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

DANILLO DA SILVA ROCHA

Validação de respostas em experimentos de Geração de Língua Natural

São Paulo

2017

DANILLO DA SILVA ROCHA

Validação de respostas em experimentos de Geração de Língua Natural

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 06 de outubro de 2017. 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

2017

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)

Rocha, Danilo da Silva

Validação de respostas em experimentos de Geração de
Língua Natural / Danilo da Silva Rocha ; orientador, Ivandré
Paraboni. – 2017.

81 f. : il.

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. Inteligência artificial. 2. Linguagem natural. I.
Paraboni, Ivandré, orient. II. Título

CDD 22.ed.– 006.35

Dissertação de autoria de Danilo da Silva Rocha, sob o título “**Validação de respostas em experimentos de Geração de Língua Natural**”, 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, na área de concentração Metodologia e Técnicas da Computação, aprovada em 06 de outubro de 2017 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Ivandré Paraboni

Universidade de São Paulo

Presidente

Profa Dra. Fátima de Lourdes dos Santos Nunes Marques

Universidade de São Paulo

Prof. Dr. Esteban Fernandez Tuesta

Universidade de São Paulo

Profa. Dra. Valéria Delisandra Feltrim

Universidade Estadual de Maringá

Dedico este trabalho à minha avó Maria José da Silva Rocha (in memoriam) por sua criação e ensinamentos que levarei para toda a vida.

Agradecimentos

Agradeço imensamente ao meu orientador Ivandré Paraboni por toda paciência, dedicação e humildade durante todos esses anos de orientação.

Agradeço as professoras Rosana Maria Traversa Palazon e Maria Márcia Matos Pinto pelos incentivos e ajuda iniciais, sem suas contribuições eu não teria iniciado este projeto.

Agradeço ao Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - IFSP, representado pelo meu colega Brunno dos Passos Alves, que possibilitou todas as ausências necessárias durante meu horário de trabalho para as disciplinas do curso e reuniões de orientação.

Agradeço, por fim, a todos os voluntários que participaram dos experimentos conduzidos nesta pesquisa.

“Quanto mais aumenta nosso conhecimento, mais evidente fica nossa ignorância”
(John F. Kennedy)

Resumo

ROCHA, Danillo da Silva. **Validação de respostas em experimentos de Geração de Língua Natural**. 2017. 81 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2017. Versão corrigida.

Este trabalho apresenta um estudo em nível de mestrado na área de Geração de Língua Natural (GLN), enfocando a experimentação na tarefa de Geração de Expressões de Referência (GER). O trabalho apresenta um levantamento bibliográfico sobre o tema, abordando principalmente o modo monólogo e diálogo de realização destes experimentos. Além disso, é apresentado um modelo computacional para a validação automática das descrições produzidas em experimentos de GER, e a sua incorporação em uma ferramenta WEB para realização de experimentos de custo mais baixo, do tipo monólogo, com os benefícios de experimentos do tipo diálogo. O Modelo é avaliado de maneira intrínseca com base em um conjunto de corpus de GER, e de maneira extrínseca em um experimento real com humanos. Resultados comprovam a hipótese inicial de que descrições coletadas em modo monólogo com validação automática das descrições são mais próximas das descrições obtidas em modo diálogo do que as obtidas em modo monólogo.

Palavras-chaves: Geração de Língua Natural. Geração de Expressões de Referência. Monólogo. Diálogo.

Abstract

ROCHA, Danillo da Silva. **The validation of responses in Natural Language Generation experiments**. 2017. 81 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2017. Revised version.

This work concerns a MSc Project in the field of Natural Language Generation (NLG), focusing on the issue of experimentation in the Referring Expression Generation (REG). The study presents a literature review on the topic, distinguishing between monologue and dialogue experiments. Moreover, a computational model for the validation of referring expressions collected in these experiments is presented. The proposed model is embedded in a WEB tool for the design of low-cost monologue experiments with the advantages of dialogue settings. The Model is assessed intrinsically based on a set of GER corpus, and extrinsically in a real experiment with humans. Results confirm the initial hypothesis that descriptions collected in monologue settings with automatic validation of the descriptions are closer to the descriptions obtained in dialog ones than those obtained in monologue settings.

Keywords: Natural language generation. Referring expression generation. Monologue. Dialogue.

Lista de figuras

Figura 1 – Exemplo de um estímulo visual com duas estrelas amarelas, um triângulo azul, e um círculo verde	17
Figura 2 – Exemplo de um estímulo visual com imagens de cones, nas cores verde e lilás, utilizado no experimento Bishop	28
Figura 3 – Exemplo de um estímulo visual utilizado pelo <i>Instruction Giver</i> durante o experimento GIVE-2	30
Figura 4 – Exemplo de um estímulo visual utilizado pelo <i>Instruction Follower</i> durante o experimento GIVE-2	31
Figura 5 – Exemplo de um estímulo visual com peças de artesanato utilizado no experimento Craft	32
Figura 6 – Exemplo de um estímulo visual utilizado pelo locutor durante o experimento ReferItGame	34
Figura 7 – Exemplo de um estímulo visual com figuras geométricas, e que estimula o uso de descrições relacionais, utilizado durante o experimento Stars2	35
Figura 8 – Exemplo de um estímulo visual com figuras geométricas, e que estimula o uso de descrições atômicas, utilizado durante o experimento Stars2	36
Figura 9 – Exemplo de dois estímulos visuais com figuras geométricas utilizados pelo locutor (à direita) e ouvinte (à esquerda) durante o experimento Stars2	37
Figura 10 – Exemplo de um estímulo visual com quadrados, números, e cores, utilizado durante o experimento Drawer	38
Figura 11 – Exemplo de um estímulo visual com fotografia de matemáticos famosos utilizado durante o experimento TUNA- <i>People</i>	39
Figura 12 – Exemplo de um estímulo visual com figuras de móveis utilizado durante o experimento TUNA- <i>Furniture</i>	40
Figura 13 – Exemplo de um estímulo visual com três figuras geométricas utilizado durante o experimento GRE3D3	41
Figura 14 – Exemplo de um estímulo visual com figuras geométricas utilizado durante o experimento GenX	42

Figura 15 – Exemplo de um estímulo visual com imagens complexas utilizado du-	
rante o experimento Wally	43
Figura 16 – Exemplo de um estímulo visual com imagens de figuras geométricas	
utilizado durante o experimento Stars	45
Figura 17 – Exemplo de um estímulo visual com imagem de um mapa utilizado	
durante o experimento Zoom	46
Figura 18 – Exemplo de um estímulo visual com imagem da fotografia de rostos	
humanos durante o experimento b5-ref	47

Lista de algoritmos

Algoritmo 1 – Interpretação	61
Algoritmo 2 – Identificação	63

Lista de quadros

Quadro 1	–	Representação semântica do estímulo visual apresentado na Figura 1	. . .	23
Quadro 2	–	Resultados estatísticos do estudo exploratório		56
Quadro 3	–	Relação entre o estudo exploratório e o experimento Stars2		80
Quadro 4	–	Relação entre a avaliação extrínseca e as imagens do córpus b5-ref	. . .	81
Quadro 5	–	Relação entre a avaliação extrínseca e as imagens do córpus GRE3D3	. . .	81

Lista de tabelas

Tabela 1 – Análise dos experimentos de GER considerados	49
Tabela 2 – Exemplos de propriedades semânticas e possíveis palavras que as realizam	59
Tabela 3 – Córpus de teste	64
Tabela 4 – Resultados da Avaliação Intrínseca	65
Tabela 5 – Frequências de propriedades no domínio GRE3D3	70
Tabela 6 – Frequências de propriedades no domínio b5-ref	71
Tabela 7 – Coeficientes Dice	71
Tabela 8 – Diferenças de frequência entre os modos <i>Feedback</i> e Monólogo em comparação ao modo Diálogo para o domínio b5-ref	72
Tabela 9 – Diferenças de frequência entre os modos <i>Feedback</i> e Monólogo em comparação ao modo Diálogo para o domínio GRE3D3	72

Sumário

1	Introdução	17
1.1	<i>Objetivos</i>	20
1.2	<i>Hipótese</i>	21
1.3	<i>Organização da monografia</i>	21
2	Conceitos Fundamentais	23
2.1	<i>A tarefa computacional de GER</i>	23
2.1.1	Avaliação de algoritmos de GER	25
2.2	<i>Considerações</i>	26
3	Revisão bibliográfica	27
3.1	<i>Experimentos conduzidos em modo diálogo</i>	27
3.1.1	COCONUT	27
3.1.2	Bishop	28
3.1.3	GIVE-2	29
3.1.4	Craft	32
3.1.5	ReferItGame	33
3.1.6	Stars2	34
3.2	<i>Experimentos conduzidos em modo monólogo</i>	37
3.2.1	Drawer	37
3.2.2	TUNA	39
3.2.3	GRE3D3 e GRE3D7	40
3.2.4	GenX	42
3.2.5	Wally	43
3.2.6	Stars	44
3.2.7	Zoom	46
3.2.8	b5-ref	47
3.3	<i>Considerações</i>	48
4	Estudo exploratório	51
4.1	<i>Hipóteses</i>	51

4.2	<i>Sujeitos</i>	54
4.3	<i>Materiais</i>	54
4.4	<i>Procedimento</i>	54
4.5	<i>Resultados</i>	55
4.6	<i>Considerações</i>	56
5	Modelo de validação automática de REs	58
5.1	<i>Interpretação de descrições textuais</i>	59
5.2	<i>Identificação do objeto-alvo</i>	62
6	Avaliação Intrínseca	64
6.1	<i>Dados</i>	64
6.2	<i>Procedimentos</i>	64
6.3	<i>Resultados</i>	65
6.4	<i>Discussão</i>	66
7	Avaliação Extrínseca	67
7.1	<i>Visão Geral</i>	67
7.2	<i>Hipótese</i>	68
7.3	<i>Sujeitos</i>	68
7.4	<i>Procedimentos</i>	69
7.5	<i>Materias</i>	70
7.6	<i>Resultados</i>	70
7.7	<i>Análise de Erros</i>	71
8	Conclusões	74
8.1	<i>Contribuições</i>	75
8.2	<i>Limitações</i>	75
	Referências¹	76
	Apêndice A – Instâncias do experimento Stars2 utilizadas no estudo exploratório	80

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

Apêndice B – Imagens dos corpos b5-ref e GRE3D3 utilizadas na avaliação extrínseca	81
---	----

1 Introdução

Sistemas de Geração de Língua Natural (GLN) transformam informações tipicamente não linguísticas em língua compreensível para seres humanos (REITER; DALE, 2000). Sistemas de GLN possuem uma ampla gama de aplicações que incluem a geração de boletins meteorológicos a partir de dados de satélite, relatórios médicos a partir de sensores (KRAHMER; DEEMTER, 2012; REITER; DALE, 2000) etc.

Figura 1 – Exemplo de um estímulo visual com duas estrelas amarelas, um triângulo azul, e um círculo verde

Fonte: Danillo da Silva Rocha, 2017

Uma das subáreas de GLN pesquisada e discutida extensivamente nas últimas duas décadas é a Geração de Expressões de Referência (GER), que é a tarefa computacional de produzir formas linguísticas, ou expressões de referência (REs), de modo a distinguir um objeto-alvo de interesse de outros objetos no mesmo estímulo visual (chamados distraidores). Exemplos de REs para o objeto-alvo C do cenário da Figura 1 são “o círculo”, “o círculo verde”, “o círculo verde à direita do triângulo azul”, dentre outras.

Em muitos casos, REs podem assumir a forma de uma descrição definida (e.g., “o homem”) ou indefinida (e.g., “algumas mulheres”). Este trabalho dará enfoque apenas para o caso de geração de expressões de referência na forma de descrições definidas como nos exemplos anteriores, desconsiderando outras formas referenciais como pronomes (“ele”), nomes próprios (“Pedro”) etc. Além disso, trataremos exclusivamente da tarefa de determinação do conteúdo semântico destas expressões (o que dizer), mas não de sua

possível realização superficial em uma língua-alvo (ou como dizer) (WHITE; RAJKUMAR; MARTINS, 2007)

A partir do trabalho apresentado em Dale e Reiter (1995) observar-se na pesquisa em GER uma crescente demanda por experimentos envolvendo sujeitos humanos, seja para a produção de descrições que possam servir de dados de treinamento ou teste para abordagens computacionais para o problema, para estudo de fenômenos psicolinguísticos variados, ou simplesmente para a construção de *córpus* de expressões de referência. Estes experimentos tipicamente assumem a forma de uma tarefa em que participantes humanos precisam descrever, em forma oral ou escrita, determinados objetos presentes em estímulos visuais que representam contextos referenciais de interesse. Por exemplo, a Figura 1 poderia ser usada como estímulo em um experimento deste tipo, no qual a tarefa do sujeito humano poderia ser descrever um determinado objeto em especial

Experimentos deste tipo podem ser conduzidos em modo monólogo ou diálogo. Em modo monólogo, a tarefa é desempenhada por um único participante (chamado locutor), que responde aos estímulos visuais fornecidos de forma independente. Exemplos de experimentos do tipo monólogo incluem Viethen e Dale (2006), Gatt, Sluis e Deemter (2007), Viethen e Dale (2008), Viethen e Dale (2011), Clarke, Elsner e Rohde (2013), FitzGerald, Artzi e Zettlemoyer (2013), Paraboni et al. (2014), Altamirano et al. (2015).

Experimentos em modo monólogo mostram-se como os mais comuns na pesquisa em GER, provavelmente pelo fato de serem mais fáceis de implementar (e.g., via WEB) e de custo relativamente baixo. No entanto, os dados coletados desta forma podem apresentar qualidade variável, com a inclusão de grandes quantidades de itens mal-formados, expressões ambíguas e outras dificuldades que podem limitar os resultados observáveis destes estudos (ALTAMIRANO et al., 2015).

Para melhor ilustrar a problemática da realização de experimentos de coleta de dados de GER em modo monólogo, suponha-se um experimento de coleta de dados no qual a tarefa do participante é descrever o objeto A na Figura 1 anterior, e para o qual um determinado indivíduo tenha respondido de forma ambígua (e.g., “a estrela”). Conforme discutido em Altamirano et al. (2015), descrições ambíguas como a deste exemplo são um dos grandes desafios em experimentos do tipo monólogo, e provavelmente o principal motivo para descarte de dados coletados e consequente desperdício de recursos humanos participantes da coleta de dados.

Em experimentos do tipo diálogo, por outro lado, considera-se a reprodução de condições mais realistas de comunicação com a inclusão de um segundo participante (dito ouvinte). Este participante é responsável por interpretar as expressões produzidas pelo locutor, de modo a garantir que as mesmas possam ser consideradas válidas para fins de comunicação (e.g., livres de ambiguidade, bem-formadas etc). Exemplos de experimentos do tipo diálogo incluem Eugenio et al. (2000), Gorniak e Roy (2004), Gargett et al. (2010), Mitchell, Deemter e Reiter (2010), Kazemzadeh et al. (2014a), Paraboni, Galindo e Iacovelli (2017).

Experimentos do tipo diálogo são a forma possivelmente ideal de coletar dados para muitos tipos de estudos em GER, já que a presença do ouvinte garante que as respostas fornecidas pelo locutor sejam válidas para fins de comunicação. No entanto, experimentos do tipo diálogo possuem custo consideravelmente mais elevado, tanto pelo recrutamento de um número maior de participantes como pela frequente necessidade de execução de forma presencial quando há instruções complexas ou quando os sujeitos precisam de treinamento específico para a tarefa.

A produção de descrições ambíguas ou mal formadas em modo monólogo, ou a exigência de mais participantes em modo diálogo, são entretanto dificuldades contornáveis, pois dizem respeito essencialmente ao custo da coleta de dados. Deixando-se estas questões de lado, observa-se que o descarte de dados (por motivo de ambiguidade ou outros) mais frequente em experimentos do tipo monólogo sugere que as REs obtidas, mesmo estando corretas, podem ser de alguma forma diferentes nestas duas modalidades.

A possível diferença entre expressões produzidas em modo monólogo e diálogo é um tema de pesquisa inexplorado na área de GER, e que no entanto pode trazer consequências importantes para estudos deste tipo. REs produzidas por participantes humanos são, em grande parte, o próprio objeto de estudo desta área de pesquisa, e experimentos deste tipo objetivam obter exemplos de uso em real da língua para então tirar conclusões sobre questões altamente específicas como a estratégia de seleção de conteúdo (DEEMTER; GOMPEL; KRAHMER, 2012; GATT et al., 2013), a superespecificação referencial (GOMPEL et al., 2012; KOOLEN; GOUDBEEK; KRAHMER, 2013; PARABONI et al., 2017) e muitos outros. Ao desconsiderar-se a eventual diferença entre expressões produzidas em monólogo e diálogo, entretanto, não é possível descartar a possibilidade de que algumas das conclusões destes estudos possam estar, de alguma forma, comprometidas.

Diante deste cenário, sugere-se a oportunidade de realização de um estudo sobre as práticas de pesquisa em GER no que diz respeito à coleta de REs em experimentos envolvendo seres humanos, e a subsequente proposta de um método alternativo de realização de coleta de dados que procura agregar as vantagens de experimentos implementados em modo diálogo (e.g., maior qualidade dos dados coletados etc.) às vantagens de experimentos em modo monólogo (e.g., o menor custo decorrente do número reduzido de participantes humanos).

O método a ser proposto, aqui denominado *Feedback*, pode ser visto como uma forma de interpretação superficial das formas textuais produzidas pelos participantes de um experimento de GER com o propósito de verificar sua validade como expressão de referência em tempo de execução do experimento e, se necessário, fornecer ao participante algum tipo de informação (no presente caso limitada a certos tipos de mensagens de erro) que permita a correção do texto fornecido. Por exemplo, no caso do experimento ilustrado anteriormente, ao receber uma RE ambígua como “a estrela” um experimento do tipo *Feedback* poderia exibir uma mensagem de erro como a seguir:

Desculpe, eu não entendi a qual estrela você se refere. Tem certeza que deseja continuar (s/n)

Com melhorias deste tipo, espera-se que um experimento de GER realizado em modo *Feedback* possa sanar alguns dos problemas típicos da coleta de dados em monólogo (e.g., participantes desatentos ou que não tenham compreendido a tarefa corretamente). Além disso, espera-se também emular um modo de interação mais próximo do que seria observado em um diálogo entre participantes humanos, e que com isso seja possível obter REs mais próximas das que seriam obtidas em um diálogo real.

1.1 Objetivos

O objetivo do presente trabalho é propor um modelo computacional, aqui denominado modo *Feedback*, de validação automática de expressões de referência em experimentos de GER. Este modelo possibilitará a coleta de expressões mais próximas daquelas que seriam obtidas por meio de um experimento em modo diálogo entre participantes humanos das expressões que seriam obtidas em um experimento em modo monólogo.

Os objetivos específicos do estudo realizado são os seguintes:

- 1 Propor e implementar um modelo computacional para validar as descrições produzidas por participantes humanos de experimentos de GER.
- 2 Validar de forma intrínseca o modelo implementado com base em corpus de GER já existentes.
- 3 Incorporar o modelo a uma ferramenta com interface WEB para a experimentação em GER em modo monólogo com validação automática das respostas fornecidas pelos participantes de experimentos de GER.
- 4 Avaliar de forma extrínseca o impacto da ferramenta com interface WEB desenvolvida em um experimento real da área de GER envolvendo participantes humanos.

1.2 Hipótese

Com base em um conjunto de descrições produzidas por sujeitos humanos em situações idênticas em três modos de condução experimental (a saber: Monólogo, *Feedback* e Diálogo), este estudo considera a seguinte hipótese de pesquisa:

Descrições coletadas em modo Feedback são mais próximas das descrições obtidas em modo Diálogo do que as obtidas em modo Monólogo.

Esta hipótese será verificada comparando-se as propriedades semânticas que representam REs produzidas por locutores humanos em um mesmo experimento em modo Monólogo e *Feedback* com as REs produzidas em modo Diálogo pelos mesmos participantes, medindo-se coeficientes Dice (DICE, 1945) de semelhança entre conjuntos (neste caso, de propriedades). Espera-se que as REs produzidas em modo *Feedback* possuam, em média, coeficientes Dice superiores aos observados em modo Monólogo, ou seja, que sejam mais próximas das REs observadas em modo Diálogo real.

1.3 Organização da monografia

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta os conceitos fundamentais da Geração de Expressões de Referência (GER), apresentando a tarefa

computacional envolvida, e como os algoritmos da área são avaliados. O Capítulo 3 aborda uma revisão bibliográfica da experimentação em GER nos modos monólogo e diálogo. O Capítulo 4 apresenta um estudo exploratório que teve como objetivo investigar indícios das diferenças entre o modo monólogo e diálogo em experimentos de GER. O Capítulo 5 discute a construção de um modelo de validação automática de REs em experimentos de GER, apresentando suas características e funcionalidades. O capítulo 6 apresenta uma avaliação intrínseca realizada com quatro corpúscos de GER, e com o objetivo de testar o modelo proposto no capítulo 5. O Capítulo 7 aborda uma avaliação extrínseca realizada com o objetivo de investigar o uso prático do modelo proposto no Capítulo 5. E por fim, o Capítulo 8 apresenta as considerações finais, e uma proposta de trabalhos futuros.

2 Conceitos Fundamentais

Este capítulo trata dos conceitos fundamentais a serem abordados no restante do trabalho. A seção 2.1 discute a tarefa computacional de Geração de Expressões de Referência (GER), e algumas formas de avaliação em GER com métricas intrínsecas. A seguir, a seção 2.2, apresenta algumas considerações sobre este tema.

2.1 A tarefa computacional de GER

Na tarefa computacional de GER, descrições podem ser vistas como conjuntos de propriedades semânticas representadas por pares atributo-valor, como tipo-estrela, cor-verde etc (KRAHMER; DEEMTER, 2012). O Quadro 1 ilustra uma possível representação deste tipo para o domínio usado na Figura 1.

Quadro 1 – Representação semântica do estímulo visual apresentado na Figura 1

Objeto	Propriedades Atômicas			Propriedades Relacionais			
	tipo	tamanho	cor	à esquerda de	à direita de	acima de	abaixo de
A	Estrela	Grande	Amarelo	-	-	B	-
B	Triângulo	Pequeno	Azul	C	-	-	A
C	Esfera	Grande	Verde	D	B	-	-
D	Estrela	Grande	Amarelo	-	C	-	-

Fonte: Danillo da Silva Rocha, 2017

A tarefa de GER é realizada de forma trivial por seres humanos, pois faz parte de seu cotidiano como, por exemplo, apontar para uma pessoa na rua ou dar instruções a respeito de como chegar a um determinado local. Entretanto, ao transferir esta tarefa para o processamento computacional, encontramos alguns desafios, sendo o principal deles relacionado a como produzir descrições que sejam semelhantes às produzidas por seres humanos (KRAHMER; DEEMTER, 2012; GATT et al., 2013; FERREIRA; PARABONI, 2017).

Na arquitetura tradicional de sistemas de GLN, a tarefa de determinar o conteúdo semântico de descrições definidas é conhecido como a tarefa computacional de Geração de Expressões de Referência (GER) (REITER; DALE, 2000). Esta tarefa consiste em, dado um estímulo visual C e um objeto-alvo r a ser descrito, produzir um conjunto de propriedades L que seja verdadeiro apenas para r, mas não para os demais objetos distraidores no mesmo estímulo visual como “o triângulo” em referência ao objeto B do exemplo da Figura 1. Se uma descrição com estas características for possível, L é dita

livre de ambiguidade, e poderia posteriormente (em outro ponto da arquitetura GLN) ser realizada em uma língua-alvo como uma descrição definida (e.g., “o triângulo”). Por outro lado, caso a expressão resultante descreva mais de um objeto em C (ou seja, pelo menos um distraidor além de r), então L é dita ambígua, e poderia ser realizada, por exemplo, como uma descrição indefinida (e.g., “uma estrela”).

Uma forma simples de descrever um objeto-alvo durante a tarefa computacional de GER é utilizar descrições do tipo atômicas, que utilizam apenas as propriedades do próprio objeto (BOHNET; DALE et al., 2005; KRAHMER; DEEMTER, 2012). Um exemplo de uma descrição atômica é “O triângulo azul”, que descreve o objeto B da Figura 1. Entretanto, existem situações em que o uso de propriedades atômicas não é o suficiente para descrever um objeto-alvo. Neste caso, é necessário considerar os objetos que estejam próximos a ele, fazendo uso de propriedades relacionais (DALE; HADDOCK, 1991; PARABONI; GALINDO; IACOVELLI, 2015). Usamos propriedades do tipo relacional para descrever um objeto-alvo com base em um ponto de referência (ou *landmark*) próximo. Assim, por exemplo, obtemos a descrição “o triângulo ao lado esquerdo da esfera” para o objeto B da Figura 1.

O principal objetivo de um algoritmo de GER é produzir descrições que sejam suficientes e adequadas para distinguir um objeto-alvo de forma única. No entanto, diferentes estratégias de GER podem produzir descrições com maior ou menor volume de informação. Quando a descrição resultante possui um número de propriedades menor do que o mínimo necessário para eliminar ambiguidade, ela é dita subespecificada, como por exemplo a descrição “a estrela” para o domínio da Figura 1. Por outro lado, quando a descrição possui um número de propriedades maior do que o mínimo necessário para a identificação do objeto-alvo, ela é dita superespecificada. Na Figura 1, um exemplo de uma descrição superespecificada para o objeto B seria “o triângulo azul ao lado esquerdo da esfera verde”, na qual a referência ao landmark (i.e., a esfera) é dita superespecificada ou redundante. Finalmente, quando a descrição possui o menor número possível de propriedades que permitem a identificação sem ambiguidade, ela é dita mínima, como em “a estrela ao lado do círculo”.

A superespecificação referencial pode ser vista, em princípio, como uma violação da Máxima de Quantidade (GRICE, 1975) que estabelece que devemos dizer “o necessário”, e somente o “necessário”. Entretanto, diversos estudos de diversas áreas correlatas têm demonstrado amplamente que a superespecificação referencial é inerente à produção

humana de língua natural (PECHMANN, 1989; ENGELHARDT; BAILEY; FERREIRA, 2006; ARTS et al., 2011b; ENGELHARDT; DEMIRAL; FERREIRA, 2011; KOOLEN et al., 2011; PARABONI et al., 2017).

2.1.1 Avaliação de algoritmos de GER

Considerando-se que o presente trabalho trata da questão da qualidade dos dados coletados em experimentos de GER envolvendo sujeitos humanos, nesta seção descrevemos brevemente as métricas utilizadas na área de GER para avaliação de expressões produzidas de forma automática com base em um corpus de referência.

Uma vez que descrições são representadas por conjuntos (de propriedades semânticas), uma das formas de avaliar os resultados de um algoritmo de GER é comparando-se sua saída com exemplos coletados em experimentos como os descritos anteriormente. Neste caso, é comum fazer uso de métricas intrínsecas para medir o grau de similaridade entre duas descrições. Na área de GER as métricas mais utilizadas são a Dice e Acurácia (DICE, 1945).

Conforme apresentado pela fórmula 1 o coeficiente Dice (DICE, 1945) realiza a avaliação de similaridade entre duas descrições representadas por conjuntos de propriedades semânticas A e B.

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad (1)$$

Os possíveis resultados obtidos com o coeficiente Dice assume valores entre 0 e 1. O valor 1 significa que as descrições são idênticas, e o valor 0 que as descrições são totalmente divergentes.

Uma métrica complementar ao coeficiente Dice é a medida de Acurácia, que consiste em atribuir o valor 1 quando duas descrições são idênticas, e o valor 0, quando as duas descrições não forem idênticas (GATT; SLUIS; DEEMTER, 2007). Uma das deficiências da medida de Acurácia é sua alta granularidade, pois mesmo uma pequena divergência entre duas descrições (e.g., “a grande bola azul” e “a bola azul”) resulta em acurácia zero, diferença esta que seria melhor representada em termos de coeficientes Dice.

2.2 Considerações

Este capítulo discutiu os conceitos fundamentais que serão utilizados nos capítulos seguintes deste trabalho. Apresentamos o problema computacional de GER destacando conceitos básicos como a questão da geração de descrições atômicas e relacionais, e de superespecificação referencial. Além disso, foi discutida brevemente a questão da avaliação de algoritmos de GER com base em um corpus de referência fazendo uso das métricas intrínsecas de Dice e Acurácia.

3 Revisão bibliográfica

Este capítulo apresenta alguns dos principais experimentos de coleta de dados de sujeitos humanos nos últimos anos na área de GER. O principal objetivo deste levantamento é verificar a prática de experimentação utilizada em cada estudo e identificar, principalmente, o modo de comunicação (em monólogo ou diálogo) empregado em cada caso.

3.1 *Experimentos conduzidos em modo diálogo*

3.1.1 COCONUT

O experimento realizado em Eugenio et al. (2000) faz a coleta do diálogo produzido por participantes humanos durante a partida de um jogo, que envolvia a ação de mobilhar dois cômodos de uma casa fictícia através da compra e venda de móveis. Os objetivos do experimento foram criar um modelo de linguagem baseado na análise das mensagens produzidas e criar um *cópus*, chamado COCONUT.

Neste experimento, a tarefa de cada dupla era negociar, por meio da compra ou venda, os itens que desejavam possuir para atingir os objetivos estabelecidos no início da partida. Estes objetivos incluíam obter itens de mobília para compor uma casa fictícia, por exemplo.

Os participante das duplas foram posicionados em salas diferentes durante o experimento e tinham à sua disposição um computador para uso individual onde o jogo era disponibilizado. Neste ambiente, os participantes trocavam mensagens como “quero comprar o sofá pequeno” etc.

A condução do experimento foi realizada em duas etapas. A primeira etapa foi destinada ao treinamento dos participantes que eram expostos ao mesmo ambiente que seria utilizado no experimento real. Durante o treinamento era possível que os participantes tirassem qualquer tipo de dúvida a respeito do jogo. A segunda etapa foi o experimento propriamente dito. Em cada partida, os jogadores se comunicavam e manipulavam os itens dos cômodos. Cada dupla participou de três partidas do jogo e, em cada partida, havia a restrição de tempo máximo de duas horas para concluí-la.

A comunicação entre os participantes do experimento era realizada apenas por meio de mensagens enviadas pelo sistema computacional disponível, não sendo permitida a comunicação verbal. Durante o experimento cada participante podia visualizar apenas a sua relação de móveis disponíveis, mas não a de seu parceiro.

Para a condução do experimento foram recrutadas doze duplas de participantes que deveriam se conhecer previamente, de forma a facilitar as negociações durante a partida do jogo. Ao todo, 482 instâncias de diálogo foram produzidas, e posteriormente anotadas.

3.1.2 Bishop

O experimento realizado em Gorniak e Roy (2004) faz a coleta de descrições de referência de forma verbal em modo diálogo. O objetivo do trabalho foi a coleta de descrições para a construção de um modelo de linguagem baseado na forma como humanos descrevem objetos em estímulos visuais tridimensionais, e que posteriormente foi implementado em uma ferramenta computacional chamada BISHOP.

O cenário utilizado no experimento era composto por 30 objetos geométricos na forma de cones, distribuídos de forma randômica. Estes cones possuíam o mesmo tamanho e eram distribuídos de forma que metade fosse da cor verde e a outra metade da cor lilás, conforme apresentado na Figura 2. A disposição dos objetos pelo cenário era igual para ambos os participantes.

Figura 2 – Exemplo de um estímulo visual com imagens de cones, nas cores verde e lilás, utilizado no experimento Bishop

Fonte: (GORNIAC; ROY, 2004)

Cada participante podia exercer o papel de locutor, responsável por escolher um objeto-alvo dentro de um estímulo visual e descrever sua localização verbalmente, ou o de ouvinte, que devia selecionar o objeto-alvo com base apenas nas descrições verbais feitas pelo locutor. Os participantes não tinham contato visual e usavam um computador e um

mouse para manipular o cenário. Não era permitido que os participantes visualizassem o cenário um do outro durante o experimento.

A execução da tarefa era sempre iniciada por um locutor que deveria selecionar um objeto de sua livre escolha dentre os disponíveis no cenário, e que se tornaria o objeto-alvo da tarefa. Em seguida, o locutor deveria dar uma instrução verbal para o sujeito ouvinte que, a partir desta instrução, deveria selecionar o objeto-alvo descrito. Durante o experimento não houve restrição de tempo para a execução da tarefa.

O único auxílio disponível para os ouvintes era as descrições verbais dadas pelos locutores. Caso o ouvinte selecionasse o objeto-alvo correto, o sistema retirava este objeto do cenário. Em contrapartida, caso o sujeito ouvinte selecionasse o objeto errado, o sistema mantinha o objeto no cenário, sinalizando ao locutor que refizesse a sua instrução até que o ouvinte conseguisse selecionar o objeto-alvo correto.

Esta sequência de execuções repetia-se até que todos os objetos disponíveis fossem utilizados pelos participantes. Ao término de uma rodada de experimento, os participantes trocavam de papéis e iniciavam imediatamente uma nova rodada.

Foram recrutados para o experimento seis participantes nativos e não nativos da língua inglesa, com idades entre 22 e 30 anos. Ao todo, foram coletadas 268 descrições que foram gravadas, segmentadas e transcritas.

3.1.3 GIVE-2

O experimento em Gargett et al. (2010) trata de uma coleta de instruções feita com a participação de humanos para a navegação em mundos virtuais. Um Exemplo de instrução possível seria “Dobre à esquerda e pressione o botão vermelho”, que inclui descrições de objetos manipuláveis do mundo virtual (como o botão vermelho).

O objetivo do experimento foi a construção de um corpus de instruções, chamado GIVE-2 e a estratégia para a coleta baseou-se na utilização de um jogo de caça ao tesouro, onde um jogador seguia instruções fornecidas por outro para navegar em um mundo virtual tridimensional. A condução do experimento foi realizada em modo diálogo.

Para realizar a coleta das instruções, foram recrutadas duplas de participantes que exerciam os papéis de *Instruction Giver* (IG) e *Instruction Follower* (IF). Exemplos de cenas visualizadas pelo IG e IF são apresentadas na Figura 3 e Figura 4.

Figura 3 – Exemplo de um estímulo visual utilizado pelo *Instruction Giver* durante o experimento GIVE-2

Fonte: (GARGETT et al., 2010)

No experimento GIVE, o IG era responsável por dar instruções de forma escrita ao outro participante da dupla e detinha o conhecimento sobre como chegar ao destino final do cenário de caça ao tesouro, porém não podia ele próprio se movimentar. O IF foi o responsável por realizar os movimentos necessários com base nas instruções fornecidas pelo IG. O IF não possuía os conhecimentos sobre como chegar ao destino final, porém exercia o papel de jogador propriamente dito, manipulando objetos e navegando o mundo virtual.

O experimento foi realizado com o auxílio de um software desenvolvimento de forma a oferecer dois tipos de visualização para cada um dos papéis. O IG possuía acesso a um mapa que mostrava um panorama completo do cenário, conforme apresentado na Figura 3. Com este mapa era possível observar os melhores caminhos para se chegar ao destino desejado, visualizando dinamicamente a localização do IF dentro do cenário. Além disso, era disponibilizada uma caixa de texto para que as instruções fossem dadas ao IF.

Em contrapartida, o IF tinha acesso a um ambiente interativo navegável em três dimensões, conforme apresentado na Figura 4. Para que o IF avançasse pelo cenário, era necessário que certos botões distribuídos pelo mundo virtual fossem pressionados, liberando assim o acesso a outros cenários do jogo. A decisão de qual botão apertar era feita com base nas instruções dadas pelo IG, as quais eram exibidas para o IF dentro do mundo virtual.

As instruções dadas eram expressas na forma textual, não sendo possível que os participantes se comunicassem verbalmente. Além disso, cada participante podia visualizar apenas sua própria representação (mapa ou mundo virtual). Caso a instrução dada pelo IG

Figura 4 – Exemplo de um estímulo visual utilizado pelo *Instruction Follower* durante o experimento GIVE-2

Fonte: (GARGETT et al., 2010)

não fosse compreendida pelo IF, o mesmo poderia solicitar que a instrução fosse repetida pelo IG.

Antes de iniciar o experimento real, os participantes realizaram um treinamento a respeito da utilização da ferramenta e foram instruídos a respeito dos controles disponíveis. A condução do experimento foi realizada de forma a oferecer para cada dupla três tipos de cenários de caça ao tesouro, apresentados de forma randômica para cada dupla, e realizando a troca de papéis a cada novo cenário. Os dois papéis do jogo permitiram que os autores avaliassem o quanto a experiência dos participantes em cada um deles influenciou no jogo como um todo, e em cada cenário em particular. Ademais, foi avaliado o quanto a habilidade dos participantes em navegar pelos cenários propostos influenciou no tempo dispendido para concluir o objetivo do experimento, e no número total de objetos que foram manipulados pelo IF. Durante o experimento não houve restrição de tempo para executar a tarefa.

O experimento GIVE contou com um total de 15 duplas de participantes fluentes na língua alemã e 21 duplas na língua inglesa. Do total de participantes de língua alemã, 17 eram mulheres e 13 homens, enquanto os de língua inglesa foram 16 mulheres e 26 homens. Ao todo foram produzidas 2763 instruções pelos participantes de língua alemã e 3417 instruções pelos participantes de língua inglesa.

3.1.4 Craft

O experimento em Mitchell, Deemter e Reiter (2010) apresenta a coleta de descrições destinadas à montagem do rosto de um boneco a partir de um estímulo visual físico, isto é, a partir de objetos reais. Na Figura 5 é apresentado um exemplo de um dos cenários que foi utilizado durante o experimento. Esta coleta foi conduzida em modo diálogo, onde duplas de participantes deveriam realizar a troca de descrições verbais até que um determinado rosto, apresentando previamente em uma imagem, fosse construído. O objetivo do experimento foi avaliar as expressões de referência obtidas com base nas descrições coletadas em uma atividade realista, onde os participantes deveriam explicar como manipular os objetos para compor as imagens de interesse. Isso possibilitou que os participantes utilizassem expressões que descrevessem partes de um objeto, fizessem comparação de tamanho sem uma definição precisa, e utilizassem analogias para descrever os diversos componentes.

Para compor o rosto proposto, eram oferecidos para os participantes 51 objetos de artesanato como cordões de vários tipos, papéis crepom, penas sintéticas, argolas de plástico, entre outros. Cada sujeito era responsável por visualizar o rosto de um boneco pronto e dar descrições para a reconstrução do mesmo. Os sujeitos possuíam à sua disposição o rosto que deveria ser reconstruído e as peças disponíveis para serem utilizadas.

Figura 5 – Exemplo de um estímulo visual com peças de artesanato utilizado no experimento Craft

Fonte: (MITCHELL; DEEMTER; REITER, 2010)

Cada participante tinha como tarefa produzir descrições verbais para construir cinco rostos, apresentados de forma randômica. O primeiro rosto do experimento era destinado apenas ao treinamento dos participantes, que podiam tirar dúvidas sobre o que

deveria ser feito. Para os rostos subsequentes, os participantes não podiam tirar dúvidas e foram simplesmente instruídos a formular descrições claras. Além disso, foi estipulada uma restrição de tempo para que os participantes completassem as suas tarefas, necessidade esta que foi identificada em um experimento piloto realizado previamente.

Foram recrutadas vinte pessoas que incluíam estudantes e profissionais nativos na língua inglesa, remunerados por sua participação. Ademais, o experimento foi realizado sem o uso de recursos computacionais, e todas as expressões de referência foram coletadas com o auxílio de um gravador de voz. Após a gravação das descrições, uma transcrição foi realizada para anotação das expressões de referência, totalizando 505 expressões.

Do total de vinte experimentos, um foi descartado devido a problemas na gravação de voz de um dos participantes, e outro pelo baixo número de objetos utilizados para compor os rostos propostos no experimento.

3.1.5 ReferItGame

O experimento apresentado em Kazemzadeh et al. (2014b) aborda a tarefa de referenciar objetos em um nicho de pesquisa ainda pouco explorado na área de GER: a descrição de objetos em fotos reais. Para alcançar este objetivo, foi desenvolvido um jogo para até dois jogadores, disponibilizado via WEB para participantes nativos da língua inglesa. Para incentivar o uso do jogo por um grande número de participantes, foi utilizado o serviço de *crowdsourcing Amazon Mechanical Turk*¹. Um exemplo de estímulo utilizado no experimento é apresentado na Figura 6.

O jogo possuía os papéis de locutor e ouvinte das expressões. O locutor era o responsável por escrever uma descrição que referenciasse o objeto-alvo da cena de forma única. Já o participante ouvinte deveria, a partir de uma descrição dada pelo locutor, identificar o objeto-alvo do cenário. Os papéis eram invertidos a cada imagem corretamente identificada ou, em caso de erro, eram mantidos. Em cada partida, os jogadores estavam geograficamente distantes e disputavam entre si para ver quem conseguiria fazer mais pontos no jogo.

Além da disputa entre dois participantes, uma forma alternativa de partida era a disputa entre um jogador humano e o próprio computador, para casos em que só um jogador estava disponível online. Nesta situação, o computador poderia assumir tanto o

¹ Serviço disponível em: <https://www.mturk.com/mturk/welcome>

Figura 6 – Exemplo de um estímulo visual utilizado pelo locutor durante o experimento ReferItGame

Fonte: (KAZEMZADEH et al., 2014b)

papel de locutor como o de ouvinte. Para oferecer esta modalidade de jogo foi necessária a automação do jogo para que o participante humano não percebesse que estava disputando contra uma máquina. A automação da tarefa de referenciar os objetos foi realizada de forma simples, pois foi necessário apenas armazenar um grande número de expressões pré-gravadas e apresentá-las durante a partida. Por outro lado, a automação da tarefa de interpretar a descrição recebida do participante humano e identificar o objeto-alvo correspondente mostrou-se mais difícil. Neste caso, a descrição enviada pelo humano era comparada com expressões que estavam pré-gravadas para então decidir se seriam aceitas ou não.

Ambas as modalidades de jogo, entre humano versus humano ou humano versus computador, possuíam uma restrição de tempo para que as expressões de referência fossem produzidas e avaliadas. Ao todo foram coletadas 130.525 expressões que referenciaram 19.894 fotos.

3.1.6 Stars2

O experimento realizado em Paraboni, Galindo e Iacovelli (2017) faz a coleta de expressões de referência em modo diálogo para a construção de um corpus chamado Stars2. Os objetivos do experimento foram avaliar questões como a geração de expressões de referência atômicas e relacionais com o uso de até três objetos, como em “A caixa ao lado

direito do cone e abaixo da esfera”, a variação humana das expressões e a superespecificação referencial. Exemplos de estímulos visuais que exigem descrições atômicas e relacionais são ilustrados na Figura 7 respectivamente.

Para executar as atividades propostas, 16 imagens foram utilizadas em cada experimento, sendo apresentadas uma-a-uma e de forma randômica. Todas as imagens possuíam 15 objetos em seu estímulo visual, sendo um objeto-alvo e outros 14 objetos distraidores, e eram compostas por figuras geométricas como cones, cubos e esfera representados na forma bidimensional.

Em cada experimento foram modelados dois tipos de condições: o primeiro tipo, aqui denominado relacional, considerava situações de referência em que o uso de uma relação espacial era altamente provável (ou mesmo necessário), e um segunda tipo, aqui denominado atômico, em que o uso destas relações não era em princípio necessário para a identificação do objeto-alvo. O experimento era formado de oito condições, sendo quatro relacionais e quatro atômicas, e cada condição foi testada duas vezes (usando-se imagens planas e com um certo grau de sobreposição de objetos).

Na Figura 7 é apresentado um exemplo de cenário utilizado para o modelo relacional. Neste cenário, o objeto alvo é representado por um cone, e como existem objetos distraidores que são idênticos ao objeto-alvo, é comum referenciar objetos próximos ao objeto-alvo como “O cone ao lado direito do cubo”.

Figura 7 – Exemplo de um estímulo visual com figuras geométricas, e que estimula o uso de descrições relacionais, utilizado durante o experimento Stars2

Fonte: (PARABONI; GALINDO; IACOVELLI, 2017)

Na Figura 8 é apresentado um exemplo de um cenário utilizado para o modelar as condições atômicas do experimento. Neste cenário, o objeto-alvo é representado por uma esfera, que é o único objeto deste tipo no cenário e neste caso uma descrição atômica como, por exemplo, “a esfera” seria suficiente para referenciá-lo.

Figura 8 – Exemplo de um estímulo visual com figuras geométricas, e que estimula o uso de descrições atômicas, utilizado durante o experimento Stars2

Fonte: (PARABONI; GALINDO; IACOVELLI, 2017)

A condução do experimento Stars2 foi realizada com duplas de participantes que exerciam os papéis de locutor e ouvinte. O locutor era responsável por elaborar as expressões de referência baseado nos estímulos visuais apresentados e enviá-las para a análise. O ouvinte era responsável por indicar o objeto-alvo referenciado pelo locutor e assim validar a descrição. O experimento foi realizado de forma presencial, na qual os participantes trabalhavam em computadores individuais e eram proibidos de se comunicar exceto via interface do experimento. Antes de iniciar o experimento real, as duplas foram submetidas a um treino que simulava a tarefa real. Este treino foi conduzido até que cada dupla sanasse todas as dúvidas sobre a tarefa que iriam realizar.

Os cenários apresentados para os sujeitos locutores possuíam uma imagem, uma caixa de texto onde deveria ser escrita a expressão de referência; e um botão “Send”, conforme a Figura 9. Além disso, o objeto-alvo era apontado por uma seta.

O estímulo visual dos sujeitos ouvintes possui as mesmas imagens apresentadas aos locutores, porém não havia a seta que apontava para o objeto-alvo, mas sim cinco rótulos, representados pelas letras A, H, I, O, M e que eram associados ao objeto-alvo e quatro objetos distraidores. Esses rótulos eram utilizados pelo ouvinte para identificar o objeto-alvo, conforme mostra a Figura 9.

Figura 9 – Exemplo de dois estímulos visuais com figuras geométricas utilizados pelo locutor (à direita) e ouvinte (à esquerda) durante o experimento Stars2

Fonte: (PARABONI; GALINDO; IACOVELLI, 2017)

Neste estímulo visual, o sujeito ouvinte deveria pressionar o botão da letra correspondente ao objeto-alvo e em seguida pressionar o botão “Send”. Para casos em que o locutor enviasse respostas que o próprio sistema identificasse como sendo palavras ilegais (por exemplo, expressões como “canto”, “borda” etc., que eram proibidas no experimento) ou quando o ouvinte pressionasse o botão “What?”, o sistema automaticamente apresentava uma nova imagem. Os experimentos foram realizados em duas rodadas e com as mesmas duplas de participantes, que deviam inverter de papel ao iniciar a segunda rodada. Não houve restrição de tempo para as tarefas do experimento.

Foram recrutados para o experimento 56 participantes nativos de Português brasileiro e estudantes de graduação. Todos foram recrutados via e-mail ou por redes sociais, e atuaram de forma voluntária. Ao todo, foram coletadas 884 expressões de referência, sendo que doze foram descartadas por não estarem de acordo com os objetivos do experimento.

3.2 Experimentos conduzidos em modo monólogo

3.2.1 Drawer

O trabalho apresentado em Viethen e Dale (2006) faz a comparação entre as expressões de referência coletadas em um experimento realizado com participantes humanos e as produzidas por três algoritmos de GER: *Full Brevity* (DALE, 1989), o algoritmo apresentado em Dale e Haddock (1991) e o Algoritmo Incremental (DALE; REITER, 1995). O objetivo do experimento foi avaliar o grau de similaridade entre as expressões de

referência produzidas por humanos e as produzidas pelos três algoritmos. O experimento foi conduzido em modo monólogo, onde cada participante realizou a sua tarefa de forma independente.

A elaboração do experimento utilizou um quadro físico onde 16 desenhos de gavetas bidimensionais eram distribuídas lado a lado, formando um conjunto de quatro linhas e quatro colunas. Para localizar as gavetas, foram utilizados números de 1 a 16 e quatro cores. Não foram utilizados meios computacionais para coleta de dados ou para a condução do experimento. A Figura 10 apresenta um exemplo de um dos cenários utilizado.

Figura 10 – Exemplo de um estímulo visual com quadrados, números, e cores, utilizado durante o experimento Drawer

1 (blue)	2 (orange)	3 (pink)	4 (yellow)
8 (blue)	7 (blue)	6 (yellow)	5 (pink)
9 (orange)	10 (blue)	11 (yellow)	12 (orange)
16 (yellow)	15 (pink)	14 (orange)	13 (pink)

Fonte: (VIETHEN; DALE, 2006)

A tarefa de cada participante era descrever verbalmente e de forma única o objeto-alvo do cenário, que lhes eram indicados com base na numeração existente. Os números passados para cada participante eram escolhidos de forma randômica, possibilitando que um mesmo participante referenciasse o mesmo objeto-alvo mais de uma vez. Ao todo, cada participante recebeu sete números de gavetas para descrever e foram instruídos a utilizar qualquer tipo de atributo que quisessem para produzir as descrições.

As expressões coletadas neste experimento não passavam por nenhum tipo de validação automática prévia de consistência. Além disso, não houve restrição de tempo para a tarefa.

Foram recrutados 20 participantes que produziram um total de 140 expressões de referência. Destas, 22 expressões foram descartadas, sendo oito por serem ambíguas e doze por utilizarem propriedades de conjuntos (e.g., “o objeto está a direita das gavetas vermelhas”).

3.2.2 TUNA

O experimento realizado em Gatt, Sluis e Deemter (2007) cria um córpus de expressões de referência chamado TUNA e faz a comparação de similaridade deste córpus com as expressões produzidas pelos algoritmos *Full Brevity*, Greedy e Incremental, apresentados em Dale e Reiter (1995). O córpus TUNA é consideravelmente influente na área de GER, e foi utilizado em uma série de três desafios computacionais (*TUNA Challenges*) no período entre 2007 e 2009 (BELZ; GATT, 2007; GATT; BELZ; KOW, 2008; GATT; BELZ; KOW, 2009). O experimento contou com o auxílio de participantes humanos, e foi realizado em modo monólogo via WEB.

A condução do experimento foi dividida em duas partes, sendo a primeira realizada com fotografias de matemáticos famosos e executada com 45 participantes (constituindo o subcórpus *TUNA-People*), conforme a Figura 11. A segunda parte foi conduzida com imagens de móveis como mesas, cadeiras etc. e executada por outros 45 participantes (constituindo o subcórpus *TUNA-Furniture*), conforme a Figura 12. Cada imagem do experimento era composta por uma grade de sete objetos dispostos em três linhas e cinco colunas, os quais foram distribuídos de forma randômica.

Figura 11 – Exemplo de um estímulo visual com fotografia de matemáticos famosos utilizado durante o experimento *TUNA-People*

Fonte: (GATT; SLUIS; DEEMTER, 2007)

Em cada experimento, vinte imagens eram apresentadas de forma randômica, e a tarefa era referenciar um ou dois objetos-alvo de forma a diferenciá-los dos demais. Assim, o córpus inclui expressões singulares e plurais. Em metade dos experimentos foi dito aos participantes que deveria ser levada em consideração a posição dos objetos em relação à grade, pois isto supostamente faria diferença para a análise dos dados. Esta instrução

motivou a produção de descrições como “A cadeira na linha do meio”. Em contrapartida, na outra metade dos experimentos os participantes foram orientados a não utilizar qualquer referência à localização dos objetos, pois isso supostamente iria invalidar a resposta. Além disso, foi esclarecido aos participantes que as expressões de referência produzidas estavam sendo analisadas pelo computador, porém na prática está análise não era realizada.

Figura 12 – Exemplo de um estímulo visual com figuras de móveis utilizado durante o experimento TUNA-*Furniture*

Fonte: (GATT; SLUIS; DEEMTER, 2007)

Durante o experimento TUNA não houve a validação automática das expressões enviadas pelos participantes e não houve restrição de tempo para a tarefa. Ao todo, foram produzidas 1800 expressões de referência.

3.2.3 GRE3D3 e GRE3D7

Os estudos realizados em Viethen e Dale (2008), Viethen e Dale (2011) apresentam dois experimentos de GER, ambos com a participação de humanos e conduzidos em monólogo via WEB. O objetivo destes experimentos foi identificar o uso de propriedades relacionais na produção de expressões de referência mesmo quando o uso de tais propriedades não era estritamente necessário para a identificação do objeto-alvo. Em Viethen e Dale (2008) foi criado um cópula, chamado GRE3D3 e em Viethen e Dale (2011) foi criado um segundo cópula chamado GRE3D7.

As imagens utilizadas nestes experimentos eram compostas por objetos geométricos como cubos e esferas. Estes objetos variavam entre duas cores e em dois tamanhos possíveis. No cópula GRE3D3 cada imagem possuía um objeto-alvo e dois objetos distraidores, conforme a Figura 13, enquanto que no cópula GRE3D7 todas as imagens eram compostas por um objeto-alvo e seis objetos distraidores.

Figura 13 – Exemplo de um estímulo visual com três figuras geométricas utilizado durante o experimento GRE3D3

Fonte: (VIETHEN; DALE, 2008)

A tarefa de cada participante, em ambos experimentos, era elaborar expressões de referência para cada imagem apresentada como estímulo, de forma a distinguir o objeto-alvo dos distraidores. No experimento GRE3D3 cada participante visualizou 10 imagens, apresentadas uma-a-uma e de forma não randômica. Já no caso do GRE3D7 foram apresentadas 32 imagens para cada participante, sendo 16 imagens de estímulo e outras 16 imagens *filler*, todas apresentadas de forma randômica. As imagens *filler* foram utilizadas com o intuito de impedir que os participantes estabelecessem padrões nas respostas e para diminuir a monotonicidade do experimento.

Em ambos os experimentos, não existiu validação automática das expressões de referência, e os participantes podiam enviar qualquer texto como resposta. Entretanto, como forma de minimizar a produção de expressões mal formadas, foi dito aos participantes que não era possível corrigir as expressões após inseri-las.

Para o corpus GRE3D3 foram recrutados 74 participantes que produziram um total de 740 descrições. Para o corpus GRE3D7 foram recrutados 318 participantes que produziram um total de 4704 descrições. Em ambos experimentos, os participantes eram nativos da língua inglesa.

No corpus GRE3D3 foram descartadas 10 expressões de referência devido ao participante ser daltônico. Outras 90 expressões foram descartadas por serem mal formadas, restando 630 expressões úteis para compor o corpus. Para o corpus GRE3D7, do total de 318 participantes apenas 294 conseguiram concluir o experimento com sucesso, sendo descartados os participantes que não conseguiram chegar ao final da tarefa. Destes 294 participantes, sete foram descartados durante a análise de dados por terem elaborado expressões de referência mal formadas, sugerindo que os mesmos talvez não tenham

compreendido as instruções fornecidas. Após o descarte restaram 4592 expressões de referência para o corpus.

3.2.4 GenX

O experimento realizado em FitzGerald, Artzi e Zettlemoyer (2013) faz a coleta de expressões de referência para a construção de um modelo computacional que reproduz a distribuição de preferências humanas durante a tarefa de referenciar conjuntos de objetos. O objetivo do trabalho foi realizar esta modelagem baseada na análise de distribuição das descrições produzidas por sujeitos humanos. A condução do experimento foi realizada em modo monólogo.

Foram utilizadas para o experimento um total de 118 imagens tridimensionais compostas por objetos geométricos como cones, esferas, cubos, cilindros etc. Para apontar o grupo de objetos a serem referenciados, foram utilizadas marcações de forma a separá-los dos objetos distraidores, conforme a Figura 14.

Figura 14 – Exemplo de um estímulo visual com figuras geométricas utilizado durante o experimento GenX

Fonte: Adaptado de FitzGerald, Artzi e Zettlemoyer (2013)

Durante o experimento, os participantes deviam referenciar imagens que eram apresentadas como estímulo em uma tarefa de *crowdsourcing* disponibilizada no serviço *Amazon Mechanical Turk*. Em cada experimento, a tarefa consistia em elaborar expressões de referência na forma textual para cada imagem apresentada e, em seguida, enviá-las como resposta para o sistema. Não foi realizado nenhum tipo de validação automática destas expressões, e não houve restrição de tempo para a tarefa.

Ao todo, 5380 expressões de referência foram coletadas, sendo que deste total, 860 expressões de referência foram descartadas por estarem fora dos padrões esperados, restando 4520 expressões de interesse.

3.2.5 Wally

O trabalho conduzido em Clarke, Elsner e Rohde (2013) faz a coleta de expressões de referência em um experimento realizado com humanos e com o uso de imagens complexas retiradas do livro infantil “Onde está Wally?”², com dezenas de objetos distraidores e apenas um objeto-alvo (o personagem Wally que dá nome à obra). O objetivo do experimento era identificar expressões de referência elaboradas por humanos em estímulos visuais mais complexos do que já utilizados em experimentos da área de GER, e identificar a influência da complexidade visual na elaboração destas expressões. A condução do experimento foi realizada em modo monólogo e disponibilizada com uma tarefa de *crowdsourcing* disponível pelo serviço *Amazon Mechanical Turk*.

O experimento utilizou imagens bidimensionais que possuíam centenas de personagens humanos com alto grau de semelhança entre si e alguns objetos não humanos como cabanas, barcos etc., conforme apresentado na Figura 15.

Figura 15 – Exemplo de um estímulo visual com imagens complexas utilizado durante o experimento Wally

Fonte: (CLARKE; ELSNER; ROHDE, 2013)

² <http://www.whereswally.co.uk/>

A execução do experimento foi dividida em duas etapas. A primeira consistiu de um treino, onde cada participante recebia duas expressões de referência já elaboradas, sendo uma com ambiguidade e a outra sem ambiguidade. A cada descrição recebida, os participantes deviam apontar para o objeto-alvo da imagem. Esta primeira etapa foi realizada com o intuito de demonstrar aos participantes a importância de expressões de referência sem ambiguidade. Durante a segunda etapa, realizada na sequência da primeira, os participantes foram expostos a 28 imagens apresentadas uma-a-uma, e deveriam elaborar expressões de referência na forma escrita, de maneira a distinguir o objeto-alvo dos outros objetos distraidores do estímulo visual.

Durante o experimento não houve validação automática das descrições coletadas e não houve restrição de tempo para a execução da tarefa. Foram recrutados 155 participantes, e foram coletadas ao todo 4256 descrições. Do total de participantes, 152 concluíram a tarefa com sucesso. Descrições de três participantes foram descartadas, dois deles por duplicaram seus experimentos e um por produzir descrições consideradas ambíguas.

3.2.6 Stars

O experimento elaborado em Paraboni et al. (2014) faz a coleta de expressões de referência para a composição de um corpus chamado Stars. Os objetivos do experimento foram investigar o uso de expressões relacionais, considerando especialmente situações de referência complexas em que era necessário referenciar até três objetos, como em “a estrela ao lado da caixa, próxima ao quadrado cinza”. Foi investigada também a questão da variação humana na seleção de conteúdo e a subespecificação de descrição relacionais, assunto este já abordado em Paraboni e Deemter (2014). A execução do experimento foi realizada com a participação de humanos pelo método via WEB, e a condução feita em modo monólogo. Um exemplo de estímulo utilizado no experimento Stars é apresentado na Figura 16.

As imagens utilizadas como estímulo visual do experimento foram representadas por cenas simples contendo figuras geométricas do tipo estrela, cubo e círculo. O objeto-alvo era sempre uma estrela apontada por uma seta em destaque. Além disso, em todas as imagens existiam objetos distraidores idênticos ao objeto-alvo, forçando assim o uso de

propriedades relacionais como em “a estrela ao lado da caixa vermelha, próxima ao círculo branco”.

Cada participante recebeu, através de um e-mail, um aplicativo desenvolvido na linguagem JAVA que deveria ser executado em seus computadores. Neste aplicativo existiam algumas instruções breves sobre o experimento, porém nenhum exemplo do tipo de descrição esperada foi disponibilizado. Em cada experimento, onze imagens eram apresentadas uma-a-uma de forma randômica, sendo seis imagens de interesse para o experimento e outras cinco imagens do tipo *filler*. Durante o experimento não houve restrição de tempo para a tarefa.

Figura 16 – Exemplo de um estímulo visual com imagens de figuras geométricas utilizado durante o experimento Stars

Fonte: (PARABONI et al., 2014)

Após produzir a descrição para cada imagem, os participantes deveriam pressionar um botão “próximo” para enviar sua resposta e para que outra imagem fosse apresentada. Ao receber estas respostas, o sistema realizava uma validação automática simples de ambiguidade para coibir o uso de expressões obviamente incorretas como “a estrela”, já que este tipo de descrição não era possível em nenhum dos estímulos visuais considerados. Nestes casos, uma mensagem era apresentada solicitando o envio de uma descrição mais detalhada. Em casos de ambiguidade mais complexos, esta validação automática não era realizada e os participantes conseguiam responder normalmente com qualquer tipo de cadeia de caracteres.

Foram recrutados para o experimento 73 estudantes Brasileiros nativos, e foram coletadas ao todo 803 expressões de referência. Do total de expressões coletadas, 99 foram descartadas devido à ambiguidade, restando 704 descrições de interesse para o experimento.

3.2.7 Zoom

O experimento realizado em Altamirano et al. (2015) faz a coleta de descrições de pontos de interesses em mapas, como restaurantes, bares, museus etc., representados em dois níveis de visualização, próximo e distante. O objetivo foi criar um corpus, chamado Zoom, composto por expressões de referência obtidas em modo monólogo e realizado via WEB. Um exemplo de estímulo utilizado no experimento é ilustrado na Figura 17.

Para realizar o experimento, foram utilizados mapas de dez regiões selecionadas das cidades de Madri e Lisboa adaptados para exibirem dois níveis de zoom. Dependendo do nível de exibição, um mesmo estímulo visual poderia apresentar maior ou menor número de distraidores a considerar e diferente grau de detalhamento geral (nomes de ruas etc).

Figura 17 – Exemplo de um estímulo visual com imagem de um mapa utilizado durante o experimento Zoom

Fonte: (ALTAMIRANO et al., 2015)

No experimento Zoom, a tarefa dos participantes era descrever na forma escrita os locais apontados por uma seta em cada um dos mapas. A cada descrição produzida, os participantes deviam pressionar o botão “próximo” e automaticamente um novo mapa era apresentado até que o experimento chegasse ao término. Os dois primeiros mapas apresentados no experimento foram utilizados para treino dos participantes. Nenhuma das descrições enviadas durante os experimentos passou por uma validação automática dos dados, ou seja, os participantes podiam enviar qualquer texto que desejassem. Além disso, não existiu restrição de tempo para a tarefa.

Foram recrutados para o experimento 173 participantes, sendo 93 nativos da língua portuguesa, formados por 66 homens e 27 mulheres e 80 nativos da língua espanhola, divididos entre 59 homens e 26 mulheres. Ao todo, foram coletadas 3560 expressões de referência, sendo que 968 foram descartadas por estarem mal formadas, restando um total de 2592 expressões.

De modo geral, as expressões coletadas no domínio Zoom são relativamente complexas, envolvendo até quatro objetos por descrição (e.g., “a Igreja na Rua da Vitoria entre os cruzamentos com a Rua do Ouro e a Rua do Crucifixo”). Por este motivo, o corpus Zoom não encontra-se ainda totalmente revisado e contém um número considerável de descrições ambíguas.

3.2.8 b5-ref

O experimento elaborado em Ramos et al. (2017) faz a coleta de descrições em experimentos com sujeitos humanos para a composição de um corpus chamado b5-ref. O objetivo do experimento foi coletar descrições para imagens que representavam fotografias de rostos humanos com diversos tipos de emoções e características físicas. Um exemplo de estímulo utilizado durante o experimento é ilustrado na figura 18. A condução do experimento foi em modo monólogo e realizado via WEB.

Figura 18 – Exemplo de um estímulo visual com imagem da fotografia de rostos humanos durante o experimento b5-ref

Fonte: (RAMOS et al., 2017)

Foram utilizados para o experimento um total de doze imagens bidimensionais. Cada imagem de estímulo possuía seis rostos humanos dispostos lado a lado, e distribuídos em

duas linhas com três rostos em cada uma das linhas. As imagens utilizadas no experimento b5-ref foram extraídas da base *Face Place*³.

Durante o experimento, cada participante recrutado foi exposto a doze imagens apresentadas uma-a-uma, e teve como objetivo elaborar descrições para cada imagem apresentada. Nestes estímulos as descrições deviam ser elaboradas de maneira a distinguir o objeto-alvo, destacado por uma borda vermelha, dos objetos distraidores.

Durante o experimento não houve restrição de tempo para a execução da tarefa, e foram coletadas ao todo 1810 descrições. Para a coleta de todas estas descrições foram recrutados 152 participantes, sendo que deste total, 86 foram do sexo feminino, e 66 do sexo masculino.

3.3 Considerações

Os estudos apresentados nesta seção demonstraram algumas das principais práticas de experimentação que têm sido seguidas na pesquisa de GER. Com o objetivo de resumir estas práticas de forma visual, é apresentada a Tabela 1. Nesta tabela, a coluna “Tarefa” indica se a condução do experimento em questão foi realizada com um tipo de tarefa real(R) ou simulada(S). Consideramos uma tarefa como sendo real quando ela envolve uma situação de comunicação em que o uso de uma RE ocorre naturalmente. Por exemplo, em uma tarefa em que participantes negociam (ainda que de forma fictícia) a compra e venda de itens, como em Eugenio et al. (2000), ou participantes de um jogo textual como em Kazemzadeh et al. (2014a). Por outro lado, consideramos a tarefa como sendo simulada quando os participantes eram explicitamente orientados a descrever um objeto sem que este ato estivesse inserido em uma situação de comunicação, como na tarefa de identificar objetos no corpus TUNA (GATT; SLUIS; DEEMTER, 2007).

A coluna modo indica se a condução de cada experimento foi realizada em modo monólogo (M) ou modo diálogo (D). A coluna “Corr” (correção) indica se os participantes podiam corrigir suas descrições após as terem produzido.

A coluna “Val” indica se houve a validação automática das descrições produzidas durante o experimento. A coluna “Lin” (linguagem) indica se as descrições foram coletadas de forma escrita (E) ou oral (O); a coluna “Suj” (sujeito) indicada a quantidade de participantes recrutados nos experimentos e a coluna Est (estímulos) indica a quantidade

³ Cortesia de: Righi, Peissig e Tarr (2012)

de estímulos visuais utilizados, ou seja, o número de imagens que cada participante tinha que descrever.

A coluna “Itens” indica o número total de descrições que foram coletadas. A coluna “Método” indica de que forma os participantes acessaram os experimentos, dividida em forma presencial (P), via serviço de *crowdsourcing Amazon Turk* (AT), e via WEB (WEB).

Por fim, a coluna “Tempo” indica se existiu restrição de tempo para os participantes realizarem as tarefas dos experimentos e a coluna “Comp” (computador) indica se foram utilizados meios computacionais para sua realização.

Tabela 1 – Análise dos experimentos de GER considerados

	Tarefa	Modo	Corr	Val	Lin	Suj	Est	Itens	Método	Tempo	Comp
Bishop	S	D	sim	não	E	6	30	268	P	não	sim
COCONUT	R	D	não	não	E	24	3	482	P	sim	sim
Craft	R	D	sim	não	O	20	5	505	P	sim	não
Drawer	S	M	não	não	O	20	7	140	P	não	não
GenX	S	M	não	não	E	na	var	5380	AT	não	sim
GIVE-2	R	D	sim	não	E	72	3	6180	P	não	sim
GRE3D3	S	M	não	não	E	74	10	740	WEB	não	sim
GRE3D7	S	M	não	não	E	318	32	4704	WEB	não	sim
ReferItGame	R	D	não	sim	E	na	var	130.525	AT	sim	sim
Stars	S	M	não	sim	E	73	11	803	WEB	não	sim
Stars2	S	D	sim	sim	E	56	16	884	P	não	sim
TUNA	S	M	não	não	E	90	20	1800	WEB	não	sim
Wally	S	M	não	não	E	155	28	4256	AT	não	sim
Zoom	R	M	não	não	E	173	20	3560	WEB	não	sim
b5-ref	S	M	não	não	E	152	12	1810	WEB	não	sim

Fonte: Danillo da Silva Rocha, 2017

Com base neste levantamento, destacamos algumas considerações pertinentes. Em primeiro lugar, a análise dos experimentos considerados nesta seção permitem observar que pouco mais da metade (nove) dos experimentos foram conduzidos em modo de condução monólogo com possível coleta de expressões ambíguas ou mal-formadas.

A validação automática dos dados foi observada em três experimentos. Nos experimentos Stars e Stars2 uma validação limitada das descrições foi realizada, produzindo uma mensagem de erro, e solicitando uma nova resposta. Em casos de extrema ambiguidade (e.g., “a bola”), entretanto, não existiu validação de dados no Stars. Validações mais complexas foram aplicadas no experimento ReferItGame, no modo de jogo entre um participante e o computador. Neste caso, foi implementado um método de comparação das descrições enviadas pelos participantes com descrições previamente capturadas. Ainda assim, entretanto, observa-se que a construção de uma base de exemplo deste tipo pode

ser um fator limitante para o tipo de descrição coletada, tornando-se não apropriado para certos tipos de estudos.

Experimentos de comunicação interativa acrescentam a possibilidade de correção das descrições, como observado nos experimentos conduzidos no modo em diálogo Bishop, Stars2, Craft e GIVE-2. Durante estes experimentos, caso um sujeito locutor enviasse uma instrução ou descrição que não fosse compreendida pelo sujeito ouvinte, o mesmo deveria refazê-la de forma textual, como nos experimentos Stars2 e GIVE-2, ou de forma verbal, como nos experimentos Bishop e Craft, e enviá-la para uma nova análise.

A restrição de tempo para a conclusão das tarefas foi adotada em três trabalhos. No experimento COCONUT havia uma restrição de duas horas para a conclusão de cada partida do jogo que foi proposto. No experimento Craft, a necessidade de uma restrição de tempo havia sido identificada previamente em um experimento piloto. No caso do experimento ReferItGame, foi imposta uma restrição de tempo tanto para as partidas entre humano-humano, quanto para as partidas entre humano-computador.

4 Estudo exploratório

Diferenças metodológicas como as apresentadas no capítulo anterior demonstram esforços da área de GER na validação das descrições produzidas em experimentos, e isso sugere uma preocupação com a qualidade dos dados obtidos pelos diferentes métodos. Em especial, é possível que estes esforços procurem, em certa medida, contornar as limitações de situações experimentais pouco realistas (como as observadas em experimentos de tipo monólogo via WEB), de modo a aproximá-las mais de situações de comunicação mais natural (como as observadas em diálogos).

Nesta seção será apresentado um estudo exploratório conduzido com participantes humanos para a geração de expressões de referência em modo monólogo. O objetivo deste estudo foi o de reproduzir em modo monólogo o experimento Stars2 (PARABONI; GALINDO; IACOVELLI, 2017), originalmente realizado em modo diálogo, de modo a investigar se há diferenças significativas na produção da língua nestas duas modalidades e, consequentemente, na interpretação dos resultados destes experimentos.

Para a condução deste novo experimento, foram selecionados 24 instâncias do corpus Stars2, ou seja, reproduzimos 24 sequências de estímulos visuais tal qual ocorreram no experimento original, porém sem a participação de um ouvinte. Estas sequências são documentadas no Apêndice A.

4.1 Hipóteses

O experimento considera quatro hipóteses relacionadas a diferenças entre descrições produzidas em modo monólogo e diálogo. Estas hipóteses abordam aspectos quantitativos da produção de descrições, definidos como a quantidade de propriedades semânticas anotadas conforme o esquema de anotação adotado no experimento Stars2.

De modo geral, as quatro hipóteses partem do pressuposto básico de que, quando há presença de um ouvinte (ou seja, em modo diálogo), locutores tendem a ser mais cuidadosos e a incluir mais informações do que fariam em monólogo (por exemplo, escrevendo para uma interface via WEB como no presente experimento). Este pressuposto é motivado, em parte, pela observação em Arts et al. (2011a) de que, em situações de comunicação consideradas críticas (ou de alguma forma importantes para o locutor), expressões de

referência tendem a ser mais superespecificadas. Assim, consideramos que as descrições coletadas em modo diálogo, em virtude da presença de um ouvinte humano, tendem a conter mais informação do que as produzidas em modo monólogo (no qual o locutor pode ser menos cuidadoso com a forma como será interpretado).

A primeira hipótese propõe uma possível explicação para diferenças no tamanho total da descrição nos dois modos:

h1. Descrições produzidas em modo diálogo são mais extensa do que as produzidas em modo monólogo.

A hipótese *h1* será testada comparando-se o número médio de propriedades presentes nas descrições em modo diálogo com o número médio de propriedades presentes nas descrições produzidas sob circunstâncias idênticas em modo monólogo, utilizando para este fim todo o conjunto de dados coletado. Espera-se que as descrições produzidas em modo diálogo sejam em média mais extensas do que as produzidas em monólogo. Por exemplo, espera-se que diálogos contenham expressões mais detalhadas como “a caixa azul”, e que os monólogos contenham expressões mais breves, como “a caixa”.

A segunda hipótese trata especificamente do uso de relações espaciais nas duas modalidades:

h2. Descrições produzidas em modo diálogo possuem maior quantidade de relações espaciais do que aquelas produzidas no modo monólogo.

A hipótese *h2* será testada comparando-se o número médio de propriedades relacionais presentes nas descrições em modo diálogo com o número médio de propriedades relacionais presentes nas descrições produzidas sob circunstâncias idênticas em modo monólogo, utilizando para este fim apenas as descrições do experimento em que esse tipo de descrição era necessária, ou seja, as condições relacionais do experimento Stars2 (cf. Capítulo 3). Espera-se que as descrições produzidas em modo diálogo possuam, em média, mais propriedades relacionais do que as produzidas em modo monólogo. Por exemplo, espera-se que diálogos contenham mais expressões relacionais como “a caixa ao lado do cone”, e que os monólogos contenham mais expressões atômicas, como “a caixa”.

A terceira hipótese diz respeito ao grau de especificação utilizado para descrever o objeto-alvo da descrição nos dois modos:

h3. Descrições produzidas em modo diálogo referenciam o objeto-alvo com maior grau de especificação do que descrições produzidas em modo monólogo.

A hipótese *h3* será testada comparando-se o grau de especificação do objeto-alvo em descrições produzidas em modo diálogo com as descrições produzidas em modo monólogo. Espera-se que as descrições produzidas em modo diálogo possuam, em média, grau de especificação do objeto-alvo maior do que as produzidas em modo monólogo. Por exemplo, espera-se que diálogos contenham mais expressões cujo alvo é superespecificado, como “a caixa *azul* ao lado do cone”, e que os monólogos contenham mais expressões em que o alvo é minimamente especificado, como “a caixa ao lado do cone”.

Finalmente, a quarta hipótese diz respeito ao grau de especificação do ponto de referência, ou seja, ao objeto auxiliar empregado para completar a descrição do objeto-alvo, como em “a bola ao lado do cone”, onde “cone” serve como ponto de referência para o alvo “bola”:

h4. Descrições produzidas em modo diálogo referenciam o ponto de referência com maior grau de especificação do que descrições produzidas em modo monólogo.

A hipótese *h4* será testada comparando-se o grau de especificação do ponto de referência em descrições produzidas em modo diálogo com as descrições produzidas em modo monólogo. Espera-se que as descrições produzidas em modo diálogo possua, em média, grau de especificação do ponto de referência maior do que as produzidas em modo monólogo. Por exemplo, espera-se que diálogos contenham mais expressões em que o ponto de referência é superespecificado, como “a caixa ao lado do cone *vermelho*”, e que os monólogos contenham mais expressões em que o ponto de referência é minimamente especificado, como em “a caixa ao lado do cone”.

4.2 *Sujeitos*

Foram recrutados 24 brasileiros nativos, que possuíam faixa etária média de 31 anos, eram predominantemente do sexo masculino (22 pessoas) e graduados. Nenhum participante possuía qualquer tipo de problema de visão que pudesse comprometer sua atuação, e todos participaram de forma voluntária.

4.3 *Materiais*

O experimento faz uso dos mesmos materiais utilizados no experimento Stars2 (PARABONI; GALINDO; IACOVELLI, 2017). As 24 instâncias selecionadas para serem reproduzidas foram escolhidas dentre aquelas em que não houve reformulação da descrição produzida, ou seja, foram instâncias do experimento Stars2 em que as 16 respostas foram produzidas corretamente na primeira tentativa.

A lista de estímulos visuais exibidos em cada uma destas instâncias é apresentada no Apêndice A. As sequências de estímulos visuais de cada instância do experimento foram disponibilizadas em uma ferramenta WEB que procurou, tanto quanto possível, ser fiel ao ambiente Stars2 original.

4.4 *Procedimento*

O experimento foi realizado presencialmente. Os participantes deveriam realizar o acesso ao sistema por meio de um nome de usuário e senha que havia sido previamente informado com o envio de um e-mail. Ao realizar o acesso, os participantes eram direcionados a um questionário com perguntas a respeito de dados pessoais como idade, gênero e mão de escrita, coletados para fins descritivos.

Após os participantes concluírem estas respostas, o sistema apresentava um conjunto de instruções a respeito do experimento, como os objetivos envolvidos e as restrições que deveriam ser observadas durante a tarefa proposta. Estas instruções eram as mesmas apresentadas no experimento Stars2 original e, assim como naquele caso, nenhum exemplo de como elaborar as expressões foi apresentado. Com a conclusão da leitura das instruções

iniciais, os participantes eram direcionados para o experimento propriamente dito. Nesta etapa um conjunto de 20 imagens eram apresentadas uma-a-uma.

As primeiras quatro imagens do experimento eram utilizadas para o treinamento dos participantes, que podiam tirar dúvidas a respeito da utilização do sistema e dos objetivos do experimento. Estas imagens não foram consideradas para a análise de dados, e foram as mesmas para todos os participantes. A partir da quinta imagem não era mais permitido que os participantes tirassem dúvidas.

Este procedimento era assim o mesmo do experimento original, porém sem o retorno de um segundo participante ouvinte para indicar se a expressão produzida permitia a identificação do objeto-alvo, ou se era ambígua.

4.5 Resultados

O experimento resultou em um conjunto de 384 descrições que foram anotadas seguindo-se o mesmo esquema de anotação do *corpus Stars2* (PARABONI; GALINDO; IACOVELLI, 2017). Antes da apresentação dos resultados propriamente ditos, foi realizado um teste para medir o grau médio de semelhança entre os dois grupos de expressões (em monólogo e diálogo) medindo-se o coeficiente Dice (DICE, 1945). Como resultado, observamos que o grau de semelhança entre estas expressões foi de 0,72, o que de imediato já sugere algum tipo de diferença entre as duas modalidades e/ou entre os grupos de participantes avaliados. Os resultados do experimento a seguir procuram assim tornar esta diferença mais explícita.

Para o conjunto de dados originais (diálogo) e os coletados no presente experimento (monólogo), o Quadro 2 sumariza o tamanho médio das expressões (*length*), a quantidade média de relações espaciais empregadas (*rel-count*), o grau médio de superespecificação do objeto-alvo (*over-tg*), e o grau médio de superespecificação do ponto de referência (*over-lm1*) conforme exigido para o teste das hipóteses: $h1$, $h2$, $h3$, e $h4$, respectivamente.

Neste trabalho, o grau de superespecificação é definido como a quantidade de propriedades (exceto o tipo do objeto) que poderiam ser removidas da descrição sem provocar ambiguidade.

Para a avaliação do grau médio de superespecificação (*over-tg*) do objeto-alvo foi utilizado metade do conjunto de descrições coletas (192), que apresentavam situações que

Quadro 2 – Resultados estatísticos do estudo exploratório

	h1-length		h2-rel-count		h3-over-tg		h4-over-lm1	
	diálogo	monólogo	diálogo	monólogo	diálogo	monólogo	diálogo	monólogo
média	3,35	2,98	0,95	0,82	0,59	0,51	0,84	0,71
var	3,57	2,60	0,68	0,58	0,47	0,32	0,38	0,32
n	384	384	384	384	192	192	192	192

Fonte: Danillo da Silva Rocha, 2017

motivam o uso de descrições superespecificadas. A outra metade do conjunto de descrições (192) foi utilizado para avaliar o grau médio de superespecificação do ponto de referência (over-lm1), pois apresentavam situações que motivavam o uso de descrições relacionais.

Os resultados sugerem que, de modo geral, expressões em monólogo são mais breves (cf. *h1*), possuem um número médio menor de relações espaciais (cf. *h2*), menor grau de especificação do objeto-alvo (cf. *h3*) e do ponto de referência (cf. *h4*). Para verificar a significância destas observações utilizamos a seguir uma análise do tipo ANOVA de fator único entre grupos de sujeitos sobre cada variável de interesse.

Para a hipótese *h1*, de que descrições produzidas em modo diálogo são em média maiores do que aquelas produzidas em modo monólogo, a diferença observada é significativa ($F(1,101)=8,52$, $MSE=26,285$ $p<0,05$). Isso oferece suporte à hipótese *h1*.

Para a hipótese *h2*, de que as descrições produzidas em modo diálogo possuem em média maior quantidade de relações espaciais do que aquelas produzidas em modo monólogo, a diferença observada também é significativa ($F(1,101)=5,15$, $MSE=3,245$ $p<0,05$). Isso oferece suporte à hipótese *h2*.

Para a hipótese *h3*, de que descrições produzidas em modo diálogo descrevem o objeto-alvo com maior grau de especificação do que descrições produzidas em modo monólogo, a diferença observada não é significativa.

Finalmente, para a hipótese *h4*, de que descrições produzidas em modo diálogo descrevem o ponto de referência do objeto-alvo com maior grau de especificação do que descrições produzidas em modo monólogo, a diferença observada é significativa ($F(1,101)=4,635$, $MSE=1,622$ $p<0,05$). Isso oferece suporte à hipótese *h4*.

4.6 Considerações

Apesar do pequeno número de participantes recrutados para este estudo exploratório, e das dificuldades de comparação entre grupos de sujeitos distintos, o experimento realizado

oferece algum indício de que descrições produzidas em modo monólogo podem ser diferentes daquelas que seriam produzidas em situações de diálogo. Estas observações motivam o modelo computacional a ser detalhada no próximo capítulo.

5 Modelo de validação automática de REs

Cópus para a Geração de Expressões de Referencias (GER) são essenciais para o estudo de fenômenos de referência e outras questões de pesquisa da área, mas possuem um elevado custo de desenvolvimento. Em especial, o estudo de questões de seleção de conteúdo em GER, que é um fenômeno de natureza altamente semântica, depende da anotação de grandes conjuntos de dados coletados. Além disso, conforme discutido no Capítulo 1, a coleta de dados via web em modo monólogo pode acarretar grande desperdício de dados (e participantes), já que participantes menos atentos tendem a produzir mais descrições ambíguas (ALTAMIRANO et al., 2015), o que não ocorre com tanta frequência em experimentos envolvendo duplas de participantes em modo diálogo (PARABONI; GALINDO; IACOVELLI, 2017).

Com base nestas observações, este capítulo apresenta a proposta de um modelo computacional de apoio à implementação de experimentos de GER de baixo custo (i.e., envolvendo o uso de participantes em modo monólogo) nos quais o segundo participante humano é substituído por um método de validação automática. O modelo proposto recebe como entrada uma descrição textual fornecida pelo participante de um experimento de GER, e indica se a descrição permite a identificação única do objeto-alvo de interesse com base no contexto fornecido. Em outras palavras, o modelo verifica se a descrição fornecida é ambígua ou não.

A informação de ambiguidade fornecida pelo modelo pode ser utilizada pelo ambiente de realização do experimento para emitir uma mensagem de erro e, se pertinente, solicitar uma nova descrição. Além disso, como a verificação de ambiguidade requer uma interpretação mínima do texto fornecido (e.g., na forma de uma análise superficial) para associar propriedades semânticas aos termos lexicais da descrição, o modelo fornece também uma anotação semântica inicial, e possivelmente incompleta, que pode ser posteriormente revisada pelo responsável pela anotação dos dados coletados.

O modelo proposto é dividido em dois módulos: um módulo que faz uma interpretação superficial da descrição fornecida com base nas suas palavras individuais, associando propriedades semânticas a elas, e um segundo módulo que, dada uma lista de propriedades (e.g., provenientes da análise automática feita no primeiro módulo), verifica

se estas propriedades constituem uma descrição livre de ambiguidade em um dado contexto. Cada um destes módulos é discutido individualmente a seguir.

5.1 Interpretação de descrições textuais

A tarefa de anotação de grandes massas de descrições definidas é um dos pontos críticos da construção de um *corpus* de GER, e uma das suas etapas de maior demanda de recursos humanos especializados. Uma vez que os resultados dos futuros estudos baseados nestes dados depende fundamentalmente da acurácia da anotação, não seria realista supor que este processo possa (ou deva) ser totalmente automatizado. No entanto, assim como em inúmeras tarefas de anotação de *corpus* típicas da pesquisa em GLN, a anotação semântica de descrições definidas pode em princípio beneficiar-se de métodos semiautomáticos para facilitar sua execução.

Com o objetivo de agilizar a tarefa de anotação de dados coletados em experimentos de GER, o modelo computacional proposto torna a tarefa de anotação semiautomática, e utiliza para isso uma base de conhecimento que faz o mapeamento entre as propriedades semânticas, e as palavras que as representam. Esta base de conhecimento é constituída por uma lista de propriedades do tipo atributo-valor, sendo que para cada propriedade um grupo de palavras é estabelecido. Existem casos em que um grupo de palavras pode representar mais de uma propriedade na base de conhecimento (por exemplo, japonês) Exemplo de uma base de conhecimento deste tipo é apresentado na Tabela 2.

Tabela 2 – Exemplos de propriedades semânticas e possíveis palavras que as realizam

Propriedade semântica	Palavra possíveis
<i>blue:colour</i>	azul, anil
<i>large:size</i>	grande, enorme, gigante
<i>light:eye</i>	olhos verdes, olhos azuis
<i>male:gender</i>	homem, rapaz
<i>ball:type, small:size</i>	bolinha, esferinha, redondinho
<i>yellow:colour, small:size</i>	amarelinha, douradinha
<i>male:gender, asian:race</i>	japonês, asiático
<i>female:gender, asian:race</i>	japonesa, asiática

Fonte: Danillo da Silva Rocha, 2017

Mapeamentos de palavras para propriedades como os ilustrados na Tabela 2 devem ser construídos pelos responsáveis pela pesquisa antes da execução do experimento de coleta de dados propriamente dito. Isso no entanto não necessariamente representa um obstáculo ao uso do método proposto, pois é inerente ao projeto destes experimentos a

definição prévia do contexto de referência, constituído do conjunto de objetos possíveis e de suas propriedades na forma atributo-valor. Com base nesta definição do domínio, a obtenção das palavras possíveis de cada propriedade pode ser obtida com relativa facilidade por meio de um certo número de experimentos-piloto com o objetivo de identificar as construções mais frequentes. Estas construções podem posteriormente ser complementadas pelo próprio pesquisador com sinônimos e variações mais prováveis, constituindo uma base representativa dos tipos de REs observados naquele domínio.

No caso específico dos domínios considerados nas avaliações do modelo proposto, que são objeto deste e do próximo capítulo, a base de mapeamentos entre propriedades e palavras possíveis foi obtida a partir de *cópus* de GER já existentes. De modo mais específico, partiu-se de conjuntos de dados para os quais as propriedades e suas possíveis palavras já são conhecidas nos domínios TUNA (GATT; SLUIS; DEEMTER, 2007), GRE3D3 (VIETHEN; DALE, 2008) e b5-ref (PARABONI; MONTEIRO; LAN, 2017).

O módulo de interpretação proposto consiste de um método baseado em heurísticas para associar propriedades às palavras do *string* fornecido como entrada. Resumidamente, o método identifica palavras constantes da base de mapeamentos descrita acima, e faz uso de regras de desambiguação simples (por exemplo, de associação de um adjetivo ao termo substantivo mais próximo) conforme descrito a seguir. Seja um domínio D composto de todos os objetos (alvos e distraidores) de um contexto de interesse, e suas propriedades possíveis na forma de pares atributo-valor. Dado um *string* s que representa uma descrição definida como uma lista de palavras, e um conjunto de mapeamentos M de propriedades-palavras no domínio D , tem-se como saída um conjunto L de propriedades anotadas que corresponde ao subconjunto de palavras de s que puderam ser identificadas nos mapeamentos de M . O Algoritmo 1 ilustra este procedimento.

Inicialmente, o algoritmo avalia qual a língua *lang* utilizada na elaboração da descrição (dada como entrada para o método). Caso a língua seja o Inglês (linha 2) o algoritmo inverte o *string* s (linha 3), de maneira a garantir que os possíveis adjetivos sejam localizados na mesma posição daquelas descrições elaboradas na língua portuguesa. Assim, descrições como “dark guy” são transformadas em “guy dark” e tratadas da mesma forma que “rapaz moreno”.

Em seguida, é criado um conjunto vazio de propriedades L (linha 4) a ser retornado como resultado, e a função *Split* (não detalhada) é invocada para dividir o *string* de entrada s em k subcomponentes $z_1, \dots, z_k$ separados por propriedades relacionais (linha

Algoritmo 1 Interpretação

```

1: procedure INTERPRETAR( $S, M, D, lang$ )
2:   if  $lang == English$  then
3:      $Reverse(S)$ 
4:    $L \leftarrow \emptyset$ 
5:    $Z \leftarrow Split(S, D)$ 
6:   for  $z_i \in Z$  do
7:     for  $w_j \in z_i$  do
8:        $np \leftarrow NearestNoun(w_j, z_i)$ 
9:        $p \leftarrow M[w_j + np]$ 
10:      if  $p \neq null$  then
11:         $L \leftarrow L \cup p$ 
12:      else
13:         $p \leftarrow M[w_j]$ 
14:        if  $p \neq null$  then
15:           $L \leftarrow L \cup p$ 
16:   return  $L$ 

```

 Fonte: Danillo da Silva Rocha, 2017

5). O objetivo deste procedimento é o de tratar individualmente cada objeto referenciado na expressão. Por exemplo, em “green ball near blue cube” a propriedade relacional “near” separa a porção esquerda do *string*, que se refere ao objeto-alvo (bola) da porção da direita, que se refere ao objeto landmark (cubo).

A tarefa de decidir se uma determinada propriedade é do tipo relacional (como *near*) ou atômica (i.e., intrínseca ao objeto, como *colour*) a partir da definição do domínio D é implementada de forma trivial pela função *Split* em virtude de que os valores destas propriedades denotam outros objetos do contexto, e não características do objeto em si. Este método de identificação de propriedades relacionais foi empregado em Ferreira e Paraboni (2017). Por exemplo, a propriedade *near-c4* é reconhecida como sendo relacional porque o valor *c4* representa o identificador de um objeto em D , e não uma característica intrínseca do objeto-alvo, como seria o caso de uma propriedade atômica como *colour-green*.

Uma vez delimitado, cada *substring* z_i é tratado individualmente (linha 6) e suas palavras são consideradas em associação ao substantivo mais próximo (8-12) e, se necessário, de forma isolada (12-15). Em um primeiro momento, o substantivo mais próximo np (que é o provável núcleo do sintagma ao qual w_j é subordinada) é localizado pela função auxiliar *NearestNoun* (não detalhada no pseudocódigo) (8). Cabe salientar que a função *NearestNoun* retorna um string vazio nos casos em que a palavra w_j for ela própria um substantivo. A seguir, a combinação da palavra atual w_j e núcleo np (por exemplo, para

“cabelo preto”) é consultada na base de mapeamentos M para verificar se esta corresponde a alguma propriedade p (9). Em caso afirmativo (10), a propriedade p correspondente é acrescentada ao conjunto de propriedades anotadas L (11).

A função *NearestNoun* (8) é a responsável por buscar o substantivo mais próximo para uma determinada palavra. Por simplicidade, a busca é feita com o auxílio de uma lista de substantivos previamente elaborada para o domínio de interesse. Na prática, entretanto, seria possível consultar um dicionário como Muniz (2004).

Caso não haja um mapeamento da combinação $w_j + n_p$ para nenhuma propriedade do domínio, a palavra w_j é então consultada individualmente, ou seja, sem considerar possíveis associações a nenhum núcleo específico (15). Caso haja um mapeamento entre a palavra e uma propriedade (e.g., “sorridente” mapeada para *smile-yes*) p (16), p é acrescentada ao conjunto de propriedades anotadas L (17). O processo é repetido até que todas as palavras de todos os *substrings* tenham sido consideradas, e então o conjunto de anotações L identificadas é retornado (21). Ao longo deste processo, palavras não identificadas são desconsideradas, e portanto a anotação L resultante pode permanecer incompleta ou mesmo vazia.

Apesar da simplicidade, o método proposto pode ser considerado suficiente para o domínio em questão, ou seja, para a interpretação de descrições definidas do tipo comumente considerado em experimentos de GER em inglês (GATT; SLUIS; DEEMTER, 2007; VIETHEN; DALE, 2008) e português (PARABONI; MONTEIRO; LAN, 2017). Em domínios deste tipo, observamos que não há necessidade real de uma análise sintática de ampla cobertura (e.g., para tratamento de cláusulas relativas e outros fenômenos), sendo de modo geral suficiente a identificação de componentes lexicais individuais.

5.2 Identificação do objeto-alvo

Uma vez que as propriedades representadas por uma forma textual sejam, no todo ou em parte, identificadas pelo método descrito na seção anterior, o próximo passo do modelo de validação proposto consiste em verificar se esta RE permite a identificação de forma única do objeto-alvo de interesse. Para este fim, foi desenvolvido um método de identificação do objeto-alvo conforme descrito a seguir.

O algoritmo proposto recebe como entrada uma lista de propriedades semânticas L e um conjunto de objetos C representando o contexto em que deseja-se verificar se L é uma expressão única ou não, retornado verdadeiro ou falso conforme o caso. O Algoritmo 2 ilustra este procedimento.

Algoritmo 2 Identificação

```

1: procedure IDENTIFICAR( $L, C$ )
2:   for  $\langle a, v \rangle \in L$  do
3:     if  $|rulesOut(\langle a, v \rangle)| > 0$  then
4:        $C \leftarrow C - rulesOut(\langle a, v \rangle)$ 
5:   if  $|C| == 1$  then
6:     return true
7:   else
8:     return false

```

Fonte: Danillo da Silva Rocha, 2017

Inicialmente, o algoritmo faz a comparação entre cada propriedade $\langle a, v \rangle$ em L com os objetos de C (2-4). Nesta comparação é utilizada a função *rulesOut* (não detalhada) e utilizada no algoritmo Incremental (DALE; REITER, 1995). Essa função é responsável por retornar o conjunto de objetos que seriam eliminados pela propriedade $\langle a, v \rangle$. Caso o retorno da função seja um conjunto não vazio (3), o conjunto C é atualizado (4) com a retirada dos objetos encontrados.

Ao final da busca de propriedades $\langle a, v \rangle$, o algoritmo realiza uma verificação da cardinalidade em $|C|$ (5). Caso exista apenas um objeto em C , a função retorna *true* (6), caso contrário retorna *false* (8).

6 Avaliação Intrínseca

Como forma de testar o modelo proposto nesta seção, foi realizada uma avaliação intrínseca com base em quatro corpú de descrição definidas em inglês já disponibilizados com anotação semântica. O objetivo desta avaliação foi o de medir o grau de proximidade entre a anotação existente em cada corpú e aquela que seria obtida de forma semiautomática com o uso do modelo proposto.

6.1 Dados

Os corpú utilizados na avaliação foram o TUNA-*Furniture* e TUNA-*People* (GATT; SLUIS; DEEMTER, 2007), GRE3D3 (VIETHEN; DALE, 2008), e GRE3D7 (VIETHEN; DALE, 2011). Para fins de avaliação do método proposto, foi considerado como sistema de baseline um método de anotação fruto de um projeto paralelo descrito em (ROCHA; LAN; PARABONI, 2017). Esse método é baseado na ferramenta nlpnet (FONSECA E. R., 2013), e funciona como em uma tarefa de anotação de *part-of-speech* através de etiquetas como V (verbo) ou S (substantivo).

Para cada um dos quatro corpú, uma pequena porção de dados (de 14 a 18 %) foi utilizada para treinamento do modelo, e o restante reservado para teste. A Tabela 3 apresenta o número de instâncias de treinamento e teste e alguns exemplos linguísticos observados em cada domínio.

Tabela 3 – Corpú de teste

Domínio	Treinamento	Teste	Exemplo
TUNA- <i>Furniture</i>	63	288	the large red couch
TUNA- <i>People</i>	54	303	the man with gray beard and glasses
GRE3D3	90	540	the small green cube
GRE3D7	624	3856	the red ball next to a large cube

Fonte: Danillo da Silva Rocha, 2017

6.2 Procedimentos

A avaliação intrínseca consistiu em aplicar os dois modelos previamente treinados ao conjunto de teste de cada domínio, comparando-se seus resultados com a anotação de referência disponibilizada em cada corpú por meio de coeficiente Dice (DICE, 1945)

e Acurácia. As palavras do conjunto de testes foram anotadas com suas propriedades semânticas no formato exigido por cada um dos sistemas sob avaliação: no caso do baseline em POS-*tagging*, e no caso do modelo proposto com o uso de regras heurísticas simples.

6.3 Resultados

A Tabela 4 representa os valores médios de coeficientes Dice e Acurácia obtidos pelos dois métodos avaliados para cada um dos corpú de teste.

Tabela 4 – Resultados da Avaliação Intrínseca

Corpú de teste	Baseline		Proposta	
	Dice	Acc.	Dice	Acc.
TUNA- <i>Furniture</i>	0,63	0,09	0,83	0,47
TUNA- <i>People</i>	0,50	0,12	0,38	0,01
GRE3D3	0,76	0,74	0,99	0,96
GRE3D7	0,95	0,92	0,97	0,86

Fonte: Danillo da Silva Rocha, 2017

De modo geral, o método proposto apresenta resultados superiores ao baseline baseado em POS-*tagging* com exceção do domínio TUNA-*People*. Observa-se resultados altamente positivos, e próximos de 100% nos domínios GRE3D3 e GRE3D7, e resultados mais modestos no caso dos domínios TUNA-*Furniture* e, de modo especial, TUNA-*People*. Os domínios TUNA são assim mais indicativos da complexidade da tarefa, diferença está que pode ser em grande parte explicada pela maior variedade lexical (especialmente no caso de TUNA-*People*), a qual não é suficientemente representada pelo pequeno conjunto de treinamento utilizado na avaliação.

A comparação entre coeficientes Dice médios dos dois métodos foi realizada com o uso do teste de Wilcoxon. O método proposto apresentou acurácia média significativamente superior à do método POS nos domínios TUNA-*Furniture* ($W=23027$, $Z=1,85$, $p < 0.001$), GRE3D3 ($W=10072$, $Z=10.15$, $p < 0.001$) e GRE3D7 ($W=17180$, $Z=3.03$, $p = 0.0024$). No domínio TUNA-*People*, um efeito contrário foi observado ($W=-14630$, $Z=-6.13$, $p < 0.001$).

Finalmente a comparação entre valores de acurácia média dos dois métodos foi realizada com o uso do teste de Qui-quadrado. O método proposto apresentou acurácia media significativamente superior à do método POS nos domínios TUNA-*Furniture* ($X^2 = 35.81$, $df = 1$, $p < 0.01$), e GRE3D3 ($X^2 = 18.98$, $df = 1$, $p = 0.00013$). No caso do domínio TUNA-*People*, um efeito contrário foi observado ($X^2 = 9.95$, $df = 1$, $p = 0.001604$), e no

caso do domínio GRE3D7, a acurácia média dos dois métodos não apresenta diferença significativa.

6.4 *Discussão*

Neste capítulo foi apresentado uma comparação entre dois métodos de anotação semiautomática de descrições definidas: um método de associação de palavras às propriedades semânticas que elas representam, e um método baseado em uma ferramenta de *POS-tagging* já existente. Os métodos avaliados são de modo geral equivalentes e apresentam desempenho próximo, servindo em ambos os casos para simplificar a tarefa de anotação de grandes massas de dados que são a saída típica de experimentos de coleta de dados para a pesquisa em GER.

Na análise apresentada, foi observada uma ligeira vantagem para o método baseado em *POS-tagging* no caso do *TUNA-People*, mas cabe destacar que o método proposto pode ser mais adequado para processamento de descrições individuais durante a execução do experimento de coleta de dados (por exemplo, para fins de validação em tempo real), já que este método não depende de treinamento prévio.

7 Avaliação Extrínseca

Em complemento à avaliação intrínseca apresentada no capítulo anterior, este capítulo descreve o uso prático do modelo computacional proposto na forma de uma ferramenta web de validação de descrições produzidas por participantes humanos em uma série de experimentos para coleta de dados de GER. O objetivo principal desta avaliação extrínseca é comparar as descrições produzidas a partir de experimentos realizados com uso da ferramenta de validação automática com descrições produzidas nas mesmas situações, porém a partir de experimentos em modo monólogo, ou seja, envolvendo apenas um participante humano que utiliza uma interface computacional sem *feedback* de modo a verificar qual destas duas modalidades seria a mais próxima de um diálogo.

7.1 Visão Geral

A avaliação proposta considera três execuções de um mesmo experimento de coleta de descrições produzidas por sujeitos humanos. Estas três execuções, ou modos de condução do experimento, são denominadas Monólogo, *Feedback* e Diálogo.

A condução em modo Monólogo consiste em um experimento normal baseado em uma interface web sem *feedback*. Este modo é semelhante ao adotado na construção de diversos corpúscos de GER como TUNA (GATT; SLUIS; DEEMTER, 2007) e GRE3D3/7 (VIETHEN; DALE, 2008; VIETHEN; DALE, 2011).

Em modo *Feedback*, a condução consiste na repetição do mesmo experimento, porém realizando a interpretação e identificação da descrição produzida com uso dos modelos descritos no Capítulo 5. Em outras palavras, o modo *Feedback* simula, ainda que de forma rudimentar, a presença de um interlocutor humano com uso da ferramenta proposta. Esta ferramenta é responsável por verificar a validade do *string* produzido e, em caso de falha, pedir uma confirmação ao participante. Situações típicas deste tipo incluem, por exemplo, casos em que o participante descreveu o alvo como “bola” porém há mais de um objeto do tipo bola no contexto. Como já mencionado, esta forma de validação é rudimentar, e objetiva apenas minimizar os casos de respostas incompletas ocasionadas pela falta de atenção do participante e outras dificuldades deste tipo.

A condução em modo Diálogo consiste de uma terceira execução do mesmo experimento, porém utilizando um segundo participante humano (o próprio responsável pela pesquisa) no papel de ouvinte. Este método corresponde ao paradigma experimental *director matcher* (KRAUSS; WEINHEIMER, 1966) no qual o papel de diretor é desempenhado pelos participantes do experimento (locutores), e o papel de assistente é desempenhado pelo autor desta pesquisa (o ouvinte).

7.2 Hipótese

Com base em um conjunto de descrições produzidas por sujeitos humanos em situações idênticas nos três modos de condução experimental (Monólogo, *Feedback* e Diálogo), este estudo considera a seguinte hipótese de pesquisa:

h5. Descrições coletadas em modo Feedback são mais próximas das descrições obtidas em modo Diálogo do que as obtidas em modo Monólogo.

Espera-se assim que o uso da ferramenta proposta (e utilizada na coleta de dados do modo *Feedback*) possua impacto positivo sobre as descrições coletadas no sentido específico de que estas descrições seriam mais próximas das obtidas em um diálogo “real” (i.e., entre humanos). A hipótese será testada comparando-se o grau de sobreposição entre os conjuntos de propriedades semânticas das descrições Monólogo, *Feedback* e Diálogo medido por coeficientes Dice. Espera-se que o coeficiente Dice entre os modos *Feedback* e Diálogo seja maior do que o coeficiente Dice entre os modos Monólogo e Diálogo.

7.3 Sujeitos

Foram recrutados 35 brasileiros nativos que participaram de forma voluntária na pesquisa. Os participantes eram predominantemente do sexo masculino (30), e possuíam faixa etária média de 34 anos (mínimo 20 e máximo 50 anos).

7.4 Procedimentos

Para a condução do experimento, um mesmo procedimento de coleta de dados foi executado três vezes, uma em cada modo de condução, e com o mesmo grupo de sujeitos. Para minimizar possíveis efeitos de memorização, foi mantido um intervalo mínimo de tempo entre a primeira e segunda execução, e também entre a segunda e a terceira.

Considerando-se que o objetivo maior da investigação é o de verificar o grau de semelhança entre descrições do tipo Monólogo e *Feedback* em relação as descrições do tipo Diálogo, a condução Diálogo foi sempre a última a ser realizada para todos os sujeitos. Além disso, como forma de evitar que a ordem das execuções tivesse influência no resultado (por exemplo, o sujeito pode repetir descrições utilizadas na execução mais recente) metade dos sujeitos realizaram primeiramente o experimento em Monólogo, e a outra metade realizou primeiramente o modo *Feedback*.

As três conduções foram realizadas na forma de experimentos de GER tradicionais semelhantes ao discutido no Capítulo 3. Após instruções básicas e um certo número de imagens de prática, os participantes eram informados de que o experimento real iria começar. Isso era seguido da exibição das imagens de estímulo, uma a uma. A tarefa do participante era descrever o objeto alvo destacado em cada imagem e então prosseguir para a próxima, até a conclusão da tarefa.

A condução em modo Monólogo não acrescentava nenhum outro procedimento ao descrito acima. Entretanto, na condução em modo *Feedback*, a descrição produzida era validada e, em caso de ambiguidade, uma mensagem era apresentada. Nesta mensagem os participantes eram informados de que sua descrição não havia sido compreendida para o estímulo apresentado, e era oferecida a possibilidade de continuar o experimento sem alterar a descrição, ou elaborar uma nova frase para o mesmo estímulo visual. Durante o experimento os participantes entretanto não eram informados de que havia uma ferramenta de validação automática em ação.

Na condução em Diálogo, o pesquisador responsável colaborava remotamente como ouvinte, informando ao participante se a descrição produzida era identificável ou não. Neste caso os participantes estavam cientes desde o começo do experimento de que estavam se comunicando com outro participante humano.

7.5 *Materias*

Como materiais para o experimento foram utilizadas 13 situações de referência do corpus GRE3D3 (VIETHEN; DALE, 2008) e mais 13 situações de referência do corpus b5-ref (PARABONI; MONTEIRO; LAN, 2017). As imagens dos dois domínios eram intercaladas para minimizar a repetição de padrões de respostas. As quatro imagens iniciais do experimento foram utilizadas para o treino dos participantes e não foram consideradas para a análise de dados, restando 22 imagens de interesse. A relação das imagens utilizadas nesta avaliação extrínseca com as imagens do corpus b5-ref e GRE3D3 é apresentada no Apêndice B.

7.6 *Resultados*

Ao todo, 96 execuções do experimento foram realizadas em um período de três meses, com um intervalo médio de 30 dias entre a primeira e segunda execução de cada participante (mínimo 13 e máximo 44 dias), e com um intervalo médio de 28 dias entre a segunda e terceira execução (mínimo 16 e máximo 45 dias).

Como resultado, foram coletadas 2112 descrições, sendo 1056 no domínio GRE3D3 e outras 1056 no domínio b5-ref. Todas descrições foram anotadas pelo responsável pela pesquisa e validadas por um segundo avaliador segundo os esquemas de anotação originais de cada corpus. As frequências dos atributos observados no modo Monólogo, *Feedback* e Diálogo são sumarizadas na Tabela 5 (domínio GRE3D3) e Tabela 6 (domínio b5-ref).

Tabela 5 – Frequências de propriedades no domínio GRE3D3

Atributo	Monólogo	<i>Feedback</i>	Diálogo
targe:type	100%	100%	100%
targe:colour	76,4%	73,0%	75,3%
targe:size	40,3%	43,2%	43,5%
targe:hpos	0,0%	0,0%	0,0%
targe:vpos	0,9%	0,0%	0,0%
below	13,9%	13,6%	14,2%
in-front-of	8,0%	8,8%	6,3%
near	0,9%	0,0%	0,6%
lm.type	21,3%	21,3%	19,6%
lm.colour	18,8%	20,2%	17,3%
lm.size	6,5%	7,7%	8,2%
lm.hpos	0,6%	0,6%	0,3%
lm.outros	6,3%	4,8%	5,4%

Fonte: Danillo da Silva Rocha, 2017

Tabela 6 – Frequências de propriedades no domínio b5-ref

Atributo	Monólogo	<i>Feedback</i>	Diálogo
gender	95,5%	98,0%	99,4%
isYoung	46,3%	41,2%	48,6%
race	44,9%	48,0%	49,7%
skin	3,4%	1,4%	1,1%
emotion	6,3%	4,5%	8,2%
smile	38,6%	42,0%	45,7%
facial.hair	4,0%	3,1%	2,3%
face	1,7%	0,9%	2,3%
hair.colour	12,5%	14,8%	13,6%
hair.length	7,7%	7,7%	8,8%
hair.style	2,3%	1,1%	1,4%
hair.others	2,8%	2,3%	3,1%
eye	1,7%	1,7%	2,6%
glasses	1,4%	2,0%	2,0%
relation	8,0%	6,8%	7,1%
relation.others	9,9%	7,1%	7,7%

Fonte: Danillo da Silva Rocha, 2017

Com relação à hipótese de pesquisa, que estabelece que descrições *Feedback* são mais próximas de Diálogo do que as descrições Monólogo, a Tabela 7 sumariza o coeficiente Dice (DICE, 1945) médio das duas comparações.

Tabela 7 – Coeficientes Dice

Cópus de teste	Monólogo		<i>Feedback</i>	
	Média	Dv.	Média	Dv.
GRE3D3	0,88	0,19	0,91	0,15
b5-ref	0,77	0,22	0,83	0,19

Fonte: Danillo da Silva Rocha, 2017

A comparação entre coeficientes Dice médios para as conduções em modo Monólogo e *Feedback* foi realizada com o uso do teste de Wilcoxon. O modo Monólogo apresentou Dice médio significativamente inferior ao modo *Feedback* para o cópus GRE3D3 ($W=3101, Z=-3.44, p=0.0006$), b5-ref ($W=8287, Z=-4.54, p<.0001$), e de maneira global (b5-ref e GRE3D3) ($W=21611, Z=-5.73, p<.0001$). Estes resultados oferecem suporte à hipótese de pesquisa *h5*.

7.7 Análise de Erros

Do total de 2112 descrições coletadas durante a avaliação extrínseca, 148 delas (7,1%) apresentaram coeficiente Dice maior em modo de condução Monólogo em comparação com o modo *Feedback* (91 descrições para o b5-ref e 57 descrições para o GRE3D3). Estes casos,

que contradizem a hipótese geral comprovada no experimento, são discutidos brevemente a seguir.

A Tabela 8 apresenta as diferenças entre os modos Monólogo e *Feedback* em relação ao modo Diálogo para o domínio b5-ref, e a Tabela 9 apresenta estas diferenças para o domínio GRE3D3.

Tabela 8 – Diferenças de frequência entre os modos *Feedback* e Monólogo em comparação ao modo Diálogo para o domínio b5-ref

Atributo	Monólogo	<i>Feedback</i>
gender	-2,2%	-1,1%
isYoung	1,1%	-18,7%
race	-0,5%	-1,6%
skin	0,0%	1,1%
emotion	-4,4%	-6,6%
smile	-2,2%	-1,1%
facial.hair	1,1%	2,2%
face	-3,3%	-3,3%
hair.colour	-2,2%	1,1%
hair.length	-1,1%	2,2%
hair.style	-2,2%	-2,2%
hair.others	-3,3%	0,0%
eye	2,2%	1,1%
glasses	-2,2%	1,1%
relation	-1,1%	-2,2%
relation.others	-1,1%	-1,1%

Fonte: Danillo da Silva Rocha, 2017

Tabela 9 – Diferenças de frequência entre os modos *Feedback* e Monólogo em comparação ao modo Diálogo para o domínio GRE3D3

Atributo	Monólogo	<i>Feedback</i>
targe:type	0,0%	0,0%
targe:colour	-3,5%	-22,8%
targe:size	1,8%	-8,8%
targe:hpos	0,0%	0,0%
targe:vpos	1,7%	0,0%
below	-3,5%	-5,27%
in-front-of	0,0%	3,5%
near	1,8%	-1,7%
lm.type	-3,5%	-3,5%
lm.colour	0,0%	0,0%
lm.size	0,0%	5,7%
lm.hpos	0,0%	-1,8%
lm.outros	0,0%	-12,3%

Fonte: Danillo da Silva Rocha, 2017

Com as diferenças de frequências apresentadas nas Tabelas 8 e 9 é possível identificar que as frequências de uso de atributos para o domínio b5-ref apresentou uma diferença considerável para o atributo “isYoung” (-18,7% para *Feedback*, e 1,1% para Monólogo em comparação a frequência obtida em Diálogo).

Para o domínio GRE3D3 observa-se uma diferença considerável para os atributos “target:colour” (-22,8% para *Feedback* e -3,5% para Monólogo em comparação a frequência obtida em Diálogo), “target:size” (-8,8% para *Feedback* e -1,8% para Monólogo em comparação a frequência obtida em Diálogo) , e “lm.outros” (-12,3% para *Feedback* e 0,0% para Monólogo em comparação a frequência obtida em Diálogo).

Uma possível explicação para estas diferenças estaria fora do escopo do presente estudo. Cabe destacar, entretanto, que a seleção de atributos é um tema central na pesquisa em GER, e que estudos deste tipo normalmente não consideram o modo de comunicação (em monólogo ou diálogo) como sendo uma variável relevante em seus experimentos. Por exemplo, o próprio uso do atributo cor, que apresentou grande diferença nas conduções do experimento no domínio GRE3D3, é o tema de estudos como (VIETHEN; GOUDBEEK; KRAHMER, 2012; TARENSKEEN; BROERSMA; GEURTS, 2014), e no entanto estudos deste tipo desconsideram a possibilidade (sugerida no presente levantamento) de que seus resultados, sejam eles obtidos em modo monólogo ou diálogo, possam ser limitados ao tipo de comunicação empregado.

8 Conclusões

Este trabalho apresentou uma pesquisa em nível de mestrado acadêmico na área de GLN enfocando o problema da experimentação em GER. O trabalho resultou em uma proposta de automatização de experimentos do tipo monólogo, chamada de modo *Feedback*.

Em um primeiro momento, um levantamento bibliográfico sobre experimentos de GER foi realizado e teve como objetivo identificar como os experimentos da área são feitos. Identificamos que os experimentos são comumente conduzidos com seres humanos nos modos monólogo e diálogo.

Diante das duas possibilidades de experimentação observadas, um estudo exploratório foi conduzido para elucidar hipóteses sobre a experimentação em GER. Descobrimos como resultado que expressões produzidas em experimentos em modo monólogo são mais breves, possuem um número médio menor de relações espaciais, e possuem um menor grau de especificação do ponto de referência (*landmark*) quando comparadas a expressões produzidas em modo diálogo. Após os resultados do estudo exploratório, identificou-se a oportunidade de construção de um modelo de validação automática de REs para a experimentação de GER que pudesse unir os benefícios do modo monólogo e do modo diálogo.

Este modelo foi avaliado de maneira intrínseca com os corpúscos TUNA-*People* e TUNA-*Furniture* (GATT; SLUIS; DEEMTER, 2007), GRE3D3 (VIETHEN; DALE, 2008), e GRE3D7 (VIETHEN; DALE, 2011). Os resultados da avaliação intrínseca demonstram que o modelo proposto apresenta resultados superiores ao modelo do baseline (ROCHA; LAN; PARABONI, 2017).

Finalmente, uma avaliação extrínseca foi realizada, e consistiu em três execuções de um mesmo experimento de coleta de descrições produzidas por sujeitos humanos. As três execuções foram denominadas Monólogo, *Feedback* e Diálogo. Os resultados da avaliação extrínseca corroboram para a hipótese de que descrições coletadas em modo *Feedback* são mais próximas das descrições obtidas em modo Diálogo do que as obtidas em modo Monólogo.

8.1 Contribuições

O trabalho realizado oferece as seguintes contribuições:

- Um levantamento bibliográfico de como é feita a experimentação na área de GER para a produção de expressões de referência com humanos em modo monólogo e diálogo.
- Uma investigação de como acontece a comunicação entre seres humanos e computadores em experimentos controlados para a produção de expressões de referência.
- Um modelo de validação automática de REs que pode ser acoplado a uma ferramenta de experimentação em GER, de maneira a validar de forma automática as descrições produzidas por participantes humanos.

8.2 Limitações

Uma limitação importante do presente trabalho está relacionada ao fato de sua validação ter sido feita apenas para os domínios TUNA-*People* e TUNA-*Furniture*, GRE3D3, GRE3D7, e b5-ref. Domínios mais complexos como o Zoom (ALTAMIRANO et al., 2015) não foram utilizados para testar o modelo de validação automática.

O trabalho proposto tratou exclusivamente de experimentos da área de GER. Outras subáreas da GLN não foram assim contempladas.

Referências¹

- ALTAMIRANO, R. et al. Zoom: a corpus of natural language descriptions of map locations. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: [s.n.], 2015. p. 69–75. Disponível em: <http://www.aclweb.org/anthology/P15-2012>. Citado 4 vezes nas páginas 18, 46, 58 e 75.
- ARTS, A. et al. Overspecification in written instructions. *Linguistics*, Tilburg University, v. 49, n. 3, p. 555–574, 2011. Citado na página 51.
- ARTS, A. et al. Overspecification facilitates object identification. *Journal of Pragmatics*, v. 43, p. 361–374, 2011. Citado na página 25.
- BELZ, A.; GATT, A. The attribute selection for GRE challenge: Overview and evaluation results. *Proceedings of UCNLG+ MT: Language Generation and Machine Translation*, Copenhagen, p. 75–83, 2007. Citado na página 39.
- BOHNET, B.; DALE, R. et al. Viewing referring expression generation as search. In: *IJCAI*. [S.l.: s.n.], 2005. p. 1004–1009. Citado na página 24.
- CLARKE, A. D.; ELSNER, M.; ROHDE, H. Where's Wally: the influence of visual salience on referring expression generation. *Frontiers in psychology*, Frontiers Media SA, v. 4, 2013. Citado 2 vezes nas páginas 18 e 43.
- DALE, R. Cooking up referring expressions. In: *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: [s.n.], 1989. (ACL '89), p. 68–75. Disponível em: <http://dx.doi.org/10.3115/981623.981632>. Citado na página 37.
- DALE, R.; HADDOCK, N. Generating referring expressions involving relations. In: *Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*. [S.l.: s.n.], 1991. p. 161–166. Citado 2 vezes nas páginas 24 e 37.
- DALE, R.; REITER, E. Computational interpretations of the Grice an maxims in the generation of referring expressions. *Cognitive science*, Wiley Online Library, v. 19, n. 2, p. 233–263, 1995. Citado 4 vezes nas páginas 18, 37, 39 e 63.
- DEEMTER, K. van; GOMPEL, A. G. R. van; KRAHMER, E. Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, v. 4, p. 166–183, 2012. Citado na página 19.
- DICE, L. R. Measures of the amount of ecologic association between species. *Ecology*, Wiley Online Library, v. 26, n. 3, p. 297–302, 1945. Citado 5 vezes nas páginas 21, 25, 55, 64 e 71.
- ENGELHARDT, P. E.; BAILEY, K. G.; FERREIRA, F. Do speakers and listeners observe the Grice an maxim of quantity? *Journal of Memory and Language*, Elsevier, v. 54, n. 4, p. 554–573, 2006. Citado na página 25.

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

- ENGELHARDT, P. E.; DEMIRAL, S. B.; FERREIRA, F. Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, v. 77, n. 2, p. 304–314, 2011. Citado na página 25.
- EUGENIO, B. D. et al. The agreement process: An empirical investigation of human–human computer-mediated collaborative dialogs. *International Journal of Human-Computer Studies*, Elsevier, v. 53, n. 6, p. 1017–1076, 2000. Citado 3 vezes nas páginas 19, 27 e 48.
- FERREIRA, T. C.; PARABONI, I. Generating natural language descriptions using speaker-dependent information. *Natural Language Engineering*, Cambridge University press, p. 1–22, 2017. Citado 2 vezes nas páginas 23 e 61.
- FITZGERALD, N.; ARTZI, Y.; ZETTLEMOYER, L. S. Learning distributions over logical forms for referring expression generation. In: *Conference on Empirical Methods on Natural Language Processing*. [S.l.: s.n.], 2013. p. 1914–1925. Citado 2 vezes nas páginas 18 e 42.
- FONSECA E. R., R. J. L. G. Mac-morpho revisited: Towards robust part-of-speech tagging. In: *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*. [S.l.: s.n.], 2013. p. 98–107. Citado na página 64.
- GARGETT, A. et al. The GIVE-2 corpus of giving instructions in virtual environments. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-2010)*. Valletta, Malta: [s.n.], 2010. p. 2401–2406. ACL Anthology Identifier: L10-1365. Disponível em: http://www.lrec-conf.org/proceedings/lrec2010/pdf/532_Paper.pdf. Citado 4 vezes nas páginas 19, 29, 30 e 31.
- GATT, A.; BELZ, A.; KOW, E. The TUNA-REG challenge 2008: Overview and evaluation results. In: *Proceedings of the Fifth International Natural Language Generation Conference*. [S.l.: s.n.], 2008. p. 198–206. Citado na página 39.
- GATT, A.; BELZ, A.; KOW, E. The TUNA-REG challenge 2009: Overview and evaluation results. In: *Proceedings of the 12th European Workshop on Natural Language Generation*. [S.l.: s.n.], 2009. p. 174–182. Citado na página 39.
- GATT, A. et al. Production of referring expressions: Preference trumps discrimination. In: *35th meeting of the cognitive science society*. [S.l.: s.n.], 2013. p. 483–488. Citado 2 vezes nas páginas 19 e 23.
- GATT, A.; SLUIS, I. V. D.; DEEMTER, K. V. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In: *Proceedings of the Eleventh European Workshop on Natural Language Generation*. [S.l.: s.n.], 2007. p. 49–56. Citado 10 vezes nas páginas 18, 25, 39, 40, 48, 60, 62, 64, 67 e 74.
- GOMPEL, R. van et al. PRO: A computational model of referential overspecification. In: *Proceedings of AMLAP-2012*. [S.l.: s.n.], 2012. Citado na página 19.
- GORNIK, P.; ROY, D. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, p. 429–470, 2004. Citado 2 vezes nas páginas 19 e 28.
- GRICE, H. P. Logic and conversation. In: COLE, P.; MORGAN, J. L. (Ed.). *Syntax and semantics*. [S.l.]: New York: Academic Press, 1975. v. 3. Citado na página 24.

KAZEMZADEH, S. et al. ReferItGame: Referring to objects in photographs of natural scenes. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.]: Association for Computational Linguistics, 2014. p. 787–798. Citado 2 vezes nas páginas 19 e 48.

KAZEMZADEH, S. et al. ReferItGame: Referring to objects in photographs of natural scenes. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: [s.n.], 2014. p. 787–798. Disponível em: <http://www.aclweb.org/anthology/D14-1086>. Citado 2 vezes nas páginas 33 e 34.

KOOLEN, R. et al. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, Elsevier, v. 43, p. 3231–3250, 2011. Citado na página 25.

KOOLEN, R.; GOUDBEEK, M.; KRAHMER, E. The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, v. 37, n. 2, p. 395–411, 2013. Citado na página 19.

KRAHMER, E.; DEEMTER, K. V. Computational generation of referring expressions: A survey. *Computational Linguistics*, MIT Press, v. 38, n. 1, p. 173–218, 2012. Citado 3 vezes nas páginas 17, 23 e 24.

KRAUSS, R. M.; WEINHEIMER, S. Concurrent feedback, confirmation and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, v. 4, p. 343–346, 1966. Citado na página 68.

MITCHELL, M.; DEEMTER, K. van; REITER, E. Natural reference to objects in a visual domain. In: *Proceedings of the 6th international natural language generation conference*. [S.l.: s.n.], 2010. p. 95–104. Citado 2 vezes nas páginas 19 e 32.

MUNIZ, M. C. M. *A construção de recursos linguístico-computacionais para o Português do Brasil: o projeto de Unitex-PB*. Dissertação (Mestrado) — ICMC / USP São Carlos, 2004. Citado na página 62.

PARABONI, I.; DEEMTER, K. van. Reference and the facilitation of search in spatial domains. *Language, Cognition and Neuroscience*, v. 29, n. 8, p. 1002–1017, 2014. Citado na página 44.

PARABONI, I.; GALINDO, M.; IACOVELLI, D. Stars2: a corpus of object descriptions in a visual domain. *Language Resources and Evaluation*, Springer, v. 51, n. 2, p. 439–462, 2017. Citado 9 vezes nas páginas 19, 34, 35, 36, 37, 51, 54, 55 e 58.

PARABONI, I.; GALINDO, M. R.; IACOVELLI, D. Generating overspecified referring expressions: the role of discrimination. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics ACL-2015*. Beijing, China: Association for Computational Linguistics, 2015. p. 76–82. Citado na página 24.

PARABONI, I. et al. Effects of cognitive effort on the resolution of overspecified descriptions. *Computational Linguistics*, v. 43, n. 2, p. 451–459, 2017. Citado 2 vezes nas páginas 19 e 25.

PARABONI, I.; MONTEIRO, D. S.; LAN, A. G. J. Personality-dependent referring expression generation. In: *Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415*. Prague, Czech Republic: Springer-Verlag, 2017. p. 20–28. Citado 3 vezes nas páginas 60, 62 e 70.

- PARABONI, I. et al. Generating underspecified descriptions of landmark objects. In: *Text, Speech and Dialogue*. [S.l.: s.n.], 2014. p. 76–83. Citado 3 vezes nas páginas 18, 44 e 45.
- PECHMANN, T. Incremental speech production and referential overspecification. *Linguistics*, v. 27, n. 1, p. 98–110, 1989. Citado na página 25.
- RAMOS, R. M. et al. O corpus b5 de textos e inventários de personalidade. relatório técnico PPgSI nr.0001/2017. In: . São Paulo, Brazil: [s.n.], 2017. Citado 2 vezes nas páginas 47 e 81.
- REITER, E.; DALE, R. *Building natural language generation systems*. [S.l.]: MIT Press, 2000. v. 33. Citado 2 vezes nas páginas 17 e 23.
- RIGHI, G.; PEISSIG, J. J.; TARR, M. J. Recognizing disguised faces. *Visual Cognition*, v. 20, n. 2, p. 143–169, 2012. Citado na página 48.
- ROCHA, D. da S.; LAN, A. G. J.; PARABONI, I. Anotação semiautomática de expressões de referência. relatório técnico PPgSI nr.0002/2017. In: . São Paulo, Brazil: [s.n.], 2017. Citado 2 vezes nas páginas 64 e 74.
- TARENSKEEN, S.; BROERSMA, M.; GEURTS, B. Referential overspecification: Colour is not that special. In: *RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production*. Edinburgh, UK: [s.n.], 2014. Citado na página 73.
- VIETHEN, J.; DALE, R. Algorithms for generating referring expressions: Do they do what people do? In: *Proceedings of INLG-2006*. [S.l.: s.n.], 2006. p. 63–70. Citado 3 vezes nas páginas 18, 37 e 38.
- VIETHEN, J.; DALE, R. The use of spatial relations in referring expression generation. In: *Proceedings of the Fifth International Natural Language Generation Conference*. Stroudsburg, PA, USA: [s.n.], 2008. (INLG '08), p. 59–67. Disponível em: <http://dl.acm.org/citation.cfm?id=1708322.1708334>. Citado 10 vezes nas páginas 18, 40, 41, 60, 62, 64, 67, 70, 74 e 81.
- VIETHEN, J.; DALE, R. GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In: *Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop*. Stroudsburg, PA, USA: [s.n.], 2011. (UCNLG+EVAL '11), p. 12–22. ISBN 978-1-937284-18-3. Disponível em: <http://dl.acm.org/citation.cfm?id=2187741.2187744>. Citado 5 vezes nas páginas 18, 40, 64, 67 e 74.
- VIETHEN, J.; GOUDBEEK, M.; KRAHMER, E. The impact of colour difference and colour codability on reference production. In: *Proceedings of CogSci-2012*. [S.l.: s.n.], 2012. p. 1084–1098. Citado na página 73.
- WHITE, M.; RAJKUMAR, R.; MARTINS, S. Towards broad coverage surface realization with CCG. In: *UCLG+MT: Language Generation and Machine Translation*. Copenhagen, Denmark: [s.n.], 2007. p. 267–276. Citado na página 18.

Apêndice A – Instâncias do experimento Stars2 utilizadas no estudo exploratório

Os identificadores da coluna “*Subject Stars2*” do Quadro 3 refere-se aos identificadores originais do corpus ¹

Quadro 3 – Relação entre o estudo exploratório e o experimento Stars2

Estudo Exploratório	Subject Stars2
Experimento 1	52
Experimento 2	89
Experimento 3	105
Experimento 4	115
Experimento 5	136
Experimento 6	307
Experimento 7	439
Experimento 8	455
Experimento 9	503
Experimento 10	538
Experimento 11	585
Experimento 12	597
Experimento 13	621
Experimento 14	704
Experimento 15	788
Experimento 16	823
Experimento 17	832
Experimento 18	858
Experimento 19	895
Experimento 20	898
Experimento 21	969
Experimento 22	972
Experimento 23	978
Experimento 24	998

Fonte: Danillo da Silva Rocha, 2017

¹ Mais informações disponíveis em: <http://ivandreparaboni.wixsite.com/research/visconde>

Apêndice B – Imagens dos corpos b5-ref e GRE3D3 utilizadas na avaliação extrínseca

Os identificadores da coluna “Imagem b5-ref” do Quadro 4 referem-se aos identificadores originais das imagens do corpo b5-ref¹

Quadro 4 – Relação entre a avaliação extrínseca e as imagens do corpo b5-ref

Estímulo avaliação extrínseca	Imagem b5-ref
Estímulo 1	01
Estímulo 3	02
Estímulo 5	03
Estímulo 7	04
Estímulo 9	05
Estímulo 11	06
Estímulo 13	07
Estímulo 15	08
Estímulo 17	09
Estímulo 19	10
Estímulo 21	11

Fonte: Danilo da Silva Rocha, 2017

Os identificadores da coluna “Imagem GRE3D3” do Quadro 5 referem-se aos identificadores originais das imagens do corpo GRE3D3²

Quadro 5 – Relação entre a avaliação extrínseca e as imagens do corpo GRE3D3

Estímulo avaliação extrínseca	Imagem GRE3D3
Estímulo 2	03
Estímulo 4	04
Estímulo 6	05
Estímulo 8	06
Estímulo 10	07
Estímulo 12	08
Estímulo 14	09
Estímulo 16	10
Estímulo 18	11
Estímulo 20	12
Estímulo 22	13

Fonte: Danilo da Silva Rocha, 2017

¹ Mais informações disponíveis em Ramos et al. (2017)

² Mais informações disponíveis em Viethen e Dale (2008)