UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

GEORGES BASILE STAVRACAS NETO

Reescrita sentencial baseada em traços de personalidade

São Paulo

2018

GEORGES BASILE STAVRACAS NETO

Reescrita sentencial baseada em traços de personalidade

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 13 de março de 2018. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Ivandré Paraboni

São Paulo

2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)

CRB-8 4936

Stavracas Neto, Georges Basile

Reescrita sentencial baseada em traços de personalidade /
Georges Basile Stavracas Neto ; orientador, Ivandré Paraboni. –
2018.

69 f.

Dissertação (Mestrado em Ciências) - Programa de Pós-
Graduação em Sistemas de Informação, Escola de Artes,
Ciências e Humanidades, Universidade de São Paulo.

Versão corrigida

1. Linguagem natural - Processamento. 2. Análise de texto. 3.
Sintaxe e semântica da linguagem natural. 4. Personalidade -
Análise. I. Paraboni, Ivandré, orient. II. Título.

CDD 22.ed.– 006.35

Dissertação de autoria de Georges Basile Stavracas Neto, sob o título “**Reescrita sentencial baseada em traços de personalidade**”, apresentado à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, como parte dos requisitos para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovado em 13 de março de 2018 pela comissão examinadora constituída pelos doutores:

Prof. Dr. Ivandré Paraboni

Presidente

Instituição: Universidade de São Paulo

Profa. Dra. Ariani Di Felippo

Instituição: Universidade Federal de São Carlos

Prof. Dr. Marcelo Lauretto

Instituição: Universidade de São Paulo

Prof. Dr. Helton Hideraldo Bísaro

Instituição: Universidade de São Paulo

Agradecimentos

Meus mais sinceros e profundos agradecimentos ao Prof^o Ivandré Paraboni, pelos 30 meses de excepcional orientação; à Daniela, pelos 30 meses de excepcional paciência; à família, pelos 30 meses de excepcional compreensão; e aos amigos, pelos 30 meses de excepcional motivação.

“Ao encontrar a própria senda, é-se encontrado por ela”
(Do Autor)

Resumo

STAVRACAS NETO, Georges Basile. **Reescrita sentencial baseada em traços de personalidade**. 2018. 69 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018. Versão corrigida.

Sistemas de Geração de Língua Natural tentam produzir textos de maneira automatizada. Em sistemas desse tipo, é desejável produzir textos de forma realista - ou psicologicamente plausível - como forma de aumentar o engajamento do leitor. Uma das formas de alcançar esse objetivo é gerando textos de modo a refletir uma personalidade-alvo de interesse. Por exemplo, uma pessoa extrovertida usaria palavras mais simples e seus textos teriam mais interjeições e traços de oralidade. Esse trabalho tem o objetivo de desenvolver um modelo de reescrita sentencial para o português brasileiro com base em traços de personalidade de um locutor-alvo. Para isso, foi coletado um *cópus* de textos e inventários de personalidade e, com base em uma análise preliminar desses dados, foram encontrados fortes indícios de correlação entre os fatores de personalidade e as características observadas dos textos em português brasileiro. Foram gerados três modelos de lexicalização, referentes à adjetivos, substantivos e verbos. Esses modelos de lexicalização, então, foram utilizados na proposta de um modelo de reescrita sentencial para selecionar as palavras mais adequadas à personalidade-alvo. Os resultados demonstram que o uso de personalidade permite que o texto gerado seja mais próximo do desempenho humano se comparado a um sistema de *baseline* que faz escolhas lexicais mais frequentes.

Palavras-chave: Processamento de linguagem natural, Texto-para-texto, Reescrita de texto, Big Five.

Abstract

NETO, Georges Basile Stavracas. **Text rewrite based on psychological profiles** 2018. 69 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018. Corrected version.

Natural Language Generation Systems attempt to produce texts in an automated fashion. In systems of this kind, it is desired to produce texts realistically - or at least psychologically plausible - as a way to increase reader's engagement. One way to achieve this goal is generating texts in such a way to reflect a target personality profile. For example, an extroverted individual would use simpler words and its texts would have more interjections and orality traces. This work proposes the development of a Brazilian Portuguese personality-based sentence rewrite model. To this end, a corpus with text samples and personality inventories has been collected, and, based on a preliminary analysis, strong correlations between personality and text features have been found. Three lexicalization models were generated, related to adjectives, nouns and verbs. These models were then used by the sentence rewrite model to select the most appropriate word for the target personality. Results show that the usage of personality allows the generated text to be closer to human performance when compared to a baseline system that makes lexical choices based on frequency.

Keywords: Natural language processing, Text to text, Text rewrite. Big Five.

Lista de figuras

Figura 1 – Exemplo de figura de estímulo da base GAPED.	16
Figura 2 – Exemplo de imagem da base <i>GAPED</i>	36
Figura 3 – Diagrama demonstrando as três fases da reescrita sentencial.	48
Figura 4 – Diagrama demonstrando a fase de modificação sentencial.	50
Figura 5 – Captura de tela da ferramenta de reescrita <i>gmafell</i>	51

Lista de quadros

Quadro 1 – Lista de parâmetros na fase de planejamento do documento.	28
Quadro 2 – Lista de parâmetros na fase de seleção da estrutura sintática.	28
Quadro 3 – Lista de parâmetros na fase de agregação.	29
Quadro 4 – Lista de parâmetros na fase de inserção de marcas pragmáticas	29
Quadro 5 – Lista de parâmetros na fase de escolha lexical.	29
Quadro 6 – Excerto do corpus gerado pela coleta de dados	37
Quadro 7 – Exemplos dos cinco conceitos mais frequentes de cada classe gramatical do corpus, o número de vezes que foi utilizado, e suas possíveis lexicalizações	45
Quadro 8 – Medida F dos cinco grupos de substantivos mais frequentes.	57
Quadro 9 – Medida F dos cinco grupos de adjetivos mais frequentes.	57
Quadro 10 – Medida F dos cinco grupos de verbos mais frequentes.	57

Lista de tabelas

Tabela 1 – Resumo dos dados textuais do corpus <i>b5</i>	36
Tabela 2 – Média da frequência de palavras de sensação agrupadas pelo fator conscienciosidade. O primeiro valor, por não tem exemplos de indivíduos com conscienciosidade igual a um, foi estimado usando regressão linear	38
Tabela 3 – Categorias LIWC mais afetadas pelo fator <i>amabilidade</i>	39
Tabela 4 – Categorias LIWC mais afetadas pelo fator <i>conscienciosidade</i>	40
Tabela 5 – Categorias LIWC mais afetadas pelo fator <i>extroversão</i>	40
Tabela 6 – Categorias LIWC mais afetadas pelo fator <i>estabilidade emocional</i>	41
Tabela 7 – Categorias LIWC mais afetadss pelo fator <i>abertura a experiências</i>	41
Tabela 8 – Número de instâncias e possíveis valores das palavras encontradas no corpus <i>b5-text</i>	46
Tabela 9 – Médias ponderadas de precisão, revocação e medida F do modelo de lexicalização de substantivos. O maior valor de medida F de cada imagem é destacado	54
Tabela 10 – Médias ponderadas de precisão, revocação e medida F do modelo de lexicalização de adjetivos. O maior valor de medida F de cada imagem é destacado	54
Tabela 11 – Médias ponderadas de precisão, revocação e medida F do modelo de lexicalização de verbos. O maior valor de medida F de cada imagem é destacado	55
Tabela 12 – Médias ponderadas de precisão, revocação e medida F dos dez CRS mais frequentes do corpus. O maior valor de medida F de cada conceito é destacado	55
Tabela 13 – médias ponderadas de precisão, revocação e medida F dos dez CRA mais frequentes do corpus. O maior valor de medida F de cada CRA é destacado	56
Tabela 14 – Médias ponderadas de precisão, revocação e medida F dos dez CRV mais frequentes do corpus. O maior valor de medida F de cada CRV é destacado.	56

Tabela 15 – Médias e desvios padrão das distâncias de edição agrupadas por imagem.

O menor valor da média de cada imagem é destacado. 60

Lista de abreviaturas e siglas

CGF	Cinco grandes fatores, ou Big Five
CRA	Conceitos relacionados a adjetivos
CRS	Conceitos relacionados a substantivos
CRV	Conceitos relacionados a verbos
GLN	Geração de língua natural
IGFP-5	Inventário dos Cinco Grande Fatores de Personalidade
NEO-FFI-R	Inventário de Cinco Fatores NEO Revisado (versão resumida)
NEO-PI-R	Inventário de Personalidade NEO Revisado
PLN	Processamento de língua natural

Sumário

1	Introdução	14
1.1	<i>Objetivo</i>	17
1.1.1	Objetivos específicos	17
1.2	<i>Hipótese</i>	17
1.3	<i>Estrutura do documento</i>	18
2	Conceitos fundamentais	19
2.1	<i>O modelo dos Cinco Grandes Fatores</i>	19
2.2	<i>Geração de língua natural</i>	21
2.3	<i>Avaliando o resultado de um sistema de GLN</i>	23
2.4	<i>O dicionário LIWC-BR</i>	24
3	Revisão bibliográfica	25
3.1	<i>Variação estilística em sistemas de GLN</i>	25
3.2	<i>CGF como parâmetros para variação estilística</i>	26
3.3	<i>Efeitos da personalidade na geração de texto</i>	28
3.4	<i>Substituição lexical</i>	30
4	Metodologia	33
5	Construção do corpus <i>b5</i>	35
5.1	<i>Visão geral</i>	35
5.2	<i>Correlação entre personalidade e texto no corpus <i>b5</i></i>	37
5.2.1	Amabilidade	39
5.2.2	Conscienciosidade	39
5.2.3	Extroversão	40
5.2.4	Estabilidade emocional	40
5.2.5	Abertura a experiências	41
5.3	<i>Considerações</i>	42
6	Modelos de lexicalização	43
6.1	<i>Preparação dos dados</i>	43

6.2	<i>Modelos propostos</i>	45
6.3	<i>Treinamento</i>	46
6.4	<i>Considerações</i>	47
7	Modelo de reescrita sentencial	48
7.1	<i>Modelo proposto</i>	48
7.1.1	Normalização	49
7.1.2	Substituição lexical	49
7.1.3	Reescrita	50
7.2	<i>Implementação</i>	51
7.3	<i>Considerações</i>	52
8	Avaliação	53
8.1	<i>Avaliação dos modelos de lexicalização</i>	53
8.1.1	Resultados	53
8.1.2	Discussão	58
8.2	<i>Avaliação do modelo de reescrita sentencial</i>	59
8.2.1	Resultados	60
8.3	<i>Conclusão</i>	61
9	Conclusão	62
	Referências¹	64

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

No campo da inteligência artificial, a área que estuda as linguagens humanas é conhecida como processamento de linguagens naturais. De particular interesse para a área, a geração em línguas naturais (GLN) investiga métodos e modelos para gerar textos em língua natural de maneira automatizada. Sistemas de GLN são empregados para os mais variados propósitos, dentre os quais podem-se citar a geração de relatórios a partir de dados, previsão do tempo, sistemas de diálogo, dentre outros (REITER; DALE, 1997). Recentemente, sistemas de diálogo com funções de assistentes pessoais, bem como as técnicas que os possibilitam, como redes neurais artificiais e sistemas de síntese de fala, tem recebido especial destaque pelo desempenho que apresentam.

A geração automática de textos em língua natural a partir de bases de dados não-linguísticos, como nos exemplos citados, é conhecida como *dados-para-texto*. Em contrapartida, quando a geração automática de textos em língua natural recebe como entrada outros textos em língua natural, é conhecida como *texto-para-texto* (REITER; DALE, 1997). Exemplos de sistemas desse tipo são tradutores automatizados, sistemas de sumarização de textos, sistema de correção automática de textos, dentre outros. No presente trabalho, será tratado o caso de geração texto-para-texto, com especial ênfase em aplicações de reescrita sentencial para fins estilísticos.

Sistemas de GLN podem ter inúmeras aplicações além das tradicionais aplicações citadas. Por exemplo, na área médica, existem exemplos de sistemas de geração de históricos de tratamento médico (EUGENIO et al., 2014; JORDAN et al., 2014), geração de diagnósticos clínicos produzidos a partir da leitura de sensores de equipamentos hospitalares (PORTET et al., 2009), de geração de relatórios de comunicação entre profissionais em plantão médico (HUNTER et al., 2012) e ainda ferramentas de apoio à comunicação de pessoas com necessidades especiais (TINTAREV et al., 2016). Outras aplicações de sistemas de GLN que também podem ser citadas são exemplos de sistemas de geração de relatórios descritivos do mercado de ações a partir de indicadores financeiros (REITER; DALE, 2000), geração de diálogos realistas em sistemas de narrativa automática (WALKER et al., 2011a; WALKER et al., 2011b), geração de texto em sistemas tutores inteligentes (EUGENIO; GLASS; TROLIO, 2002) e até mesmo geração de cartas personalizadas em iniciativa contra o tabagismo (REITER; ROBERTSON; OSMAN, 2003).

Sistemas de GLN procuram produzir textos de forma natural, com desempenho comparável ao humano. No entanto, esses sistemas tendem a produzir o mesmo texto quando executados com os mesmos parâmetros de entrada, enquanto que seres humanos podem produzir textos distintos para uma mesma tarefa. Dos vários fatores que afetam a produção textual humana, um fator que se destaca é a personalidade do autor (MAIRESSE, 2008; IVANOV et al., 2011).

A personalidade humana pode ser representada por vários modelos que procuram descrever o comportamento de um indivíduo (ANDRADE, 2008). Dentre os vários modelos de personalidade propostos pela Psicologia, o modelo de traços é considerado particularmente expressivo.

O modelo de traços é descrito por Allport e Allport (1921) como uma organização dinâmica dos sistemas psicológicos e psicofísico de um indivíduo que determina seus pensamentos e comportamentos pessoais. Dessa forma, a personalidade humana é composta por traços, por vezes chamados de *dimensões*, invariantes de personalidade.

Um dos modelos de traços com grande difusão atualmente é modelo dos Cinco Grandes Fatores (CGF), também conhecido como o modelo *Big Five* (HUTZ et al., 1998). O modelo CGF é um modelo de personalidade amplamente difundido na Psicologia e também na computação, por sua facilidade de ser representado computacionalmente (MAIRESSE, 2008). Esse modelo propõe que a personalidade humana pode ser mensurada por cinco dimensões: extroversão, estabilidade emocional, amabilidade, conscienciosidade e abertura a experiências (GOLDBERG, 1990). O modelo CGF é fortemente baseado nas escolhas lexicais feitas por um indivíduo e, conseqüentemente, pessoas de personalidades distintas tendem a se expressar de forma também distinta. Por exemplo, o contexto da figura 1 de estímulo poderia ser descrito como “Em frente ao fogão” e “Vovó da estrada”, respectivamente, por indivíduos introvertidos e extrovertidos.

Sabe-se que indivíduos tendem a apresentar maior engajamento e interações mais intensas com outros indivíduos que tenham personalidades próximas. De modo geral, isso também pode ser considerado quando os agentes envolvidos são computacionais; ou seja, um indivíduo apresenta maior engajamento quando os diálogos aparentam ter sido gerados por agentes com a personalidade próxima a do leitor (MAIRESSE, 2008). Indo além, indivíduos aparentam tratar seus computadores e outros aparelhos eletrônicos de utilização pessoal como pessoas, e atribuir a esses objetos características humanas como personalidade e sentimentos (REEVES; NASS, 1996).

Figura 1 – Exemplo de figura de estímulo da base GAPED.

Fonte: Dan-Glauser e Scherer (2011)

Percebe-se, então, que a habilidade de poder gerar ou adaptar textos com base em traços de personalidade pode ser benéfica para, por exemplo, tutores virtuais, dado que o objetivo é ter forte engajamento dos alunos. Sistemas de assistentes pessoais também seriam amplamente beneficiados desta habilidade, pois é de seu interesse que os usuários interajam e se engajem com sistemas desse tipo. Outro exemplo de aplicação de sistemas capazes de gerar ou adaptar textos com base em traços de personalidade é na modelagem de personagens mais realistas por produtores e roteiristas de personagens de jogos, onde seria possível adaptar o texto escrito para a personalidade do personagem. Em outras palavras, seria em princípio possível escrever um texto de forma genérica e então utilizar uma ferramenta que implementa o modelo de reescrita automática para impor um determinado estilo de interesse de forma automática.

Apesar do papel central que a personalidade desempenha na tarefa de geração de textos em língua natural, os sistemas de GLN existentes de modo geral não levam em conta a personalidade de um autor-alvo de interesse e, nos raros casos em que um modelo de personalidade é considerado, a geração é do tipo dados-para-texto enfocando a língua inglesa (MAIRESSE, 2008; WALKER et al., 2008; BROCKMANN, 2009). Não foram identificados na literatura exemplos de sistema de GLN do tipo texto-para-texto que levem em conta a personalidade humana.

Em sistemas desse tipo, parte-se de uma sentença de entrada para então serem alteradas as características de estilo de escrita, de forma que o texto gerado pelo sistema

se aproxime de um texto gerado por um humano com a personalidade de interesse. Estas observações motivam o estudo de sistemas de GLN do tipo texto-para-texto para o português brasileiro com base em traços de personalidade. O estudo e desenvolvimento de um modelo computacional de reescrita sentencial capaz de reproduzir tais diferenças nas escolhas lexicais a partir de traços de personalidade é o foco da presente proposta de pesquisa.

1.1 Objetivo

O objetivo desta pesquisa é desenvolver um modelo de reescrita sentencial baseado em traços de personalidade para o português brasileiro capaz de fazer escolhas lexicais mais próximas do desempenho humano se comparado à simples escolha de termos lexicais mais frequentes observados em *cópus*.

1.1.1 Objetivos específicos

1. Construir, validar e analisar um *cópus* de dados composto por pares de valores de traços de personalidade de diversos autores e textos produzidos por eles.
2. Utilizar técnicas de aprendizagem de máquina supervisionada para treinar modelos de escolha lexical para conceitos realizados como substantivos, adjetivos e verbos com base no *cópus*.
3. Propor um modelo de reescrita sentencial com base em traços de personalidade baseado nos três modelos de escolha lexical.
4. Validar os modelos propostos com os métodos descritos na seção 2.3.

1.2 Hipótese

Levando-se em conta a personalidade de interesse na tarefa computacional de reescrita sentencial, é possível gerar textos mais próximos do desempenho humano do que os que seriam produzidos por métodos tradicionais de GLN baseados em estimativas de frequência em *cópus*.

Esta hipótese será testada comparando-se sentenças geradas pelo modelo proposto - que considera a personalidade-alvo de interesse - com estas mesmas sentenças geradas por métodos de *baseline* que tomam as decisões de geração mais frequentes observadas no domínio, mediante uso de métricas tradicionais de aprendizagem de máquina e de comparação de texto como distância de edição e outras.

Espera-se que o modelo de reescrita sentencial baseado em traços de personalidade apresente resultados superiores aos dos sistemas de *baseline*.

1.3 Estrutura do documento

Os conceitos fundamentais necessários para o entendimento desta obra são expostos no capítulo 2. Em especial, os conceitos relacionados ao modelo CGF são explicados na seção 2.1 e os relacionados ao processamento de línguas naturais são expostos nas seções 2.2 e 2.3.

O corpus de textos e inventários de personalidade gerado e utilizado no decorrer desta obra é descrito em detalhes no capítulo 5, em conjunto com uma análise exploratória feita para verificar possíveis correlações entre traços de personalidade e textos no corpus.

Em seguida, os modelos de lexicalização são descritos no capítulo 6. Esses modelos são utilizados pelo modelo de reescrita sentencial descrito no capítulo 7. A avaliação e resultados obtidos pelos modelos de lexicalização e o resultado do modelo de reescrita sentencial baseado em traços de personalidade são discutidos no capítulo 8. A conclusão desta obra se encontra no capítulo 9.

2 Conceitos fundamentais

2.1 O modelo dos Cinco Grandes Fatores

Desde antigos registros históricos a sofisticados modelos propostos recentemente, filósofos e psicólogos tentam compreender os elementos que compõem a psique humana. Começando com Theophrastus (371 A.C. - 287 A.C.) da Grécia antiga em sua obra *O Covarde*, até Freud e Jung com seus modernos posicionamentos, a personalidade humana foi modelada de diferentes formas. Dentre esses modelos, destacam-se (MAIRESSE, 2008):

- O modelo humanístico, que determina a personalidade como “um conjunto de metas de um indivíduo”.
- O modelo psicanalista, que define a personalidade como uma complexa cadeia de interações entre o consciente, o inconsciente e o subconsciente.
- O modelo biológico, cuja premissa principal é que a personalidade é a consequência da herança genética de um indivíduo.
- O modelo dos traços, onde a personalidade humana é composta por traços invariantes de personalidade.

Nesta última categoria, o modelo CGF é um modelo de traços de personalidade que identifica cinco traços fundamentais da personalidade humana. Foi proposto após a análise e constatação de que outros modelos, quando submetidos à análise fatorial, convergiam seus atributos para aproximadamente cinco atributos, e que outros atributos eram uma combinação desses cinco atributos essenciais. Como descrito por Hutz et al. (1998), o modelo CGF é “[...] uma generalização empírica, replicada independentemente inúmeras vezes”. Por ser um modelo empírico, não há uma explicação prévia do porquê de serem cinco traços de personalidade, e não por exemplo quatro ou seis. Dado que a língua é uma das principais interfaces de comunicação interpessoal, é natural que as pesquisas utilizadas para a geração do modelo fossem, na maioria das vezes, baseadas na análise de textos de diferentes gêneros, tais como dissertações, entrevistas e questionários (HUTZ et al., 1998).

O modelo CGF propõe que a personalidade humana pode ser mensurada por um conjunto conciso de cinco fatores escalares, o que o torna possivelmente adequado para uso em aplicações computacionais. Os cinco fatores propostos conforme Goldberg (1990) são:

1. **Extroversão:** mensura a tendência dos indivíduos a serem mais ativos, comunicativos, positivos e dominantes.
2. **Estabilidade emocional:** mensura a instabilidade emocional do indivíduo. Indivíduos neuróticos tendem a ser mais nervosos, preocupados, instáveis e sensíveis. Indivíduos pouco neuróticos, por outro lado, tendem a ser mais calmos e satisfeitos.
3. **Amabilidade:** mensura as habilidades de socialização de um indivíduo. Altas pontuações em amabilidade estão relacionadas a indivíduos agradáveis, cooperativos, afáveis e afetuosos.
4. **Conscienciosidade:** também rotulado como “falta de impulsividade”, mensura a capacidade de organização, responsabilidade e a cautela de um indivíduo.
5. **Abertura a experiências:** também “intelecto” ou “cultura” ou mesmo “imaginação”, mensura a profundidade e complexidade de um indivíduo. Indivíduos com alta pontuação no fator são mais imaginativos, artísticos e originais.

Muitos estudos sobre os CGF foram feitos utilizando dados textuais e linguísticos, como inventários, textos e entrevistas (MAIRESSE, 2008). Esse fato é relevante à discussão pois os diferentes fatores do modelo CGF causam variações consideráveis – uns mais do que outros – em diferentes aspectos do texto. Indivíduos com alto índice de extroversão, por exemplo, tendem a ter maior *output* verbal que indivíduos com baixo índice de extroversão, além de comumente apresentarem maior taxa de fala, silêncios mais curtos, maior taxa de palavras positivas e uma língua menos formal. Indivíduos com menor índice de extroversão, por outro lado, apresentam maior formalidade linguística, palavras em geral mais neutras e maior riqueza léxica, dentre outros (SIEGMAN; POPE, 1965; SCHERER, 1979; OBERLANDER; GILL, 1992; GREENE; WEINER, 2008). Da mesma forma que a extroversão, os outros fatores também se relacionam a variações das estruturas e estilos de escrita e fala dos indivíduos. As marcas de personalidade são mais evidentes de maneira proporcional a complexidade da tarefa e a pressão que o indivíduo é submetido (GOLDBERG, 1990).

Cada fator do modelo está relacionado a um conjunto de adjetivos. Por exemplo, os adjetivos “tímida”, “quieta”, “calada” e “inibida” estão fortemente relacionados a indivíduos com baixa extroversão (HUTZ et al., 1998). É importante notar, entretanto, que cada adjetivo pode estar relacionado a mais de um fator, e nem sempre é claro o quanto cada adjetivo está relacionado a cada fator (GOLDBERG, 1981). Os adjetivos

e suas correlações para cada fator de personalidade foram identificados por Hutz et al. (1998) para o português brasileiro.

A identificação dos cinco fatores de personalidade em um indivíduo é normalmente feita via inventários, dentre os quais podem ser citados o inventário de personalidade NEO revisado (NEO-PI-R), o inventário dos cinco fatores NEO revisado (NEO-FFI-R) e o inventário dos cinco grandes fatores de personalidade (IGFP-5) (JOHN; DONAHUE; L, 1991; JOHN; NAUMANN; SOTO, 2008). O último, que contém 44 questões, foi inicialmente proposto por John, Donahue e L (1991) e posteriormente validado para o português brasileiro por Andrade (2008). Também investigado por Andrade (2008), o questionário IGFP-5 produz resultados satisfatórios para o português brasileiro, dadas as referências da área. Por ser um inventário validado com base em um grande número de indivíduos, de rápida aplicação e reduzido em tamanho, o inventário IGFP-5 será o inventário adotado nesse trabalho.

2.2 Geração de língua natural

Na pesquisa em Ciência da Computação, a área que estuda as aplicações da linguística é conhecida como a área de processamento de língua natural (PLN). Seu campo de atuação abrange tanto a interpretação de textos como a geração de textos. Dentro da subárea de geração de texto, existe uma sequência bem estabelecida de passos que são normalmente usados para criar e modificar um plano inicial do documento, que é o objetivo comunicativo do documento a ser gerado. Cada passo gera uma representação intermediária, até chegar ao texto em língua natural. Esta sequência de passos, como representada pela arquitetura *pipeline* proposta por Reiter e Dale (1997), divide-se em:

- **Planejamento do conteúdo:** recebe como entrada um objetivo comunicativo, e opera sobre o mesmo para refiná-lo e para selecionar e estruturar o conteúdo a ser apresentado.
- **Planejamento sentencial:**
 - Seleção da estrutura sintática:** seleciona qual é a estrutura sintática mais adequada para representar a elocução.

Agregação: define como aglomerar elocuições de modo que seja capaz de expressar diferentes relações retóricas, como por exemplo contrastar duas sentenças, justificar uma sentença por outra, dentre outros.

Inserção de marcas pragmáticas: coloca marcas para transformar a representação sintática do enunciado.

- **Escolha lexical:** seleciona a forma superficial de cada palavra.
- **Realização superficial:** converte a representação intermediária em um texto em língua natural.

Um sistema de geração de língua natural, então, recebe dados (linguísticos ou não) como entrada e gera como saída um texto em língua natural. É comum que a entrada de um sistema de GLN normalmente seja um conjunto de dados não-linguísticos, configurando assim um sistema de dados-para-texto. Dentre as aplicações clássicas de sistemas desse tipo, podem ser citadas a geração de relatórios de previsões do tempo (GOLDBERG; KITTEDGE; DRIEDGER, 1994), sumarização de dados de uma base de dados (IORDANSKAJA et al., 1992), dentre outros.

Quando os dados de entrada de um sistema de geração de língua natural são também textos em língua natural, o sistema é conhecido como um sistema texto-para-texto. Menos comuns que sistemas do tipo *dados-para-texto*, sistemas *texto-para-texto* são tradicionalmente aplicados em áreas como traduções automáticas, sumarização de textos, dentre outras (NOVAIS; TADEU; PARABONI, 2010).

Recentemente, novas aplicações estão sendo desenvolvidas, como por exemplo a utilização de redes neurais artificiais para a tarefa de compreensão da fala (LI et al., 2016), extração de emoções (TABODA et al., 2011), mineração de dados estruturados a partir de conteúdo desestruturado (FELDMAN; SANGER, 2007), dentre outros. Essas novas aplicações mesclam técnicas de aprendizagem de máquina, processamento de sinais, processamento de línguas naturais e gerenciamento de conhecimento estruturado.

Um tipo de sistema de GLN de especial interesse para o presente trabalho é o de sistemas capazes de expressar diferentes estilos. Um sistema desse tipo, como descrito na seção 3.1, poderia usar os cinco fatores de personalidade do modelo CGF para gerar alterações no estilo de escrita do texto. A personalidade de um indivíduo se reflete na forma como o indivíduo usa a língua, e produz traços linguísticos que são reproduzíveis computacionalmente (MAIRESSE, 2008). O sistema PERSONAGE (MAIRESSE, 2008),

por exemplo, utiliza esses fatores para gerar descrições de restaurantes, como será descrito na seção 3.2.

2.3 Avaliando o resultado de um sistema de GLN

Para avaliar o desempenho de um sistema GLN, uma série métodos podem ser usados. Esses métodos são geralmente classificados como intrínsecos ou extrínsecos. Métodos de avaliação intrínsecos são aqueles baseados em propriedades inerentes e internas a um sistema, e tipicamente comparam o texto gerado com um *gold standard*, como por exemplo um *cópus* (JONES; GALLIERS, 1995).

Dentre os métodos intrínsecos tradicionais, pode-se citar o *Bleu* (PAPINENI et al., 2002), *NIST* (DODDINGTON, 2002), a avaliação pelas distâncias de edição, dentre outros. Os métodos *Bleu* e *NIST* são utilizados especialmente para verificar o desempenho de sistemas de tradução, enquanto que o método de avaliação por distâncias de edição tem aplicação mais genérica.

A avaliação por distâncias de edição é uma maneira de avaliar o quão similares são dois excertos de textos. Seu cálculo é feito pela quantidade de operações que devem ser feitas em uma sentença para que esta seja igual a outra. Tradicionalmente, considera-se as operações da distância de edição de Levenshtein: inserção, remoção e substituição (LEVENSHTEIN, 1966). Por exemplo, a distância entre os textos “criança” e “criado” é 3, pois é necessário substituir “n” por “d”, “ç” por “o” e remover o último caractere “a”.

O método de avaliação *Bleu* avalia a qualidade de textos traduzidos de uma língua natural para outra língua natural, comparando o texto gerado automaticamente com o mesmo texto gerado por um humano. A avaliação em si é feita quebrando os textos traduzidos em elementos textuais - geralmente sentenças - e comparando-os por n-gramas. Os n-gramas são comparados entre si usando distâncias de edição. Esse método de avaliação descarta a correteude dos elementos textuais (PAPINENI et al., 2002).

O método *NIST* é baseado no *Bleu*, com a diferença de que também considera a frequência em que o n-grama aparece nos textos. Quanto menor a frequência de um determinado n-grama, maior a influência que tem comparado aos outros n-gramas (DODDINGTON, 2002).

De especial interesse para esse trabalho, o método de avaliação por distâncias de edição será utilizado na avaliação dos textos gerados pelo modelo de reescrita sentencial proposto no capítulo 7.

2.4 O dicionário LIWC-BR

O LIWC (*Linguistic Inquiry and Word Count*) (PENNEBAKER; FRANCIS; BOOTH, 2001) é um *software* de análise de texto que utiliza um dicionário lexical, o dicionário LIWC, como recurso. O dicionário LIWC-BR é a versão traduzida para português brasileiro do original inglês e validada no domínio da análise de sentimentos (FILHO; ALUÍSIO; PARDO, 2013).

O dicionário LIWC-BR é uma base lexical que agrupa palavras em categorias que podem ser utilizadas para executar análises psicolinguísticas sobre os textos (PENNEBAKER; FRANCIS; BOOTH, 2001). Possui 64 categorias e 127,161 entradas. Por exemplo, as palavras “abraçar”, “acolher” e “viver” são etiquetadas com a categoria *emoção positiva*, enquanto que as palavras “aborrecer”, “iludido” e “imaturo” são etiquetadas com a categoria *emoção negativa*.

No caso do presente trabalho, o dicionário LIWC-BR será empregado para investigar possíveis correlações entre personalidade dos autores dos textos do corpus e as categorias das palavras. Esta análise é apresentada no capítulo 5.

3 Revisão bibliográfica

Nesse capítulo são revisados os estudos que tratam da variação estilística em GLN e de modelos de personalidade nesses sistemas. Em especial, os estudos de Paiva e Evans (2005) e Mairesse (2008) constituem os representantes mais expressivos do problema computacional tratado no presente trabalho - a geração de texto com base em traços de personalidade do modelo CGF - e por isso são discutidos de forma mais detalhada nas seções 3.1 e 3.2.

3.1 *Variação estilística em sistemas de GLN*

Sistemas de geração de língua natural geram sentenças e textos aplicados a um determinado cenário, de modo que o texto gerado seja apropriado para o contexto. Quando a descrição desse contexto é imprecisa ou insuficiente, entretanto, o sistema gerador deve lidar com um espaço grande de soluções válidas, mas nem sempre ótimas (PAIVA; EVANS, 2005).

Para solucionar o problema do espaço de soluções válidas muito grande, é possível, por exemplo, adaptar o gerador de língua natural para a aplicação, o que acarretaria em um sistema gerador que dificilmente poderia ser portado para outras aplicações (PAIVA; EVANS, 2005). Outra possível solução para o problema consiste em gerar um conjunto de possíveis textos, avaliá-los com uma técnica de comparação possivelmente sensível ao contexto, e então selecionar a realização mais adequada (LANGKILDE; KNIGHT, 1998; LANGKILDE-GEARY, 2002; CHAMBERS; ALLEN, 2004). Esse método tem sido usado na geração de língua natural com base em dados empíricos como avaliação humana (WALKER et al., 2007), ou estatísticas de uma base de dados previamente populada (LANGKILDE-GEARY, 2002).

A geração de textos com base em dados empíricos, entretanto, possui algumas desvantagens. Por ter que gerar muitas possíveis realizações textuais, muitas vezes milhares delas, o desempenho desses sistemas fica seriamente comprometido e podem torná-los inapropriados para aplicações em tempo real (PAIVA; EVANS, 2005).

Desde a publicação do primeiro dos modelos estatísticos de língua natural (MELN) por Langkilde e Knight (1998), esses modelos têm sido aplicados em tarefas variadas e

produzem resultados de maneira satisfatória, o que explica o crescimento rápido em seu uso (LAPATA, 2003). A principal característica dos MELN é a técnica de *overgeneration*, em que um conjunto de planos sentenciais é gerado e ordenado por um classificador na fase de planejamento sentencial.

Para que seja possível controlar o processo de decisão de um sistema GLN, é necessário ter acesso às decisões que o sistema faz até chegar ao texto final. Com a possibilidade de observar os pontos em que houve decisões, é possível então fazer uma análise de correlação estatística para detectar quais decisões afetam quais traços estilísticos (PAIVA, 2004).

O mecanismo que detecta as decisões de um sistema GLN definido por Paiva (2004) é composto de uma árvore decisória em que os nós representam as decisões que foram tomadas pelo gerador, e cada nível da árvore representa uma fase da geração do texto no sistema GLN. Como existe a possibilidade do sistema GLN ter que fazer um número grande de decisões, cada texto gerado pelo sistema gera uma árvore decisória separada, ao invés de uma única árvore com todas as possíveis decisões (PAIVA, 2004).

A partir da árvore de decisões que resultaram no texto realizado, é possível contar a quantidade de características que foram geradas. As características de estilo são qualquer aspecto do texto que é afetado pelo processo de geração, como por exemplo a quantidade de pronomes, verbos auxiliares, frases proposicionais, dentre outros. Com as informações contextuais da árvore de decisão, a contagem de decisões e o texto realizado, é feita então a análise de correlação estatística para poder detectar quais decisões afetam as características de estilo do texto e, ao final do experimento, foi gerado um modelo estatístico de geração de textos no qual os traços de estilo podem ser mapeados para parâmetros que podem ser ajustados pelo usuário final do sistema (PAIVA, 2004).

3.2 CGF como parâmetros para variação estilística

Sabendo-se da possibilidade de reproduzir e controlar variações estilísticas no processo de geração de texto em língua natural, como descrito por Paiva e Evans (2005), é de especial interesse nesta obra compreender a forma com que os fatores do modelo CGF podem ser usado como parâmetro para a variação estilística.

O estudo de Mairesse (2008) fez um levantamento das decisões de geração que podem ser manipuladas para fins estilísticos em cada fase da arquitetura de GLN, e definiu parâmetros a serem posteriormente computados com base em uma personalidade-alvo a ser imposta ao texto de saída. Nesta seção, é apresentado um levantamento dos parâmetros de GLN selecionados por Mairesse (2008), e na seção seguinte será discutida a relação entre esses parâmetros e a tarefa de geração de textos.

A reprodução das alterações dos fatores de personalidade do modelo CGF no texto requer um mapeamento de um-para-muitos, onde que um único fator pode afetar várias marcas linguísticas. Dessa forma, como exposto por Mairesse (2008), é necessário investigar e evidenciar como cada um dos traços de personalidade do modelo CGF afeta as marcas linguísticas de um determinado texto.

A tarefa de expressão de dados em língua natural requer que toda elocução linguística seja gramaticalmente e semanticamente correta. Um gerador de língua natural só é considerado válido quando atende a esse requisito. Dessa forma, a tarefa de geração de língua natural com base em traços de personalidade é mais restritiva que a tarefa de detecção de traços de personalidade, dado que a última pode detectar traços em textos gramaticalmente incorretos (MAIRESSE, 2008).

Cada um dos componentes de um sistema de GLN, descritos por Paiva e Evans (2005) na seção 3.1, podem ser controlados por uma série de parâmetros. No trabalho de (MAIRESSE, 2008), tais parâmetros são controlados de acordo com os valores dos fatores de personalidade dados como entrada. É importante notar que cada parâmetro é afetado por um ou mais traços de personalidade, e que o resultado final é considerado priorizando-se esses traços. Por exemplo, a extroversão é o fator que influencia mais fortemente a verbosidade (isto é, o tamanho) do texto; outros traços não influenciam tão fortemente. Portanto, apenas a extroversão afetaria a verbosidade.

Com cada traço devidamente observado, é possível fazer uma análise fatorial para encontrar a correlação final entre cada traço de personalidade e cada parâmetro considerado no sistema de GLN. Dado que alguns traços foram estudados mais extensivamente que outros no que se refere a comunicação, em Mairesse (2008), foram deduzidas hipóteses sobre os fatores menos estudados, de forma a minimizar a tendência do modelo de geração de língua natural em gerar melhores resultados para os traços mais estudados, como extroversão e a estabilidade emocional.

Os valores de cada parâmetro de geração do sistema de GLN, *a priori*, assumem um valor binário. Alternativamente, é possível atribuir para cada parâmetro um valor real entre $[0, 1]$ calculado pelo método de análise fatorial.

No quadro 3.2, os parâmetros da fase de planejamento do documento são descritos. No quadro 3.2, encontram-se os parâmetros de geração de texto para a fase de seleção da estrutura sintática. No quadro 3.2, encontram-se os parâmetros para a fase de agregação. No quadro 3.2, encontram-se os parâmetros da fase de inserção de marcas pragmáticas e no quadro 3.2, encontram-se os parâmetros de geração de texto da fase de escolha lexical.

Quadro 1 – Lista de parâmetros na fase de planejamento do documento.

Parâmetro	Descrição
Concessão	Dá ênfase a um atributo.
Conteúdo positivo antes	Determina se elocuições positivas aparecem antes.
Mitigação de competência	Demonstra uma avaliação negativa do ouvinte por parte do orador.
Pedido de confirmação	Determina se a elocução começa com uma confirmação do pedido do ouvinte.
Polaridade da concessão	Enfatiza atributos negativos ou positivos.
Polaridade do conteúdo	Controla a polaridade das expressões de conteúdo.
Polaridade da repetição	Controla a polaridade das reafirmações.
Polarização	O conteúdo é neutro ou extremista.
Reafirmação	Paráfrase de uma proposição existente.
Rejeição inicial	Começa a elocução com uma evasão ou rejeição.
Repetição	Repetição de uma proposição existente.
Verbosidade	Número de proposições na elocução.

Fonte: Mairesse (2008)

Quadro 2 – Lista de parâmetros na fase de seleção da estrutura sintática.

Parâmetro	Descrição
Autorreferências	Número de pronomes em primeira pessoa.
Complexidade sintática	A complexidade sintática da elocução.
Polaridade do modelo	Se a conotação do modelo é positiva ou negativa.

Fonte: Mairesse (2008)

3.3 Efeitos da personalidade na geração de texto

Conforme investigado por Mairesse (2008), sistemas de GLN do tipo dado-para-texto que levam em conta a personalidade de um locutor-alvo possuem parâmetros tais como os descritos nos quadros 3.2, 3.2, 3.2, 3.2 e 3.2.

Quadro 3 – Lista de parâmetros na fase de agregação.

Parâmetro	Descrição
Cláusula relativa	Mescla duas elocuições com uma cláusula relativa.
Conjunção	Mescla duas preposições com uma conjunção, ou vírgula caso sejam mais de duas.
Elipse do objeto	Substitui parte de uma elocução repetida por elipse.
Mesclar	Mescla duas proposições com o mesmo sujeito e verbo.
Mesclar com vírgula	O mesmo que <i>mesclar</i> , mas usando uma vírgula.
Período	Faz com que duas proposições sejam expostas em suas próprias sentenças.
Traço <i>com</i>	Mescla duas elocuições com o traço <i>com</i> .
Traço de concessão	Faz a concessão de duas proposições com os operadores <i>apesar</i> , <i>mesmo que</i> , <i>mas</i> ou <i>embora</i> .
Traço de contraste	Contrasta duas proposições com os operadores <i>com</i> , <i>mas</i> <i>entretanto</i> , <i>enquanto</i> ou <i>por outro lado</i> .
Traço de justificativa	Justifica uma proposição inserindo os operadores <i>porque</i> , <i>dado que</i> ou <i>visto que</i> .
Traço <i>também</i>	Mescla duas proposições usando <i>também</i> .

Fonte: Mairesse (2008)

Quadro 4 – Lista de parâmetros na fase de inserção de marcas pragmáticas .

Parâmetro	Descrição
Confirmadores	Insere uma palavra de confirmação no início da frase.
Enfatizadores	Insere elementos sintáticos que enfatizam uma determinada proposição.
Exclamação	Insere um ponto de exclamação.
Expletivos	Insere um expletivo.
Gaguez	Duplica partes de uma palavra de conteúdo.
Implicidade do sujeito	Faz com que um objeto fique implícito ao mover seus atributos ao sujeito.
Marcador de grupo	Faz uma referência ao ouvinte como parte do mesmo grupo social, indicando intimidade.
Negação	Nega um verbo e substitui o modificador do objeto pelo seu antônimo.
Partícula questionativa	Coloca uma partícula questionativa no final da elocução.
Pausas preenchidas	Insere elementos sintáticos que expressam hesitação.
Pronominalização	Substitui referências do objeto por pronomes.
Quasi-expletivos	Insere um quasi-expletivo.
Suavizadores	Insere elementos sintáticos como <i>meio que</i> , <i>parece que</i> , <i>acredito que</i> , <i>penso que</i> , etc, que suavizam a força da preposição.

Fonte: Mairesse (2008)

Quadro 5 – Lista de parâmetros na fase de escolha lexical.

Parâmetro	Descrição
Frequência dos léxicos	Controla o uso médio de um léxico para cada palavra de conteúdo.
Tamanho dos léxicos	Controla o número médio de letras de cada palavra de conteúdo.
Força do verbo	Controla a intensidade dos verbos usados.

Fonte: Mairesse (2008)

A extroversão é o traço de personalidade que mais afeta a comunicação, e é um dos traços com maior quantidade de estudos relacionados. A extroversão afeta o tamanho e a repetição do texto, a positividade e a polarização (extremização) do conteúdo. Está relacionada também a marcas de oralidade e a riqueza do vocabulário.

A estabilidade emocional é o segundo traço de personalidade com maior quantidade de estudos. Está relacionado a ansiedade, negatividade (ainda mais que a extroversão) e polarização do conteúdo, bem como a repetição do conteúdo. Pessoas instáveis tendem a falar a mesma informação repetidamente e tendem a colocar conteúdos negativos à frente dos conteúdos positivos, e seus textos contém mais autoreferências e artigos, e menos pronomes (OBERLANDER; GILL, 1992).

A conscienciosidade está relacionada a autodisciplina e a autopercepção. Indivíduos com alto grau de conscienciosidade tendem a ser mais diretos, evitam repetições e são mais assertivos nas suas afirmações, evitando marcadores de negação. Tal como a amabilidade, esse traço emergiu com o modelo CGF e, portanto, não foi tão extensivamente estudado quanto a extroversão e a estabilidade emocional.

O traço da amabilidade está relacionado a positividade, empatia e interesse em outras pessoas. Como é um traço que foi inicialmente exposto pelo modelo CGF, assim como a conscienciosidade, seus efeitos na comunicação não são tão estudados quanto os dois primeiros. De modo geral, indivíduos pouco agradáveis tendem a produzir sentenças mais curtas, e em maior quantidade.

A abertura a experiências, também chamado de cultura ou de intelectualidade, é o fator relacionado a criatividade, imaginação e resistência a mudanças. Está associado a maior quantidade de palavras positivas e relacionadas a exclusão como “mas” e “sem” (PENNEBAKER; KING, 1999). É o traço de personalidade que recentemente mais causa controvérsias e é o que aparentemente menos afeta a comunicação (MAIRESSE, 2008).

3.4 *Substituição lexical*

Em sistemas de GLN do tipo dados-para-texto como os discutidos nas seções anteriores, os conteúdos a serem representados em forma textual efetuam um processo de escolha lexical na fase de realização superficial. Sistemas do tipo texto-para-texto, por

outro lado, partem de um texto fornecido como entrada para então efetuar operações de substituição lexical, adaptando-o assim a um formato específico.

A substituição lexical é a tarefa de substituir palavras em um texto de acordo com um determinado critério. É utilizada para, por exemplo, simplificar textos (SHARDLOW, 2014) ou sumarizar documentos (LLORET, 2008). A tarefa de substituição lexical, de maneira geral, envolve duas grandes etapas:

1. A busca de de palavras sinônimas, onde o sistema de GLN busca candidatos a substituição lexical.
2. A classificação das palavras candidatas a substituição.

Na primeira etapa é feita a identificação de palavras que são candidatas válidas a substituição em um determinado contexto ou domínio. Esta etapa normalmente também engloba o processo de detecção da classe gramatical da palavra, e de desambiguação do contexto a palavra, caso a palavra tenha mais de uma classe gramatical possível.

Após a seleção das palavras candidatas, a segunda etapa é o ranqueamento dessas palavras de acordo com uma métrica de adequação dependente do contexto e do domínio (RIEDL; GLASS; GLIOZZO, 2014).

Para o domínio geral, o trabalho iniciado por McCarthy e Navigli (2007) usa a base de dados WordNet (MILLER, 1995) para buscar palavras candidatas a substituição lexical adequadas, e então utiliza n-gramas para ranquear essas palavras candidatas. Esse e outros trabalhos relacionados, entretanto, foram feitos para a língua inglesa, e foi observado também que pouco foi feito em relação à tarefa de substituição lexical para domínios específicos (RIEDL; GLASS; GLIOZZO, 2014), ainda na língua inglesa. Há indícios de que a substituição lexical como descrita por McCarthy e Navigli (2007) é aplicável a outras línguas.

A tarefa de substituição lexical em sistemas de GLN do tipo texto-para-texto é o tema central do presente trabalho. Diferentemente dos critérios de seleção adotados em estudos desse tipo, entretanto, será tratado o caso da substituição lexical com base na personalidade de um locutor-alvo de interesse. No caso do presente trabalho, as adjetivos e verbos sinônimos são encontradas com o auxílio do dicionário TeP 2.0 (MAZIERO et al., 2008), enquanto que substantivos utilizam uma anotação manual. A seleção da palavra mais adequada é feita usando a personalidade-alvo como critério de seleção.

Um exemplo de aplicação que faz uso extensivo de substituição lexical é o caso dos sistemas de simplificação textual (ALUÍISIO et al., 2008). Em sistemas desse tipo, é comum a substituição de palavras por sinônimos adequados ao nível de escolaridade ou habilidade linguística de um leitor-alvo de interesse. O critério de substituição nesse caso pode, por exemplo, fazer uso de métricas de qualidade textual como em Flesch (1949).

4 Metodologia

Com base na revisão bibliográfica feita no capítulo anterior, percebe-se que há grande necessidade de desenvolver estudos de sistemas de geração ou reescrita sentencial com base em traços de personalidade para o português brasileiro. Com base nesta observação, retomamos o objetivo geral desta pesquisa, tal como enunciado no capítulo 1:

O objetivo desta pesquisa é desenvolver um modelo de reescrita sentencial baseado em traços de personalidade para o português brasileiro capaz de fazer escolhas lexicais mais próximas do desempenho humano se comparado à simples escolha de termos lexicais mais frequentes observados em *cópus*.

O método adotado para a realização desse projeto consiste em três grandes partes a serem descritas nos capítulos subsequentes: a coleta de um *cópus* de textos associados aos valores dos cinco fatores de personalidade de seus autores, o desenvolvimento de modelos de substituição lexical que levam em conta uma personalidade-alvo de interesse, e, por fim, o desenvolvimento de um modelo de reescrita sentencial baseado em personalidade que utiliza os modelos de substituição lexical para executar a tarefa de reescrita.

O *cópus* coletado, denominado *cópus b5*, consiste de uma coleção de textos produzidos por participantes humanos, junto com seus perfis psicológicos, em um experimento onde o ambiente e os estímulos eram controlados. Dois componentes desse *cópus* são de especial interesse para esta pesquisa: o sub*cópus b5-text*, com descrições multisentenciais; e o sub*cópus b5-caption*, com legendas monossentenciais. Esses dois componentes do *cópus b5*, bem como o procedimento de coleta desses dados, são descritos no capítulo 5, juntamente com uma breve análise de correlação entre os textos coletados e os fatores de personalidade do modelo CGF.

O modelo de lexicalização baseado em traços de personalidade faz uso de métodos de aprendizagem de máquina para reproduzir os efeitos da personalidade na escolha lexical. O modelo é treinado com base no sub*cópus b5-text*, e incluem três classificadores que implementam as tarefas de substituição lexical de substantivos, adjetivos e verbos. Esse modelo é descrito no capítulo 6.

O modelo de reescrita sentencial faz uso do modelo de substituição lexical proposto para reformular uma sentença fornecida como entrada, com base em uma personalidade-alvo de interesse, também fornecida como entrada. Nesse processo, todos os substantivos,

adjetivos e verbos da sentença de entrada são examinados e possivelmente adaptados de acordo com a personalidade-alvo dada como entrada. Esse modelo é descrito no capítulo 7.

Finalmente, os modelos de lexicalização e reescrita sentencial propostos são avaliados de forma intrínseca usando o subcorpus de legendas *b5-caption*, e de forma extrínseca com auxílio de sujeitos humanos. Os resultados das avaliações desses modelos são apresentados no capítulo 8.

5 Construção do *cópus b5*

Esse capítulo descreve um trabalho de coleta de dados e formação de um *cópus* de textos em português brasileiro e inventários de personalidade (SANTOS; PARABONI; SILVA, 2017; RAMOS et al., 2018), bem como uma análise preliminar desses dados com o objetivo de identificar correlações entre os valores dos fatores de personalidade do modelo CGF e diferentes características do texto. O *cópus*, denominado *cópus b5*, foi coletado conforme descrito na seção 5.1 e analisado conforme descrito na seção 5.2. Os resultados obtidos fornecem indícios de que os fatores do modelo CGF afetam de maneira significativa e observável os textos produzidos por brasileiros em língua portuguesa.

5.1 *Visão geral*

Em colaboração com projetos de computação de personalidade relacionados a esta pesquisa, foi realizada a coleta de um *cópus* de textos e inventários de personalidade de seus respectivos autores do *cópus b5*. A coleta do *cópus* foi realizada de forma presencial e *online*, e incluiu três tarefas de produção de texto. Destas, duas são de especial interesse para esse projeto: as tarefas de descrição multi e monossentencial de imagens que resultaram nos sub*cópus* *b5-text* e *b5-caption*.

Juntamente com os sub*cópus* *b5-text* e *b5-caption*, foram coletados dados de expressões de referências que resultaram no sub*cópus* denominado *b5-ref*. Em um estudo preliminar, observou-se que a utilização de traços de personalidade como parâmetros para a escolha lexical supera em desempenho um modelo de lexicalização de base, que faz apenas mapeamentos de significado-para-texto (LAN; PARABONI, 2018).

As bases *b5-text* e *b5-caption* contém respectivamente descrições longas e multissentenciais e legendas de um conjunto de imagens de estímulo provenientes da base *GAPED* (DAN-GLAUSER; SCHERER, 2011) de imagens classificadas por valência e significância normativa. Estas imagens foram designadas de modo a despertar diferentes graus de emoção positiva e negativa. Um exemplo de imagem disponibilizada pela base *GAPED* (DAN-GLAUSER; SCHERER, 2011) é ilustrado na figura 2.

Foram utilizadas 10 imagens como a desse exemplo, descritas por 150 participantes. Para cada imagem, os participantes escreveram uma descrição longa, com limite mínimo

Figura 2 – Exemplo de imagem da base *GAPED*

Fonte: Dan-Glauser e Scherer (2011)

de 200 caracteres, e uma legenda, com limite máximo de 30 caracteres. Esses textos foram submetidos a correção ortográfica e análise sintática utilizando a versão online do *parser* PALAVRAS (BICK, 2000). Mesmo com possível perda de informações sobre personalidade, a correção ortográfica é necessária dado que o modelo de reescrita sentencial proposto não tem como objetivo reproduzir erros, gerando apenas sentenças corretas.

O resultado da coleta de dados foi a construção de três subcorpú (*b5-text*, *b5-caption* e *b5-ref*) do corpú *b5*, com 1500 descrições longas e mais 1500 legendas, e com aproximadamente 89.000 palavras. Um resumo das informações textuais do corpú *b5* se encontra na tabela 5.1.

Tabela 1 – Resumo dos dados textuais do corpú *b5*

Subcorpú	Sujeitos	Sentenças	Itens	Tipos
post	1 019	194 382	2 219 585	866 243
text	151	1 510	84 463	37 210
caption	151	1 510	4 896	4 121
ref	152	4 558	64 518	18 700

Fonte: Ramos et al. (2018)

Exemplos de textos produzidos em resposta ao estímulo da figura 2 e os valores das cinco dimensões de personalidade do locutor que as produziu são apresentados no quadro 5.1.

Quadro 6 – Excerto do corpus gerado pela coleta de dados

Extr.	Cons.	Neur.	Estab.	Abert.	Exemplo de texto
4,1	3,2	3,1	3,5	3,4	“trabalho escravo com condições desumanas onde os trabalhadores estão confinados em áreas cercadas de arame farpado e parece estarem sendo abusados profissionalmente, olhar não transmite esperança ou crença em algo, parece perdido.’
4,2	3,7	4,7	2,5	3,5	figura representada por um homem negro atrás de um cerca de madeira com arame farpado, ele parece estar observando algo, ou a espera de algo. aparentemente não está sozinho, está acompanhado pelo homem que está ao seu lado de boné azul.
4,6	3,6	3,4	3,4	3,6	Homem negro escorado em uma madeira que está envolta de arames farpados. formando uma cerca, o homem olha fixamente para um lugar não determinado na imagem. O homem está sem camisa e ao seu lado é possível ver parte de outro homem vestido com roupas verdes e boné azul.
4,1	3,8	3,8	1,9	3,4	A imagem mostra um homem de pele escura segurando uma cerca de arame e tabuas ao lado dele tem um outro homem que não aparece totalmente na imagem mas está bem próximo trazendo assim a sensação de estarem trabalhando na construção da cerca.
4,0	3,6	3,3	4,7	3,2	Só é possível ver metade de cima do rosto de um homem negro, pois diante dele há uma ripa de madeira sobre a qual as mãos dele estão apoiadas e com arames farpados ao redor. Claramente o homem está olhando para algo do outro lado da cerca e em seu olhar é visível o sofrimento. No canto direito da imagem é possível ver que há outro homem ao lado e o fundo está desfocado, não sendo possível determinar o que há.

Fonte: Georges Basile Stavracas Neto, 2018

5.2 Correlação entre personalidade e texto no corpus b5

Os corpus *b5-text*, *b5-caption* foram coletados de forma controlada, a partir de estímulos visuais predefinidos. O corpus *b5-text*, então, foi submetido a uma análise para verificar a existência de uma possível influência da personalidade na escolha lexical das palavras presentes, com o objetivo de verificar a viabilidade do uso desse conjunto de dados na pesquisa.

De especial interesse para a análise é o fato de que as categorias das palavras presentes no dicionário LIWC-BR (FILHO; ALUÍSIO; PARDO, 2013) podem ser usados como indicativos de personalidade (ANDRADE, 2008). Por isso, as ocorrências das categorias lexicais do LIWC-BR são aqui utilizadas como possíveis indicadores de personalidade.

Para ilustrar o possível efeito de um fator de personalidade sobre uma categoria LIWC específica, considere-se o exemplo da relação entre o fator de personalidade *conscienciosidade* e palavras da categoria de palavras de "sensação" do dicionário LIWC. Para esse caso, a tabela 2 ilustra a frequência média de palavras de sensação observadas no corpus para cada faixa de valor do fator *conscienciosidade*.

Tabela 2 – Média da frequência de palavras de sensação agrupadas pelo fator conscienciosidade. O primeiro valor, por não tem exemplos de indivíduos com conscienciosidade igual a um, foi estimado usando regressão linear

Conscienciosidade	Frequência média
1	2,40 (estimado)
2	2,25
3	2,16
4	2,07
5	1,95

Fonte: Georges Basile Stavrakas Neto, 2018

No exemplo, é possível perceber que quanto maior o valor de *conscienciosidade*, menor é a quantidade média de palavras de sensação no texto. De fato, a correlação calculada entre o fator *conscienciosidade* e a característica *palavras de sensação* é de $cor(consc, feel) = -0.99$. Igualmente, a correlação entre o fator *abertura a experiências* e *palavras de sensação* é de $cor(open, feel) = 0.99$.

Com base nestas observações, foi realizada uma análise dos dados do corpus *b5-text* com o objetivo de verificar se há correlações entre as palavras dos textos, representados pelas categorias LIWC-BR (FILHO; ALUÍSIO; PARDO, 2013) que expressam, e os valores dos fatores de personalidade.

Para determinar quais medidas de personalidade mais afetam uma dada característica do texto, os valores dos fatores de personalidade - que são números reais - foram discretizados em 5 grupos usando o arredondamento para o número inteiro mais próximo. Em seguida, para cada fator de personalidade, foi calculada a média de cada uma das características citadas.

Como os fatores de personalidade do modelo CGF são independentes entre si (ANDRADE, 2008), a princípio não é necessário considerar as combinações dos fatores de personalidade para calcular as correlações para com as características de estilo observadas nos textos.

O uso da média dos valores ao invés dos valores diretamente é necessário pois a análise busca determinar qual o valor médio esperado - e, conseqüentemente, o desvio padrão - de uma determinada característica, ao invés de determinar o valor exato de cada característica.

Os valores gerados a partir da análise da correlação estatística forneceram indícios para determinar quais fatores de personalidade do modelo CGF afetam com maior intensidade cada uma das características observadas. Apenas os fatores cujo valor absoluto da correlação é maior ou igual a 0.9 foram considerados na análise, pois são indícios de uma correlação muito forte. Os resultados obtidos são explicados observando como cada fator de personalidade individual afeta cada característica do texto que foram levados em consideração nesta análise exploratória.

5.2.1 Amabilidade

Há indícios de que o fator *amabilidade* afeta a quantidade de verbos e a quantidade de vezes que o interlocutor contrasta elementos das orações. As Categorias LIWC mais afetadas pela dimensão amabilidade podem ser vistas na tabela 3.

Tabela 3 – Categorias LIWC mais afetadas pelo fator *amabilidade*.

Categorias LIWC	Correlação
Verbos	-0.985
Discrepâncias	-0.944

Fonte: Georges Basile Stavracas Neto, 2018

5.2.2 Conscienciosidade

O fator *conscienciosidade* afeta principalmente as características estruturais do texto, como quantidade de conjunções, palavras funcionais e verbos auxiliares. Condizente com os resultados obtidos em outras pesquisas (MAIRESSE, 2008), pessoas com valores altos de conscienciosidade tendem a ser mais objetivas e, conseqüentemente, suas produções textuais são menores. As correlações encontradas podem ser vistas na tabela 4.

Tabela 4 – Categorias LIWC mais afetadas pelo fator *conscienciosidade*.

Categoria LIWC	Correlação
Sensações	-0.998
Conquistas	0.979
Trabalho	0.975
Inclinações	-0.969
Corpo	-0.963
Terceira pessoa (plural)	-0.935
Palavras funcionais	-0.984
Morte	0.928
Conjunções	-0.927
Processos biológicos	-0.923
Futuro	0.911

Fonte: Georges Basile Stavracas Neto, 2018

5.2.3 Extroversão

O fator *extroversão* é a dimensão da personalidade que mais recebeu atenção dos estudos linguísticos, por três grandes razões: a extroversão está relacionada a maior variabilidade entre os adjetivos que estão relacionados aos fatores de personalidade do modelo CGF (GOLDBERG, 1981), está presente na maior parte dos modelos de personalidade, e porque a extroversão é a dimensão que afeta com mais intensidade a comunicação de um indivíduo (MAIRESSE, 2008). Os resultados podem ser observados na tabela 5.

Tabela 5 – Categorias LIWC mais afetadas pelo fator *extroversão*.

Categoria LIWC	Correlação
Inclinações	-0.977
Tempo	-0.950
Preposições	-0.937
Negação	0.972
Assentimentos	0.906

Fonte: Georges Basile Stavracas Neto, 2018

5.2.4 Estabilidade emocional

O fator *estabilidade emocional* é o segundo fator mais estudado dentre as dimensões de personalidade do modelo CGF (MAIRESSE, 2008). Como pode ser observado na tabela

6, o fator *estabilidade emocional* teve correlações fortes com apenas quatro categorias LIWC.

Tabela 6 – Categorias LIWC mais afetadas pelo fator *estabilidade emocional*.

Traço de estilo	Correlação
Causa	0.968
Artigos	0.958
Movimento	0.952
Religiosidade	-0.920

Fonte: Georges Basile Stavracas Neto, 2018

5.2.5 Abertura a experiências

O fator *abertura a experiências* é um dos fatores de personalidade que mais causam controvérsias. De acordo com a literatura existente, é o fator que menos impacta nas produções linguísticas (MAIRESSE, 2008). Entretanto, como pode ser observado na tabela 7, foi o fator com maior número de fortes correlações encontradas com as categorias do dicionário LIWC.

Tabela 7 – Categorias LIWC mais afetadas pelo fator *abertura a experiências*.

Categoria LIWC	Correlação
Tempo	-0.997
Certeza	0.996
Sensações	-0.992
Discrepâncias	0.986
Humanidade	0.983
Advérbios	0.976
Futuro	0.975
Terceira pessoa (singular)	0.967
Percepções	-0.952
Visão	-0.951
Família	0.950
Inclusão	-0.949
Espaço	-0.928
Trabalho	0.920
Casa	0.915
Falta de fluência	0.912
Palavras sociais	0.911
Sexualidade	0.901
Movimento	-0.901
Terceira pessoa (plural)	0.901

Fonte: Georges Basile Stavracas Neto, 2018

5.3 Considerações

Nesse capítulo foi apresentado o *cópus b5*, um *cópus* de textos em português brasileiro onde há informações de personalidade dos autores de cada texto. Também foi feita uma análise dos textos, com o objetivo de verificar possíveis correlações entre as personalidades dos autores e os textos do *cópus b5-text*.

Os resultados da análise vão ao encontro do que está presente na literatura (MAIRESSE, 2008; WALKER et al., 2008; MAIRESSE; WALKER, 2011) e, assim como ocorre em outros idiomas, os fatores do modelo CGF também se refletem em características de estilo de escrita de textos na língua portuguesa.

Dentre os resultados encontrados, destaca-se a grande quantidade de correlações com o fator *abertura a experiências*. Mesmo sendo um fator de personalidade que afeta com pouca intensidade as características do texto (MAIRESSE, 2008), o fator afeta uma maior quantidade de características textuais. Outro fator que se destaca é o fator *extroversão*, que afeta poucas características textuais, mas com maior intensidade.

Esses resultados a princípio indicam a viabilidade do *cópus b5* para os estudos subsequentes, tanto pela distribuição de personalidades presente, quanto pela forte correlação entre os textos coletados e os fatores de personalidade do modelo CGF.

6 Modelos de lexicalização

Nesse capítulo são apresentados três modelos de lexicalização baseada na personalidade de um locutor-alvo de interesse. Os modelos propostos tratam da escolha lexical de conceito a serem representados por substantivos, verbos e adjetivos encontrados no *cópus b5-text* e *b5-caption* (RAMOS et al., 2018). Esses conceitos são aqui denominados conceitos relacionados a adjetivos (CRA), a substantivos (CRS) e conceitos relacionados a verbos (CRV).

Esses modelos recebem como entrada um conceito - que pode ser realizado por meio de diferentes formas lexicais sinônimas - e informações de personalidade de um locutor-alvo de interesse, e produzem como saída uma palavra adequada à entrada fornecida.

6.1 Preparação dos dados

Para o desenvolvimento dos modelos de lexicalização, foi necessário inicialmente realizar a anotação manual dos elementos visuais das dez figuras do *cópus b5-text* (RAMOS et al., 2018), mapeando os conceitos representados por substantivos (CRS) que esses elementos visuais representam. Conceitos representados por verbos (CRV) e conceitos representados por adjetivos (CRA) foram posteriormente rotulados de forma automática, utilizando o dicionário de sinônimos TeP 2.0 (MAZIERO et al., 2008) conforme descrito a seguir.

A anotação manual dos CRS foi feita da seguinte forma:

1. Enumeração de todos os elementos visuais das dez imagens do *cópus b5-text* (RAMOS et al., 2018). Utilizando a figura 2 do capítulo 5 como exemplo, foram enumerados “balde” (*2o500*), “criança” (*2h1*), “barraco” (*2o501*), dentre outros.
2. Cômputo dos substantivos presentes no *cópus*, utilizando a ferramenta PALAVRAS (BICK, 2000).
3. Descarte de substantivos com menos de cinco instâncias.
4. Extração da forma canônica dos substantivos, também utilizando a ferramenta PALAVRAS (BICK, 2000).

5. Associação de todos os substantivos restantes com os conceitos identificados no passo 1. Por exemplo, o objeto que representa o balde da figura 2 (conceito *2o*) pode ser representado como “balde”, “recipiente”, “pote”, dentre outros.

O resultado desse procedimento é um conjunto de anotações semânticas, organizadas na forma de pares de CRS e as palavras que representam. Por exemplo:

2o500 → { “balde”, “recipiente”, “pote” }

No total, foram encontrados 151 CRS associados a 191 substantivos únicos com cinco ou mais ocorrências. Substantivos que faziam referência a uma mesma entidade de cada contexto foram assim rotulados com um identificador único. Às palavras que podem representar mais de um conceito (por exemplo, na cena da figura 2, “plástico” pode fazer referência tanto ao balde quanto a lona no fundo da cena) foram atribuídos todos os conceitos a que elas poderiam ser referentes. Esta anotação foi realizada por dois anotadores, e casos ambíguos foram decididos por um terceiro.

O procedimento para rotulação de verbos e adjetivos foi realizado de forma automática utilizando o dicionário de sinônimos TeP 2.0 (MAZIERO et al., 2008). Esse procedimento gerou um mapeamento de conceitos similar ao dos substantivos. A rotulação de verbos e adjetivos foi feita do seguinte modo:

1. Cômputo dos verbos e adjetivos presentes no córpus, utilizando a ferramenta PALAVRAS (BICK, 2000).
2. Descarte de verbos e adjetivos com menos de cinco instâncias.
3. Extração da forma canônica dos verbos e adjetivos, também utilizando a ferramenta PALAVRAS (BICK, 2000).
4. Associação de todos os substantivos restantes com os *synsets* do dicionário de sinônimos TeP 2.0 (MAZIERO et al., 2008).
5. Remoção das palavras do dicionário de sinônimos TeP 2.0 (MAZIERO et al., 2008) ausentes do córpus *b5* (RAMOS et al., 2018).
6. Remoção dos *synsets* vazios e duplicados.

O resultado desses passos também é um conjunto de anotações semânticas, similar ao gerado para substantivos e organizadas da mesma forma, consistindo de pares de CRV ou CRA e as palavras que representam. Por exemplo:

$s3265 \rightarrow \{ \text{“cansar”, “debilitar”} \}$

$s4565 \rightarrow \{ \text{“escuro”, “denso”, “forte”, “intenso”} \}$

Apenas conceitos com mais de um verbo ou adjetivo foram considerados. Foram encontrados 198 CRV e 114 CRA associados a, respectivamente, 352 verbos e 191 adjetivos únicos com cinco ou mais ocorrências.

O quadro 7 mostra os cinco conceitos mais frequentes de cada classe gramatical (adjetivos, substantivos e verbos) do corpus *b5-text* (RAMOS et al., 2018), o número de vezes que suas lexicalizações são utilizadas no corpus, e as possíveis lexicalizações do conceito. Apenas conceitos com mais de uma lexicalização foram considerados.

Quadro 7 – Exemplos dos cinco conceitos mais frequentes de cada classe gramatical do corpus, o número de vezes que foi utilizado, e suas possíveis lexicalizações

Classe gramatical	Conceito	Instâncias	Lexicalizações
Adjetivo	s4559	989	Claro, branco
	s8514	823	Escuro, preto, negro
	s8461	674	Pálido, desmaiado, claro
	s8195	660	Claro, sereno, limpo
	s8223	359	Grande, longo, comprido
Substantivo	2h1	1399	Criança, menino, garoto, bebê
	2o513	761	Imagem, cena, cenário, foto
	2o501	437	Terra, barro, areia
	2o532	338	Local, área, lugar, ambiente
	2o500	312	Balde, plástico, recipiente, vasilha, pote
Verbo	s2342	8024	Haver, ser, estar, existir
	s2702	4592	Querer, relevar, estar, assentar
	s2501	3144	Ser, servir
	s2448	1600	Desfrutar, possuir, viver, ter
	s3118	1574	Incluir, conter, ter, possuir

Fonte: Georges Basile Stavrakas Neto, 2018

6.2 Modelos propostos

Os modelos de lexicalização foram implementados com uso de métodos supervisionados de aprendizagem. Três classificadores foram construídos: um para substantivos, um para adjetivos e um para verbos. Os três classificadores são do tipo multiclasse, onde o objetivo é aprender a palavra que representa cada conceito (de adjetivo, verbo ou substantivo).

Para cada palavra de interesse dos textos do corpus, as instâncias de aprendizagem são compostas pelo identificador do conceito (CRA, CRV ou CRS) e pelos cinco fatores de personalidade do modelo CGF do autor de cada texto. As classes a serem aprendidas são as próprias palavras.

Esses modelos foram denominados *b5-CRA*, *b5-CRS* e *b5-CRV*. Para fins de avaliação, foram também construídas versões reduzidas dos modelos onde as informações de personalidade foram omitidas, aos quais foram dados o nome de *baseline-CRA*, *baseline-CRS* e *baseline-CRV*.

6.3 Treinamento

Para a geração dos modelos de lexicalização de substantivos, verbos e adjetivos, foram utilizados os textos multisentenciais do corpus *b5-text* (RAMOS et al., 2018) processados pela ferramenta PALAVRAS (BICK, 2000), e os inventários de personalidade da base *b5-subject*, como descrito na seção 6.1. Um resumo da quantidade de palavras por classe gramatical encontra-se na tabela 8.

Tabela 8 – Número de instâncias e possíveis valores das palavras encontradas no corpus *b5-text*.

Classe	Instâncias	Valores possíveis
Adjetivo	5741	191
Substantivo	10428	191
Verbo	18905	352

Fonte: Ramos et al. (2018)

O conjunto de instâncias de aprendizagem foi criada utilizando-se dos materiais citados anteriormente. No corpus *b5-caption*, o número de possíveis escolhas lexicais para cada conceito variou de 2 até 12.

A seguir, foi executado o algoritmo de máquina de vetores de suporte (SVM) com *kernel* de base radial disponibilizado pelo pacote WEKA (WITTEN; FRANK; HALL, 2011). Para a construção do modelo *baseline*, os cinco fatores de personalidade foram omitidos, restando assim apenas o identificador do conceito e a palavra utilizada para representar o conceito.

O resultado desse procedimento são os modelos de lexicalização baseado em dados de personalidade (modelo *b5*) e sem dados de personalidade (modelo *baseline*) para as

três classes gramaticais. Os valores de γ e C foram adequados aos modelos utilizando a técnica de *grid search*.

6.4 Considerações

Nesse capítulo foram apresentados três modelos de lexicalização: o modelo de lexicalização de substantivos, de adjetivos e verbos. Para fins de comparação, também foram gerados três modelos alternativos sem acesso a informações de personalidade, chamados de modelos *baseline*. Esses modelos são empregados na construção de um modelo de reescrita sentencial a ser discutida no capítulo 7 e avaliada no capítulo 8.

7 Modelo de reescrita sentencial

Nesse capítulo é apresentado um modelo de reescrita sentencial que recebe como entrada os cinco valores de personalidade do modelo CGF de uma personalidade-alvo de interesse e uma sentença, e utiliza dos modelos de lexicalização apresentados no capítulo 6 para alterar a sentença de acordo com a personalidade-alvo. Os resultados da avaliação do modelo de reescrita sentencial podem ser observados no capítulo 8.

7.1 Modelo proposto

O modelo de reescrita sentencial proposto recebe como entrada os valores das cinco dimensões de personalidade do modelo CGF de um locutor-alvo, e uma sentença. O modelo, então, é definido por uma série de fases na forma de um *pipeline* onde a sentença dada como entrada é analisada e modificada, e a saída de cada fase é a entrada da próxima, até que a sentença seja efetivamente reescrita. O diagrama da figura 3 demonstra as fases. Estas fases podem ser agrupadas em três grandes grupos:

1. **Normalização:** A sentença original passa por dois tratamentos: a tokenização e a lematização dos termos constituintes. Esse procedimento gera uma lista de palavras na sua forma canônica.

Figura 3 – Diagrama demonstrando as três fases da reescrita sentencial.

Fonte: Georges Basile Stavrakas Neto, 2018

2. **Substituição lexical:** Utiliza os modelos de lexicalização para potencialmente alterar uma ou mais das palavras recebidas como entrada. Esse procedimento gera uma nova lista de palavras em suas formas canônicas, potencialmente diferente da lista de palavras dada como entrada.
3. **Reescrita:** Os lemas obtidos no processo da fase de normalização são flexionados para a obtenção da forma lexical correta, e uma sentença reescrita é gerada por esse procedimento.

Estas três fases serão discutidas individualmente nas seções a seguir. Ao longo da discussão, será utilizada a sentença da figura 1 do corpus *b5-caption*, que possui a seguinte forma:

(i) “*Família sentada sobre grama*”

7.1.1 Normalização

Na fase de normalização, a sentença recebida como entrada é tokenizada e cada palavra é convertida para sua forma canônica, com exceção de substantivos. Como a forma canônica pode potencialmente mudar a classe do substantivo, e a mudança de classe pode alterar o CRS da palavra (algo que, pelo menos no caso do corpus *b5*, não ocorre com adjetivos e verbos), lematizar os substantivos potencialmente geraria resultados errados.

Tomando por base o exemplo em (i), a frase é convertida após a a fase de normalização para a seguinte lista de palavras: {“família”, “sentada”, “sobre”, “grama”}.

7.1.2 Substituição lexical

Durante a fase de substituição lexical, para cada palavra recebida como entrada, os modelos de lexicalização são invocados para escolher a palavra mais adequada para representar os conceitos que representam. Cada modelo é invocado dependendo da classe gramatical da palavra a ser reescrita.

Uma vez que a reescrita é realizada com base nas predições dos classificadores de substantivos, adjetivos e verbos, e uma vez que esses classificadores necessariamente apresentam uma margem de erro, um procedimento adicional foi implementado com a finalidade de evitar escolhas lexicais inválidas. Por exemplo, o classificador de adjetivos

pode, em casos extremos, sugerir que “claro” seja lexicalizado como “rural”, o que obviamente representa um erro. Para evitar esta classe de problemas, que é inerente ao método de aprendizagem de máquina empregado, uma consistência adicional é realizada para garantir que somente escolhas lexicais válidas sejam de fato reescritas. Palavras sugeridas pelo modelo de lexicalização escolhido são verificadas de forma que, quando isto acontece, estas são descartadas e não ocorre substituição lexical.

Um diagrama representando a fase de substituição lexical sentencial pode ser observado na figura 4.

Figura 4 – Diagrama demonstrando a fase de modificação sentencial.

Fonte: Georges Basile Stavrakas Neto, 2018

Na sentença usada como exemplo, uma das possibilidades de modificações é {“pessoa”, “sentada”, “sobre”, “vegetação”}.

7.1.3 Reescrita

Com a lista de palavras potencialmente substituídas que foi gerada na fase de modificação, são reaplicadas as informações de gênero, número e grau extraídas durante a fase de normalização. Para isto, foi utilizado o dicionário Unitex-PB (MUNIZ, 2004).

É importante notar que a ferramenta gerada para implementar o modelo não executa a tarefa de reaplicação das flexões de gênero, número e grau dos lemas produzidos como saída da fase anterior. Entretanto, como todos os modelos são penalizados de forma equivalente, esta característica não afeta a avaliação do modelo de reescrita. A sentença utilizada como exemplo, portanto, seria reescrita como “*Pessoas sentadas sobre vegetação*”.

7.2 Implementação

Para a execução do modelo de reescrita proposto, foram geradas duas ferramentas: uma biblioteca compartilhada com abstrações sintáticas e estruturas de dados otimizadas de manipulação de textos (*libgwords*), bem como uma ferramenta que utiliza essa biblioteca e implementa o sistema de reescrita sentencial baseado em traços de personalidade proposto (*gmafell*). Uma captura de tela da ferramenta pode ser vista na figura 5.

Figura 5 – Captura de tela da ferramenta de reescrita *gmafell*.

Fonte: Georges Basile Stavrakas Neto, 2018

A modelagem da ferramenta foi feita de forma genérica, de forma que a reescrita sentencial pudesse ser feita com qualquer estratégia de reescrita. O uso de informações de personalidade é, então, apenas uma das estratégias de reescritas possíveis, mas outras estratégias poderiam ser implementadas também. Por exemplo, seria em princípio possível reescrever sentenças de acordo com qualquer classe de interesse (gênero, grau de escolaridade, etc) de um locutor-alvo específico.

A ferramenta também faz a avaliação do modelo de reescrita de maneira automatizada. A avaliação do modelo de reescrita é feito comparando uma sentença original com sua versão reescrita com base na distância de edição, conforme será discutido no próximo capítulo.

7.3 Considerações

Nesse capítulo foi descrito um modelo de reescrita sentencial que usa os modelos de lexicalização gerados no capítulo 6 para alterar sentenças de acordo com um personalidade-alvo. O modelo de reescrita sentencial foi projetado de forma que possui três fases: a normalização da sentença tomada como entrada, a substituição lexical dos lemas com o uso dos modelos de lexicalização, e por fim a reescrita sentencial. Foram geradas uma biblioteca de abstração sintática *libgwords*, e uma ferramenta auxiliar que consome suas abstrações, *gmafell*. As ferramentas geradas para a execução desse experimento são licenciadas sob GPLv3 e estão disponíveis para estudo, uso e modificação em repositório público¹. Isso compôs uma ferramenta potencialmente reutilizável, e os principais aspectos desse trabalho serão avaliados no capítulo seguinte.

¹ <https://gitlab.gnome.org/users/feaneron/projects>

8 Avaliação

Nesse capítulo são apresentados a avaliação e os resultados dos modelos de lexicalização e do modelo de reescrita sentencial anteriormente propostos.

8.1 Avaliação dos modelos de lexicalização

O objetivo desta avaliação é comparar as duas variações (*baseline* e *b5*) dos modelos de lexicalização de substantivos, adjetivos e verbos gerados como descrito no capítulo 6.

A hipótese a ser investigada nesta avaliação é a de que os modelos de lexicalização de adjetivos, verbos e substantivos que têm acesso a informações de personalidade (modelos *b5*) têm desempenho melhor que os modelos sem acesso a essas informações (modelos *baseline*).

Foram utilizados os textos processados pela ferramenta PALAVRAS (BICK, 2000) do corpus *b5-text* (RAMOS et al., 2018). Os modelos foram avaliados por meio de validação cruzada com 10 *folds*.

8.1.1 Resultados

Os resultados dos modelos *baseline* e *b5* de lexicalização de substantivos, verbos e adjetivos são apresentados a seguir. Na apresentação desses resultados, os dados obtidos serão visualizados de três formas: pelo desempenho médio por imagem, pelo desempenho por conceito e pelo desempenho por grupo de conceitos.

Tabela 9 – Médias ponderadas de precisão, revocação e medida F do modelo de lexicalização de substantivos. O maior valor de medida F de cada imagem é destacado

Imagem	baseline			b5		
	Precisão	Revocação	Medida F	Precisão	Revocação	Medida F
1	0,379	0,573	0,442	0,750	0,769	0,728
2	0,459	0,649	0,529	0,783	0,796	0,741
3	0,291	0,509	0,362	0,698	0,648	0,575
4	0,625	0,748	0,668	0,804	0,835	0,807
5	0,414	0,609	0,482	0,718	0,721	0,682
6	0,498	0,658	0,552	0,752	0,745	0,700
7	0,445	0,645	0,518	0,709	0,737	0,692
8	0,663	0,803	0,722	0,698	0,836	0,781
9	0,400	0,584	0,458	0,663	0,695	0,654
10	0,395	0,582	0,457	0,702	0,683	0,645

Fonte: Georges Basile Stavrakas Neto, 2018

Tabela 10 – Médias ponderadas de precisão, revocação e medida F do modelo de lexicalização de adjetivos. O maior valor de medida F de cada imagem é destacado

Imagem	baseline			b5		
	Precisão	Revocação	Medida F	Precisão	Revocação	Medida F
1	0,760	0,720	0,713	0,811	0,798	0,787
2	0,723	0,700	0,695	0,865	0,793	0,809
3	0,762	0,801	0,761	0,819	0,828	0,806
4	0,802	0,825	0,797	0,859	0,838	0,831
5	0,784	0,748	0,744	0,860	0,786	0,795
6	0,884	0,873	0,866	0,894	0,881	0,876
7	0,808	0,802	0,786	0,867	0,843	0,834
8	0,812	0,822	0,799	0,861	0,814	0,819
9	0,760	0,794	0,757	0,793	0,831	0,801
10	0,733	0,756	0,733	0,819	0,791	0,784

Fonte: Georges Basile Stavrakas Neto, 2018

Tabela 11 – Médias ponderadas de precisão, revocação e medida F do modelo de lexicalização de verbos. O maior valor de medida F de cada imagem é destacado

Imagem	baseline			b5		
	Precisão	Revocação	Medida F	Precisão	Revocação	Medida F
1	0,742	0,730	0,703	0,766	0,764	0,742
2	0,703	0,638	0,620	0,738	0,677	0,667
3	0,798	0,795	0,762	0,806	0,818	0,792
4	0,696	0,718	0,680	0,740	0,733	0,703
5	0,769	0,795	0,795	0,778	0,804	0,771
6	0,780	0,809	0,774	0,817	0,826	0,796
7	0,788	0,788	0,770	0,797	0,788	0,772
8	0,752	0,788	0,753	0,730	0,789	0,742
9	0,755	0,771	0,739	0,770	0,793	0,760
10	0,806	0,824	0,794	0,827	0,843	0,843

Fonte: Georges Basile Stavrakas Neto, 2018

Já em relação ao desempenho médio por conceito, as tabelas 12, 13 e 14 ilustram resultados semelhantes aos resultados em relação ao desempenho médio por imagem. Os modelos *b5*, que levam em consideração as informações de personalidade, tem resultados sistematicamente superiores aos modelos *baseline*.

Tabela 12 – Médias ponderadas de precisão, revocação e medida F dos dez CRS mais frequentes do corpus. O maior valor de medida F de cada conceito é destacado

CRS	baseline			b5		
	Precisão	Revocação	Medida F	Precisão	Revocação	Medida F
5h7	0,580	0,723	0,644	0,607	0,678	0,640
10h3	0,664	0,550	0,501	0,653	0,591	0,612
9h1	0,520	0,531	0,525	0,586	0,514	0,541
8h1	0,356	0,367	0,361	0,610	0,433	0,459
10h2	0,710	0,503	0,468	0,612	0,584	0,588
10h1	0,434	0,589	0,494	0,548	0,582	0,522
4a0	0,610	0,690	0,647	0,614	0,666	0,639
3h1	0,605	0,817	0,695	0,624	0,763	0,686
6h1	0,578	0,774	0,662	0,674	0,712	0,684
7h1	0,198	0,229	0,213	0,430	0,249	0,268

Fonte: Georges Basile Stavrakas Neto, 2018

Tabela 13 – médias ponderadas de precisão, revocação e medida F dos dez CRA mais frequentes do corpus. O maior valor de medida F de cada CRA é destacado

CRA	baseline			b5		
	Precisão	Revocação	Medida F	Precisão	Revocação	Medida F
s10484	0,755	0,750	0,722	0,837	0,764	0,791
s4650	0,399	0,500	0,444	0,732	0,664	0,665
s4148	0,276	0,500	0,356	0,487	0,491	0,466
s8365	0,721	0,750	0,687	0,719	0,767	0,727
s4645	0,306	0,500	0,380	0,923	0,834	0,859
s4490	0,424	0,500	0,459	0,923	0,749	0,791
s8514	0,445	0,500	0,422	0,624	0,550	0,545
s4565	0,897	0,500	0,642	0,849	0,557	0,670
s9761	0,300	0,334	0,316	0,422	0,445	0,431
s9782	0,300	0,334	0,316	0,850	0,889	0,869

Fonte: Georges Basile Stavracas Neto, 2018

Tabela 14 – Médias ponderadas de precisão, revocação e medida F dos dez CRV mais frequentes do corpus. O maior valor de medida F de cada CRV é destacado.

CRV	baseline			b5		
	Precisão	Revocação	Medida F	Precisão	Revocação	Medida F
s2448	0,409	0,500	0,450	0,910	0,504	0,458
s1678	0,510	0,511	0,459	0,646	0,544	0,551
s1897	0,696	0,631	0,621	0,696	0,631	0,621
s3107	0,812	0,584	0,676	0,927	0,470	0,582
s3933	0,412	0,373	0,380	0,528	0,413	0,449
s4026	0,565	0,543	0,526	0,636	0,557	0,590
s2227	0,441	0,500	0,469	0,645	0,558	0,562
s2177	0,740	0,500	0,597	0,732	0,480	0,580
s3716	0,411	0,466	0,408	0,762	0,674	0,669
s2076	0,485	0,611	0,536	0,706	0,739	0,708

Fonte: Georges Basile Stavracas Neto, 2018

Por fim, os resultados do desempenho por grupo de conceitos são ilustrados nos quadros 8, 9 e 10. No quadro 8, é possível notar que o modelo *baseline* escolhe a forma superficial mais frequente para um conceito, enquanto que o modelo *b5* faz decisões melhores e consegue alcançar taxas de acerto relevantes em outras formas superficiais além da mais frequente. Nos resultados apresentados pelo quadro 9, é possível ver que o desempenho do modelo *baseline* é eventualmente superior ao do modelo *b5* no adjetivo mais frequente de cada *synset*. Entretanto, nos adjetivos menos frequentes, o modelo *b5* tem um desempenho melhor.

Quadro 8 – Medida F dos cinco grupos de substantivos mais frequentes.

Substantivo	baseline	b5
foto	0,453	0,526
imagem	0,559	0,621
cena	0,000	0,325
criança	0,640	0,641
jovem	0,000	0,000
casaco	0,691	0,667
jaqueta	0,000	0,330
senhora	0,689	0,682
mulher	0,638	0,622
calça	0,861	0,821
jeans	0,000	0,345

Fonte: Georges Basile Stavracas Neto, 2018

Quadro 9 – Medida F dos cinco grupos de adjetivos mais frequentes.

Adjetivo	baseline	b5
alto	0,825	0,850
forte	0,648	0,706
bravo	0,692	0,818
simples	0,888	0,885
singelo	0,000	0,444
rural	0,711	0,632
rústico	0,000	0,300
velho	0,629	0,706
idoso	0,744	0,748
preocupado	0,759	0,917
apreensivo	0,000	0,800

Fonte: Georges Basile Stavracas Neto, 2018

Quadro 10 – Medida F dos cinco grupos de verbos mais frequentes.

Verbo	baseline	b5
mostrar	0,727	0,740
dar	0,496	0,599
revelar	0,000	0,667
dividir	0,333	0,750
destacar	0,800	0,810
separar	0,476	0,564
carregar	0,582	0,556
passar	0,611	0,603
tirar	0,930	0,923
separar	0,476	0,564
afastar	0,000	0,200
espalhar	0,556	0,581
dar	0,496	0,599

Fonte: Georges Basile Stavracas Neto, 2018

8.1.2 Discussão

Como resultado do experimento de geração de modelos de lexicalização, duas variações de três modelos de lexicalização foram gerados e avaliados. Ao analisar o desempenho dos modelos, com dados de personalidade (modelo *b5*) e sem dados de personalidade (modelo *baseline*), os modelos com informações de personalidade tiveram um desempenho sistematicamente superior aos dos modelos sem acesso a essas informações. Esse melhor desempenho é observado tanto ao avaliar os conceitos de todas as figuras quando ao avaliar os conceitos de cada figura individualmente, um indício de que não é um fenômeno específico do *córpus b5* (RAMOS et al., 2018). O resultado também vai ao encontro ao que é citado na literatura (MAIRESSE, 2008).

Nas tabelas 9, 10 e 11, é possível perceber que os modelos *b5* tiveram um resultado médio melhor que os modelos *baseline*. É interessante notar que os modelos *baseline* quase sempre tem valores de revocação maiores que os modelos *b5*, enquanto que os modelos *b5* possuem maior precisão que os modelos *baseline*.

Os maiores valores de revocação no modelo *baseline* são causados pelo desbalanceamento da base de dados. Como o *córpus b5-text* (RAMOS et al., 2018) não possui uma distribuição uniforme de palavras, é comum a existência de formas superficiais de um determinado conceito com um número de instâncias muito maior que outras palavras do mesmo determinado conceito. Portanto, ao selecionar a palavra com maior número de instâncias, é natural que o valor de revocação seja maior também, em comparação ao modelo *b5* que toma decisões mais informadas.

É possível observar pelas tabelas 12, 13 e 14 que os modelos de lexicalização *b5* tiveram um desempenho superior ao dos modelos *baseline* na maioria dos conceitos. Nas três classes gramaticais analisadas, os modelos *b5* tem o valor da medida F em geral maior que os modelos *baseline*. Esses resultados são de especial interesse para esta análise por serem os grupos de palavras mais frequentes do *córpus b5-caption* e, portanto, tem maior relevância para a avaliação do desempenho dos modelos.

Por fim, de acordo com os quadros 8, 9 e 10, que agrupam as palavras pelos conceitos, um fenômeno do modelo *baseline* é o desempenho alto para a palavra mais frequente de um conceito, mas falhar em todas as outras palavras de um dado conceito (CRS, CRA ou CRV). O modelo *b5*, entretanto, tem um desempenho ligeiramente mais

baixo nas palavras mais frequentes de cada conceito, mas consegue alcançar uma taxa de acerto para as outras palavras do mesmo conceito. Esse fenômeno ocorre com substantivos, adjetivos e verbos.

Como as palavras podem representar vários conceitos diferentes, é possível que uma determinada palavra seja o mais frequente de um dado conceito, mas não seja a mais frequente de outros. Por exemplo, o verbo “dar” é o verbo mais frequente do conceito *s3151*, mas também aparece nos conceitos *s4026* e *s3716*. Nesses casos, com o modelo *baseline*, o desempenho do verbo é o mesmo em todos os CRVs ao qual pertence.

De modo geral, os modelos *b5*, treinados com acesso a informações de personalidade, tiveram desempenhos melhores que os modelos *baseline*. A diferença de desempenho é similar ao analisar os textos de cada imagem individualmente, um possível indicador de que os resultados são fenômenos intrínsecos a manifestação de personalidade na escolha de palavras, e não um fenômeno específico do corpus *b5-text*. Os resultados se mantêm positivos aos modelos *b5* quando observados os conceitos mais frequentes do corpus. Ao observar as palavras mais frequentes agrupadas, é possível observar o comportamento dos modelos *b5* em comparação aos modelos *baseline*.

8.2 Avaliação do modelo de reescrita sentencial

Como complemento a avaliação dos modelos de lexicalização individuais discutida na seção anterior, nesta seção apresentamos os resultados da avaliação da tarefa de reescrita sentencial que utiliza esses modelos para substituir palavras por candidatos mais adequados de acordo com uma personalidade-alvo. Tal como a avaliação dos modelos de lexicalização, o objetivo da avaliação do modelo de reescrita sentencial é comparar seu desempenho quando utilizados os modelos de lexicalização com acesso a informações de personalidade (*b5*), e os sem acesso a essas informações (*baseline*).

A hipótese investigada nesta avaliação é a de que as frases geradas pelo modelo de reescrita proposto com base na personalidade do locutor-alvo serão mais próximas das frases que esse autor teria escrito do que as frases geradas por um modelo *baseline* que não leva em conta a informação de personalidade. Esta hipótese foi testada comparando-se as frases geradas pelo modelo proposto *b5* e *baseline* com frases (humanas) provenientes do corpus *b5-caption*, e medindo-se distância de edição entre elas. Esperava-se que o modelo

de reescrita com informações de personalidade fizesse escolhas lexicais tais que as sentenças geradas sejam idênticas ou próximas às sentenças originais.

Os modelos de lexicalização foram treinados com os dados do corpus *b5-text*, e as legendas do corpus *b5-text* (RAMOS et al., 2018) foram utilizadas como dados de teste. O avaliação do modelo de reescrita é feito comparando uma sentença original com sua versão reescrita.

8.2.1 Resultados

Os resultados das distâncias de edição para as duas variantes dos modelos de lexicalização (*b5* e *baseline*) encontradas podem ser observadas na tabela 15.

Tabela 15 – Médias e desvios padrão das distâncias de edição agrupadas por imagem. O menor valor da média de cada imagem é destacado.

Imagem	baseline		b5	
	Média	Desv.	Média	Desv.
1	3,284	3,005	2,814	3,292
2	3,523	1,930	2,231	2,408
3	3,019	3,692	2,913	3,733
4	0,450	3,276	0,562	3,511
5	1,105	3,390	1,264	3,250
6	1,072	2,652	1,298	3,058
7	2,887	2,696	2,278	3,138
8	0,324	1,470	0,364	1,721
9	3,337	4,224	2,701	3,744
10	2,271	3,607	1,569	2,918
Total	3,365	3,362	2,529	3,364

Fonte: Georges Basile Stavrakas Neto, 2018

A reescrita sentencial que usa modelos com informações de personalidade tem menor distância de edição em mais da metade das imagens. Observa-se também que, quanto mais reescrita ocorre, maior a diferença entre os modelos, e maior também é a vantagem da reescrita que usa informações de personalidade.

O desvio padrão foi em geral maior no modelo de reescrita *b5*, o que indica uma variação maior nas distâncias de edição em comparação ao modelo de reescrita *baseline*.

Conforme observado na tabela 15, quando o sistema utiliza os modelos *b5*, apresenta distância de edição média menor do que quando utiliza os modelos *baseline*. A diferença é significativa de acordo com um teste ANOVA de fator único ($F(1, 1624)=13,23$, $MSE=11,39$, $p<0,001$). Esse resultado confirma a hipótese de pesquisa.

8.3 Conclusão

Nesse capítulo, foram avaliados os modelos de lexicalização (de adjetivos, substantivos e verbos), e também o modelo de reescrita sentencial. Foi observado que, em todas as avaliações, os modelos que consideravam as informações de personalidade tiveram desempenho melhor - de acordo com as métricas adotadas em cada avaliação - que os modelos sem acesso a informações de personalidade.

Esses resultados, tomados em conjunto, oferecem suporte à hipótese principal desse trabalho de que, levando-se em conta a personalidade-alvo de interesse na tarefa computacional de reescrita sentencial, é possível gerar textos mais próximos do desempenho humano do que os que seriam produzidos por métodos tradicionais de GLN baseados em estimativas de frequência em *cópus*.

9 Conclusão

Esta monografia apresentou um estudo em nível de mestrado que teve como foco o problema computacional da reescrita sentencial com base em traços de personalidade. Esse estudo apresentou os seguintes resultados:

- Os corpúscos *b5-text* e *b5-caption*, integrantes do corpúscos *b5* (RAMOS et al., 2018).
- Os modelos de lexicalização de substantivos, adjetivos e verbos, gerados a partir do corpúscos *b5-text* e testados com o corpúscos *b5-caption*.
- O modelo de reescrita sentencial que utiliza os traços de personalidade de um locutor-alvo.

O estudo iniciou-se com a coleta de um corpúscos de textos e inventários de personalidade. Com base em uma análise preliminar dos subcorpúscos *b5-text* e *b5-caption*, foram encontrados indícios de correlação entre os fatores de personalidade e as categorias do dicionário LIWC-BR (FILHO; ALUÍSIO; PARDO, 2013) em textos em português brasileiro.

Com base nos dados coletados, foram gerados três modelos de lexicalização de substantivos, adjetivos e verbos baseados nos corpúscos *b5-text*. Esses modelos de lexicalização fizeram parte, então, do modelo de reescrita sentencial que recebe como entrada uma sentença em português e valores de personalidade-alvo, e produz como saída um texto reescrito com base nesta personalidade-alvo.

O trabalho realizado foi avaliado tanto sob uma perspectiva de aprendizagem de máquina, considerando-se cada um dos modelos de lexicalização individuais (para substantivos, adjetivos e verbos), como uma perspectiva de GLN, considerando-se o modelo de reescrita sentencial aplicado à geração de sentenças do corpúscos coletado. Em ambos os casos, resultados sugerem que o uso de informação de personalidade permite a geração de texto mais próximo do desempenho humano quando comparados a um *baseline* que toma decisões baseado apenas na frequência de utilização das palavras.

Apesar dos resultados positivos alcançados, o trabalho realizado possui uma série de limitações. Em especial, o modelo de reescrita sentencial com base em traços de personalidade trata apenas da lexicalização, e não considera outras formas de modificação sentencial, como por exemplo modificar a voz da sentença, a estrutura dos parágrafos,

dentre outras formas. Além disso, esse estudo utilizou apenas os textos do *cópus b5*, e limitado a um conjunto reduzido de sujeitos humanos.

Como trabalho futuro, sugere-se investigar a modificação das estruturas sentenciais como voz (passiva ou ativa), a ordem das sentenças, e possivelmente conectivos. Outra possível melhoria aos modelos de lexicalização propostos é a criação e utilização de um *cópus* mais volumoso e com mais amostras de textos gerados por humanos. Adicionalmente, pode ser feita também uma avaliação extrínseca dos resultados da ferramenta, possivelmente baseando-se em estudos similares feitos por Mairesse (2008).

Referências¹

- ALLPORT, F. H.; ALLPORT, G. W. Personality Traits: Their Classification and Measurement. *The Journal of Abnormal Psychology and Social Psychology*, v. 16, n. 1, p. 6–40, 1921. ISSN 0145-2347. Disponível em: <http://prx.library.gatech.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1926-05605-001&site=ehost-live>. Citado na página 15.
- ALUÍLIO, S. M. et al. Towards Brazilian Portuguese Automatic Text Simplification Systems. In: *The ACM Symposium on Document Engineering*. [S.l.: s.n.], 2008. p. 240–408. Citado na página 32.
- ANDRADE, J. M. d. *Evidências de Validade do Inventário dos Cinco Grandes Fatores de Personalidade para o Brasil*. 153 p. Tese (Doutorado) — Universidade de Brasília, 2008. Citado 4 vezes nas páginas 15, 21, 37 e 38.
- BICK, E. *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. 505 p. Tese (Doutorado) — Aarhus University, 2000. Disponível em: <http://www.lavoisier.fr/notice/frYWOOR6SAS3WLLS.html>. Citado 5 vezes nas páginas 36, 43, 44, 46 e 53.
- BROCKMANN, C. *Personality and alignment processes in dialogue: towards a lexically-based unified model*. Tese (Doutorado) — The University of Edinburgh, 2009. Disponível em: <http://www.era.lib.ed.ac.uk/handle/1842/3980>. Citado na página 16.
- CHAMBERS, N.; ALLEN, J. Stochastic Language Generation in a Dialogue System: Toward a Domain Independent Generator. *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, p. 9–18, 2004. Citado na página 25.
- DAN-GLAUSER, E. S.; SCHERER, K. R. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior research methods*, v. 43, n. 2, p. 468–77, 2011. ISSN 1554-3528. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/21431997>. Citado 3 vezes nas páginas 16, 35 e 36.
- DODDINGTON, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the second international conference on Human Language Technology Research*, p. 138–145, 2002. Citado na página 23.
- EUGENIO, B. D.; GLASS, M.; TROLIO, M. The DIAG experiments: Natural language generation for tutoring systems. In: *Proceedings of the Second International Conference on Natural Language Generation (INLG-2002)*. Harriman, NY: [s.n.], 2002. p. 120–127. Citado na página 14.
- EUGENIO, B. di et al. PatientNarr: Towards generating patient-centric summaries of hospital stays. In: *Proceedings of the 8th International Natural Language Generation Conference (INLG)*. Philadelphia, Pennsylvania, U.S.A.: Association for Computational Linguistics, 2014. p. 6–10. Citado na página 14.

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

- FELDMAN, R.; SANGER, J. The text mining handbook: advanced approaches in analyzing unstructured data. *Imagine*, v. 34, p. 410, 2007. ISSN 14653133. Citado na página 22.
- FILHO, P. P. P. B.; ALUÍSIO, S.; PARDO, T. T. An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. *9th Brazilian Symposium in Information and Human Language Technology – STIL*, p. 215–219, 2013. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/stil/2013/0029.pdf>. Citado 4 vezes nas páginas 24, 37, 38 e 62.
- FLESCHE, R. *The Art of Readable Writing*. New York: Harper Brothers, 1949. Citado na página 32.
- GOLDBERG, L. Language and individual differences: the search for universals in personality lexicons. *Rev Personal Soc Psychol*, v. 2, n. 1, p. 142–165, 1981. Citado 2 vezes nas páginas 20 e 40.
- GOLDBERG, L.; KITTREDGE, R. I.; DRIEDGER, N. Using Natural-Language Processing to Produce Weather Forecasts. *IEEE Expert-Intelligent Systems and their Applications*, v. 9, n. 2, p. 45–53, 1994. ISSN 08859000. Citado na página 22.
- GOLDBERG, L. R. An alternative description of personality: The Big-Five factor structure. *Journal of Personality and Social Psychology*, v. 59, n. 6, p. 1216–1229, 1990. ISSN 1939-1315. Disponível em: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.59.6.1216>. Citado 3 vezes nas páginas 15, 19 e 20.
- GREENE, R. L.; WEINER, I. B. *Handbook of personality assessment*. [S.l.: s.n.], 2008. 519 p. ISSN 01918869. ISBN 9780471228813. Citado na página 20.
- HUNTER, J. et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial intelligence in medicine*, Elsevier, v. 56, n. 3, p. 157–172, 2012. Citado na página 14.
- HUTZ, C. S. et al. O desenvolvimento de marcadores para a avaliação da personalidade no modelo dos Cinco Grandes Fatores. *Psicologia: Reflexão e Crítica*, v. 11, n. 2, p. 395–411, 1998. ISSN 01027972. Citado 4 vezes nas páginas 15, 19, 20 e 21.
- IORDANSKAJA, L. et al. Generation of Extended Bilingual Statistical Reports. *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, p. 23–82, 1992. Citado na página 22.
- IVANOV, A. V. et al. Recognition of personality traits from human spoken conversations. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. [S.l.: s.n.], 2011. p. 1549–1552. ISBN 19909772 (ISSN). ISSN 19909772. Citado na página 15.
- JOHN, O. P.; DONAHUE, E. M.; L, K. R. *The Big Five Inventory–Versions 4a and 5a*. [s.n.], 1991. v. 37. 504–528 p. ISSN 00926566. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0092656603000461>. Citado na página 21.
- JOHN, O. P.; NAUMANN, L. P.; SOTO, C. J. Paradigm shift to the integrative big-five trait taxonomy: History, measurement, and conceptual issues. In: *Handbook of personality: Theory and research*. [S.l.: s.n.], 2008. p. 114–158. ISBN 9781593856502. Citado na página 21.

JONES, K. S.; GALLIERS, J. R. Evaluating Natural Language Processing Systems: An Analysis and Review. *Computational Linguistics*, v. 24, n. 2, p. 336–338, 1995. Citado na página 23.

JORDAN, P. et al. TBI-Doc: Generating Patient & Clinician Reports from Brain Imaging Data. In: *Proceedings of the 8th International Natural Language Generation Conference (INLG)*. Philadelphia, Pennsylvania, U.S.A.: Association for Computational Linguistics, 2014. p. 143–146. Citado na página 14.

LAN, A. G. J.; PARABONI, I. Definite Description Lexical Choice: taking speaker's Personality into account. In: *11th International Conference on Language Resources and Evaluation (LREC-2018) (to appear)*. Miyasaki, Japan: ELRA, 2018. Citado na página 35.

LANGKILDE-GEARY, I. An empirical verification of coverage and correctness for a general-purpose sentence generator. *Proceedings of the 12th International Natural Language Generation Workshop*, p. 17–24, 2002. Disponível em: <http://www.cs.rutgers.edu/~mdstone/inlg02/112.pdf>. Citado na página 25.

LANGKILDE, I.; KNIGHT, K. Generation that exploits corpus-based statistical knowledge. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, v. 1, p. 704–710, 1998. Disponível em: <http://dl.acm.org/citation.cfm?id=980963>. Citado na página 25.

LAPATA, M. Probabilistic text structuring: experiments with sentence ordering. *Annual Meeting of the ACL*, p. 545, 2003. Disponível em: <http://portal.acm.org/citation.cfm?id=1075165>. Citado na página 26.

LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, v. 10, n. 8, p. 707–710, 1966. ISSN 00385689. Citado na página 23.

LI, J. et al. Fundamentals of speech recognition. In: *Robust Automatic Speech Recognition*. [s.n.], 2016. p. 9–40. ISBN 0130151572. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/B9780128023983000027>. Citado na página 22.

LLORET, E. *Text Summarization: An Overview*. Alicante, Spain, 2008. v. 1, 1–24 p. Disponível em: <http://www.dlsi.ua.es/~elloret/publications/TextSummarization.pdf>. Citado na página 31.

MAIRESSE, F. *Learning to Adapt in Dialogue Systems: Data-driven Models for Personality Recognition and Generation*. 285 p. Tese (Doutorado) — University of Sheffield, 2008. Disponível em: <http://www.dcs.shef.ac.uk/intranet/research/phdtheses/Mairesse2008.pdf>. Citado 16 vezes nas páginas 15, 16, 19, 20, 22, 25, 27, 28, 29, 30, 39, 40, 41, 42, 58 e 63.

MAIRESSE, F.; WALKER, M. A. Controlling User Perceptions of Linguistic Style: Trainable Generation of Personality Traits. *Computational Linguistics*, v. 37, n. 3, p. 455–488, 2011. ISSN 0891-2017. Citado na página 42.

- MAZIERO, E. G. et al. A Base de Dados Lexical e a Interface Web do TeP 2.0 – Thesaurus Eletronico para o Portugues do Brasil. *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana – TIL*, n. August, p. 390–392, 2008. Citado 3 vezes nas páginas 31, 43 e 44.
- MCCARTHY, D.; NAVIGLI, R. SemEval-2007 Task 10: English Lexical Substitution Task. In: *Proceedings of SEMEVAL 2007*. [s.n.], 2007. p. 48–53. ISBN 1057900990. ISSN 1574020X. Disponível em: <http://www.aclweb.org/anthology/S/S07/S07-1009>. Citado na página 31.
- MILLER, G. A. WordNet: a lexical database for English. *Communications of the ACM*, v. 38, n. 11, p. 39–41, 1995. ISSN 00010782. Disponível em: <http://portal.acm.org/citation.cfm?doid=219717.219748>. Citado na página 31.
- MUNIZ, M. C. M. *A construção de recursos linguístico-computacionais para o Português do Brasil: o projeto de Unitex-PB*. Tese (Doutorado) — ICMC / USP São Carlos, 2004. Citado na página 50.
- NOVAIS, E. M. D.; TADEU, T. D.; PARABONI, I. Text-to-text surface realisation using dependency-tree replacement. In: *IBERAMIA-2010, Lecture Notes in Artificial Intelligence*. Bahia Blanca, Argentina: Springer-Verlag, 2010. v. 6433 LNAI, p. 326–335. ISBN 3642169511. ISSN 03029743. Citado na página 22.
- OBERLANDER, J.; GILL, A. J. Individual differences and implicit language : personality , parts-of-speech and pervasiveness. *26th Annual Conference of the Cognitive Science Society*, p. 1035–1040, 1992. Citado 2 vezes nas páginas 20 e 30.
- PAIVA, D. S. *Using Stylistic Parameters to Control a Natural Language Generation System*. 176 p. Tese (Doutorado) — University of Brighton, 2004. Citado na página 26.
- PAIVA, D. S.; EVANS, R. Empirically-based control of natural language generation. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, n. June, p. 58–65, 2005. Disponível em: <http://dl.acm.org/citation.cfm?id=1219848>. Citado 3 vezes nas páginas 25, 26 e 27.
- PAPINENI, K. et al. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, n. July, p. 311–318, 2002. ISSN 00134686. Disponível em: <http://dl.acm.org/citation.cfm?id=1073135>. Citado na página 23.
- PENNEBAKER, J. W.; FRANCIS, M. E.; BOOTH, R. J. *Linguistic Inquiry and Word Count*. Austin, Texas, USA: LIWC.net, 2001. Citado na página 24.
- PENNEBAKER, J. W.; KING, L. a. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, v. 77, n. 6, p. 1296–1312, 1999. ISSN 0022-3514. Citado na página 30.
- PORTET, F. et al. Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. *Artificial Intelligence*, v. 173, p. 789–816, 2009. Citado na página 14.
- RAMOS, R. M. S. et al. Building a Corpus for Personality-dependent Natural Language Understanding and Generation. In: *11th International Conference on Language Resources and Evaluation (LREC-2018) (to appear)*. Miyasaki, Japan: ELRA, 2018. Citado 10 vezes nas páginas 35, 36, 43, 44, 45, 46, 53, 58, 60 e 62.

REEVES, B.; NASS, C. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York, NY, USA: Cambridge University Press, 1996. v. 34. ISSN 0018-9235. ISBN 1575860538. Citado na página 15.

REITER, E.; DALE, R. Building Applied Natural Language Generation Systems. *Natural Language Engineering*, v. 3, n. 1, p. 57–87, 1997. ISSN 1351-3249. Disponível em: <http://dx.doi.org/10.1017/S1351324997001502>. Citado 2 vezes nas páginas 14 e 21.

REITER, E.; DALE, R. *Building natural language generation systems*. New York, NY, USA: Cambridge University Press, 2000. ISBN 0-521-62036-8. Citado na página 14.

REITER, E.; ROBERTSON, R.; OSMAN, L. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, v. 144, p. 41–58, 2003. Citado na página 14.

RIEDL, M.; GLASS, M. R.; GLIOZZO, A. Lexical Substitution for the Medical Domain. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 610–614, 2014. Citado na página 31.

SANTOS, V. G. dos; PARABONI, I.; SILVA, B. B. C. Big Five Personality Recognition from Multiple Text Genres. In: *Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415*. Prague, Czech Republic: Springer-Verlag, 2017. p. 29–37. Citado na página 35.

SCHERER, K. R. Personality Markers in Speech. In: *Social markers in speech*. The Edinburgh Building, Cambridge, UK: Cambridge University Press, 1979. p. 147–209. ISBN 878-0-521-22321-8. Disponível em: <http://dx.doi.org/10.1525/aa.1981.83.2.02a00440>. Citado na página 20.

SHARDLOW, M. A Survey of Automated Text Simplification. *(IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Natural Language Processing*, p. 58–70, 2014. ISSN 2156-5570. Citado na página 31.

SIEGMAN, A. W.; POPE, B. Effects of question specificity and anxiety-producing messages on verbal fluency in the initial interview. *Journal of personality and social psychology*, v. 2, n. 4, p. 522–30, 1965. ISSN 0022-3514. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/5841357>. Citado na página 20.

TABOADA, M. et al. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, v. 37, n. 2, p. 267–307, 2011. ISSN 0891-2017. Disponível em: http://www.mitpressjournals.org/doi/10.1162/COLI_a.00049. Citado na página 22.

TINTAREV, N. et al. Personal storytelling: Using Natural Language Generation for children with complex communication needs, in the wild... *International Journal of Human-Computer Studies*, v. 92–93, p. 1–16, 2016. Citado na página 14.

WALKER, M. et al. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research*, v. 30, p. 413–456, 2007. ISSN 10769757. Citado na página 25.

WALKER, M. et al. Trainable Generation of Big-Five Personality Styles through Data-driven Parameter Estimation. In: *Proceedings of ACL-08: HLT*. Columbus, Ohio, USA: Association for Computational Linguistics, 2008. p. 165–173. Citado 2 vezes nas páginas 16 e 42.

WALKER, M. A. et al. Perceived or Not Perceived: Film Character Models for Expressive NLG. In: SI, M. et al. (Ed.). *Lecture Notes in Computer Science 7069*. [S.l.]: Springer Berlin Heidelberg, 2011. p. 109–121. Citado na página 14.

WALKER, M. A. et al. Murder in the Arboretum: Comparing Character Models to Personality Models. In: *Intelligent Narrative Technologies IV, Papers from the 2011 AIIDE Workshop*. [S.l.: s.n.], 2011. Citado na página 14.

WITTEN, I. H.; FRANK, E.; HALL, M. a. *Data Mining: Practical Machine Learning Tools and Techniques*. 3. ed. San Francisco, USA: Morgan Kaufman Publishers, 2011. v. 54. 664 p. ISSN 14337851. ISBN 9780123748560. Citado na página 46.