	Os avanços tecnológicos	As mudanças trazidas
Hiato	636	Para alcançar a	Determina a autorização
Plebeísmo	22	Muitas das vezes	Troço
Pleonasmo	85	Ha muito ano atras	Si mesmo
Cliche	42	Abrir mão	Via de regra
Marcas de oralidade	300	De primeiro	Dai
Repetição exagerada	134	-	-
Sentença Longa	540	-	-

Observa-se que pequenos desvios podem comprometer a qualidade de um texto descritivo formal. Por se tratar de problemas que exigem um nível maior de atenção, estes desvios passam muitas vezes despercebidos pelos corretores humanos. Importante salientar que a maioria desses desvios não são detectados pelos editores comerciais.

Existem poucas ferramentas disponíveis na língua portuguesa para a detecção de desvios de linguagem. Uma delas é a LanguageTool [Mi‡kowski 2010], uma ferramenta de código aberto adaptada para vários idiomas, incluindo o português. Apesar de não utilizar uma gramática totalmente formalizada nem um analisador profundo [Mi‡kowski 2010], pode detectar erros de ortografia, bem como erros gramaticais e estilísticos. Verificou-se que ela detecta alguns desvios, tais como pleonasmo, marcas de oralidade, barbarismo e sentença longa. Entretanto, além de arcaísmos e cacófatos, outros desvios com muitas ocorrências, tais como eco, hiato e colisão não foram detectados. Verificou-se também que a ferramenta possui limitações na detecção de erros gramaticais, além de significativa quantidade de falsos positivos.

5. Considerações Finais

A solução se mostrou apta a encontrar vícios e desvios de linguagem, e com isso, tem

potencial para apoiar avaliadores humanos e reduzir o tempo e esforço empregado para correção. Além da correção dos textos, a ferramenta poderá ser utilizada em outras soluções, como por exemplo, a avaliação da Competência 1 do ENEM (norma culta da língua portuguesa) na correção automática de redações.

A ferramenta será incorporada a um aplicativo em desenvolvimento no NIASI para auxiliar a correção de textos descritivos formais, tais como TCCs, artigos, trabalhos acadêmicos e textos em geral. Além de erros ortográficos, em testes preliminares constatou-se que o aplicativo poderá identificar considerável quantidade de erros gramaticais não detectados pelos corretores e editores em uso na língua portuguesa.

Referências

Belinky, T. (2010) Cacoliques, Editora Melhoramentos

Broder, A. Z. et al. (1997) Syntactic Clustering of the Web. Journal Computer Networks and ISDN Systems, Amsterdam, p. 1157-1166, Sep. 1997.

Chrupala, G. et al. (2008) Learning morphology with Morfette. In: Language Resources and Evaluation, 2008, Marrakech. Proceedings of LREC, 2008, p. 2362-2367.

Gobbes, A.; Medeiros, J. B. (2009) Dicionário de Erros Correntes da Língua Portuguesa - Conforme Nova Ortografia. Editora Atlas, 5ª Ed.

Krivochein, N. (2015) A Senhorita Redundância e o Senhor Pleonasmo, Editora Brasil

Leal, E. S. (2012) Vícios de Linguagem e Idiotismos: A Fala Como Unidade de Estudos nas Gramáticas Normativas Brasileiras em Língua Portuguesa Revista Uniletras.

Negreiros, G. R. C. (2009) Marcas de Oralidade na Poesia de Manuel Bandeira, Editora Paulistana.

Maior, M. S. (2010) Dicionário do Palavrão e Termos Afins. Belo Horizonte: Editora Leitura, 2010.

Miłkowski, M. (2010) Developing an open-source, rule-based proofreading tool. Journal of Software Practice and Experience, New York, p. 543-566. Jun. 2010.

Pinheiro, G. M. (2007) "Redações do ENEM: estudo dos desvios da norma padrão sob a perspectiva de corpus". Dissertação apresentada à Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. São Paulo, 2007.

Pliger, D. B. (2009) "Vícios de Linguagem e a Norma Culta". Trabalho de Conclusão de Curso apresentado à Faculdade de Educação São Luís. São Paulo, 2009.

Rino, L. H. M. et al. (2002) Aspectos da Construção de um Revisor Gramatical Automático para o Português. Revista Estudos Linguísticos. 2002.

Viterbo, J. S. R. (1993) Elucidário das Palavras, Termos e Frases Que em Portugal antigamente Se Usaram e Que hoje regularmente Se Ignoram, Civilização Editora.

Valente, A. et al. (2004) Homem Chavão, Panda Books.