UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES

**EM BUSCA DO PLAUSÍVEL: DESAFIOS E
CONTRIBUIÇÕES DA GERAÇÃO
AUTOMÁTICA DE LÍNGUA NATURAL**

IVANDRÉ PARABONI

São Paulo
Setembro de 2017

UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES

**EM BUSCA DO PLAUSÍVEL: DESAFIOS E
CONTRIBUIÇÕES DA GERAÇÃO
AUTOMÁTICA DE LÍNGUA NATURAL**

IVANDRÉ PARABONI

Texto sistematizado apresentado à Escola de Artes,
Ciências e Humanidades da Universidade de São
Paulo, como parte dos requisitos para a obtenção
do título de Livre Docente na área de concentração
Computação - Especialidade: Estruturas de Dados.

São Paulo
Setembro de 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)
CRB 8 - 4936

Paraboni, Ivandré

Em busca do plausível: desafios e contribuições da geração automática de língua natural / Ivandré Paraboni. – 2017
145 f. : il

Texto sistematizado (Livre docência em Computação) -
Escola de Artes, Ciências e Humanidades, Universidade de
São Paulo, em 2018.

Especialidade : Estruturas de dados

1. Processamento de linguagem natural. 2. Linguagem
natural. 3. Inteligência artificial. I. Título.

CDD 22.ed.- 006.35

Texto sistematizado apresentado à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Livre Docente na área de concentração Computação - Especialidade: Estruturas de Dados, aprovada em 16 de março de 2018 pela comissão julgadora constituída pelos doutores:

Profa. Dra. Sarajane Marques Peres

Universidade de São Paulo

Presidente

Prof. Dr. Luciano Antonio Digiampietri

Universidade de São Paulo

Prof. Dr. Hélio Pedrini

Universidade Estadual de Campinas

Prof. Dr. Gustavo Enrique de Almeida Prado Alves Batista

Universidade de São Paulo

Prof. Dr. João Luís Garcia Rosa

Universidade de São Paulo

Aos meus pais.

AGRADECIMENTOS

- A todos alunos e alunas - do passado, presente e futuro - pelo esforço e dedicação.
- Aos colegas do Bacharelado em Sistemas de Informação e do PPgSI, pela constante aprendizagem.
- Aos colegas de Aberdeen e Córdoba, pelas oportunidades de colaboração.
- Às agências FAPESP, CNPq e CAPES pelo apoio financeiro ao conjunto de trabalhos aqui apresentados.

RESUMO

Sistemas de Geração de Língua Natural (GLN) – que tipicamente produzem descrições textuais a partir de uma entrada de dados não linguística – são empregados quando o uso de texto predefinido não é suficiente, ou seja, quando é necessária uma maior variação linguística ou estilística nos documentos gerados e/ou maior proximidade em relação ao desempenho humano. Esta monografia discute a geração de língua natural dita *plausível* sob duas perspectivas - aqui denominadas a perspectiva do locutor e a do ouvinte - através de uma série de exemplos de estudos que trataram de problemas computacionais como o de seleção de conteúdo, realização superficial e outras, e destacando seus desafios e suas contribuições científicas. Inicialmente, a discussão é baseada em exemplos de estudos que seguem uma definição mais convencional (ou talvez mais intuitiva) de plausibilidade baseada na perspectiva do locutor. A seguir, é destacado que estudos deste tipo tendem a apresentar necessidade constante de novos conjuntos de dados para que novas questões de pesquisa possam ser investigadas. Como alternativa a esta abordagem, são então apresentados exemplos de estudos que consideram a perspectiva do ouvinte, os quais apresentam grande potencial de reuso e, em certos casos, contribuições de outra forma difíceis de obter com base puramente em estudos baseados na perspectiva do locutor. Finalmente, a discussão sobre a plausibilidade na tarefa computacional de GLN é ampliada de modo a considerar a questão da variação humana na produção da língua, apresentando exemplos de estudos que consideram tanto modelos de locutores individuais como de grupos de indivíduos e, de modo especial, baseados em personalidade humana. Como resultado deste trabalho, espera-se que novos pesquisadores e estudantes (e.g., de Ciência da Computação e áreas relacionadas) sintam-se motivados a abordar o problema computacional de GLN de modo plausível - seja com base na perspectiva do ouvinte, do locutor, ou perfis destes - e a desenvolver suas próprias metodologias, auxiliando assim no crescimento da linha de pesquisa em GLN de modo geral, e de forma mais específica para o caso da geração de texto em português do Brasil.

Palavras-chave: Geração de Língua Natural, Seleção de Conteúdo, Realização Superficial

ABSTRACT

Natural Language Generation (NLG) systems - which typically generate text descriptions from non-linguistic input data - are employed when canned text is not sufficient, that is, when greater linguistic or stylistic variation are expected in the output text, or when the application requires closer proximity to human performance. This work addresses the computational problem of plausible NLG based on two perspectives - hereby called the speaker's and the hearer's perspectives - by presenting a number of examples of studies on content selection, surface realisation and other NLG tasks, and by focusing on their research challenges and contributions. The initial discussion is based on examples of studies that follow a rather traditional (or perhaps more intuitive) view on plausibility based on the speaker's perspective. Next, it is argued that in speaker-oriented studies of this kind further research questions will generally require collecting new data. As an alternative to this view, examples of hearer-oriented studies are discussed, which may turn out to be highly reusable and, in some cases, provide contributions that are otherwise difficult to obtain by means of speaker-oriented studies. Finally, the speaker and hearer perspectives are generalised so as to cover the issue of human variation in language production by discussing examples of studies that taken into account models of individual speakers or groups and, in particular, using models of human personality. As a result, this text is expected to help motivate new researchers and students (e.g., in Computer Science and related fields) to address the computational problem of plausible natural language generation - either from a speaker's or hearer's perspective - and to develop their own methodologies, hence contributing to the expansion of the NLG research field in general, and particularly so for the case of the Brazilian Portuguese text generation.

Keywords: Natural Language Generation, Content Selection, Surface Realisation

LISTA DE FIGURAS

1	A arquitetura <i>pipeline</i> de GLN.	2
2	Um contexto visual contendo quatro entidades referenciáveis <i>e1..e4</i>	14
3	Exemplo de estímulos visuais do córpus TUNA <i>Furniture</i> e <i>People</i>	25
4	Exemplo de estímulo utilizado no córpus GRE3D7.	26
5	Um exemplo de contexto visual do sistema GIVE.	26
6	Exemplo de estímulo utilizado no córpus ReferItGame.	27
7	Exemplo de especificação de entrada.	29
8	Exemplo com três mesas brancas e grandes, e uma mesa preta e pequena. . . .	43
9	Arquitetura do modelo de realização superficial.	48
10	Um mundo virtual do tipo GIVE contendo quatro botões interativos (da esquerda para direita: azul, verde, vermelho e verde).	53
11	Arquitetura do modelo de geração de relações espaciais.	56
12	Situação de referência a um objeto-alvo (estrela) via <i>landmark</i> (caixa) no córpus Stars (rótulos adicionados para fins ilustrativos).	65
13	Duas situações de referência ao alvo <i>e2</i> via <i>landmark q6</i> (rótulos adicionados para fins ilustrativos).	66
14	Fragmento de representação XML do córpus Stars.	68
15	Exemplo de contexto referencial no córpus Stars2.	71
16	Situações de referência com uso de descrições relacionais (rótulos adicionados para fins ilustrativos).	72
17	Telas do locutor (esquerda) e ouvinte (direita).	75

18	Interface do experimento Zoom.	78
19	Interface do experimento, incluindo quatro pontos de referência de alto grau de saliência (plantas e cadeiras) e objetos do tipo botão de baixa saliência (ocultos).	85
20	Mapa esquemático representando o contexto da figura 19, com quatro botões ocultos (três azuis e um verde). O ouvinte está posicionado em [start] e o alvo está localizado em [target].	86
21	Exemplo de contexto com estrelas de 7/8 pontas contendo letras ou números (objeto-alvo destacado para fins ilustrativos).	91
22	Exemplo de contexto baseado em letras e números (objeto-alvo destacado para fins ilustrativos).	93
23	Componentes do córpus <i>b5</i>	104
24	Exemplo de estímulo usado na coleta dos córpus <i>b5-text</i> e <i>b5-caption</i>	105
25	Exemplos de estímulo usado na coleta do córpus <i>b5-ref</i>	106
26	Um contexto de referência do córpus <i>b5-ref</i>	110

LISTA DE TABELAS

1	Projetos financiados.	7
2	Artigos em periódicos.	8
3	Artigos em conferências e workshops.	10
4	Orientações de Iniciação Científica.	11
5	Orientações de Mestrado.	12
6	Dimensões de avaliação consideradas.	59
7	Córpus de GER.	81

Plausível: Razoável, aceitável, admissível.

Dicionário Aurélio

SUMÁRIO

CAPÍTULO 1 – INTRODUÇÃO	1
1.1 Visão geral	1
1.2 Objetivos	5
1.3 Justificativas	5
1.4 Trabalhos considerados	5
1.4.1 Projetos financiados	6
1.4.2 Publicações	6
1.4.3 Orientações	8
1.5 Organização da monografia	8
 CAPÍTULO 2 – CONCEITOS FUNDAMENTAIS	 13
2.1 Geração de Expressões de Referência	14
2.1.1 A tarefa computacional de GER	15
2.1.2 Algoritmos de GER	18
2.1.3 Discussão	21
2.2 Córpus para GER/GLN	23
2.2.1 Visão geral	23
2.2.2 Córpus de expressões de referência	24
2.2.3 Discussão	28
2.3 Realização superficial	29

2.3.1	Especificação de entrada	30
2.3.2	Formalismo gramatical	30
2.3.3	Paradigma computacional	31
2.3.4	Discussão	32
2.4	Avaliação de sistemas de GLN	33
2.4.1	Avaliação baseada na efetividade da tarefa	33
2.4.2	Avaliação por usuários	35
2.4.3	Avaliação baseada em córpus	35
2.4.4	Discussão	36
2.5	Língua natural e personalidade	36
2.5.1	O modelo dos Cinco Grandes Fatores (CGF)	37
2.5.2	O tratamento computacional da personalidade	38
2.5.2.1	O reconhecimento de personalidade a partir de texto	38
2.5.2.2	Geração de texto com base em traços de personalidade	39
2.5.3	Discussão	40
2.6	Considerações finais	40

CAPÍTULO 3 – A PERSPECTIVA DO LOCUTOR **41**

3.1	Visão geral	41
3.2	Experiências de GER com a série de desafios TUNA	42
3.2.1	Primeiras iniciativas: o desafio TUNA 2008	43
3.2.2	Refinamentos para o desafio TUNA 2009	44
3.2.3	Contribuições	45
3.3	Realização superficial em português	46
3.3.1	O uso de modelos fatorados para linearização sentencial	46
3.3.2	Arquitetura	47
3.3.3	Especificação de entrada	48

3.3.4	Modelos propostos	51
3.3.5	Avaliação	51
3.3.6	Contribuições	52
3.4	Geração de relações espaciais em mundos virtuais	53
3.4.1	O problema da geração de relações espaciais	54
3.4.2	Modelos propostos	55
3.4.3	Avaliação	58
3.4.4	Contribuições	60
3.5	Considerações finais	60
CAPÍTULO 4 – A NECESSIDADE DE NOVOS CONJUNTOS DE DADOS		63
4.1	Visão geral	63
4.2	O cópup Stars	65
4.2.1	Materiais e método	67
4.2.2	Resultados	68
4.2.3	Contribuições	69
4.3	O cópup Stars2	69
4.3.1	Materiais e método	74
4.3.2	Resultados	76
4.3.3	Contribuições	77
4.4	O cópup Zoom	77
4.4.1	Materiais e método	78
4.4.2	Resultados	79
4.4.3	Contribuições	80
4.5	Considerações finais	80
CAPÍTULO 5 – A PERSPECTIVA DO OUVINTE		83

5.1	Visão geral	83
5.2	Efeitos da estratégia referencial e memória do ouvinte sobre a busca em domínios espaciais	84
5.2.1	Materiais e método	88
5.2.2	Resultados	88
5.2.3	Contribuições	89
5.3	Esforço cognitivo e superespecificação	89
5.3.1	Experimento 1: Propriedades relacionais	91
5.3.2	Experimento 2: Propriedades atômicas	93
5.3.3	Contribuições	94
5.4	Considerações finais	94
CAPÍTULO 6 – PLAUSÍVEL PARA QUEM?		97
6.1	Visão geral	97
6.2	Modelos de locutores individualizados	98
6.2.1	Modelos propostos	99
6.2.2	Resultados	101
6.2.3	Contribuições	101
6.3	Modelos baseados em personalidade humana	102
6.3.1	O <i>cópus b5</i> de textos e inventários de personalidade	103
6.3.1.1	Visão geral	104
6.3.1.2	Reconhecimento de personalidade no <i>cópus b5</i>	107
6.3.2	GER com base em traços de personalidade	109
6.4	Considerações finais	112
CAPÍTULO 7 – CONCLUSÕES		113
REFERÊNCIAS		116

Capítulo 1

INTRODUÇÃO

1.1 Visão geral

O simples objetivo de comunicar um determinado fato em língua natural de modo que seja compreensível para uma audiência humana suscita diversas questões de interesse para a Ciência da Computação. Como os seres humanos se expressam em língua natural? Por que se expressam de uma determinada maneira? Por que pessoas diferentes se expressam de forma diferente, mesmo quando falando do mesmo assunto? Questões deste tipo são a motivação fundamental dos estudos relatados ao longo desta monografia, que apresentam propostas de modelos computacionais de Geração de Língua Natural (GLN) (GATT; KRAHMER, 2017) sob diferentes pontos de vista.

A visualização de conteúdos de bases de dados complexos e/ou em grandes volumes frequentemente é feita com uso de documentos de textos em linguagem natural. Sistemas de GLN – que tipicamente produzem descrições textuais a partir de uma entrada de dados não linguística – são assim empregados quando o uso de texto predefinido não é suficiente, ou seja, quando é necessária uma maior variação linguística ou estilística nos documentos gerados e/ou maior proximidade em relação ao desempenho humano.

As primeiras aplicações de GLN de maior visibilidade foram possivelmente os sistemas de geração de relatórios descritivos do mercado de ações a partir de indicadores financeiros (REITER; DALE, 2000) e os boletins de previsão do tempo gerados a partir de dados de satélites (BELZ, 2008; SRIPADA et al., 2014). Em anos recentes, entretanto, o rol de aplicações de GLN observou grande expansão, especialmente na área médica. Sistemas deste tipo incluem a geração de diagnósticos clínicos produzidos a partir da leitura de sensores de equipamentos hospitalares (PORTET et al., 2009), de históricos de tratamento (EUGENIO et al., 2014; JORDAN et al.,

2014), de relatórios para comunicação entre plantonistas (HUNTER et al., 2012), para suporte a emergências médicas em zonas rurais (SCHNEIDER et al., 2015) e campanhas antitabagismo (REITER; ROBERTSON; OSMAN, 2003). Sistemas de GLN têm sido empregados também na geração de diagnósticos de equipamentos de alta tecnologia como turbinas de gás (REITER et al., 2005), na geração de diálogos realistas em sistemas de narrativa automática (WALKER et al., 2011a, 2011b), em sistemas tutores inteligentes e aplicações educacionais em geral (EUGENIO; GLASS; TROLIO, 2002; WILLIAMS; REITER, 2008), em ferramentas de apoio à comunicação de pessoas com necessidades especiais (TINTAREV et al., 2016), na geração automática de poesia (ZHANG; LAPATA, 2014) e legendas automáticas para imagens (KARPATHY; FEI-FEI, 2015; XU et al., 2015), dentre muitos outros.

Embora consideravelmente mais recente do que estudos de interpretação da linguagem - e talvez por este motivo ainda pouco explorada pela comunidade científica brasileira - GLN é uma ativa linha de pesquisa dentro da grande área de Processamento de Língua Natural (PLN). O desenvolvimento de sistemas deste tipo é um empreendimento interdisciplinar de grande escala, frequentemente envolvendo conceitos de diversas áreas do conhecimento, tais como Inteligência Artificial, Linguística, Psicologia e outros.

De forma simplificada, a arquitetura de um sistema de GLN pode ser vista como um *pipeline* de três estágios (REITER; DALE, 2000) como na figura 1.

Figura 1: A arquitetura *pipeline* de GLN.

Fonte: Adaptado de Reiter e Dale (2000).

Tendo como ponto de partida um objetivo de comunicação de alto nível (no exemplo, o objetivo de descrever um objeto do domínio identificado como 'joao'), um sistema de GLN constrói gradativamente um plano para representar este objetivo em forma textual (no exemplo, uma sentença em português). Ao longo do processo de geração, o plano inicial sofre uma série de transformações, cada qual constituindo uma forma de representação intermediária do conhecimento comunicado, até o ponto em que o texto em língua natural é obtido.

A etapa de planejamento do documento ou macroplanejamento estabelece qual o conteúdo e a estrutura a serem representados no documento final com base nos dados de entrada. A entrada desta etapa é a própria entrada do sistema, ou seja, dados provenientes do domínio da aplicação. O sistema então determina o conteúdo que se deseja comunicar na forma de coleções de mensagens representando o significado pretendido, e.g., na forma de proposições lógicas organizadas em uma estrutura discursiva como a *Rhetorical Structure Theory* (RST) (MANN; THOMPSON, 1988).

A etapa seguinte, planejamento de sentenças (ou microplanejamento) estabelece a organização das mensagens na forma sentencial, ou seja, determina quais construções sintáticas, escolhas lexicais, expressões de referência etc. serão utilizadas para comunicar o conteúdo selecionado. A entrada desta etapa é um plano do documento (e.g., um conjunto de relações RST e mensagens de conteúdo), e a saída é uma especificação de como o texto será realizado de forma tão independente quanto possível da língua-alvo.

Finalmente, a terceira etapa da geração, a realização superficial, recebe como entrada a especificação textual fornecida pelo microplanejamento e utiliza regras gramaticais ou formalismo equivalente para estabelecer mapeamentos entre a especificação abstrata de uma sentença e o texto correspondente em língua natural. Este texto pode então ser acrescido de atributos visuais (cores, formatação etc.) ou convertido em áudio.

Dentre os múltiplos requisitos que um texto gerado automaticamente precisa atender, o mais fundamental é que seja fiel ao plano inicial do que deve ser comunicado, e que seja correto do ponto de vista gramatical. Estas condições preliminares podem no entanto ser asseguradas com relativa robustez pelas diversas etapas da arquitetura de GLN (por exemplo, com o uso de gramáticas na fase de realização superficial), e não são necessariamente representativas do verdadeiro desafio da pesquisa na área: para muitas (ou talvez todas) aplicações práticas, não é suficiente apenas gerar um texto correto. É preciso que o texto seja natural, e que se assemelhe tanto quanto possível ao que seria observado em uma situação de comunicação envolvendo participantes humanos naquelas mesmas circunstâncias. Sistemas que produzem texto com este tipo de naturalidade são ditos *plausíveis*, e a plausibilidade da língua gerada automaticamente tende assim a ser um dos principais objetos de estudo nesta área.

Gerar língua natural de forma que seja plausível não é entretanto apenas uma questão estética ou uma forma de evitar construções artificiais ou repetitivas. Usuários de sistemas computacionais tendem a tratar computadores e outros meios como sendo pessoas reais (REEVES; NASS, 1996), e alimentam expectativas de que sistemas computacionais se comportem de forma humanamente plausível. Sistemas que se comunicam de forma ‘robótica’ correm o risco de não

despertar engajamento por parte do usuário, e podem assim ser subutilizados.

A importância da plausibilidade para a GLN pode também ser ilustrada traçando-se um paralelo entre os objetivos perseguidos na pesquisa da área e os normalmente considerados na pesquisa da grande área de PLN. Enquanto a pesquisa em PLN se ocupa do problema de extrair o significado correto de uma determinada forma linguística, a pesquisa em GLN pode ser vista como a tarefa de escolher uma forma linguística adequada - ou plausível - para expressar um determinado conteúdo *dentre múltiplas alternativas igualmente corretas*. Assim, podemos dizer que em aplicações práticas de GLN não há a noção de texto ‘ideal’, mas apenas variações de maior ou menor plausibilidade.

Decidir o que constitui uma estratégia plausível de geração de língua natural é assim a chave para desenvolver bons sistemas deste tipo, e constitui a própria medida de qualidade destes sistemas. Neste sentido, podemos em princípio imaginar o problema computacional de GLN como uma tarefa de reproduzir o comportamento (linguístico) de locutores humanos da forma mais fiel possível. No entanto, uma vez que a comunicação é um processo colaborativo (CLARK; WILKES-GIBBS, 1986) entre locutor e ouvinte¹, o conceito de plausibilidade não pode ser considerado apenas sob a perspectiva do locutor: estudos da teoria da mente como em Keysar, Lin e Barr (2003) sugerem a existência de um forte efeito *egocêntrico* na produção humana da língua, isso é, locutores apresentam pouca habilidade em distinguir seu próprio conhecimento do conhecimento detido pelo ouvinte, e tendem assim a se expressar sem considerar todas as possíveis dificuldades de interpretação. Em outras palavras, a língua produzida por um locutor humano não necessariamente é a ideal do ponto de vista do ouvinte.

Como base nestas observações, podemos afirmar que há no mínimo um certo fator de risco em modelar um sistema de GLN apenas com base no comportamento do locutor, pois a perspectiva do ouvinte pode ser tão ou mais importante para que a comunicação seja efetiva. Um sistema de GLN pode assim ser desenvolvido e avaliado tanto pela sua habilidade em espelhar o comportamento de locutores humanos, como pela sua capacidade de suscitar algum tipo de reação (e.g., compreensão, engajamento etc.) por parte dos ouvintes.

Ambas perspectivas - do locutor e ouvinte - apresentam desafios e possibilidades de contribuição científica distintas, e são sob muitos aspectos conflitantes. Conhecer estas alternativas é um requisito fundamental para o projeto bem-sucedido de sistemas e artefatos de GLN, e por este motivo a questão da plausibilidade em GLN sob as perspectivas do locutor e do ouvinte será o tema central desta monografia.

¹ Ao longo desta monografia usamos os termos ‘locutor’ e ‘ouvinte’ no sentido geral de emissor e receptor de língua falada ou escrita, embora os estudos aqui relatados tratem exclusivamente da língua escrita.

1.2 Objetivos

O objetivo deste trabalho é apresentar exemplos de estudos de Geração de Língua Natural escrita (i.e., textual) desenvolvidos sob a perspectiva do locutor e do ouvinte, discutindo o conceito de plausibilidade adotado em cada caso, seus desafios e suas contribuições científicas.

A presente monografia procura também atingir os seguintes objetivos secundários:

- Apresentar uma visão geral da área de GLN e, de forma mais detalhada, do problema computacional da seleção de conteúdo.
- Sistematizar e apresentar de forma crítica a produção científica do autor e seus colaboradores na área de GLN decorrente de estudos posteriores à defesa de doutorado deste.
- Ilustrar as perspectivas do locutor e do ouvinte na pesquisa em GLN através de exemplos práticos de desenvolvimento e avaliação de modelos e artefatos de GLN.
- Discutir iniciativas de pesquisa em GLN específicas para o português brasileiro desenvolvidas pelo autor e seus colaboradores.
- Discutir possíveis generalizações das perspectivas do locutor e ouvinte aqui apresentadas.

1.3 Justificativas

Apesar de agregar um conjunto de técnicas sofisticado e com uma ampla gama de aplicações práticas, e pesquisa em GLN ainda é relativamente pouco explorada pela comunidade científica brasileira. Esta carência, que seria motivo de atenção em qualquer área do conhecimento, torna-se especialmente relevante em se tratando de tecnologias da língua humana que, em maior ou menor grau, podem ser dependentes do idioma, e que exigem assim contribuição de pesquisadores locais. Por este motivo, espera-se que esta monografia possa atrair a atenção de novos pesquisadores e estudantes brasileiros da Ciência da Computação e áreas correlatas, e a eventualmente motivá-los a produzir suas próprias contribuições, sejam elas de caráter geral ou específicas para o português brasileiro.

1.4 Trabalhos considerados

Dada a interdisciplinaridade inerente ao problema de GLN, os exemplos de estudos discutidos ao longo desta monografia apresentam contribuições de natureza variada, enfocando tanto

aspectos computacionais como cognitivos deste tipo de pesquisa. De um lado, observa-se estudos de natureza mais computacional e aplicada, nos quais o objetivo é o desenvolvimento de um sistema ou artefato específico para GLN. Em outro extremo, observa-se estudos de natureza cognitiva e com maior apelo teórico, nos quais o objetivo é um melhor entendimento de aspectos específicos do fenômeno de produção de língua natural por humanos, sem necessariamente considerar uma aplicação computacional imediata deste conhecimento. É importante observar entretanto que não há um rigor absoluto nesta divisão - seja entre pesquisa aplicada ou teórica, ou entre Ciência da Computação e Psicolinguística - e que muitos dos trabalhos aqui considerados podem ser classificados de mais de uma forma.

Os estudos de GLN exemplificados ao longo da presente monografia são decorrentes de uma série de atividades de pesquisa desenvolvidas no período de 2006 até o presente, em colaboração com alunos de graduação e pós-graduação da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (USP-EACH), e colaboradores internos e externos à instituição. Estas atividades incluem projetos de pesquisa financiada sob responsabilidade deste autor, publicações em periódicos e eventos científicos, e também orientações de mestrado, iniciação científica e trabalhos de conclusão de graduação, dentre outros. Esta monografia não considera nenhuma publicação decorrente do projeto de doutorado (concluído no ano de 2003, embora publicações desta natureza tenham sido produzidas de forma esporádica até 2007).

Uma vez que toda a produção científica aqui relatada se encontra publicada ou aceita para publicação em sua forma definitiva, a descrição destes estudos é necessariamente sucinta, com mais ênfase nos desafios e contribuições científicas de cada trabalho, e com menor atenção aos aspectos técnicos ou científicos de cada um. Estes detalhes podem no entanto ser consultados nas respectivas publicações originais, conforme citado ao longo do texto da monografia.

1.4.1 Projetos financiados

O conjunto de trabalhos considerados nesta monografia foi produzido em grande parte no contexto de seis projetos de pesquisa financiada sob responsabilidade deste autor. Estes projetos são relacionados na tabela 1 com a indicação das seções da monografia onde serão mencionados.

1.4.2 Publicações

Esta monografia faz referência a um conjunto de artigos publicados em periódicos de diferentes portes - das áreas de Computação, Linguística Computacional e/ou Ciências Cognitivas - no período de 2010 a 2017. Estes estudos são relacionados na tabela 2 com indicação da seção

Tabela 1: Projetos financiados.

Período	Descrição	Seção
Nov/06 - Out/08	<i>Construção de um Córpus Paralelo com Informações de Referência para Tradução Automática.</i> Pesquisa regular FAPESP nro. 2006/03941-7.	3.3
Jan/08 - Dez/09	<i>Resolução e Interpretação de Expressões de Referência na Geração Automática de Textos.</i> Pesquisa regular CNPq nro. 484015/2007 9 (Edital Universal 15/2007).	3.2
Set/09 - Ago/11	<i>Estudo e Desenvolvimento de Modelos Híbridos de Geração de Linguagem Natural.</i> Pesquisa regular FAPESP nro. 2009/08499-9.	3.3
Ago/11 - Fev/12	<i>Extensão de Algoritmos de Geração de Expressões de Referência para Domínios Físicos.</i> Pós-doutorado CAPES nro. 6415/10-5 (University of Aberdeen, Escócia).	5.2
Jun/14 - Mai/16	<i>Geração de Língua Natural para Descrição de Objetos em Contextos Visuais.</i> Pesquisa regular FAPESP nro. 2013/23169-0.	4.2, 4.3, 4.4, 5.3, 6.2
Mar/17 - Fev/19	<i>Tratamento Computacional da Personalidade Humana para Aplicações de Processamento de Língua Natural.</i> Pesquisa regular FAPESP nro. 2016/14223-0.	6.3

Fonte: Ivandré Paraboni, 2017.

da monografia onde serão posteriormente citados.

Além de publicações em periódico, a monografia também faz referência a uma série de artigos completos e resumidos que foram publicados em conferências e *workshops* da área de Computação no período de 2007 a 2017. Estes artigos são relacionados na tabela 3, também com indicação da seção do texto onde serão posteriormente citados.

Finalmente, observamos que a produção científica no período coberto por esta monografia inclui ainda pesquisa em quatro áreas fora do escopo deste trabalho, a saber: resolução anafórica (CUEVAS et al., 2008; CUEVAS; PARABONI, 2008a, 2008b; LUCENA; PARABONI, 2009a, 2010), tradução automática (PEREIRA; PARABONI, 2007b; AZIZ; PARDO; PARABONI, 2007, 2008a, 2008b, 2009a, 2009b), planejamento de documentos (NOVAIS et al., 2009; OLIVEIRA et al., 2009; ARAUJO et al., 2010) e recomendação de conteúdo textual (CASIMIRO; PARABONI, 2014). Embora não discutidas na presente monografia, estas publicações também são fruto de pesquisas e orientações desenvolvidas em período posterior ao doutoramento, e de forma independente deste. Os estudos ligados à resolução anafórica e tradução automática foram desenvolvidos no contexto de um projeto de apoio FAPESP no período 2006-2008.

Tabela 2: Artigos em periódicos.

Referência	Veículo	Título	Seção
Lucena, Paraboni e Pereira (2010)	Inteligencia Artificial	<i>From semantic properties to surface text: The generation of domain object descriptions</i>	3.2
Novais e Paraboni (2012)	JBCS	<i>Portuguese text generation using factored language models</i>	3.3
Silva Junior, Paraboni e Novais (2013)	RITA	<i>Um Sistema de Realização Superficial para Geração de Texto em Português</i>	3.3
Silva e Paraboni (2014)	Linguamática	<i>Geração de Expressões de Referência em Ambientes Virtuais Interativos</i>	3.4
Paraboni e Deemter (2014)	Lang. Cog. and Neuro.	<i>Reference and the Facilitation of Search in Spatial Domains</i>	5.2
Silva e Paraboni (2015)	Spatial Cog. & Comp.	<i>Generating Spatial Referring Expressions in Interactive 3D Worlds</i>	3.4
Paraboni, Galindo e Iacovelli (2017)	Language Res. and Eval.	<i>Stars2: a corpus of object descriptions in a visual domain</i>	4.2
Paraboni et al. (2017)	Computational Linguistics	<i>Effects of cognitive effort on the resolution of overspecified descriptions</i>	5.3
Ferreira e Paraboni (2017a)	Natural Language Engin.	<i>Generating Natural Language Descriptions using Speaker-dependent Information</i>	6.2

Fonte: Ivandré Paraboni, 2017.

1.4.3 Orientações

Em sua maior parte, as publicações relacionadas na seção anterior são fruto de pesquisas realizadas em colaboração com orientandos de iniciação científica e mestrado desta instituição. Estes projetos são relacionados na tabela 4 e 5.

1.5 Organização da monografia

Esta monografia apresenta uma sistematização de diversas publicações da área de GLN desenvolvidas por este autor e seus colaboradores. De modo geral, o foco da discussão é concentrado nos desafios e contribuições dos diversos estudos apresentados, e na perspectiva (locutor versus ouvinte) adotada em cada caso. A exceção é a seção 6.3, que apresenta os primeiros resultados de um projeto de pesquisa recentemente iniciado (março / 2017) e que trata da generalização de certos conceitos discutidos nos capítulos anteriores tanto para fins de geração como de interpretação de língua natural.

O capítulo 2 apresenta conceitos fundamentais da área de Geração de Língua Natural (GLN) e temas correlatos e, de forma mais específica, do problema computacional de seleção

de conteúdo para Geração de Expressões de Referência (GER) que é retomado em diversos capítulos subsequentes. Além deste, é abordada de forma mais breve a tarefa computacional de realização superficial, a questão da avaliação de sistemas de GLN e de modelos de personalidade humana.

O conteúdo principal da monografia é distribuído em quatro capítulos. O capítulo 3 aborda o desenvolvimento de sistemas e artefatos de GLN sob a perspectiva do locutor (ou produtor da mensagem em língua natural). Como forma de ilustrar esta perspectiva, são empregados exemplos de estudos que cobrem o período inicial da pesquisa aqui relatada, tratando principalmente de aspectos computacionais dos problemas das tarefas de seleção de conteúdo, realização superficial e geração de relações espaciais em ambientes virtuais tridimensionais interativos.

A principal limitação da perspectiva do locutor - a necessidade recorrente de criação de novos conjuntos de dados - é o tema do capítulo 4. Neste capítulo são discutidas iniciativas de construção de conjuntos de dados deste tipo, ou *corpus* para GLN, e suas dificuldades de criação, anotação e reutilização. Em especial, procura-se demonstrar que a criação destes recursos é uma tarefa que vai além da mera coleta de dados, podendo em certos casos assumir proporções de uma contribuição científica independente do próprio uso que a pesquisa venha a fazer deles.

Como alternativa às dificuldades de construção de *corpus* para GLN, o capítulo 5 aborda o problema de GLN sob a perspectiva do ouvinte (ou receptor da mensagem em língua natural). Embora sem perder de vista a aplicação computacional, esta perspectiva é ilustrada com exemplos de estudos de teor fortemente psicolinguístico, nos quais a ênfase está mais concentrada nos aspectos cognitivos da produção da língua, e que evidenciam a natureza verdadeiramente interdisciplinar da pesquisa em GLN.

Finalmente, o capítulo 6 retoma a questão da plausibilidade em GLN e propõe extensões do modelo tradicional de GLN orientado ao comportamento médio de um grupo de indivíduos (seja de acordo com a perspectiva de um grupo de locutores ou ouvintes) considerando duas alternativas: modelos baseados em preferências de locutores individuais e, de forma mais geral, modelos de personalidade humana. A primeira alternativa é discutida à luz de resultados de pesquisa já concluída sobre este tema. A segunda, por outro lado, é proposta de modo mais especulativo, na forma de resultados preliminares de um projeto de pesquisa recém iniciado.

Tabela 3: Artigos em conferências e workshops.

Citação	Evento	Título	Seção
(PEREIRA; PARABONI, 2007a)	TIL	<i>A Language Modelling Tool for Statistical NLP</i>	3.2
(LUCENA; PARABONI, 2008a)	Iberamia	<i>Combining frequent and discriminating attributes in the generation of definite descriptions</i>	3.2
(LUCENA; PARABONI, 2008b)	INLG	<i>USP-EACH frequency-based greedy attribute selection for referring expressions generation</i>	3.2
(PEREIRA; PARABONI, 2008a)	INLG	<i>From TUNA attribute sets to Portuguese text: a first report</i>	3.2
(PEREIRA; PARABONI, 2008b)	Gotal	<i>Statistical surface realisation of Portuguese referring expressions</i>	3.2
(SANTOS; PEREIRA; PARABONI, 2008)	TIL	<i>Rule-based vs. probabilistic surface realisation of definite description</i>	3.2
(LUCENA; PARABONI, 2009b)	ENLG	<i>USP-EACH: Improved frequency-based greedy attribute selection</i>	3.2
(NOVAIS; TADEU; PARABONI, 2010a)	Iberamia	<i>Improved text generation using n-gram statistics</i>	3.3
(NOVAIS; TADEU; PARABONI, 2010b)	NAACL	<i>Text generation for Brazilian Portuguese: the surface realization task</i>	3.3
(NOVAIS; TADEU; PARABONI, 2010c)	Iberamia	<i>Text-to-text surface realisation using dependency-tree replacement</i>	3.3
(TADEU; NOVAIS; PARABONI, 2010)	LREC	<i>Extracting Surface Realisation Templates from Corpora</i>	3.3
(NOVAIS; PARABONI; FERREIRA, 2011)	CICLing	<i>Highly-inflected language generation using factored language models</i>	3.3
(NOVAIS; PARABONI; Silva Junior, 2012)	LREC	<i>Portuguese Text Generation from Large Corpora</i>	3.3
(PARABONI; DEEMTER, 2012)	SIPI	<i>Using overspecified descriptions to help people find their way in spatial domains</i>	5.2
(PEREIRA et al., 2012)	LREC	<i>Corpus-based Referring Expressions Generation</i>	3.3
(Silva Junior; NOVAIS; PARABONI, 2013)	STIL	<i>Realização superficial baseada em regras</i>	3.3
(SANDRIM; FREIRE; PARABONI, 2013)	STIL	<i>Geração de instruções em mundos virtuais: primeiros passos</i>	3.4
(SILVA; PARABONI, 2013)	STIL	<i>Geração de Expressões de Referência usando Relações Espaciais</i>	3.4
(FERREIRA; PARABONI, 2014a)	CICLing	<i>Classification-based Referring Expression Generation</i>	6.2
(FERREIRA; PARABONI, 2014b)	TSD	<i>Referring Expression Generation: taking Speakers' Preferences into account</i>	6.2
(IACOVELLI; GALINDO; PARABONI, 2014)	Propor	<i>Lausanne: a Framework for Collaborative online NLP Experiments</i>	4.2
(PARABONI et al., 2014)	TSD	<i>Generating Underspecified Descriptions of Landmark Objects</i>	4.2
(TEIXEIRA et al., 2014)	CICLing	<i>Generating Relational Descriptions Involving Mutual Disambiguation</i>	4.2
(PARABONI; GALINDO; IACOVELLI, 2015)	ACL	<i>Generating overspecified referring expressions: the role of discrimination</i>	4.2
(ALTAMIRANO et al., 2015)	ACL	<i>Zoom: a corpus of natural language descriptions of map locations</i>	4.4
(LAN; PARABONI; COUTINHO, 2015)	STIL	<i>Avaliação de algoritmos de seleção de conteúdo utilizando Rastreamento de Olhar</i>	5.3
(FERREIRA; PARABONI, 2017b)	INLG	<i>Improving the generation of personalised descriptions</i>	6.2
(SANTOS; PARABONI; SILVA, 2017)	TSD	<i>Big Five Personality Recognition from Multiple Text Genres</i>	6.3
(PARABONI; MONTEIRO; LAN, 2017)	TSD	<i>Personality-dependent Referring Expression Generation</i>	6.3

Fonte: Ivandré Paraboni, 2017.

Tabela 4: Orientações de Iniciação Científica.

Orientando	Período	Projeto
Ramon R. M. Cuevas	2007-2007	<i>Aperfeiçoamento de Técnicas Estatísticas de Tradução Automática pelo Tratamento de Correferências</i>
Daniel B. Pereira	2007-2008	<i>Avaliação Automática de Textos em Sistemas de Tradução Estatística</i>
Wilker F. Aziz	2007-2008	<i>Investigação de Técnicas de Alinhamento Textual para a Tradução Automática Estatística</i>
Diego J. Lucena	2007-2008	<i>Investigação de Técnicas de Aprendizagem de Máquina para o Processamento de Línguas Naturais</i>
Ramon R. M. Cuevas	2008-2009	<i>Investigação de Técnicas Computacionais de Resolução de Referências Pronominais</i>
Daniel B. Pereira	2008-2009	<i>Desenvolvimento e Avaliação de Modelos Estatísticos de Língua</i>
Wilker F. Aziz	2008-2009	<i>Desenvolvimento de um Protótipo de Sistema de Tradução Automática Estatística</i>
Diego J. Lucena	2008-2009	<i>Estudo de Métodos Empíricos no Processamento de Linguagem Natural</i>
Rafael L. Oliveira	2008-2009	<i>Planejamento de Documentos em Aplicações de Geração Automática de Linguagem Natural</i>
Eder M. Novais	2008-2009	<i>Seleção de Conteúdo para Geração Automática de Textos em Linguagem Natural</i>
Thiago D. Tadeu	2008-2009	<i>Planejamento de Documentos na Geração Automática de Textos: Preparação para Realização Textual</i>
Roberto P. A. Araújo	2009-2010	<i>Geração Automática de Estruturas Discursivas</i>
Thiago D. Tadeu	2009-2010	<i>Realização Textual baseada em Templates</i>
Eder M. Novais	2009-2010	<i>Um Protótipo de Realização Textual baseada em Geração Permissiva e Seleção Estatística</i>
Douglas F. P. S. Júnior	2010-2011	<i>Estudo de Técnicas Básicas de Geração de Linguagem Natural: Realização Textual Operacional</i>
André C. Mariotti	2010-2011	<i>Estudo de Técnicas Básicas de Geração de Linguagem Natural: Geração baseada em Templates</i>
Eder M. Novais	2010-2011	<i>Geração de Expressões de Referência com Restrições de Realização Superficial</i>
Hilder V. L. Pereira	2010-2011	<i>Recuperação de Estratégias de Referência para Geração de Linguagem Natural</i>
Douglas Iacovelli	2013-2014	<i>Estudo e desenvolvimento de experimentos de geração de língua natural</i>
Michelle R. Galindo	2013-2014	<i>Avaliação de Algoritmos de GER em Mundos Virtuais Interativos</i>
Matheus M. Sant'Ana	2015-2015	<i>Desenvolvimento de uma Ferramenta de Experimentação em Processamento de Língua Natural usando Rastreamento de Olhar</i>
Alex G. J. Lan	2015-2016	<i>Avaliação de Algoritmos de Seleção de Conteúdo usando Gaze Tracking</i>
Vitor G. Santos	2016-2016	<i>Tratamento de dados para reconhecimento de traços de personalidade em textos</i>

Fonte: Ivandré Paraboni, 2017.

Tabela 5: Orientações de Mestrado.

Orientando	Período	Projeto
Diego S. Silva	2010-2013	<i>Geração de referências espaciais em ambientes interativos tridimensionais</i>
Thiago C. Ferreira	2013-2014	<i>A variação humana na geração de expressões de referência</i>
Caio R. Casimiro	2012-2015	<i>Aspectos temporais na recomendação de conteúdo em microblogs</i>
André C. Mariotti	2015-2017	<i>GER em situações de comunicação com restrição de tempo</i>
Danillo S. Rocha	2015-2017	<i>Validação de respostas em experimentos de geração de língua natural</i>
Barbara B. C. Silva	2015-2018	<i>Reconhecimento de traços de personalidade com base em textos</i>
Georges B. S. Neto	2015-2018	<i>Reescrita sentencial baseada em traços de personalidade</i>
Ricelli M. S. Ramos	2016-2018	<i>Determinação de conteúdo para GLN baseada em personalidade</i>
Danielle S. Monteiro	2016-2018	<i>Seleção de conteúdo referencial com base em traços de personalidade</i>

Fonte: Ivandré Paraboni, 2017.

Capítulo 2

CONCEITOS FUNDAMENTAIS

A pesquisa em GLN contempla uma ampla gama de questões pertinentes às três etapas da arquitetura *pipeline* (macroplanejamento, microplanejamento e realização superficial, cf. capítulo anterior), porém não de modo uniforme, já que uma parte considerável dos estudos da área tende a concentrar-se na etapa intermediária de microplanejamento. Uma possível explicação para esta predominância de estudos de microplanejamento é a observação de que, dentre as três etapas do *pipeline* de GLN, o microplanejamento é a que apresenta o maior potencial de generalização de resultados. Ou seja, enquanto a tarefa de macroplanejamento é altamente dependente do domínio da aplicação, e a tarefa de realização superficial é essencialmente guiada pelas restrições da língua-alvo de interesse, a tarefa de microplanejamento pode ser tratada de forma relativamente independente de ambos, e não por acaso concentra alguns dos desafios de pesquisa mais tradicionais da área.

Este capítulo apresenta uma série de conceitos fundamentais da pesquisa de GLN e temas correlatos que serão empregados ao longo do trabalho. Estes conceitos, que enfocam especialmente questões de microplanejamento e realização superficial, são organizados da seguinte forma. Primeiramente, será discutida a questão da seleção de conteúdo de descrições definidas, também conhecida como a tarefa computacional de Geração de Expressões de Referência (GER). Esta questão é introduzida na seção 2.1 a seguir, e constitui o tema de boa parte dos estudos discutidos ao longo desta monografia. Além desta, será discutida de forma mais breve também a questão da realização superficial na seção 2.3, tema este que será retomado nos capítulos seguintes com foco específico no português brasileiro. Finalmente, a seção 2.4 discute técnicas de avaliação de sistemas de GLN aplicáveis à tarefa de GER e à realização superficial, e a seção 2.5 introduz alguns conceitos básicos de personalidade humana que também serão referenciados no decorrer do trabalho.

2.1 Geração de Expressões de Referência

Na comunicação humana, locutores e ouvintes precisam estabelecer claramente sobre o que estão falando. A base deste processo é o fenômeno de *referência*, no qual formas linguísticas - ou *expressões de referência* - são escolhidas para identificar objetos do discurso tais como pessoas, coisas, ações ou eventos em um contexto de interesse. Considere-se a cena na figura 2, na qual desejamos apontar uma pessoa em particular, ou seja, fazer referência a um determinado indivíduo.

Figura 2: Um contexto visual contendo quatro entidades referenciáveis *e1..e4*.

Fonte: Criado via Second Life (<https://secondlife.com/>).

Neste exemplo, podemos fazer uso de diferentes tipos de expressões para identificar um objeto-alvo de interesse. Para objetos que possuem nomes próprios, ou que foram recentemente citados no discurso, produzir uma expressão de referência é relativamente simples: basta fazer uso do nome adequado (e.g., ‘João’) ou de uma forma pronominal compatível com a saliência do objeto-alvo na estrutura do discurso (e.g., ‘ele’) (GROSZ; WEINSTEIN; JOSHI, 1995). Para as situações de referência que não atendem estes critérios (ou seja, para descrever objetos sem um nome próprio conhecido, e que não foram citados recentemente), a estratégia referencial a ser adotada costuma ser o uso de uma descrição definida (e.g., ‘o homem de blusão escuro’) ou indefinida (e.g., ‘um rapaz mais ao centro’).

A produção de expressões de referência desempenha um papel fundamental na comunicação humana. Embora um discurso em língua natural seja composto de muitos outros elementos além de expressões de referência, o emprego de formas de referência adequadas ao contexto é essencial tanto para a coerência como para a coesão do discurso (HALLIDAY; HASAN, 1976), e sem uma escolha adequada deste tipo a comunicação simplesmente não atingiria seus objetivos.

A tarefa de produção de expressões de referência é geralmente desempenhada sem muito

esforço por seres humanos. Considere-se entretanto esta mesma tarefa sendo desempenhada por um agente de software em um ambiente envolvendo comunicação humano-computador, ou seja, em uma aplicação típica de GLN. Neste tipo de aplicação, alguns estudos da área têm tratado de questões como a escolha da forma (e.g., nominal ou pronominal) adequada a um determinado contexto (FUKUMURA; GOMPEL, 2010), ou da geração de nomes próprios (DEEMTER, 2016b). A maior parte das iniciativas de pesquisa em GLN, Psicolinguística e áreas afins, entretanto, têm sido direcionados para o estudo do problema de geração de descrições definidas (KRAHMER; DEEMTER, 2012).

Descrições definidas podem ser empregadas para descrever pessoas (e.g., ‘o homem de azul’), objetos ou coisas em geral (e.g., ‘o bar da esquina’) e uma ampla gama de entidades abstratas (e.g., ‘a segunda vez que ele passou aqui’, ‘a opinião dela’, ‘a falta de água em São Paulo’), dentre muitas outras possibilidades. Assim como em outras subtarefas da GLN, a produção de descrições deste tipo compreende dois passos razoavelmente distintos: decidir *o que* comunicar, e *como* comunicar. A decisão sobre o que comunicar é conhecida como a tarefa computacional de *seleção de conteúdo*, e consiste em escolher um conjunto de características (ou propriedades semânticas) que irá compor uma expressão tal que permita ao ouvinte identificar de forma inequívoca o objeto-alvo em questão. Já a decisão de como comunicar o conteúdo selecionado é denominada *realização superficial*, e consiste em mapear propriedades semânticas para formas superficiais (e.g., palavras) na língua-alvo de interesse.

Descrições definidas são amplamente empregadas no discurso humano, e a seleção de conteúdo para geração destas expressões constitui uma subárea própria dentro da pesquisa em GLN. Embora seja obviamente apenas um dos muitos aspectos da tarefa de produzir formas referenciais, a seleção de conteúdo de descrições definidas é frequentemente tratada como sinônimo de Geração de Expressões de Referência (GER) (KRAHMER; DEEMTER, 2012). Não havendo risco de confusão, no presente trabalho também utilizaremos a denominação GER neste sentido mais restrito do termo, ou seja, para falar especificamente da tarefa computacional de seleção de conteúdo de descrições definidas.

2.1.1 A tarefa computacional de GER

A geração automática de expressões de referência pode parecer à primeira vista uma tarefa altamente especializada, e talvez apenas um pequeno detalhe diante do desafio de produzir textos completos formando sentenças, parágrafos e documentos inteiros de forma fluente. Em termos práticos, entretanto, cabe observar que boa parte do conteúdo de uma sentença gerada automaticamente é determinado pelos objetivos da aplicação. Considere-se por exemplo o caso

de um sistema que gera instruções em mundos virtuais como o da figura 2 anterior. Em sistema deste tipo, um módulo de GLN seria responsável por produzir instruções do tipo ‘Agora converse com o rapaz de braços cruzados, à esquerda’.

Instruções deste tipo seriam originadas no módulo de planejamento do sistema, que precisa informar ao usuário o objetivo de realizar uma determinada ação (conversar) sobre um objeto específico (um personagem da cena). Assim, parte desta instrução - em especial, a ação desejada - poderia em muitos casos ser considerada fixa ou pré-definida e, se não houver necessidade de maior variação estilística, poderia em última análise ser realizada simplesmente com uso de um *template* do tipo ‘Agora converse com *X*’, onde *X* corresponde à porção variável do texto a ser gerado. O problema, entretanto, é justamente como descrever *X*.

A descrição de *X* neste exemplo é necessariamente dependente do contexto da comunicação, e poderia variar em virtude do ângulo ou perspectiva do usuário/ouvinte, de propriedades dinâmicas do objeto-alvo (e.g., se o rapaz descruzar os braços, mover-se, ou sofrer oclusão por parte de outro objeto ou personagem da cena), e pode ainda depender do que foi dito anteriormente (tanto pelo sistema como pelo usuário, se for o caso), dentre muitas outras possibilidades. A geração de descrições deste tipo não é assim facilmente implementável com uso de *templates* de texto pré-definido, e pode representar um obstáculo inclusive para métodos de aprendizagem de máquina (em especial, se a construção de um *corpus* de treinamento de ampla cobertura não for factível).

De forma simplificada, a geração de expressões de referência pode ser vista como a tarefa computacional de decidir quais propriedades semânticas de um objeto-alvo devem ser realizadas em uma descrição definida (KRAHMER; DEEMTER, 2012). Dado um objeto-alvo ao qual deseja-se fazer referência e uma base de dados contendo todos os fatos relevantes sobre os elementos do domínio (e.g., propriedades semânticas como gênero, cor do cabelo, distâncias, posições etc.), o objetivo mais fundamental de um algoritmo de GER é assim o de produzir uma lista de propriedades do objeto-alvo tal que esta seja capaz de distingui-lo de todos os demais objetos - chamados *distraidores* - no mesmo contexto.

Produzir descrições meramente livres de ambiguidade é, entretanto, uma tarefa em princípio trivial (DALE, 2002), e o maior desafio para um algoritmo de GER é decidir o que exatamente constitui uma descrição adequada, ou seja, decidir quantas e quais propriedades do objeto-alvo ela deve expressar. Entre os muitos objetivos conflitantes a considerar nesta tomada de decisão, o projeto de algoritmos deste tipo costuma levar em conta a máxima de quantidade de H. P. Grice ‘*dizer somente o necessário, e não mais do que o necessário*’ (GRICE, 1975). Esta decisão é justificada pela necessidade de evitar falsas implicações lógicas. Por exemplo, na expressão ‘a

mesa *de madeira*’ em um contexto no qual existe apenas um objeto do tipo mesa, o acréscimo da propriedade - dita *superespecificada* cf. Pechmann (1989) - ‘de madeira’ pode gerar ao ouvinte dúvidas sobre a real intenção do locutor, já que esta informação é supérflua para fins de identificação do objeto-alvo, e portanto poderia (ou deveria) haver algum motivo específico para ter sido mencionada.

Para evitar situações deste tipo, é desejável assim que a descrição produzida não seja excessivamente longa (e.g., ‘o rapaz de braços cruzados, à esquerda, e que está ao lado do homem de blusão escuro’). De forma análoga, entretanto, a descrição também não deve ser excessivamente breve (e.g., ‘a de verde’, em referência à moça mais ao fundo), já que isso pode dificultar a identificação do objeto-alvo em domínios extensos ou estruturalmente complexos (PARABONI; DEEMTER, 2014). O que constitui uma descrição ‘excessivamente’ longa ou breve é uma questão de pesquisa em aberto na área.

A questão da superespecificação referencial é um dos problemas fundamentais de GER, e será retomado em diversos pontos deste trabalho. Se por um lado a geração de expressões superespecificadas pode ocasionar falsas implicações lógicas, é por outro lado amplamente aceito que locutores humanos acrescentam informação superespecificada a praticamente todas as instâncias da linguagem, e pelos mais variados motivos (PECHMANN, 1989; ENGELHARDT; BAILEYAND; FERREIRA, 2006; ARTS et al., 2011; ENGELHARDT; DEMIRAL; FERREIRA, 2011; KOOLEN et al., 2011). Dentre os inúmeros desafios da geração de língua natural de forma plausível, o uso judicioso de informação superespecificada é seguramente uma questão de destaque (DEEMTER, 2016a).

Além destes aspectos ditos quantitativos, entretanto, a tarefa computacional de GER deve considerar também o tipo de informação a selecionar, decisão esta que é tipicamente tomada com base na *saliência relativa* das propriedades disponíveis para fins de referência. Por exemplo, considere-se a escolha - evidentemente sutil - entre descrever a pessoa *e3* do exemplo na figura 2 anterior com base na posição dos braços (e.g., ‘o homem com os braços à cabeça’) ou com base na cor da calça (e.g., ‘o homem de calça marrom’). Embora ambas opções sejam em certo sentido equivalentes (e.g., são ambas livres de ambiguidade, possuem aproximadamente o mesmo tamanho etc.) a escolha pode ser fortemente influenciada pela *saliência relativa* das propriedades em questão. Por exemplo, supondo-se que a propriedade que representa a cor da calça seja menos perceptível (e consequentemente menos acessível para referência) do que a propriedade que representa a posição dos braços, é possível que o uso de cor exija maior esforço de interpretação e, a depender dos objetivos do algoritmo de GER, seja assim menos desejável.

Tomadas em conjunto, restrições e preferências de natureza quantitativa e qualitativa compõe

o problema computacional de GER, e possuem forte afinidade com questões observadas na tarefa de GLN como um todo, ou seja, representam a necessidade de gerar língua natural de forma não apenas correta, mas também *plausível*. Esta afinidade obviamente não é motivo de surpresa, já que GER nada mais é do que uma das várias subtarefas de um sistema de GLN. Conforme já mencionado, entretanto, o problema de GER possui destaque dentro da pesquisa em GLN por diversos motivos. Além do papel central do fenômeno de referência na produção de língua natural, a tarefa de GER tende a ser menos dependente do domínio da aplicação (ao contrário de questões de pesquisa típicas da etapa inicial da arquitetura de GLN, ou macroplanejamento) e também menos dependente da língua-alvo a ser gerada (ao contrário de questões de realização superficial). Estas características, e o fato de que a tarefa de GER encapsula muitos dos problemas inerentes de GLN em um escopo mais restrito (e possivelmente mais manuseável), ajudam a explicar o porquê de a tarefa de GER ser uma questão importante na área de GLN, e uma de suas linhas de pesquisa mais ativas.

2.1.2 Algoritmos de GER

A tarefa de seleção de conteúdo para GER foi formalizada em uma série de estudos culminando na proposta da abordagem Incremental em Dale e Reiter (1995). Embora universalmente conhecido como um ‘algoritmo’, entretanto, esta abordagem não representa uma solução completa para o problema de GER, tratando apenas de estabelecer a entrada e saída esperadas, e apresentando uma estratégia geral de seleção de conteúdo motivada por estudos da área de Psicologia. A seguir são descritos de forma simplificada os conceitos mais importantes desta abordagem como forma de ilustrar o problema computacional de GER e alguns de seus desafios, tomando como exemplo prático mais uma vez o contexto ilustrado na figura 2 anterior.

A entrada de um algoritmo de GER típico é um contexto C formado pelo objeto-alvo ou referente r que se deseja descrever e por um grupo de objetos denominados distraidores dos quais deseja-se distinguir r . Tanto o objeto-alvo como os distraidores são representados por conjuntos de propriedades semânticas na forma de pares (*atributo - valor*), como em (*hair - dark*), representando o fato de que um determinado objeto do domínio possui a propriedade de ‘ter cabelos escuros’. Propriedades intrínsecas a um objeto são ditas *atômicas* (e.g., a propriedade de ter olhos castanhos). Propriedades que representam uma relação com outros objetos do contexto são ditas relacionais (e.g., a propriedade de estar próximo de uma determinada pessoa, de ser mais jovem do que ela etc.).

O conjunto de fatos ou características relevantes de um contexto é, por definição, ilimitado. Do ponto de vista prático (i.e., computacional), entretanto, a construção de uma base deste tipo

seria limitada ao conjunto de propriedades consideradas acessíveis - ou salientes - para fins de referência aos participantes do discurso. Deste modo, por exemplo, uma característica que seja pouco saliente - como a cor dos olhos de uma das pessoas da cena do exemplo anterior - normalmente não seria incluída na base de conhecimento.

Uma possível representação do contexto de referência ilustrado pela cena da figura 2 anterior é apresentada a seguir¹.

1. e1: {(gender - male), (arms - crossed), (hair - blonde), (direction - back)}
2. e2: {(gender - female), (arms - crossed), (hair - dark), (direction - left)}
3. e3: {(gender - male), (arms - onhead), (hair - dark), (direction - front)}
4. e4: {(gender - female), (arms - crossed), (hair - dark), (direction - right)}

Por simplicidade, neste exemplo consideramos apenas quatro propriedades atômicas de cada um dos objetos do tipo pessoa observados na cena. Na prática, entretanto, esta base de conhecimento poderia ser estendida com a inclusão de muitos outros objetos referenciáveis (e.g., mesas, cadeiras, partes do corpo de cada indivíduo, peças de vestuário etc.) e com a definição de diversas propriedades relacionais. Por exemplo, o fato que um determinado objeto (e.g., e4) está ao lado de e3 poderia ser representado por uma propriedade relacional do tipo (*next-to* - e3). Nota-se entretanto que propriedades relacionais não são limitadas a este caso (talvez mais óbvio) de relação espacial, podendo representar qualquer tipo de relação entre dois objetos. Por exemplo, a observação de que uma determinada pessoa é mais jovem do que e3 poderia ser representada como (*youngerthan* - e3).

Algoritmo 1: A estratégia básica de GER (DALE; REITER, 1995).

```

1 MakeReferringExpression(r, C, P)
2   L ← {}
3   C ← C − {r}
4   for Ai ∈ P do
5     V ← value(r, Ai)
6     if |rulesOut(⟨Ai, V⟩)| > 0 then
7       L ← L ∪ ⟨Ai, V⟩
8       C ← C − rulesOut(⟨Ai, V⟩)
9     end
10    if |C| == 0 then
11      return L
12    end
13  end
14  return L

```

¹Neste exemplo, os nomes dos atributos e valores foram mantidos em inglês apenas para enfatizar a distinção entre seleção de conteúdo (semântico) e realização superficial (texto, não tratado pelo algoritmo), ou seja, o fato de que estes termos representam conceitos, e não palavras.

A estratégia básica de GER é ilustrada pelo Algoritmo 1. Dado um contexto C e um objeto-alvo r que se deseja identificar, a seleção de conteúdo para GER consiste em produzir um (sub)conjunto L de propriedades de r tal que L seja capaz de distinguir r de todos os outros objetos em C . Para este fim, uma função auxiliar *RulesOut* é definida de modo a retornar, para uma dada propriedade (a, v) , o conjunto de objetos do contexto C para os quais o atributo a possui um valor diferente de v , ou seja, o conjunto de objetos que seriam eliminados caso a propriedade venha a ser incluída na expressão de referência. Por exemplo, a inclusão de uma propriedade (*hair - dark*) eliminaria todos distraidores cuja cor de cabelo não é escura.

O algoritmo Incremental funciona da seguinte forma. As propriedades são incluídas (linha 7) em L incrementalmente (de onde provém o nome do algoritmo) segundo uma ordem P predefinida (linha 4), desde que contribuam (linhas 5-6) para a desambiguação do referente (i.e., excluindo pelo menos um objeto do contexto C). A cada seleção, os objetos distraidores são excluídos com a atualização do contexto C (8).

O procedimento é repetido (4-13) até que um conjunto único (i.e., livre de ambiguidade) de propriedades seja obtido, ou seja, até que o conjunto de distraidores em C esteja vazio (10), ou até que todas as propriedades possíveis em P tenham sido consideradas (13). No primeiro caso, L poderia ser posteriormente realizada como uma descrição definida. No segundo caso L permaneceria ambígua, e poderia ser realizada, por exemplo, como uma descrição indefinida.

Um exemplo prático de funcionamento deste algoritmo pode ser descrito da seguinte forma. Suponha-se uma ordem de preferência $P = \{ \textit{gender}, \textit{arms}, \textit{hair}, \textit{direction} \}$, e o objeto-alvo $r = e2$. Neste caso, o algoritmo consideraria primeiro o atributo *gender*, o qual seria incluído na descrição resultante L porque seu valor *female* elimina dois elementos do contexto ($e1$ e $e3$, cujo atributo *gender* possui o valor *male*). A seguir, o atributo *arms* é considerado para inclusão, mas este não será selecionado porque seu valor (*crossed*) não descarta o único elemento restante no contexto ($e4$), que também possui este mesmo valor. O atributo *hair* é também desconsiderado por este mesmo motivo, e finalmente o atributo *direction* é incluído porque seu valor (*left*) descarta $e4$ (cuja direção é à direita). Assim, a expressão resultante $L = \{ \textit{gender}=\textit{female}, \textit{direction}=\textit{left} \}$ descreve de forma única o objeto-alvo $e2$, e poderia ser realizada, por exemplo, como ‘a moça voltada para a esquerda’. De forma análoga, $e1$ poderia ser descrito como ‘o rapaz de braços cruzados’, $e3$ poderia ser descrito como ‘o rapaz com os braços à cabeça’ e $e4$ como ‘a moça voltada para a direita’.

2.1.3 Discussão

Apesar de evidentemente simples, a estratégia básica de GER ilustrada pelo algoritmo Incremental (DALE; REITER, 1995) e outros tende a apresentar resultados suficientemente positivos em domínios restritos e/ou aplicações em que a variação da língua a ser gerada não seja uma questão crucial. No entanto, esta estratégia não contempla diversas questões que passaram a ser consideradas importantes em modelos de GER em anos subsequentes, como a da naturalidade da expressão (ou *humanlikeness* em Gatt e Belz (2007)). Algumas destas questões são destacadas de forma não exaustiva a seguir.

Em primeiro lugar, observa-se que a ordem preferencial P - cuja definição é deixada em aberto em Dale e Reiter (1995) justamente por ser dependente de domínio - tem grande impacto sobre o tipo de expressão produzida pelo algoritmo. A depender de como a lista P for ordenada, o algoritmo pode produzir descrições muito distintas. No exemplo da seção anterior, se fosse considerada a ordem $P = \{gender, hair, arms, direction\}$, $e2$ continuaria sendo descrito como ‘a moça voltada para a esquerda’. No entanto, esta ordenação levaria $e1$ a ser descrito como ‘o rapaz loiro’, $e3$ como ‘o rapaz de cabelos pretos’ etc.

Em segundo lugar, observa-se que, em sua proposta original, a abordagem Incremental manipula apenas propriedades atômicas do próprio objeto-alvo, como tipo, cor etc. Para tratamento de propriedades *relacionais*, por outro lado, seriam necessárias diversas adaptações de especial importância para o uso de relações espaciais, por exemplo, como em ‘a moça à direita do rapaz com os braços na cabeça’. Relações espaciais desempenham um papel central na descrição de objetos em contextos visuais, e são comuns em aplicações baseadas em mundos virtuais (BYRON et al., 2007) e outros. Entretanto, o conhecimento sobre quando ou como produzir descrições deste tipo é ainda escasso, e apesar da aparente similaridade com propriedades atômicas, relações deste tipo colocam uma série de questões adicionais para o projeto de algoritmos de GER.

Por exemplo, seria incomum (mas certamente não impossível) selecionar mais de uma relação espacial para uma mesma expressão, como em ‘a moça à direita do rapaz com os braços na cabeça, que está à direita da moça mais ao fundo’. Além disso, há a questão de quando utilizar ou não relações deste tipo, já que elas acarretam esforço cognitivo adicional para o ouvinte (KELLEHER; COSTELLO, 2009). Alguns estudos como Dale e Haddock (1991) sugerem o uso de relações apenas como último recurso, ou seja, quando não é possível distinguir o objeto-alvo usando apenas suas propriedades atômicas. Por outro lado, estudos como Viethen e Dale (2011) sugerem que o uso de propriedades relacionais é comum mesmo quando estas propriedades não são estritamente necessárias para desambiguação.

Em terceiro lugar, o problema computacional de GER apresenta também uma série de questões referentes ao volume de informação incluído na descrição. Em especial, observa-se que locutores humanos normalmente não se limitam a dizer o que é estritamente necessário para a identificação do objeto-alvo. Pelo contrário, muitas das expressões que produzimos são do tipo superespecificado, ou seja, contendo propriedades redundantes que poderiam em tese ser omitidas sem prejuízo à identificação do alvo. A superespecificação é inerente à produção de língua natural, e se torna especialmente problemática para um algoritmo de GER que se pretenda plausível uma vez que os critérios para inclusão de informação redundante são pouco conhecidos e, conseqüentemente, de difícil formalização. Enquanto atributos altamente discriminatórios (como a informação de gênero no exemplo anterior) são facilmente selecionados por muitos tipos de algoritmos de GER, o mesmo não pode ser dito de atributos que não sejam estritamente necessários para a desambiguação.

Por exemplo, ‘o homem loiro de braços cruzados’ é uma descrição perfeitamente plausível do alvo *e1*, do tipo que muitos locutores humanos poderiam produzir. No entanto, a depender da ordem de preferência *P* escolhida, este tipo de redundância não seria reproduzida por um algoritmo de GER como em Dale e Reiter (1995) uma vez que violaria a Máxima de Quantidade de Grice (GRICE, 1975).

Finalmente, conforme observado em Viethen e Dale (2011), não existe uma estratégia de produção de língua dita ‘universal’, ou nem mesmo uma estratégia única para cada locutor. Pessoas diferentes descrevem um mesmo objeto de formas diferentes, e a mesma pessoa pode descrever um objeto de formas diferentes em momentos distintos (i.e., mesmo em situações idênticas). Assim, um algoritmo de GER que produza sempre a mesma saída fixa para uma determinada entrada está destinado a apresentar grau de sucesso variável, a depender do locutor ou ouvinte considerados.

Estas quatro questões - a escolha de propriedades preferenciais, a geração de descrições relacionais, o fenômeno de superespecificação referencial e a questão da variação humana - figuram em diversos estudos discutidos nesta monografia, a saber:

- O uso de listas de preferências na tarefa de GER é abordado no seu sentido literal (LUCENA; PARABONI; PEREIRA, 2010), analisado de forma crítica em situações de referência nas quais esta abordagem é particularmente ineficaz (PARABONI; GALINDO; IACOVELLI, 2017), remodelado com o uso de técnicas de aprendizagem de máquina (FERREIRA; PARABONI, 2017a), e discutido à luz de evidências psicolinguísticas de como diferentes propriedades afetam a identificação do objeto-alvo (PARABONI; DEEMTER, 2014; PARABONI et al., 2017).

- O uso de propriedades relacionais é tratado em estudos de geração de relações espaciais em mundos virtuais interativos (SILVA; PARABONI, 2015), e de superespecificação de propriedades deste tipo (PARABONI; GALINDO; IACOVELLI, 2017; PARABONI; DEEMTER, 2014; PARABONI et al., 2017).
- A superespecificação referencial é um tema recorrente ao longo desta monografia, discutido em praticamente todos estudos que tratam do problema de GER aqui relatados. Estes estudos incluem a questão da relação entre superespecificação e saliência (PARABONI; GALINDO; IACOVELLI, 2017) e de seus efeitos sobre a identificação de expressões de referência (PARABONI; DEEMTER, 2014; PARABONI et al., 2017).
- A questão da variação humana na tarefa de GER é tratada no contexto de modelos de locutor individualizados (FERREIRA; PARABONI, 2014b, 2017a) e de personalidade humana (PARABONI; MONTEIRO; LAN, 2017).

2.2 Córpus para GER/GLN

A evolução da pesquisa em GLN, concomitante à popularização dos métodos de Aprendizagem de Máquina (AM), tanto nesta como na grande área de PLN, levou ao desenvolvimento de uma série de córpus dedicados ao estudo e desenvolvimento de técnicas de GLN em anos recentes. Esta seção discute brevemente a natureza de recursos deste tipo e, de forma mais específica, os desafios de construção de córpus para o estudo de fenômenos de referência e para o desenvolvimento de modelos computacionais de GER.

2.2.1 Visão geral

À primeira vista, córpus para GER/GLN parecem guardar forte semelhança com córpus textuais comumente utilizados em diversas linhas de pesquisa do PLN. Estes córpus no entanto são recursos linguístico-computacionais altamente especializados, e a analogia com córpus textuais de uso geral é sob vários aspectos limitada. Algumas diferenças fundamentais entre a construção de córpus para PLN e GLN são discutidas a seguir.

Em primeiro lugar, observamos que, ao contrário de tarefas de interpretação de língua natural que exploram a ampla gama de córpus textuais existentes (ou facilmente obtidos a partir da WEB, por exemplo), poucas tarefas de GLN podem de fato ser implementadas com base apenas no texto resultante do processo de produção da língua. Para muitos estudos de GLN - incluindo o caso da tarefa de GER - faz-se necessário conhecer não apenas os tipos de texto que se pretende

gerar, mas *as condições iniciais* ou *contextos* que os motivaram (e.g., o discurso subjacente, o contexto visual etc.). Assim, um *córpus* para GLN tipicamente não contém apenas exemplos de texto a ser gerado, mas pares de textos e contextos, ambos geralmente acompanhados de anotação semântica. Como resultado, o processo de criação de recursos deste tipo é de custo elevado e, não por acaso, o número de *córpus* de GLN publicamente disponível para pesquisa é reduzido.

No caso específico da pesquisa em GER, a construção de *córpus* é normalmente implementada na forma de experimentos controlados envolvendo sujeitos humanos, de modo a estabelecer possíveis relações entre o texto produzido e os fatores que o motivaram. De modo mais específico, experimentos deste tipo fornecem estímulo textual ou visual aos participantes, e registram suas reações manifestas na forma oral, escrita, ou por ações de interação com o ambiente computacional. Após a coleta, processamento e anotação dos dados, o resultado final - ou *córpus* para GER - é assim formado pela representação semântica dos contextos de referência e das descrições neles produzidas.

Como forma de obter máximo controle sobre o processo de produção da língua (e.g., minimizando inferências indesejáveis, questões relacionadas à memória de leitura, conhecimento do domínio etc.) o uso de *imagens* representando um contexto referencial de interesse é a prática mais frequente.² Independentemente do tipo de estímulo empregado, entretanto, geralmente espera-se que as questões de pesquisa tratadas sejam suficientemente genéricas para aplicação tanto em contextos visuais como discursivos. Alguns exemplos de *córpus* construídos de acordo com esta metodologia, voltados predominantemente para a pesquisa em GER, são descritos a seguir.

2.2.2 *Córpus de expressões de referência*

Um exemplo proeminente de *córpus* para GER é o resultado do projeto TUNA (GATT; SLUIS; DEEMTER, 2007), desenvolvido especificamente para o estudo de fenômenos de referência e algoritmos de GER. O *córpus* TUNA contempla situações de referência em dois domínios distintos: peças de Móvel (*Furniture*), e fotos de Pessoas (*People*). Um exemplo destes domínios é ilustrado na figura 3.

O *córpus* TUNA contém 2280 descrições definidas (780 singulares e 1500 plurais) e seus respectivos contextos. As descrições TUNA foram geradas a partir de experimentos controlados realizados com 60 participantes que usavam o inglês como idioma nativo, ou possuíam fluência no

²Além disso, há evidências de que o uso de estímulos visuais para produção de língua natural é superior ao uso de representações semânticas (NOVIKOVA; LEMON; RIESER, 2016).

Figura 3: Exemplo de estímulos visuais do *córpus TUNA Furniture e People*.

Fonte: Adaptado de Koolen et al. (2011).

mesmo. Assim como em outras iniciativas do gênero, a tarefa dos participantes do experimento TUNA consistia em identificar de forma única (i.e., livre de ambiguidade) um determinado objeto-alvo em destaque em cada cena. No exemplo da figura 3 o alvo do primeiro contexto (*Furniture*) seria a cadeira no canto esquerdo, e o alvo do segundo contexto (*People*) seria o homem de barba branca na porção centro-superior da cena.

Sendo o primeiro recurso de maior escala deste tipo a ser disponibilizado publicamente para pesquisa, o *córpus TUNA* foi utilizado para treinamento e teste de algoritmos de GER em uma grande variedade de projetos, incluindo três competições (ou *shared tasks*) da área (GATT; BELZ, 2007; GATT; BELZ; KOW, 2008, 2009). No entanto, este *córpus* possui algumas limitações importantes, como a ausência de propriedades relacionais, a simplicidade do domínio *Furniture* e a acentuada esparsidade de atributos no domínio *People*.

Outro recurso amplamente utilizado na área é o *córpus GRE3D3* e sua extensão *GRE3D7* (DALE; VIETHEN, 2009; VIETHEN; DALE, 2011), em ambos os casos tratando da questão do uso de relações espaciais em contextos visuais tridimensionais simplificados. Em dois experimentos independentes realizados via WEB, participantes foram instruídos a descrever objetos geométricos do tipo esfera e cubo. Um exemplo deste tipo de domínio é ilustrado na figura 4.

Juntos, os *córpus GRE3D3* e *GRE3D7* somam cerca de 5110 descrições produzidas por 350 participantes, e são provavelmente o maior conjunto de dados do gênero disponível para fins de pesquisa, e também um dos poucos a contemplar o fenômeno de referência espacial. Apesar destas vantagens, o domínio *GRE3D* é ainda excessivamente simplificado, especialmente pelo número reduzido de atributos atômicos e relacionais possíveis.

O método de construção dos *córpus TUNA* e *GRE3D3/7* é também empregado na construção de *córpus* de diálogos, como no caso do *córpus iMap* (GUHE, 2009) de instruções de rota em mapas bidimensionais. O *córpus iMap* consiste de 256 diálogos construídos com base em mapas

Figura 4: Exemplo de estímulo utilizado no córpus GRE3D7.

Fonte: Adaptado de Viethen e Dale (2011).

fictícios representando cenas com diversos tipos de obstáculos que precisam ser contornados para atingir um determinado destino. Em cada diálogo, uma dupla de participantes revezava-se nos papéis de instrutor e receptor das instruções. Ambos participantes tinham acesso ao mapa, mas apenas o mapa do instrutor exibia o traçado do caminho a ser seguido. A tarefa do instrutor consistia assim em descrever este caminho de modo que o receptor pudesse desenhá-lo em seu próprio mapa. O conjunto de dados formado pelo córpus *iMap* (i.e., diálogos, imagens etc.) constitui um recurso potencialmente valioso para estudos em GLN. Entretanto, como estes dados não estão disponíveis publicamente por razões de confidencialidade (os experimentos incluíam pesquisas na área médica), houve pouco uso prático além do estudo em Guhe (2009).

O córpus GIVE-2 (GARGETT et al., 2010) de instruções em mundos virtuais foi construído por meio de experimentos envolvendo pares de participantes de língua inglesa e alemã que se alternavam nas tarefas de instrutor e jogador. Um exemplo de contexto utilizado na construção deste córpus é ilustrado na figura 5.

Figura 5: Um exemplo de contexto visual do sistema GIVE.

Agora pressione o botão azul uma única vez.

Fonte: Adaptado de Byron et al. (2007).

O córpus GIVE-2 foi produzido por 36 duplas de participantes em condições semelhantes às do modelo instrutor-receptor empregado no córpus de diálogos *iMap* (GUHE, 2009), e contém todas instruções fornecidas pelo instrutor, e as respectivas decisões tomadas pelo jogador (e.g., movimentos, ações de pressionar botões etc.). O conjunto de dados multimodal resultante pode ser visualizado na forma de animação com uso da ferramenta *Replay* em Gargett et al. (2010).

Mesmo constituindo um domínio mais próximo de uma aplicação computacional real, entretanto, o córpus GIVE-2 também apresenta algumas dificuldades de reuso. Em especial, a representação do córpus em formato de animação acompanhada de diálogos torna as tarefas de anotação e processamento de expressões de referência significativamente complexas, e provavelmente não justificáveis dada a pequena proporção de exemplos linguísticos disponíveis em relação à ampla gama de contextos possíveis.

A necessidade de construção de recursos contendo exemplos mais realistas de situações de referência (geralmente envolvendo fotografias ou imagens detalhadas) motivou a construção de diversos outros córpus em anos recentes, com graus variáveis de aplicabilidade à pesquisa em GER. Estes incluem o córpus Craft (MITCHELL; DEEMTER; REITER, 2010) de descrições de máscaras construídas a partir de objetos de uso cotidiano como pedras, penas e fitas, o córpus GenX (FITZGERALD; ARTZI; ZETTLEMOYER, 2013) de descrições de fotografias envolvendo objetos geométricos tridimensionais, o córpus Wally (CLARKE; ELSNER; ROHDE, 2013) de descrições de cenas do livro ‘Onde está Wally?’³, e o córpus ReferItGame (KAZEM-ZADEH et al., 2014) de descrições de fotografias reais de paisagens, pessoas etc. Um exemplo de estímulo do córpus ReferItGame é ilustrado pela figura 6, no qual o objetivo da tarefa era o de descrever o objeto (no caso, um prédio) destacado em amarelo à esquerda.

Figura 6: Exemplo de estímulo utilizado no córpus ReferItGame.

Fonte: Adaptado de Bartie et al. (2016).

Apesar do inegável apelo destes domínios mais realistas, experimentos deste tipo implicam menor controle sobre os estímulos e, conseqüentemente, sobre as expressões produzidas,

³ <http://www.whereswally.co.uk/>

reduzindo-se assim o escopo das conclusões possíveis sobre como e porque a língua humana é produzida. Por exemplo, a tarefa de descrição do *córpus ReferItGame* suscitou respostas do tipo esperado para pesquisa em GER como ‘o prédio de colunas, no lado esquerdo’ mas também referências a uma variedade de outros objetos possíveis, como ‘a região esquerda da fotografia’, ‘sequência de barras verticais marcadas em amarelo’ etc.

Independentemente de como são projetados, entretanto, experimentos de construção de *córpus* para GER são de modo geral de custo elevado, seja pela complexidade dos materiais envolvidos, pela necessidade de grandes números de participantes, ou simplesmente pelo tratamento e anotação subsequentes à coleta de dados. Estas e outras questões serão retomadas no capítulo 4, onde serão discutidos recursos linguístico-computacionais deste tipo desenvolvidos por este autor e seus colaboradores, e contemplando expressões linguísticas em português.

2.2.3 **Discussão**

Alguns dos recursos discutidos na seção anterior foram utilizados para fins de teste e validação de diversos artefatos de GLN que serão discutidos ao longo desta monografia. O *córpus* TUNA foi utilizado em estudos de seleção de conteúdo (LUCENA; PARABONI; PEREIRA, 2010) e em diversos estudos derivados, e também para avaliação de algoritmos de GER de propósito geral (FERREIRA; PARABONI, 2017a). Os *córpus* GRE3D3/7 foram utilizados especialmente para treinamento de modelos de referência espacial (SILVA; PARABONI, 2015) e para a validação de algoritmos de GER (FERREIRA; PARABONI, 2017a). O *córpus* GIVE-2 foi utilizado em um estudo sobre geração de referências espaciais (SILVA; PARABONI, 2015), e o sistema GIVE que deu origem a este *córpus* foi adaptado para a execução de um novo experimento enfocando questões de superespecificação referencial e memória (PARABONI; DEEMTER, 2014).

Para investigação de questões de pesquisa não observáveis em *córpus* existentes, entretanto, foi em certos casos necessária a construção de recursos próprios. Três estudos desta natureza serão relatados ao longo desta monografia, a saber: o *córpus* Stars de descrições sub e superespecificadas (TEIXEIRA et al., 2014; PARABONI et al., 2014), o *córpus* Stars2 de contextos com manipulação de saliência (PARABONI; GALINDO; IACOVELLI, 2017) e o *córpus* Zoom de descrições em um domínio de aplicação real (ALTAMIRANO et al., 2015). Estes recursos tratam especificamente da produção de descrições em português e, no caso do *córpus* Zoom, incluem também dados em espanhol.

2.3 Realização superficial

A realização superficial (ou textual) é essencialmente a tarefa computacional de mapear uma representação abstrata do texto – frequentemente já contendo algum tipo de informação estrutural – para uma representação linear em língua natural, e como tal é por vezes confundida com a própria tarefa de GLN. É a tarefa que normalmente segue a seleção de conteúdo (como no caso da geração de expressões de referência discutida nas seções anteriores), e normalmente encerra a geração de língua natural em uma arquitetura do tipo *pipeline* (REITER; DALE, 2000). Em analogia à tarefa de análise sintática de PLN (que trabalha com hipóteses de interpretação do texto), a tarefa de realização textual consiste em modelar decisões de geração de texto (RAJKUMAR; WHITE, 2014). Um exemplo de entrada típica para um realizador textual é apresentado na figura 7, que poderia ser realizada como ‘Um homem comprou dois bilhetes’.

Figura 7: Exemplo de especificação de entrada.

Fonte: Adaptado de Reiter e Dale (2000).

A implementação computacional da tarefa de realização superficial pode ser dividida em dois componentes relativamente independentes: um componente tático, mais dependente da aplicação, encarregado de mapear conceitos de domínio para algum tipo de estrutura abstrata de motivação linguística como no exemplo anterior, e um componente operacional encarregado de impor restrições gramaticais e efetuar a linearização de sentenças na língua-alvo. Embora a maioria dos sistemas atuais de alguma forma desempenhe ambas as tarefas, em alguns casos apenas o componente operacional é implementado. Este é o caso por exemplo dos sistemas YAG (MCROY; CHANNARUKUL; ALI, 2003) e SimpleNLG (GATT; REITER, 2009), que partem do princípio de que o mapeamento entre conhecimento semântico e sintático já tenha de alguma forma sido feito, e a preocupação então passa a ser a tarefa de linearização propriamente dita.

De forma conjunta, os componentes tático e operacional constituem o que é entendido como um *realizador superficial*, que pode ser visto como um modelo de conversão de estruturas semântico-sintáticas profundas (sem flexão ou ordem) em *strings* de tal forma que estas comu-

niquem o sentido pragmático desejado livre de ambiguidade. Esta definição entretanto deixa margem para um grande número de interpretações (e implementações) possíveis, que incluem desde a definição do limite exato entre a tarefa de realização superficial e outros componentes da arquitetura de GLN, até questões de representação do conhecimento. Em virtude da ampla gama de possibilidades, no restante desta seção a discussão será assim limitada a três questões específicas, relacionadas à especificação de entrada, ao formalismo gramatical e ao paradigma computacional. Um exemplo concreto de modelo de realização superficial que ilustra as questões aqui discutidas será apresentado na seção 3.3.

2.3.1 Especificação de entrada

Uma primeira questão a ser considerada no projeto de um realizador superficial é sua especificação de entrada: evidentemente, quanto mais próxima esta representação for da forma superficial, menor será o esforço de realização. Ao simplificarmos a tarefa desta forma, entretanto, as decisões de realização estão apenas sendo transferidas para a etapa anterior da arquitetura de GLN, isso é, o microplanejamento. Não há de fato um limite claro entre microplanejamento e realização superficial, e a decisão sobre o que esperar como entrada para a realização superficial é ditada pela aplicação subjacente e o grau de especificidade semântica que esta pode fornecer (REITER; DALE, 2000). Por exemplo, se a aplicação for capaz de fornecer instruções já detalhadas sobre a lexicalização e a estrutura sintática desejada, então a tarefa de realização superficial consistirá basicamente em aplicar um conjunto de regras gramaticais adequado para impor concordância, resolver dependências de longa distância etc., que poderia ser implementado com o uso de um realizador operacional simples (GATT; REITER, 2009). Por outro lado, a inserção deste tipo de processamento linguístico na aplicação pode ter um custo elevado, e que se repetirá a cada novo sistema desenvolvido.

2.3.2 Formalismo gramatical

Uma segunda questão a ser considerada na tarefa de realização superficial, e que é diretamente ligada à definição da especificação de entrada, é a questão da granularidade da representação gramatical empregada pelo realizador. Em um extremo, podemos considerar um realizador superficial que efetua mapeamentos de conceitos de domínio diretamente para constantes do tipo *string* predefinidas (ou *canned text*). Um recurso deste tipo seria de implementação muito simples porém sem nenhuma flexibilidade, exigindo manutenção a cada nova variação linguística a ser produzida (além de, é claro, frustrar o próprio objetivo de um sistema de GLN).

No outro extremo, podemos considerar um realizador que estabelece mapeamentos entre conceitos de domínio e regras de uma gramática de ampla cobertura da língua-alvo (KING; WHITE, 2016). Esta alternativa proporcionaria máxima flexibilidade e robustez a um grande número de aplicações, mas pode ser de custo elevado, tanto do ponto de vista do desenvolvimento (pois requer o conhecimento de especialistas na língua em questão) quanto da sua utilização prática (já que envolve busca em um espaço de formas linguísticas possíveis).

Cabe observar ainda que o uso de uma gramática de ampla cobertura recai no problema anteriormente citado, relacionado ao grau de especificidade de entrada: realizadores com este grau de sofisticação só podem ser plenamente aproveitados se a aplicação puder fornecer uma entrada em nível igualmente detalhado de especificação linguística.

2.3.3 Paradigma computacional

Finalmente, quanto ao paradigma computacional a ser utilizado no projeto de um realizador superficial, observa-se em Belz (2005) que a maioria dos sistemas de GLN existentes têm sido desenvolvido através de codificação de regras de geração (frequentemente do tipo *if-then*) provenientes de uma cuidadosa análise de requisitos manual baseada em *córpus*, e com o auxílio de especialistas do domínio. Como as regras que expressam o conteúdo tendem a ser diferentes em cada domínio, na prática o potencial de reutilização de componentes de GLN costuma ser baixo. Neste sentido, a introdução de métodos empíricos aplicados à GLN foi um avanço muito significativo: em anos recentes, o conhecimento proveniente de *córpus* passou a ser utilizado não apenas na análise de requisitos (como originalmente sugerido em Reiter e Dale (2000)), mas também na tomada de decisões que guiam o próprio processo de GLN (PAN; SHAW, 2004; MARCINIAK; STRUBE, 2004, 2005; ZHONG; STENT, 2005; BELZ, 2008)), motivando assim um grande número de abordagens baseadas em aprendizagem de máquina e, de forma mais expressiva, modelos estatísticos de geração de língua natural.

Sistemas de realização superficial ditos estatísticos separam o espaço de geração (possíveis entradas e/ou saídas) do mecanismo de controle de decisões sobre o que produzir: são as chamadas abordagens ‘gerar e selecionar’ ou ‘geração em dois estágios’, introduzidas pelo sistema Nitrogen em Langkilde (1998) e expandidas na família de sistemas HALOGEN (LANGKILDE, 2000; LANGKILDE-GEARY, 2002) e outros (VARGES, 2006). Sistemas deste tipo realizam um processamento predominantemente simbólico para realizar geração permissiva, ou seja, produzindo um grande número de soluções possíveis (e.g., a partir de um *template* básico), incluindo até mesmo alternativas não-gramaticais, e que são filtradas com uso de modelos de *n*-gramas em uma fase de seleção estatística. A técnica em dois estágios permite o desenvolvimento

de aplicações de GLN treináveis a partir de *cópus* e fazendo uso de um volume reduzido de conhecimento linguístico.

Após o estudo pioneiro de Langkilde (1998), o uso de técnicas estatísticas tem sido ampliado a outros aspectos da tarefa de GLN. Em geral, entretanto, o volume de geração de alternativas é intenso e de alto custo computacional, além de não ser linguisticamente informado. Por exemplo, mesmo a abordagem em Belz (2008) já considerava até 1040 alternativas de realização de cada sentença a ser gerada, o que pode representar um desempenho computacional insatisfatório para certas aplicações práticas como no caso de sistemas interativos de diálogo humano-computador. Além disso, métodos estatísticos baseados em modelos de *n*-gramas apresentam conhecidas dificuldades de tratamento de dependências de longa distância e, por definição, tendem a favorecer alternativas de realização superficial mais breves: por exemplo, em um modelo deste tipo a descrição ‘o homem de preto’ não pode ser menos provável do que ‘o homem de preto, à esquerda’, e assim tende a ser preferida, o que é claramente inadequado do ponto de vista de uma aplicação de GLN.

Assim como em outras áreas do PLN, estudos mais recentes começam a substituir a classificação de *n*-gramas por formas de representação distribuída de palavras (ou *word embeddings*) associadas a técnicas de aprendizagem profunda ou *deep learning* (MIKOLOV et al., 2011). Conforme demonstrado em Ghodsi e DeNero (2016), modelos deste tipo possuem uma capacidade estrutural ausente em modelos baseados em *n*-gramas, a qual se torna especialmente evidente no tratamento de dependências intrasentenciais, embora ainda apresentem certa dificuldade de tratamento de distribuições mais extensas, como na maioria dos casos de referência pronominal, por exemplo. Vários dos estudos que implementam a tarefa de realização textual utilizando técnicas de aprendizagem profunda tratam também de outras tarefas de GLN (ZHANG; LAPATA, 2014; KARPATHY; FEI-FEI, 2015; XU et al., 2015) - especialmente em sistemas de geração de diálogos (WEN et al., 2015; MEI; BANSAL; WALTER, 2016; DUSEK; JURCICEK, 2016) - fugindo assim ao escopo da presente discussão.

2.3.4 Discussão

Os diversos aspectos da tarefa de realização superficial aqui apresentados são abordados por meio de uma ampla gama de estratégia distintas na literatura em GLN. Em especial, observamos que não há uma razão intrínseca para que uma solução de GLN (ou de outra área relacionada) tenha de ser necessariamente limitada a apenas um único paradigma e, não por acaso, sistemas bem-sucedidos costumam aplicar uma abordagem híbrida. Realizadores textuais tidos como o ‘estado-da-arte’ tendem assim a fazer uso de regras gramaticais codificadas manualmente ou

extraídas automaticamente de um banco de árvores sintáticas (*tree bank*). As diferentes escolhas linguísticas são representadas como escolhas entre conjuntos de regras, tipicamente fazendo uso de um algoritmo de *chart parsing*, e podem ser puramente simbólicas (em realizadores que produzem uma saída única com base nas regras empregadas) ou, mais comumente, fazer uso de geração permissiva e *ranking* estatístico com uso de modelos de língua. Um exemplo proeminente de realizador superficial baseado em regras para a língua inglesa é OpenCCG (WHITE; RAJKUMAR; MARTINS, 2007), que utiliza gramáticas categoriais combinatórias induzidas a partir da base Penn Treebank (HOCKENMAIER; STEEDMAN, 2007) e usa um *ranking* baseado em estimativas de n-gramas e diversos outros atributos.

Algumas das questões aqui levantadas serão ilustradas na presente monografia por meio de exemplos de técnicas de realização superficial de expressões de referência (LUCENA; PARABONI; PEREIRA, 2010) e manchetes de notícias (NOVAIS; PARABONI, 2012), em ambos os casos tratando especificamente do caso de geração de texto em português do Brasil.

2.4 Avaliação de sistemas de GLN

A avaliação de sistemas de GLN faz uso de técnicas semelhantes às adotadas em diversas outras áreas do PLN (JONES; GALLIERS, 1995), podendo ser de natureza *intrínseca* (medindo-se propriedades específicas do texto gerado automaticamente) ou *extrínseca* (medindo-se os efeitos do texto gerado sobre outras tarefas, que podem ser desempenhadas por usuários humanos ou no contexto de outros sistemas computacionais). No caso específico das duas tarefas de GLN discutidas nas seções anteriores - a seleção de conteúdo referencial e realização superficial - podemos dividir as técnicas de avaliação em três categorias principais (REITER; BELZ, 2009): baseadas na efetividade da tarefa, avaliações feitas por usuários e baseadas em córpus. Estas alternativas são discutidas a seguir.

2.4.1 Avaliação baseada na efetividade da tarefa

A efetividade da tarefa é o critério de avaliação final e mais importante de um sistema de GLN. Independentemente do que tenha sido desenvolvido, da perspectiva adotada, ou da empatia dos usuários com o sistema, em última análise o que se deseja saber é se o sistema atinge ou não seus objetivos de comunicação, ou seja, se ele efetivamente contribui para a tarefa a qual se destina. A avaliação baseada na efetividade da tarefa é assim uma forma de avaliação extrínseca (JONES; GALLIERS, 1995), e é especialmente importante em aplicações nas quais é necessário convencer usuários (como especialistas de outras áreas) do seu valor (REITER; BELZ, 2009).

Um exemplo de avaliação baseada na efetividade da tarefa é o estudo em Carenini e Moore (2006). Este estudo apresenta a avaliação de um sistema de geração de publicidade persuasiva para venda de imóveis. Neste caso, examinou-se como o público-alvo avaliava os imóveis anunciados - e não o texto em si - com base em textos produzidos por humanos e pelo sistema em questão. O método de avaliação adotado, que fez uso de uma escala de pontuação - é assim baseado na efetividade publicitária, com o objetivo de demonstrar que imóveis divulgados por meio de textos automáticos recebiam melhor avaliação dos potenciais compradores.

Outro exemplo de uso da técnica baseada na efetividade é o observado na avaliação do sistema SKILLSUM em Williams e Reiter (2008), que é um sistema para geração de *feedback* textual para estudantes matriculados em disciplinas de nível superior. Este sistema implementa um teste básico de habilidades linguísticas e matemáticas e, a partir dos dados coletados, gera um relatório descrevendo as habilidades observadas, sua adequação à disciplina pretendida e, quando necessário, sugere ações que o estudante pode tomar para aprimorar eventuais lacunas no seu conhecimento. A avaliação do sistema SKILLSUM foi realizada com base na sua efetividade educacional, na qual 200 candidatos recebiam *feedback* gerado pelo sistema ou textos de controle, procurando demonstrar que o texto produzido pelo sistema melhorava a autoavaliação destas habilidades e auxiliava nas escolhas feitas pelos estudantes.

O sistema BT45 em Portet et al. (2009), que é uma dentre várias aplicações de um projeto de geração de texto a partir de dados extraídos de sensores e equipamentos médicos de unidades de tratamento intensivo de recém-nascidos, foi também avaliado com base na sua efetividade na tomada de decisões. De modo mais específico, foi realizada uma comparação entre as decisões tomadas por médicos e enfermeiras a partir de diagnósticos textuais gerados automaticamente com as decisões tomadas a partir de gráficos com os quais estavam habituados a trabalhar. O objetivo neste caso foi o de demonstrar a maior efetividade das decisões tomadas com base nos diagnósticos textuais gerados de forma automática.

Apesar dos benefícios da avaliação com base na efetividade da tarefa, é observado em Reiter e Belz (2009) que o investimento (financeiro ou de tempo) pode torná-la proibitiva, especialmente se considerarmos que este tipo de avaliação depende em grande parte da cooperação de terceiros (muitas vezes especialistas do domínio, com pouca disponibilidade de tempo etc.) para recrutar participantes (e.g., pacientes, estudantes etc.). Além disso, o fato de que um sistema apresenta boa correlação com o desempenho em uma tarefa específica não significa que apresente a mesma correlação em outras tarefas.

Finalmente, observamos que a avaliação com base na efetividade da tarefa é de modo geral mais aplicável a sistemas completos de GLN, mas possivelmente menos adequada para a

avaliação isolada de artefatos ou componentes específicos de GLN. Para estes casos, técnicas de avaliação baseadas em propriedades intrínsecas do sistema ou artefato podem ser necessárias.

2.4.2 Avaliação por usuários

Sistemas de GLN podem também ser avaliados de forma intrínseca com base na opinião de usuários. Neste tipo de avaliação, usuários atribuem notas (tipicamente em uma escala de 1 a 5) a aspectos relevantes do texto gerado. Costuma-se utilizar para este fim três critérios principais de avaliação: a facilidade de leitura (ou fluência), a acurácia e a utilidade da informação apresentada. Estes critérios, entretanto, não são mutuamente exclusivos (REITER; BELZ, 2009). Uma variação desta técnica utilizada em Reiter et al. (2005) consiste em mostrar diversas versões de texto (produzidos por diferentes sistemas etc.) e solicitar aos usuários ordená-los por preferência segundo um critério de interesse.

Embora tenha menor alcance do que a avaliação de efetividade discutida na seção anterior, o uso de escalas de satisfação de usuários é uma forma rápida e barata de avaliar sistemas de GLN, e talvez não por acaso seja a forma mais comum de avaliação de sistemas deste tipo (REITER; BELZ, 2009). Além disso, cabe observar que, em certo sentido, a avaliação por usuários considera o sistema como um todo, e não apenas seu desempenho em tarefas isoladas.

2.4.3 Avaliação baseada em corpus

A avaliação de artefatos ou sistemas de GLN baseada em corpus é também uma forma intrínseca de avaliação. Esta técnica consiste tipicamente em comparar o texto gerado com exemplos reais de língua natural produzidos por autores humanos, normalmente organizados na forma de um corpus (ou *gold standard*). Nesta comparação, são utilizadas métricas específicas para o nível de conhecimento linguístico (e.g., semântico ou superficial) que está sendo avaliado.

No caso particular da tarefa computacional de GER, os conjuntos de propriedades semânticas (que formam a saída convencional de um módulo de seleção de conteúdo, cf. seção 2.1.2) são normalmente avaliadas com medidas do grau de semelhança entre conjuntos. Estas medidas incluem os coeficientes Dice (DICE, 1945) ou MASI (PASSONNEAU, 2006) e a acurácia global (ou seja, o número de coincidências totais entre o conjunto de propriedades gerado e o esperado). Estas métricas foram consagradas na pesquisa em GER em iniciativas de avaliação conjunta (ou *shared tasks*) da série de desafios TUNA introduzida em Gatt e Belz (2007) e são hoje consideradas padrão da área.

No caso da tarefa de realização superficial, o uso de um corpus textual como *gold standard*

é especialmente adequado para avaliação da cobertura e correção gramatical do texto gerado de forma automática. Formas superficiais (i.e., *strings* de texto, que são a saída de um modelo deste tipo) são normalmente avaliadas com base na distância de edição de Damerau-Levenshtein (DAMERAU, 1964) ou equivalente. No caso de avaliação de conjuntos de textos, são utilizadas também métricas desenvolvidas para áreas específicas do PLN, como as medidas de contagem de n-gramas BLEU (PAPINENI et al., 2002) e NIST (DODDINGTON, 2002), ambas amplamente utilizadas na avaliação de sistemas de tradução automática. Entretanto, é observado em Reiter e Belz (2009) que métricas como BLEU são mais adequadas para avaliação durante a fase de desenvolvimento do sistema, mas não tanto para fins de avaliação final, já que a correlação com a efetividade da tarefa é geralmente muito baixa.

Assim como em outras áreas do PLN, o uso de métricas automáticas baseadas em *cópus* constitui uma técnica de baixo custo e de forte apelo para a avaliação de sistemas e artefatos de GLN. Esta prática possui no entanto algumas desvantagens bem conhecidas, como a dificuldade de avaliação de propriedades estruturais do texto (SCOTT; MOORE, 2007). Além disso, há a questão de que locutores nem sempre produzem texto ótimo do ponto de vista do ouvinte (OBERLANDER, 1998), e que métricas deste tipo - que são tipicamente orientadas ao locutor⁴ - podem assim reforçar um viés de perspectiva.

2.4.4 Discussão

As técnicas de avaliação aqui discutidas são empregadas em graus variados nos estudos a serem descritos ao longo desta monografia. No entanto, dada a natureza da pesquisa tratada - fortemente voltada para problemas específicos de GLN, e não para o desenvolvimento de sistemas completos deste tipo - há uma forte predominância de técnicas intrínsecas baseadas em *cópus* e com uso de métricas automáticas (LUCENA; PARABONI; PEREIRA, 2010; NOVAIS; PARABONI, 2012; SILVA; PARABONI, 2015; FERREIRA; PARABONI, 2017a).

2.5 Língua natural e personalidade

O tratamento computacional da personalidade humana está no centro do projeto de sistemas ditos inteligentes e, como interesse adicional para o presente trabalho, possui forte relação com a tarefa de produção de língua natural. Esta seção apresenta alguns conceitos fundamentais sobre este tema.

⁴Uma exceção notável é o caso da medida Flesch (FLESCH, 1949), amplamente utilizada em ferramentas de apoio à escrita, mas ainda pouco explorada na avaliação de sistemas de GLN (REITER; BELZ, 2009).

2.5.1 O modelo dos Cinco Grandes Fatores (CGF)

Conforme Allport e Allport (1921), as escolhas linguísticas de um indivíduo - que refletem uma ampla gama de comportamentos humanos como suas emoções e sentimentos - são decorrentes de um conjunto mais estável e reduzido de características ou traços individuais. Estes traços, que representam padrões do comportamento humano em grande parte previsíveis, constituem o que entendemos por personalidade humana.

Traços essenciais de personalidade podem ser determinados com base em diversos métodos propostos pela literatura em Psicologia. Dentre estes, os mais difundidos são os baseados na hipótese lexical, que estabelece que os traços relevantes para caracterização da personalidade são observáveis nas palavras que empregamos ao nos comunicar. Esta abordagem foi refinada a partir de um levantamento inicial de 4.500 traços identificados na década de 30 até produzir, de forma independente e simultânea em diversos estudos, um *framework* estável conhecido como o modelo dos Cinco Grandes Fatores (CGF) ou Big Five (GOLDBERG, 1990).

O modelo CGF contempla cinco dimensões fundamentais da personalidade humana – Abertura à experiência, Meticulosidade, Extroversão, Agradabilidade e Neuroticismo - que são amplamente aceitas como a base adequada para a representação da personalidade humana (ANDRADE, 2008). As dimensões de personalidade CGF são modeladas na forma de valores escalares representando o grau em que um indivíduo expressa ou não determinados traços de personalidade. Assim, por exemplo, um valor muito alto para Extroversão indica um indivíduo muito extrovertido, e um valor muito baixo para esta mesma dimensão indica um indivíduo introvertido.

Os valores das dimensões de personalidade de um indivíduo podem ser estimados por diversos métodos consagrados em Psicologia, sendo o mais comum o uso de inventários de traços de personalidade. Inventários deste tipo são questionários de personalidade aplicados por um profissional da área, ou respondidos de forma autônoma pelo próprio sujeito avaliado.

De especial interesse para o presente trabalho, as dimensões de personalidade do modelo CGF podem ser obtidas de forma relativamente ágil por meio do inventário BFI (JOHN; NAUMANN; SOTO, 2008). O inventário BFI é composto de 44 itens na forma de frases curtas contendo adjetivos que representam cada traço a ser avaliado, como ‘é, às vezes, rude (grosseiro) com os outros’. Cada item do inventário é respondido com um número inteiro variando de 1 (discordo totalmente) a 5 (concordo totalmente), e as respostas de determinados itens são posteriormente combinadas - por meio de uma média com pesos positivos e negativos - para compor um valor escalar representando cada uma das cinco dimensões CGF.

A versão em inglês do inventário BFI foi desenvolvida a partir de análise fatorial de grandes massas de dados e, apesar do número reduzido de itens, é considerada uma medida segura dos atributos mais importantes do modelo CGF sem prejuízo à cobertura ou às boas qualidades psicométricas observadas em inventários mais extensos (JOHN; NAUMANN; SOTO, 2008). Embora haja diversos inventários CGF mais detalhados do que este (e.g., da ordem de centenas de itens, e que permitem a subdivisão dos cinco fatores fundamentais em múltiplas facetas), o inventário BFI é especialmente útil para aplicações computacionais, já que inventários mais detalhados só costumam ter viabilidade prática (e.g., para fins de aprendizagem de máquina etc.) quando é possível coletar dados de um grande número de sujeitos com ampla disponibilidade de tempo e experiência em testes deste tipo (JOHN; NAUMANN; SOTO, 2008).

O inventário BFI tem sido replicado em dezenas de idiomas, incluindo alguns estudos dedicados ao português. Em especial, destaca-se o estudo em Andrade (2008), que validou uma versão em português brasileiro do inventário BFI - denominada IGFP-5 - por meio de uma análise fatorial envolvendo uma amostra de 5.089 respondentes das cinco regiões brasileiras.

2.5.2 O tratamento computacional da personalidade

Em virtude da sua fundamentação linguística, o modelo CGF constitui uma base teórica sólida para uma ampla gama de estudos de PLN/GLN, e o tratamento computacional de traços de personalidade - seja para reconhecimento destes traços a partir de texto, ou para geração de texto adaptado a um conjunto de traços específico - é assim um tema central para desenvolvimento de aplicações deste tipo.

Conhecer os traços de personalidade de um indivíduo (por exemplo, a partir de suas publicações em redes sociais) permite a produção de conteúdo personalizado de várias formas, como para a apresentação de um website de modo a se tornar mais atraente, para a geração de propaganda mais eficaz etc. De forma análoga, saber como um indivíduo com determinados traços de personalidade se expressa em língua natural permite a reprodução deste comportamento em aplicações de GLN, como para a modelagem de personagens realistas de videogames, tutores inteligentes etc. Estas duas frentes de pesquisa do tratamento computacional da personalidade humana - com enfoque em PLN e GLN, respectivamente - são discutidas brevemente a seguir.

2.5.2.1 O reconhecimento de personalidade a partir de texto

O problema computacional de reconhecimento de personalidade humana a partir de texto é uma linha de pesquisa estabelecida do PLN (MAIRESSE et al., 2007; IACOBELLI et al., 2011;

NOWSON; GILL, 2014; NAJIB; CHEEMA; ADEELNAWAB, 2015; PLANK; HOVY, 2015). Estudos deste tipo costumam seguir uma metodologia tradicional de aprendizagem de máquina supervisionada (OBERLANDER; NOWSON, 2006) ou semi-supervisionada (CELLI, 2012) para reconhecimento das dimensões CGF ou de outras características autorais. Este reconhecimento pode assumir a forma de um problema de classificação (e.g., binária, decidindo se um indivíduo é extrovertido ou não), de um problema de regressão (e.g., determinando o valor escalar da dimensão Extroversão do indivíduo) ou ainda de um problema de *ranking* (e.g., ordenando um conjunto de indivíduos segundo uma dimensão de interesse). De modo geral, entretanto, o problema tem sido frequentemente abordado de forma limitada (e.g., considerando apenas um ou dois fatores do modelo CGF), e mesmo assim com resultados relativamente modestos.

O problema computacional do reconhecimento de personalidade a partir de texto é bastante desenvolvido para o caso do idioma inglês. Usando dados como os provenientes de bases públicas como *myPersonality* (KOSINSKI et al., 2015), a área já conta inclusive com alguns eventos científicos dedicados, como a série de ‘*shared tasks*’ CLEF-PAN (RANGEL et al., 2015). Nenhuma destas iniciativas, entretanto, encontra similar para o português brasileiro no contexto do modelo CGF.

2.5.2.2 Geração de texto com base em traços de personalidade

Diversos estudos psicolinguísticos abordam a relação entre personalidade e produção da língua. Do ponto de vista computacional, entretanto, é observado em (MAIRESSE; WALKER, 2011) que a maioria dos sistemas de GLN existentes trata prioritariamente ou exclusivamente da tarefa de produzir uma saída única e gramaticalmente correta que satisfaça os objetivos da comunicação. Nos casos em que variação estilística é considerada (PAIVA; EVANS, 2004; REITER; WILLIAMS, 2010), normalmente não é considerado o objetivo de modelar um locutor específico.

Dentre os poucos estudos de GLN que levam em conta traços de personalidade de um locutor-alvo de interesse, destaca-se o sistema PERSONAGE (MAIRESSE; WALKER, 2010, 2011). Neste estudo é apresentado um método de GLN parametrizável motivado por correlações observadas entre traços de personalidade do modelo CGF e uma ampla gama de decisões de geração. O sistema PERSONAGE é treinado a partir de um *corpus* de textos e inventários de personalidade, e é capaz de expressar uma série de diferenças de personalidade que são perceptíveis por leitores humanos. De especial interesse para a pesquisa a ser discutida na presente monografia, entretanto, cabe destacar que estudos deste tipo não apresentam nenhum tratamento específico para o problema de GER, ou de sua relação com traços de personalidade.

2.5.3 Discussão

O modelo CGF - estimado com base no inventário IGFP-5 (ANDRADE, 2008) - motivará uma interpretação específica do conceito de plausibilidade da língua natural a ser discutida em alguns dos estudos exemplificados ao longo desta monografia. Estes estudos contemplam a questão da relação entre texto e traços de personalidade (SANTOS; PARABONI; SILVA, 2017) e o uso de modelos de personalidade na tarefa computacional de GER (PARABONI; MONTEIRO; LAN, 2017).

2.6 Considerações finais

Neste capítulo foi realizado um levantamento de conceitos fundamentais que serão empregados no restante desta monografia. Em especial, foi abordado o problema computacional da seleção de conteúdo semântico para GER, seus algoritmos, os tipos de corpus disponíveis para estudos desta natureza, e os desafios de pesquisa em aberto.

De forma mais breve, foi discutido também o problema computacional da realização superficial, e técnicas de avaliação intrínseca e extrínseca de sistemas e artefatos de GLN que estão diretamente ligadas ao conceito de plausibilidade da língua gerada de forma automática. Além disso, foi apresentada uma breve introdução à questão da personalidade humana e sua relação com língua natural. Estes conceitos serão retomados no restante desta monografia por meio de uma série de exemplos práticos de desenvolvimento e avaliação de modelos computacionais de GER, de realização superficial, e de construção de corpus de propósito específico para GLN.

Capítulo 3

A PERSPECTIVA DO LOCUTOR

3.1 Visão geral

Uma forma comum - e possivelmente a mais óbvia - de abordar a questão de como gerar língua natural de forma plausível é pela observação de como locutores humanos desempenham a tarefa. Em outras palavras, com base em exemplos de uso real da língua tal qual produzida por locutores humanos, torna-se em princípio possível desenvolver modelos computacionais capazes de reproduzi-la com certo grau de fidelidade.

Esta metodologia - aqui denominada perspectiva do locutor - é tipicamente baseada em análise de córpus para GLN, e guarda certa semelhança com o uso de córpus na pesquisa em PLN. Conforme discutido na seção 2.2, entretanto, o uso de córpus para GLN requer acesso não apenas a exemplos de texto produzidos por locutores humanos, mas também aos contextos originais que motivaram a produção destes exemplos. A depender do objetivo da pesquisa, isso pode assim acarretar menor capacidade de generalização de resultados para domínios para os quais um córpus de GLN não esteja disponível, como frequentemente é o caso.

Apesar destas dificuldades, entretanto, a perspectiva do locutor tem sido popular na pesquisa em GLN, e foi também empregada em diversas ocasiões no trabalho desenvolvido por este autor e seus colaboradores, tanto em estudos de seleção de conteúdo como de realização superficial. Neste capítulo, três questões de pesquisa investigadas sob esta perspectiva serão destacadas: as abordagens computacionais de seleção de conteúdo e realização superficial desenvolvidas à luz da série de competições TUNA de algoritmos de GER, o problema da lexicalização e linearização sentencial (que são as etapas finais da realização superficial) e o problema de geração de referências espaciais em ambientes virtuais interativos tridimensionais.

Os estudos aqui discutidos foram em grande parte desenvolvidos no contexto de dois projetos

de pesquisa regular, a saber: o projeto ‘*Resolução e Interpretação de Expressões de Referência na Geração Automática de Textos*’ (CNPq nro. 484015/2007-9, Edital Universal 15/2007) e o projeto ‘*Estudo e Desenvolvimento de Modelos Híbridos de Geração de Linguagem Natural*’ (Pesquisa regular FAPESP nro. 2009/08499-9). O primeiro projeto é uma colaboração com alunos de graduação, e o segundo inclui uma orientação em nível de mestrado.

3.2 Experiências de GER com a série de desafios TUNA

A perspectiva do locutor em GLN é amplamente representada em estudos de seleção de conteúdo e realização superficial para GER que fazem uso de abordagens baseadas em corpus. Como exemplo destas iniciativas, consideramos nesta seção o caso do corpus TUNA (GATT; SLUIS; DEEMTER, 2007) descrito na seção 2.2, e seu emprego na série de desafios - ou *TUNA shared tasks* - descritos em Gatt e Belz (2007), Gatt, Belz e Kow (2008, 2009).

O corpus TUNA foi o primeiro corpus publicamente disponibilizado para estudo e desenvolvimento de modelos de GER, servindo de base para três competições de algoritmos de GER fortemente orientadas a dados. Nestas competições, o objetivo era desenvolver modelos computacionais capazes de reproduzir expressões de referência semelhantes às dos corpus de exemplos, com grande ênfase no resultado numérico das métricas de avaliação intrínseca adotadas, como Dice (DICE, 1945) e MASI (PASSONNEAU, 2006), e com pouco espaço para discussão dos fundamentos psicolinguísticos destes modelos. Ainda assim, os desafios TUNA atingiram vários objetivos importantes, como a especificação do problema de seleção de atributos que viria a se consolidar na área, o uso de métricas automáticas de avaliação, e uma maior conscientização sobre a necessidade de gerar descrições plausíveis, e não meramente breves ou livres de ambiguidade como era até então enfatizado na pesquisa em GER.

Em uma abordagem de pesquisa típica da época, foram conduzidos por este autor e seus colaboradores diversos estudos baseados no corpus TUNA - incluindo participações nas edições 2008 e 2009 da competição - voltados tanto à tarefa de seleção de conteúdo como de realização superficial em português. Estes estudos, que representam contribuições incrementais para a área de GER, foram apresentadas em Lucena e Paraboni (2008a), Santos, Pereira e Paraboni (2008), Pereira e Paraboni (2008b, 2008a), Lucena e Paraboni (2008b, 2009b) e, de forma mais completa, em Lucena, Paraboni e Pereira (2010), são discutidos a seguir.

3.2.1 Primeiras iniciativas: o desafio TUNA 2008

A participação no desafio TUNA 2008 ocorreu com a colaboração de alunos de iniciação científica em duas modalidades: no desafio de seleção de atributos conforme descrito em Lucena e Paraboni (2008b), e na modalidade aberta da competição, na qual a equipe abordou a tarefa de realização superficial em português conforme descrito em Pereira e Paraboni (2008a).

A abordagem de seleção de atributos apresentada foi essencialmente baseada em uma observação não inteiramente consistente com princípios estabelecidos na área, como as preferências inerentes por certos tipos de atributo (cf. seção 2.1.3). Em linhas gerais, o estudo partiu da observação de que certas propriedades, quando de alguma forma destacadas no contexto, tinham maior probabilidade de serem selecionadas de forma mais ou menos independentemente do que outros tipos de preferências ou restrições poderiam ditar. Um exemplo desta questão é ilustrado na figura 8.

Figura 8: Exemplo com três mesas brancas e grandes, e uma mesa preta e pequena.

Fonte: Ivandré Paraboni, 2017.

Neste exemplo, dado o objetivo de descrever a menor mesa do grupo, pode-se argumentar que há uma certa preferência pelo uso do atributo *tamanho* (e.g., ‘a mesa pequena’), já que seu valor (pequeno) se destaca do valor deste mesmo atributo para os outros objetos da cena (que são grandes). A possível seleção de *tamanho* neste caso, como de fato observada com certa frequência no corpus TUNA, contradiz princípios estabelecidos em estudos de Ciências Cognitivas como os da preferência inerente por cor (PECHMANN, 1989) e, não menos importante para este exemplo específico, contradiz estudos como Gompel et al. (2014), segundo o qual atributos relativos como *tamanho* seriam de uso menos frequente.

Assim como em outros estudos deste gênero, fortemente influenciados pela interpretação contrastiva de Olson (1970) implementada em inúmeras abordagens de GER - incluindo o próprio algoritmo Incremental (DALE; REITER, 1995) - foi estabelecida a hipótese de que esta noção de ‘destaque’ de uma determinada propriedade poderia ser modelada em termos de poder discriminatório¹. Em outras palavras, partindo-se da observação empírica de que propriedades altamente contrastivas tendem a ser favorecidas por locutores humanos, a busca por descrições

¹Uma caracterização mais precisa deste fenômeno em termos de *saliência* seria posteriormente proposta em Paraboni, Galindo e Iacovelli (2017), conforme será discutido na seção 4.2.

plausíveis levou, neste primeiro momento, a priorizar a seleção de propriedades discriminatórias, como o caso de uso do atributo *tamanho* neste exemplo.

Como esperado, os resultados iniciais desta abordagem - divulgados no relatório final da competição em Gatt, Belz e Kow (2008) - foram relativamente modestos, mas ainda assim suficientes para colocá-la no grupo de sistemas classificados em terceiro lugar na competição (de um total de 15 participantes). Os sistemas de melhor desempenho, por outro lado, eram fortemente orientados a dados, fazendo uso de métodos de aprendizagem de máquina para reprodução dos exemplos fornecidos (o que é de certa forma o esperado em se tratando de um evento deste tipo), e talvez com menor ênfase na investigação dos princípios de produção humana da língua.

Com relação à tarefa de realização superficial de expressões TUNA, o trabalho em Pereira e Paraboni (2008a) apresentou uma primeira proposta de uso de modelos de n-gramas aplicados à geração destas expressões em português, obtidas por meio da tradução do corpus TUNA original. Este trabalho figurou como participante da categoria aberta do evento, sem de fato participar da competição já que não utilizava os mesmos dados (originalmente em inglês). A abordagem neste caso era bastante tradicional - baseada na geração de todas combinações possíveis de sequências lexicais - posteriormente filtradas com uso de um modelo de n-gramas.

3.2.2 Refinamentos para o desafio TUNA 2009

De forma conjunta, os modelos de seleção de conteúdo e realização superficial desenvolvidos para o Desafio TUNA 2008 foram posteriormente refinados em Lucena e Paraboni (2008a) com a inclusão de novas restrições e preferências. Na edição subsequente do desafio TUNA em Gatt, Belz e Kow (2009), esta mesma abordagem integrada - então denominada USP-EACH - foi reapresentada com pequenos aprimoramentos. Por exemplo, a seleção de atributos passou a explorar uma condição experimental adicional do corpus TUNA que não havia sido tratada na versão do ano anterior, a qual dividia o conjunto de dados entre situações em que os locutores foram informados que podiam usar as coordenadas da tela ou não (e.g., ‘a mesa no canto superior esquerdo’), em uma tentativa de fornecer uma solução mais pragmática (a portanto mais próxima dos sistemas competidores) do que a apresentada no evento anterior.

Os benefícios desta nova implementação não são muito evidentes na avaliação intrínseca realizada no desafio TUNA 2009, análise que é complicada pelo fato de envolver um total de 12 medidas automáticas para o conjunto como um todo, e individualizada para os subdomínios *Furniture* e *People* - e também pelo fato de que diferentes participantes apresentam grande variação de desempenho dependendo da métrica ou do domínio em consideração. De modo

geral, entretanto, o sistema USP-EACH ocupa uma posição mediana entre os resultados da competição, relatados em Gatt, Belz e Kow (2009), e teve resultados especialmente positivos para o subdomínio *Furniture*, sendo neste caso classificado como o melhor da competição de acordo com a medida NIST, e segundo lugar segundo as métricas de Acurácia e BLEU (PAPINENI et al., 2002).

Resultados mais expressivos foram, por outro lado, observados na avaliação extrínseca (com auxílio de juízes humanos) na mesma competição, que levou em conta os quesitos de adequação e fluência das expressões geradas. Considerando-se os dois subdomínios em conjunto, o sistema USP-EACH foi o segundo colocado geral em ambos os quesitos, e manteve esta posição no caso do subdomínio *People*. No caso do subdomínio *Furniture*, o resultado foi ainda melhor: além da segunda colocação no quesito adequação, o sistema USP-EACH foi o primeiro colocado da competição na categoria fluência. Este resultado é especialmente importante se considerarmos o valor da avaliação extrínseca em GLN (REITER; BELZ, 2009) e a correlação relativamente baixa entre métricas intrínsecas e preferências humanas (GATT; BELZ; KOW, 2009).

3.2.3 Contribuições

Refinamentos subsequentes das estratégias de seleção de conteúdo e realização superficial aqui discutidas foram sumarizados em Lucena, Paraboni e Pereira (2010). Embora esta abordagem seja claramente limitada sob vários aspectos importantes - em especial, por ter sido validada apenas nos dois subdomínios TUNA - este estudo representa, até o momento, uma das poucas iniciativas da área a combinar as duas tarefas em uma solução única, e é possivelmente o primeiro e único estudo a considerar a questão da realização superficial de expressões de referência no português Brasileiro.

Do ponto de vista do problema científico de que trata o presente trabalho - a questão da geração de expressões de referência plausíveis - o estudo em Lucena, Paraboni e Pereira (2010) é representativo da visão dominante na época, segundo a qual a plausibilidade era medida principalmente pela comparação com exemplos produzidos por locutores humanos na forma de *cópus* para GER. A relação entre o desempenho do sistema e o desempenho humano é explorada tanto de forma indireta, comparando-se propriedades semânticas selecionadas por algoritmos de GER com as propriedades que representam as expressões produzidas por humanos (e que são provenientes de anotação manual destas expressões), como de forma direta, ou seja, comparando-se *strings* gerados na tarefa de realização superficial com os exemplos produzidos por locutores humanos.

O principal legado das experiências realizadas no contexto dos desafios TUNA foi o envolvi-

mento de alunos de graduação na linha de pesquisa em GLN em um período em que a própria instituição ainda estava em fase de implantação. Estas iniciativas proporcionaram a inserção do grupo em dois eventos que movimentaram a comunidade científica internacional da área e auxiliaram na consolidação da linha de pesquisa em GER nesta instituição, com resultados que podem ser considerados positivos dadas as circunstâncias e os recursos então disponíveis.

3.3 Realização superficial em português

Talvez pela sua maior proximidade com o resultado final do sistema de GLN (ou seja, com o texto propriamente dito), o problema computacional de realização superficial é frequentemente abordada sob a perspectiva do locutor. Em outras palavras, uma vez que o objetivo de um realizador textual é o de gerar *strings* textuais, é razoável supor que este texto deva ser semelhante ao texto que um locutor humano produziria nas mesmas circunstâncias. Assim, a tarefa de realização superficial torna-se um exemplo bastante típico da perspectiva do locutor em GLN, e será tratada aqui com maior detalhamento do que na seção anterior, onde era limitada ao caso específico de expressões de referência.

A partir dos resultados preliminares com modelos estatísticos aplicados à tarefa de realização superficial em estudos como Pereira e Paraboni (2007a), Pereira e Paraboni (2008b) e Lucena, Paraboni e Pereira (2010) descritos na seção anterior, evidenciou-se a oportunidade de desenvolver modelos computacionais mais sofisticados para esta tarefa, e considerando a realização de texto livre (i.e., tratando da geração de sentenças completas). Assim, diversos experimentos preliminares desta natureza foram apresentados em Tadeu, Novais e Paraboni (2010), Novais, Tadeu e Paraboni (2010a, 2010b, 2010c), Novais, Paraboni e Ferreira (2011), Novais, Paraboni e Silva Junior (2012) e, de forma mais completa, em Novais e Paraboni (2012). Este último estudo será aqui utilizado para ilustrar a perspectiva do locutor na tarefa computacional de realização superficial ou, de forma mais específica, na etapa de linearização sentencial que trata da ordenação dos itens lexicais que compõe a sentença.

3.3.1 O uso de modelos fatorados para linearização sentencial

Partindo de experiências de construção de *cópus* em português do Brasil em um projeto anterior², o estudo em Novais e Paraboni (2012) trata especificamente da produção de texto neste idioma, e constitui a única contribuição puramente dedicada à tarefa computacional de realização

²‘Construção de um *Cópus* Paralelo com Informações de Correferência para Tradução Automática’ (Pesquisa regular FAPESP nro. 2006/03941-7).

superficial a ser tratada no presente trabalho. Como frequentemente é o caso em se tratando de sistemas deste tipo, o estudo em Novais e Paraboni (2012) exemplifica a perspectiva do locutor na pesquisa em GLN em seu sentido mais literal. Neste caso, a plausibilidade da língua gerada automaticamente é medida pela comparação direta das formas superficiais (sentenças) geradas pelo sistema com sentenças produzidas por autores humanos tal qual disponibilizadas em um corpus de referência em manchetes jornalísticas em português.

O estudo em Novais e Paraboni (2012) consiste de um protótipo de sistema híbrido de realização superficial para o português Brasileiro que combina *templates* sintaticamente estruturados e modelos estatísticos de língua fatorados (*Factored Language Models*, ou FLMs) (BILMES; KIRCHHOFF, 2003). Estes modelos, que foram responsáveis por uma parte considerável do sucesso dos sistemas de tradução automática estatística na última década, podem ser vistos como uma generalização da noção de frequências de n-gramas e *back-off* para múltiplos fatores paralelos e customizáveis.

De forma sucinta, em um FLM uma palavra é vista como um vetor de k atributos (ou fatores) na forma $f_1, f_2, \dots, f_K$. Um fator pode representar qualquer tipo de conhecimento, incluindo classes morfológicas, lemas ou quaisquer atributos de línguas morfológicamente ricas como o português, ou ainda atributos semânticos úteis também para línguas menos ricas, como o inglês. Para detalhes sobre a fundamentação teórica de FLMs e *back-off* paralelo generalizado, ver Bilmes e Kirchhoff (2003), e também a literatura em tradução automática estatística (KOEHN; OCH; MARCU, 2003; KOEHN et al., 2007).

3.3.2 Arquitetura

A arquitetura utilizada em Novais e Paraboni (2012) é um exemplo tradicional de geração de dois estágios dividida em um núcleo central para processamento dos dados de entrada e geração permissiva, e um módulo para consulta a modelos estatísticos de língua. Esta arquitetura é ilustrada na figura 9, e será apresentada em detalhes na seção a seguir como forma de ilustrar o papel dos modelos do tipo FLM, que são o verdadeiro ponto de interesse no estudo em Novais e Paraboni (2012).

Nesta proposta, o módulo de geração permissiva recebe como entrada a especificação de uma sentença na forma de um *template* da realização desejada (a ser discutido a seguir) e, consultando um dicionário desenvolvido a partir da base em Muniz (2004) e possíveis valores *default* para cada termo, determina o gênero e número de cada sintagma e, no caso dos sintagmas verbais, o modo e tempo a ser empregado.

Figura 9: Arquitetura do modelo de realização superficial.

Fonte: Novais e Paraboni (2012).

Uma vez feitas estas definições globais, o sistema determina a classe gramatical de cada item da sentença e produz a forma superficial correspondente considerando o gênero, número, modo e tempo verbal a serem empregados. Além disso, regras adicionais associadas ao *template* selecionado garantem concordância entre sujeito e predicado. Finalmente, cada palavra (W) é também acrescida de seu lema (L), *part-of-speech* (P), gênero (G) e número (N) para definição de fatores a serem considerados pelo modelo estatístico, informações estas obtidas também a partir do dicionário implementado em Muniz (2004).

O resultado desta etapa são listas não ordenadas de palavras e fatores adicionais correspondendo a cada sintagma da sentença a ser gerada. Estas listas são então permutadas de forma exaustiva, ou seja, produzindo-se todas as suas linearizações possíveis, e o resultado é submetido a um filtro estatístico. Mais especificamente, o módulo FLM submete cada permutação da sentença a um modelo estatístico treinado em uma grande massa de dados, e seleciona a alternativa de maior probabilidade para constituir a sentença de saída do sistema, ao passo que as demais alternativas são descartadas.

3.3.3 Especificação de entrada

Conforme discutido na seção 2.3, a questão de qual tipo de especificação de entrada adotar é crucial para o projeto de um artefato de realização superficial. Como forma de tornar o sistema potencialmente mais adaptável a futuras aplicações, foi proposto o uso de um formato de

entrada simplificado semelhante ao padrão de entrada dito ‘superficial’ adotado na competição *Generation Challenges 2011 Surface Realisation Shared Task*³, embora o sistema em questão não tenha participado da competição propriamente dita em virtude de não considerar a língua-alvo do evento (inglês)⁴.

Neste formato de entrada, assume-se que a escolha lexical já foi realizada, e que a forma de apresentação (e.g., seleção de um *template* de geração em voz ativa, passiva etc.) já foi definida, restando basicamente efetuar as operações de concordância entre os diversos termos da sentença e, o mais importante, a ordenação linear destes termos. Além destas operações, o presente sistema é responsável ainda pelo fornecimento de valores *default* em caso de entrada subespecificada.

Na proposta apresentada, foi implementado apenas um *template* básico para geração de sentenças simples em voz ativa (e.g., manchetes de jornal, conforme as empregadas na avaliação do sistema descrito a seguir). Este *template* consiste de uma árvore sintática contendo os termos *agent*, *action* e, opcionalmente, *patient*, sempre nesta mesma ordem fixa. Os termos *agent* e *patient* de uma sentença são conceitos a serem realizados como sintagmas nominais, enquanto que o termo *action* representa um sintagma verbal.

Dado o propósito de gerar texto de forma permissiva (e sem uso de regras gramaticais), sintagmas nominais e verbais são modelados simplesmente como conjuntos (não ordenados) de palavras, cabendo ao modelo subjacente a tarefa de impor a ordenação correta a cada termo. Assim, sintagmas nominais são conjuntos de substantivos, nomes próprios, adjetivos e sintagmas preposicionais, enquanto que sintagmas verbais são conjuntos de verbos, preposições e advérbios. Em ambos os casos a estrutura inclui a especificação de um termo núcleo (*head*) do sintagma representando o seu conceito central (nome ou verbo principal) ao qual os demais termos estão subordinados.

O Exemplo 1 a seguir apresenta um sintagma nominal e sua especificação segundo este formato, e cujo núcleo foi definido como *escritor*.

Ex.1: A escritora sul-africana Nadine Gordimer

concept(*s1,obj*, [*s*, [‘sul-africano’], *head*(‘escritor’), [‘nadine=gordimer’], *def*, *f*]).

Nesta especificação observa-se que o conceito (ou objeto *obj*) é representado por uma lista não ordenada de palavras de conteúdo, com indicação do núcleo do sintagma (*head*) e informações adicionais sobre o tipo de determinante a ser empregado (e.g., *def* corresponde a

³<http://www.nltg.brighton.ac.uk/research/sr-task/>

⁴Uma versão estendida deste formalismo é discutida em King e White (2016).

uma descrição definida) e informações de gênero e número (*f* para feminino e *s* para singular, respectivamente) referentes ao núcleo, e que no caso do sintagma nominal serão impostas aos demais componentes da estrutura.

Exceto pela declaração do núcleo, todos os demais termos são opcionais. Na falta de uma instrução sobre o tipo de determinante a empregar, o sistema gera uma descrição definida. Se as informações de gênero ou número forem omitidas, o sistema tentará inferi-las com base nos traços morfológicos dos outros termos do sintagma. Nota-se assim que neste exemplo não seria necessário informar o gênero feminino (*f*), já que um de seus componentes é o nome próprio ‘Nadine Gordimer’, que consta na base lexical do sistema apenas na forma feminina, impondo-se a todos os componentes do sintagma que tenham flexão de gênero (em especial, o termo ‘escritor’ será corretamente realizado como ‘escritora’).

Por razões de eficiência, esta especificação de entrada assume que sintagmas preposicionais inseridos na expressão principal com uso de preposições já estejam em sua forma final. O Exemplo 2 a seguir ilustra uma definição de conceito representado por sintagma nominal contendo um termo aninhado deste tipo.

Ex.2: *programa de rádio*

concept(p2,obj,[m,s,rel(['de','rádio']),head('programa')]).

Expressões como ‘programa de rádio’ acima possuem assim um núcleo a ser transformado em *string* (efetuando-se flexão de gênero e número, que no exemplo acima resultaria na mesma palavra ‘programa’) e um subcomponente já realizado (‘de rádio’) representados por cláusulas do tipo *rel* (relação). Cabe observar entretanto que estes subcomponentes poderiam também ser gerados pelo próprio sistema de forma recursiva.

Finalmente, um exemplo de especificação de sintagma verbal é apresentado a seguir. Sintagmas verbais podem incluir advérbios e preposições. Caso o tempo ou modo verbal não sejam especificados, são utilizados como valores *default* o pretérito simples do modo indicativo, já que o domínio considerado foi o de notícias jornalísticas (via de regra relatando eventos já ocorridos).

Ex.3: *ainda não terá*

concept(v3,act,[ind,['ainda'],fut,head(['ter']),vfin,['não']]).

Com base nestas definições, o trabalho do sistema consiste em receber uma especificação de sentença como entrada e produzir o *string* correspondente de saída. Um exemplo de especificação de sentença em voz ativa (agente - ação - paciente) é apresentado a seguir. Os identificadores *s1*, *v2* e *p3* referem-se aos conceitos definidos nos exemplos 1-3 anteriores, que no caso do *template* de voz ativa são tratados como agente, ação e paciente da sentença, respectivamente:

Ex.4: A escritora sul-africana Nadine Gordimer ainda não terá programa de rádio.
sentence(u123,[agent(s1),action(v2),patient(p3)]).

3.3.4 Modelos propostos

Como forma de melhorar o desempenho observado na fase experimental do projeto, os modelos estatísticos empregados em Novais e Paraboni (2012) foram treinados com base em uma grande massa de dados composta do *corpus* NILC (NUNES et al., 1996) ampliado com os artigos online do período de janeiro de 2006 a fevereiro de 2011 da Folha de São Paulo⁵ e da revista *Veja*⁶. Após uma série de tarefas de pré-processamento para remoção de ruído (e.g., código HTML, trechos de artigos em outros idiomas etc.), duplicidades e erros diversos, foi obtido um *corpus* de treinamento de cerca de 141 milhões de palavras.

O *corpus* de treinamento foi etiquetado com a ferramenta MXPOST disponibilizada pelo projeto LACIO-WEB⁷ e teve parte de suas etiquetas parcialmente corrigidas de forma semiautomática com base em uma amostragem de erros frequentes (em especial, para correta identificação de nomes próprios). A seguir, o *corpus* foi enriquecido com a inclusão automática dos fatores palavra (*W*) e *part-of-speech* (*P*) extraídos dos próprios dados etiquetados, e os fatores adicionais de lema (*L*), gênero (*G*) e número (*N*) obtidos através do dicionário disponibilizado em Muniz (2004).

O conjunto de dados assim preparado foi empregado no treinamento de diversos modelos estatísticos de língua, incluindo modelos tradicionais baseados em contagens de *n*-gramas e modelos fatorados (FLMs). Cabe observar entretanto que, na prática, a distinção entre estas duas categorias é irrelevante, já que o modelo de *n*-gramas tradicional nada mais é do que o caso especial de fator único *W*.

3.3.5 Avaliação

A avaliação apresentada em Novais e Paraboni (2012) levou em conta sete implementações do sistema proposto - baseadas em modelos de *n*-gramas ou FLMs usando bigramas, trigramas e quadrigramas com diferentes combinações de fatores - e três sistemas de *baseline*, incluindo um realizador superficial baseado em regras (Silva Junior; PARABONI; NOVAIS, 2013) que constitui uma versão do sistema *SimpleNLG* em Gatt e Reiter (2009) para o português Brasileiro.

⁵<http://www.folha.uol.com.br/>

⁶<http://veja.abril.com.br/>

⁷<https://sites.google.com/site/linguacorporus/lacioweb>

O conjunto de dados de teste foi composto de 4.297 manchetes de jornal online do ano de 2009 considerando-se uma porção distinta da utilizada como *córpus* de treinamento. Estas sentenças sofreram um processo de abstração semelhante ao adotado na preparação dos dados adotado no já citado *Surface Realisation Shared Task*, no qual cada palavra de conteúdo foi substituída pela sua forma-base, e a estrutura da sentença foi representada na forma de árvores de dependências com ordem linear aleatória.

Cada uma das 4.297 entradas foi fornecida ao sistema para geração permissiva e seleção estatística segundo os vários modelos considerados e sistemas de *baseline*. Para todos os sistemas baseados em FLM, os resultados em Novais e Paraboni (2012) são significativamente superiores aos obtidos pelo uso de modelos de *n*-gramas tradicionais (incluindo o mais completo, baseado em quadrigramas). Entretanto, não houve benefício em considerar informações de gênero (*G*) e número (*N*) nestes modelos, e o conjunto de sistemas baseados em FLMs como um todo apresentou desempenho estatisticamente equivalente (embora houvesse uma melhoria discreta proporcional à complexidade do modelo). Assim, concluiu-se em Novais e Paraboni (2012) que o FLM mais simples (do tipo bigrama e considerando apenas a palavra *W* e seu lema *L*) seria suficiente para a presente tarefa, e portanto preferível aos modelos de ordem superior e/ou levando em conta o fator adicional de *part-of-speech P*.

3.3.6 Contribuições

O trabalho em Novais e Paraboni (2012) apresentou como contribuição principal um protótipo de sistema de geração de sentenças completas em português, e constitui uma das poucas iniciativas do gênero até o presente. Este estudo é um exemplo típico de abordagem de GLN sob a perspectiva do locutor, e talvez o que mais se aproxime das técnicas baseadas em *córpus* normalmente empregadas na pesquisa em PLN. Cabe destacar, entretanto, que houve neste caso uma etapa inicial de construção do *córpus* nos moldes exigidos para GLN, ou seja, contendo não apenas os exemplos de texto a serem gerados, mas uma representação simplificada do conhecimento a partir do qual estes textos poderiam ser gerados. No presente caso, esta representação foi obtida por um processo de abstração das sentenças-alvo do *córpus* textual original, e então tomada como entrada para a tarefa de realização superficial. Não é claro, entretanto, se ou como este procedimento poderia ser aplicado a outras tarefas de GLN que façam uso de conhecimento mais profundo.

Finalmente, observa-se que estudos de realização superficial deste tipo incorrem na questão da dependência da língua e, conseqüentemente, na dificuldade de comparação com estudos similares desenvolvidos para o inglês ou outros idiomas. Assim, como forma de tentar produzir

pesquisa em GLN com maior potencial de inserção internacional, esta linha de pesquisa foi, salvo algumas iniciativas isoladas, pouco explorada nos anos posteriores à conclusão do projeto-base, e diversas oportunidades de melhoria - a começar pelo próprio desenvolvimento completo de um sistema deste tipo - permanecem em aberto.

3.4 Geração de relações espaciais em mundos virtuais

Com a evolução natural dos desafios de GER baseados em cenas estáticas bidimensionais como as do *cópus TUNA* (GATT; SLUIS; DEEMTER, 2007) para situações de comunicação mais realistas, a série de desafios GIVE (BYRON et al., 2007) proporcionou uma infraestrutura computacional completa para estudo e desenvolvimento de sistemas para geração de instruções de navegação em mundos virtuais tridimensionais (cf. seção 2.2). O sistema GIVE implementa uma aplicação do tipo ‘caça ao tesouro’ na qual o usuário navega em um mundo virtual interagindo com certos objetos (em especial, botões para abrir portar e destravar alarmes etc.) até atingir o objetivo final (que era abrir um cofre e recolher a recompensa). Este sistema também foi empregado em três desafios computacionais ou *shared tasks* da série GIVE (BYRON et al., 2009; KOLLER et al., 2010; STRIEGNITZ et al., 2011).

Um exemplo de mundo virtual GIVE é ilustrado pela figura 10 contendo quatro botões interativos.

Figura 10: Um mundo virtual do tipo GIVE contendo quatro botões interativos (da esquerda para direita: azul, verde, vermelho e verde).

Fonte: Silva e Paraboni (2015).

A infraestrutura GIVE permite a criação de mundos virtuais customizáveis, e faz uso de um módulo planejador para gerar uma sequência de passos que o usuário deve seguir para alcançar o objetivo do jogo. Ao desenvolvedor de GLN cabe assim transformar o plano fornecido pelo sistema em instruções de navegação em língua natural, como ‘Aperte o botão azul’, ‘Dobre à

direita e abra a porta ao fundo’ etc.

O conjunto de ferramentas oferecido pelo ambiente GIVE permite que sistemas geradores de instruções de navegação sejam projetados de diversas formas. Dentre estas, uma alternativa natural seria adotar a mesma perspectiva em discussão neste capítulo, ou seja, considerar a perspectiva do locutor e produzir instruções que se aproximem daquelas que seriam produzidas por locutores humanos em situações semelhantes. No caso do domínio GIVE, esta perspectiva seria inclusive favorecida pela disponibilidade do *cópus* GIVE-2 de instruções humanas (GARGETT et al., 2010) descrito na seção 2.2. No entanto, a natureza de aplicações deste tipo - baseadas em contextos de comunicação interativos e altamente dinâmicos - implica que qualquer *cópus* deste tipo seja necessariamente limitado a um número relativamente reduzido de exemplos de expressões produzidas em contextos muito específicos. Diferentes mundos virtuais podem motivar a produção de expressões distintas, e mesmo dentro de um ambiente único o mesmo objeto-alvo pode ser visualizado sob múltiplas perspectivas (e.g., com mais ou menos distraidores visíveis etc.), características estas que dificultam a generalização de observações feitas a partir de *cópus*.

Com base nestas considerações, observamos que o desenvolvimento de sistemas de geração de instruções de navegação em mundos virtuais sob a perspectiva do locutor não pode ser facilmente desenvolvido apenas com base na reprodução de exemplos de uso da língua. Uma possível solução neste caso seria a combinação de conhecimento proveniente de um *cópus* no domínio em questão (tratado assim como referência ou *gold standard* da mesma forma que nos estudos das seções anteriores) com conhecimento proveniente de outras fontes. Este tipo de perspectiva híbrida, que considera o comportamento linguístico de locutores humanos e modelos ‘externos’, é a abordagem adotada em um conjunto de estudos apresentados de forma incremental em Silva e Paraboni (2013, 2014) e, na sua forma mais completa, em Silva e Paraboni (2015). Estes estudos serão aqui tomados como exemplo da perspectiva do locutor em GLN para o caso específico do problema computacional da geração de relações espaciais em mundos virtuais interativos.

3.4.1 O problema da geração de relações espaciais

O objetivo do estudo em Silva e Paraboni (2015) foi o desenvolvimento de modelos computacionais de referência espacial para produção de expressões como ‘Pressione o botão verde, *ao lado da planta*’, nas quais poderia haver ou não referência a outros objetos do contexto com uso de relações espaciais (e.g., ‘ao lado de’). Apesar da aparente similaridade com o problema de GER, entretanto, é importante observar que o problema de geração de expressões envolvendo

relações espaciais não se limita a uma aplicação de algoritmos deste tipo. Em situações de comunicação como esta, relações espaciais são empregadas tanto no sentido mais tradicional de GER - para auxiliar a identificação de um objeto-alvo de outra forma ambíguo, como no exemplo anterior - como também para reduzir o espaço de busca por um objeto-alvo já previamente identificado. Neste segundo caso, a relação espacial seria superespecificada, como em ‘o botão azul *ao lado do verde*’, o que torna o problema de GER potencialmente complexo já que não é óbvio quais propriedades podem ser acrescentadas desta forma sem prejuízo à compreensão ou geração de falsas implicações lógicas (GRICE, 1975).

Uma forma de entender estas dificuldades é considerando-se o objetivo do algoritmo que faz uso de relações espaciais, e o tipo de conhecimento disponível para a tomada de decisão. No caso de um algoritmo de GER tradicional, o objetivo é o de identificar o objeto-alvo. Neste caso, o algoritmo pode estimar com certa facilidade se uma propriedade (espacial ou outra qualquer) é útil ou não para atingir este objetivo examinando considerando, por exemplo, o poder discriminatório ou saliência relativa da propriedade em questão. Por outro lado, o propósito de facilitar a busca pelo objeto-alvo (por exemplo, em um mundo virtual estruturalmente complexo) coloca inúmeras questões sobre como selecionar estas propriedades e de que forma utilizá-las. Neste segundo tipo de problema há assim uma combinação de elementos de planejamento da comunicação, e também aspectos tradicionais de GER, ao ponto de delimitar linhas de pesquisa de comunidades relativamente distintas em computação e psicologia (TENBRINK, 2005; BARCLAY, 2010).

Além destas diferenças, o próprio uso de algoritmos de GER em aplicações deste tipo evidencia diversas questões que normalmente não são abordadas pela pesquisa na área de GER, como a definição de quais objetos constituem o contexto da referência a ser considerado em um determinado instante da navegação, de como as relações espaciais entre estes objetos podem ser computadas, e de quais relações espaciais, dentre um amplo espectro de possibilidades, são de fato plausíveis em uma situação de comunicação com um interlocutor humano.

3.4.2 Modelos propostos

Como forma de avançar algumas das questões de pesquisa sugeridas na seção anterior, o estudo em Silva e Paraboni (2015) propõe um modelo computacional de geração de relações espaciais para identificação de objetos em mundos virtuais do tipo GIVE. Assim como em outras abordagens de GER, o objetivo do estudo foi a produção de expressões livres de ambiguidade (por exemplo, para que não haja confusão com o outro botão verde da cena) e plausíveis (observando-se os diversos tipos de relação espacial produzidos por locutores humanos em

diferentes situações).

A abordagem proposta em Silva e Paraboni (2015) consiste de uma combinação de modelos computacionais de referência espacial propostos em Kelleher e Costello (2009) adaptados para mundos virtuais do tipo GIVE, e uma abordagem de aprendizagem de máquina baseada no corpus GIVE-2 (GARGETT et al., 2010). O princípio fundamental desta abordagem é a distinção entre relações espaciais *possíveis* - que são todas aquelas que observamos entre todos pares de objetos em uma cena - e as relações espaciais que são de fato *plausíveis* para fins de referência em uma determinada situação de comunicação. Por exemplo, no contexto da figura 10 anterior, talvez seja possível afirmar que o botão azul (na extrema esquerda) está de frente para a planta. Trata-se assim de uma relação espacial possível, e que pode ser computada a partir da disposição física dos objetos no mundo virtual. No entanto, como a distância entre estes dois elementos é considerável, e também afetada pela presença de outros objetos do tipo botão localizados mais próximos da mesma planta, observamos em exemplos práticos extraídos de corpus GIVE-2 que este tipo de relação espacial não é frequente no uso comum da língua. Dizemos assim que, embora possível, a relação espacial entre este botão e a planta não é válida - ou plausível - para fins de referência.

A separação entre descrições possíveis e descrições possíveis e plausíveis é refletida na arquitetura da solução proposta ilustrada pela figura 11.

Figura 11: Arquitetura do modelo de geração de relações espaciais.

Fonte: Silva e Paraboni (2015).

Nesta proposta, o módulo 1 (Extração) é responsável por computar uma ampla gama de relações espaciais possíveis entre todos pares de objetos do contexto em questão. Para este fim,

são utilizados princípios gerais de referência espacial propostos em Kelleher e Costello (2009) e as informações de localização física dos objetos disponibilizadas pelo modelo do mundo virtual.

O módulo 2 (Classificação) é encarregado de selecionar, dentre todas relações espaciais observadas em uma dada cena, aquelas que de fato ocorrem no uso comum da língua. Como forma de manter um certo grau de independência do domínio, a solução proposta faz uso de uma abordagem de aprendizado de máquina baseada no *córpus* GIVE-2 (GARGETT et al., 2010). Esta abordagem é semelhante à adotada em Viethen (2011) com base nos *córpus* GRE3D3/7, mas fazendo uso de atributos de aprendizagem distintos e de interesse específico para o presente domínio. A estratégia adotada consiste de um classificador binário que determina, para um dado objeto-alvo e um ponto de referência (ou *landmark*) candidato em um mesmo contexto, se há ou não uma relação espacial válida para fins de referência entre estes dois objetos.

Na construção do modelo de classificação, foram utilizadas como exemplos de uso da língua as referência a objetos do tipo botão encontradas no *córpus* GIVE-2 descrito na seção 2.2. Botões são os únicos elementos manipuláveis pelo usuário do ambiente GIVE e, conseqüentemente, constituem o tipo de objeto-alvo com maior número de referências nas instruções do *córpus*. Neste procedimento, foram extraídas todas as instruções do *córpus* contendo a palavra ‘button’ (823 casos) e formas equivalentes como ‘this’ (119 casos) e ‘box’ (50 casos) inferidas pelo uso de verbos como ‘press’ e ‘hit’. No total, foram identificadas 992 descrições de interesse. Destas, 467 (47,07%) utilizam algum tipo de relação espacial, sendo 248 (25%) do tipo topológica (e.g., ‘O botão *próximo* à planta’), e 219 (22,07%) do tipo projetiva (e.g., ‘O botão *à direita* da porta’) (KELLEHER; COSTELLO, 2009).

As descrições coletadas foram anotadas de forma automática com propriedades semânticas obtidas da própria representação do mundo virtual (como cores, tamanhos etc.), e acrescidas de anotação manual para os tipos de objetos que não possuem este tipo de informação disponível (e.g., a cor de objetos exceto do tipo botão). Finalmente, o conjunto de dados resultante foi utilizado para treinamento do modelo que determina se cada relação espacial considerada é de fato ‘válida’ naquele contexto de comunicação. Nos casos em que todas relações espaciais possíveis são suficientemente referenciadas no *córpus*, a saída deste módulo pode ser o mesmo conjunto de entrada, indicando que todas relações possíveis são também válidas. Por outro lado, relações espaciais pouco representativas tendem a ser descartadas neste processo, reduzindo-se assim o número de possibilidades a serem consideradas na tarefa de GER propriamente dita.

O módulo 3 (Seleção) recebe como entrada as relações espaciais válidas fornecidas pelo módulo anterior, e estabelece um *ranking* de preferência entre elas. O objetivo desta seleção é o de eleger uma única relação espacial como preferencial a ser considerada pelo módulo de

GER subsequente. Esta estratégia é motivada pela observação prática de que expressões mais complexas, embora corretas do ponto de vista linguístico, podem ser de difícil interpretação, como ‘o botão vermelho à direita do botão verde e à esquerda da planta’, além de não ocorrerem com frequência no corpus GIVE-2 (GARGETT et al., 2010).

Ao contrário do módulo anterior, a seleção da relação espacial preferencial neste caso não faz uso de métodos de aprendizagem de máquina, pois o número de candidatos tende a ser reduzido neste estágio da execução, e não viabilizaria o uso de uma abordagem orientada a dados. Ao invés disso, o módulo de Seleção faz uso das equações de proximidade absoluta e relativa propostas em Kelleher e Costello (2009), as quais foram adaptadas à escala do mundo virtual GIVE para implementar a tarefa de seleção de uma única relação espacial preferencial, que constitui assim a saída deste módulo.

Finalmente, o módulo 4 (GER) recebe como entrada um objeto-alvo r , o conjunto de propriedades atômicas extraído do contexto e a relação espacial definida como preferencial para aquele contexto, e produz então uma descrição L não-ambígua de r , a qual pode ou não incluir a relação espacial sugerida. Este módulo é uma extensão do algoritmo Incremental em Dale e Reiter (1995) para manipulação de propriedades relacionais (e não apenas atômicas, como no algoritmo original).

3.4.3 Avaliação

A abordagem em Silva e Paraboni (2015) suscita diversas questões sobre como propriedades atômicas e relacionais podem ou devem ser combinadas de modo a produzir uma descrição bem-formada (e.g., que pode ser realizada na forma de uma descrição linguística simples e fluente). Em especial, faz-se necessário definir uma *estratégia de seleção* específica (por exemplo, como a sugerida pelo modelo proposto, ou talvez simplesmente selecionar a relação espacial mais frequente no corpus de treinamento), decidir a *ordem* em que as propriedades possíveis serão consideradas pelo algoritmo de GER (cf. seção 2.1.2), decidir a *prioridade da relação espacial* dentro da ordem estabelecida, e decidir como tratar relações espaciais com maior ou menor *poder discriminatório*, ou seja, decidir se uma relação espacial deve ser incluída mesmo que seja redundante (e.g., ‘o botão vermelho *próximo ao verde*’), ou apenas quando necessária para desambiguação (e.g., ‘o botão verde *próximo à planta*’).

Para investigação sistemática destas questões, e assim oferecer respostas conclusivas para cada uma delas, foram consideradas 54 combinações de parâmetros para o algoritmo de GER proposto, as quais foram objeto de uma extensa análise em Silva e Paraboni (2015). Estas alternativas, que constituem sistemas de *baseline* para a solução proposta, são sumarizadas na

tabela 6, categorizadas de acordo com quatro dimensões de avaliação: estratégia de seleção, método de ordenação das propriedades, prioridade da relação espacial e uso discriminatório ou redundante desta.

Tabela 6: Dimensões de avaliação consideradas.

Seleção	Ordenação	Prioridade	Uso
<i>Proposta</i>	<i>Por frequência</i>	<i>Padrão</i>	<i>Discriminatório</i>
Mais frequente	Gulosa	Máxima	Redundante
Aleatória	Aleatória	Mínima	

Fonte: Silva e Paraboni (2015).

Com relação às estratégias de seleção de relações espaciais, foram avaliadas estratégias em que a seleção é implementada tal qual proposto pelo modelo discutido anteriormente (*Proposta*), e alternativas que selecionam a relação mais frequente (*Mais frequente*) encontrada no corpus, e que selecionam uma relação espacial ao acaso (*Aleatória*); quanto ao método de ordenação, foram avaliadas estratégias em que a lista de propriedades é ordenada por frequência (*Por frequência*), por poder discriminatório (*Gulosa*) e ao acaso (*Aleatória*); quanto à prioridade atribuída à relação espacial, foram avaliadas estratégias em que a relação é considerada segundo a ordem padrão definida pelo método de ordenação anterior (*Padrão*), em que a relação é considerada antes de qualquer outra propriedade (*Máxima*), ou somente em último caso (*Mínima*); finalmente, quanto ao uso da relação espacial na composição da expressão de referência, foram avaliadas estratégias em que a relação espacial só é considerada quando possui poder discriminatório (*Discriminatório*), e estratégias em que a inclusão de relação espacial ocorre também nos casos em que esta é redundante naquele contexto (*Redundante*).

Para avaliação destas alternativas - em que cada uma constitui uma estratégia de GER com características próprias - considerou-se inicialmente a possibilidade de avaliação extrínseca, baseada no uso prático do algoritmo em apoio à tarefa de navegação em mundos virtuais GIVE. O procedimento neste caso seria semelhante ao adotado por Paraboni e Deemter (2014), envolvendo o monitoramento do comportamento de busca e navegação de sujeitos humano em resposta a instruções contendo expressões de referência produzidas por diferentes estratégias de GER. No entanto, além do grande número de alternativas em consideração, observou-se que as métricas de avaliação possíveis no ambiente GIVE são aquelas baseadas em distância ou tempo de busca, e que estas métricas não permitiriam capturar diferença mais sutis entre as alternativas avaliadas. Por exemplo, seria pouco provável que alternativas como ‘pressione o botão azul’ e ‘pressione o botão ao lado do botão verde’, apresentassem diferenças significativas em termos de distância ou tempo de busca da forma como podem ser monitoradas no ambiente GIVE.

Por este motivo, a avaliação das 54 estratégias consideradas foi feita de forma intrínseca,

com base em um conjunto de dados de teste extraído do *cópus GIVE-2* (GARGETT et al., 2010). Os resultados desta avaliação são discutidos em detalhes em Silva e Paraboni (2015), e confirmam a preferência pela estratégia proposta medida em termos de coeficientes Dice, MASI, e acurácia de seleção de propriedades gerais e de relações espaciais.

3.4.4 Contribuições

O estudo em Silva e Paraboni (2015) apresenta uma série de contribuições potencialmente relevantes para sua área de pesquisa, que incluem a inserção de uma tarefa de GER complexa em situações de comunicação baseadas em um contexto dinâmico, e uma combinação inédita de métodos de aprendizagem de máquina e modelos computacionais de referência espacial. É também digna de nota a avaliação exaustiva de todas as propriedades sistêmicas individuais e de suas interações, conforme destacado em Silva e Paraboni (2015), e os resultados positivos obtidos pela solução proposta dentre todas as alternativas consideradas.

No que diz respeito à questão de como a plausibilidade do texto produzido é interpretada, observamos neste caso um exemplo de abordagem parcialmente baseada na perspectiva do locutor, considerando não apenas um *cópus* de referência mas também modelos matemáticos ‘externos’ a este. Esta solução foi motivada pela natureza dinâmica do contexto de comunicação, que inviabiliza a coleta de um *cópus* cobrindo todos casos possíveis do fenômeno linguístico de interesse. Além disso, observa-se que a relação entre o desempenho do sistema e o desempenho humano foi explorada de forma ainda mais restrita do que na tarefa de GER discutida na seção 3.2, enfocando especialmente a produção (humana ou automática) de relações espaciais, sem nenhum tipo de comparação entre formas superficiais (que de resto não eram produzidas na presente abordagem).

Finalmente, observamos que a perspectiva do locutor não é a única forma de projetar sistemas de GLN deste tipo. Uma consideração fundamental no projeto e avaliação de sistemas de natureza interativa como o exemplo aqui discutido é a questão do impacto que o texto gerado sobre o comportamento de navegação do usuário. Esta questão será abordada no contexto em um segundo estudo, também baseado no ambiente GIVE, a ser discutido na seção 5.2.

3.5 Considerações finais

Neste capítulo foram apresentados três exemplos de como a questão da plausibilidade em GLN pode ser entendida pela perspectiva do locutor. Nestas abordagens, que em grande parte

se assemelham ao uso de *córpus* de referência para a avaliação em diversas áreas do PLN, modelos de seleção de conteúdo e de realização superficial foram propostos tomando-se por base exemplos de língua natural tal qual produzida por locutores humanos. Estas iniciativas entretanto diferem na forma como o *córpus* de referência foi utilizado ou, na falta deste, em como um recurso deste tipo foi emulado.

No caso da tarefa de realização superficial em Novais e Paraboni (2012), o *córpus* utilizado era do tipo convencionalmente adotado em PLN, contendo sentenças extraídas de veículos de notícias online, sem que houvesse, *a priori*, como determinar o estímulo ou contexto de comunicação que motivou a produção destas sentenças. Como forma de transformar este recurso em algo próximo de um *córpus* de GLN, foi assim realizado um procedimento de abstração de sentenças, transformadas em uma representação do tipo que poderia ser tomado como entrada por um módulo típico de realização superficial. Neste processo, que fez uso de uma combinação de procedimentos automatizados e anotação manual, a plausibilidade da língua gerada foi medida de forma direta comparando-se as formas superficiais geradas com as sentenças do *córpus* original. Em outras palavras, a plausibilidade da língua foi entendida no seu sentido mais tradicional de coincidência entre as versões computacional e humana da língua.

No caso da seleção de conteúdo de expressões TUNA em Lucena, Paraboni e Pereira (2010) e da geração de relações espaciais em ambientes GIVE em Silva e Paraboni (2015), por outro lado, foram utilizados *córpus* específicos de GLN, ou seja, conjuntos de dados formados por pares de texto (expressões de referência ou instruções de navegação em mundos virtuais, respectivamente) e os estímulos que lhes deram origem (i.e., contextos estáticos representados na forma de imagens bidimensionais, ou instantâneos de um contexto dinâmico representando um momento específico da navegação no mundo virtual). Em ambos os casos, a plausibilidade da língua gerada foi medida apenas com base no seu conteúdo semântico, ou seja, tomando-se por base a anotação manual presente em cada *córpus*, e não as formas superficiais propriamente ditas (que era disponibilizada nos idiomas inglês e/ou alemão). Além disso, no caso da geração de relações espaciais, uma forma adicional - e ainda mais específica - de avaliação foi considerada ao comparar-se a estratégia de referência espacial adotada pelo algoritmo proposto e aquelas observadas no *córpus* sem levar em conta o restante do conteúdo semântico destas expressões. Em outras palavras, a plausibilidade da língua foi entendida especificamente no sentido de similaridade semântica entre as versões computacional e humana da relação espacial produzida.

Os estudos aqui discutidos demonstram que, no projeto de sistemas e artefatos de GLN, a perspectiva do locutor pode ser adotada com ênfase variada, e com base em diferentes tipos de conhecimento. Esta variação pode depender dos objetivos da pesquisa, dos recursos disponíveis

e da natureza do problema tratado. No caso dos três exemplos específicos que foram discutidos, podemos afirmar que a perspectiva do locutor esteve mais presente no estudo de realização superficial e (de certo modo em menor grau) no estudo de seleção de conteúdo para GER, e que no caso da geração de relações espaciais esta perspectiva foi ainda mais suavizada pelo uso combinado de conhecimento de *cópus* e modelos externos a este.

Capítulo 4

A NECESSIDADE DE NOVOS CONJUNTOS DE DADOS

4.1 Visão geral

Estudos de GLN sob a perspectiva do locutor, como os exemplos discutidos no capítulo anterior, constituem um método efetivo de investigação de diversos problemas de pesquisa da área. Abordagens deste tipo, entretanto, exigem recursos linguístico-computacionais especializados, tipicamente representados na forma de *córpus* paralelos de formas linguísticas e seus respectivos contextos, em muitos casos com anotação adicional dos fenômenos de interesse.

Em razão destas peculiaridades, a construção de *córpus* para pesquisa em GLN é frequentemente implementada por meio de experimentos controlados envolvendo sujeitos humanos. Experimentos deste tipo constituem investigações científicas completas e com contribuições específicas, e acarretam assim um custo de desenvolvimento consideravelmente superior ao que seria observado em uma simples tarefa de coleta de dados, por exemplo, a partir da WEB.

Talvez mais importante do que a questão do custo envolvido, entretanto, é importante destacar que *córpus* para GLN são tipicamente limitados ao estudo de fenômenos altamente específicos, e a um determinado domínio no qual se concentram. Um exemplo deste tipo de limitação é o caso do *córpus* TUNA (GATT; SLUIS; DEEMTER, 2007) descrito na seção 2.2. Embora seja um recurso de razoável extensão e complexidade, na atualidade este *córpus* dificilmente apresentaria novas oportunidade de investigação (suficientemente relevantes para a área) além daquelas já exploradas de forma exaustiva nos primeiros anos de seu lançamento. Isso ocorre porque, para viabilizar a observação de fenômenos de interesse para um determinado projeto, recursos deste tipo costumam ser projetados sob condições controladas, ou seja, considerando-se um conjunto muito específico de relações entre textos e os contextos que os motivaram, e uma vez esgotadas

as oportunidades de investigação destas poucas relações possíveis, resta pouco a fazer exceto construir um novo *corpus*. Em outras palavras, novas questões de pesquisa de GLN demandam, quase que invariavelmente, a construção de novos recursos, necessidade esta que coloca limites práticos consideráveis a abordagens baseadas em *corpus*.

Deixando momentaneamente de lado estas dificuldades, entretanto, este capítulo discute a construção de *corpus* para GLN enfatizando seus desafios, as questões de pesquisa tratadas e suas contribuições. Esta discussão será ilustrada predominantemente com base em três exemplos de iniciativas de construção de *corpus* para pesquisa em GLN sob a perspectiva do locutor, a saber: o *corpus* Stars (TEIXEIRA et al., 2014; PARABONI et al., 2014) na seção 4.2, Stars2 na seção 4.3 e o *corpus* Zoom em Altamirano et al. (2015), tratado na seção 4.4.

Em linhas gerais, Stars e Stars2 são recursos produzidos sob condições controladas para estudo de problemas específicos de GER (em especial, a questão da superespecificação referencial, da variação humana e do uso de propriedades relacionais). Estes *corpus* seguem uma linha de investigação semelhante à que motivou a construção de *corpus* de expressões relacionais como GRE3D3/7 em Dale e Viethen (2009) e Viethen e Dale (2011), porém tratando de questões de pesquisa distintas e considerando fenômenos linguísticos de maior complexidade do que os tratados em trabalho prévio. *Corpus* desse tipo são os mais comumente utilizados na pesquisa em GER e Psicolinguística.

O *corpus* Zoom, por outro lado, é um recurso produzido sob condições menos restritivas do que recursos existentes na área, e representa uma iniciativa de estudo do fenômeno de referência em domínios mais próximos de uma aplicação computacional real, neste caso enfocando o problema de referência a pontos de interesse em mapas geográficos. Este *corpus* foi construído a partir de textos produzidos nos idiomas português e espanhol, e destina-se tanto à pesquisa de GER como a estudos de realização superficial nestes dois idiomas, apresentando maior apelo linguístico do que os anteriores.

Os três *corpus* aqui discutidos foram desenvolvidos em colaboração com alunos de graduação e, no caso do *corpus* Zoom, contempla também uma orientação de mestrado. O projeto do *corpus* Zoom contou ainda com a colaboração adicional da Universidade de Córdoba (Argentina) no que diz respeito à porção em espanhol do conjunto de dados. Todos estes recursos foram desenvolvidos no contexto do projeto de pesquisa regular intitulado '*Geração de língua natural para descrição de objetos em contextos visuais*' (FAPESP nro. 2013/23169-0).

4.2 O *córpus Stars*

Diversas questões de pesquisa relacionadas ao grau de especificação de descrições relacionais, e que não podiam ser observadas em número significativo em *córpus* do gênero, como GRE3D3/7 (DALE; VIETHEN, 2009; VIETHEN; DALE, 2011), levaram ao projeto de um novo experimento para coleta de descrições produzidas por sujeitos humanos. Este experimento resultou no *córpus Stars* de expressões de referência relacionais, apresentado em Paraboni et al. (2014) e utilizado em um experimento de GER em Teixeira et al. (2014).

O domínio *Stars* foi planejada de modo a induzir descrições relacionais relativamente complexas, contendo até três referentes - aqui denominados objeto-alvo principal, ponto de referência e segundo ponto de referência, como em ‘a estrela ao lado da caixa branca, ao lado de um círculo branco’. Além disso, as cenas consideradas apresentavam forte motivação para o uso de relações espaciais, que eram invariavelmente a forma mais provável (senão única) de obter descrições livres de ambiguidade. Neste sentido, o domínio *Stars* modela situações de referência significativamente distintas daquelas observadas no domínio GRE3D (DALE; VIETHEN, 2009; VIETHEN; DALE, 2011), onde o uso de propriedades relacionais era sempre opcional para fins de desambiguação.

Um exemplo de estímulo utilizado na construção do *córpus Stars* é ilustrado pela figura 12, na qual os participantes do experimento recebiam a tarefa de descrever de forma não-ambígua o objeto apontado pela seta (uma estrela).

Figura 12: Situação de referência a um objeto-alvo (estrela) via *landmark* (caixa) no *córpus Stars* (rótulos adicionados para fins ilustrativos).

Fonte: Adaptado de Paraboni et al. (2014).

Dada a estrutura da cena neste exemplo, uma maneira altamente provável de distinguir o objeto-alvo dos demais objetos neste contexto é fazendo-se referência ao objeto mais próximo, ou seja, usando o objeto caixa como ponto de referência, como em ‘a estrela ao lado da caixa’. O foco do experimento *Stars* era assim justamente a forma de descrever este ponto de referência

(no exemplo, a caixa), e não o objeto-alvo principal (i.e., estrela).

Nos tipos de contextos considerados, a descrição do ponto de referência poderia ser (a) subespecificada (resultando em uma expressão global mínima, como em ‘a estrela ao lado de *uma caixa*’), (b) exatamente especificada (resultando em uma expressão global superespecificada, como em ‘a estrela ao lado da *caixa branca*’) ou ainda (c) superespecificada (resultando em uma expressão global ainda mais superespecificada, como em ‘a estrela ao lado da *caixa branca, ao lado de um círculo branco*’).

O principal objetivo do experimento foi investigar situações em que a subespecificação do ponto de referência podia ocorrer, verificando em que situações os participantes humanos favoreceriam descrições de certa forma ‘incompletas’ do tipo (a) acima. Esta questão foi motivada por resultados apresentados em Paraboni e Deemter (2014), a ser discutido na seção 5.2, que sugerem que descrições relacionais deste tipo podem dificultar a busca e identificação de referentes em domínios espaciais tridimensionais ou estruturalmente complexos. No experimento Stars procurou-se verificar se efeitos deste tipo seriam mantidos no caso de situações de referência mais simples representadas por imagens estáticas bidimensionais.

O experimento faz uso de pares de estímulos complementares, distintos apenas no que diz respeito às propriedades do ponto de referência mais próximo do objeto-alvo em questão. Dois exemplos de estímulos utilizados são apresentados na figura 13.

Figura 13: Duas situações de referência ao alvo *e2* via *landmark* *q6* (rótulos adicionados para fins ilustrativos).

Fonte: Paraboni et al. (2014).

Neste exemplo, a diferença entre as duas versões é o atributo cor do ponto de referência *q6*. Na imagem da esquerda, este atributo é único para todos os objetos do mesmo tipo que o ponto de referência (ou seja, o ponto de referência é a única caixa de cor branca na cena), mas não na imagem da direita (onde todas as caixas possuem a mesma cor branca).

Com base em estímulos como os deste exemplo, foram definidas seis condições experimentais críticas, sendo três do tipo *abs* (absoluto) e três do tipo *proj* (projetiva), em referência ao

tipo de atributo disponível para especificação. As condições previstas objetivavam comparar situações em que um ponto de referência podia ser totalmente especificada com uso de um atributo absoluto do tipo cor (e.g., ‘a estrela próxima da caixa *branca*’, na imagem esquerda), ou apenas com uso de um atributo projetivo (e.g., uma relação espacial com um segundo ponto de referência, como em ‘a estrela próxima da caixa *ao lado de um círculo*’, na imagem da direita).

O experimento Stars considerou duas hipóteses de pesquisa, a saber:

- h1 A subespecificação do ponto de referência é mais frequente quando não é possível distinguir um ponto de referência de outros objetos do mesmo tipo com base em um atributo absoluto (cor) do que quando esta distinção é possível.
- h2 Descrições subespecificadas são menos frequentes do que descrições superespecificadas em todas situações consideradas.

Detalhes referentes à motivação destas hipóteses fogem ao escopo desta monografia, e são apresentados em Paraboni et al. (2014). A discussão a seguir trata assim prioritariamente da metodologia, resultados e contribuições deste estudo.

4.2.1 Materiais e método

O experimento Stars contou com 73 participantes voluntários, e foi realizado à distância e sem nenhum tipo de supervisão adicional. Os participantes foram solicitados a executar um pequeno aplicativo JAVA (.jar) enviado por e-mail, e a seguir as instruções exibidas na tela. Ao executar o aplicativo, uma breve página de instruções era apresentada, seguida das imagens de estímulo exibidas uma a uma, em ordem aleatória. Ao fornecer uma resposta, o participante era direcionado à próxima cena, até o final do experimento. Ao final do experimento, um arquivo com as respostas criptografadas era gerado e o participante era solicitado a enviá-lo por e-mail aos responsáveis pela pesquisa.

O conjunto de estímulo consistia de 11 imagens representando as seis condições críticas e cinco *fillers*. Embora sem função no experimento original, as imagens do tipo *filler* também foram usadas como estímulo para produção de expressões de referência válidas, e que possuem interesse para a pesquisa em GER de modo geral. O número proporcionalmente elevado de *fillers* foi necessário para evitar respostas repetitivas, já que em todas condições críticas era possível descrever o objeto-alvo com uso de uma descrição subespecificada do tipo ‘a estrela ao lado do cubo’. Esta alternativa não era possível no caso dos *fillers*.

4.2.2 Resultados

Foi coletado um total de 803 expressões de referência. Uma vez que o experimento foi realizado sem supervisão (i.e., via WEB), e considerando-se que os participantes não receberam exemplos de como proceder, o conjunto de dados foi examinado manualmente para garantir sua correção. Neste processo, nove participantes (12%) foram identificados como *outliers* e seus dados foram excluídos. Estes casos representam situações em que o participante pareceu não entender o que estava sendo solicitado, produzindo expressões não referenciais ou ambíguas (e.g., ‘a estrela branca’). Após a remoção destes dados, o *cópus* final foi constituído de 704 descrições produzidas por 64 participantes.

Três anotadores trabalharam de forma independente na anotação das imagens de estímulo e das descrições coletadas. Posteriormente, a anotação foi consolidado por um quarto juiz. Tanto imagens de estímulo como descrições foram transcritas em uma extensão do formato XML utilizado no *cópus* TUNA (GATT; SLUIS; DEEMTER, 2007) para representar descrições relacionais (e não apenas atômicas). Este formato seria reutilizado em estudos posteriores, a serem discutidos nas seções subsequentes deste capítulo.

Um exemplo desta representação é ilustrado na figura 14.

Figura 14: Fragmento de representação XML do *cópus Stars*.

```
<CONTEXT ID="abs1">
  <ATTRIBUTE-SET TARGET="e2" LANDMARK="q6"
    STRING=" a estrela à esquerda do quadrado branco "
    LENGTH="4" REL-COUNT="1">
    <ATTRIBUTE NAME="type" VALUE="star" />
    <ATTRIBUTE NAME="landmark-type" VALUE="cube" />
    <ATTRIBUTE NAME="landmark-colour" VALUE="white" />
    <ATTRIBUTE NAME="left" VALUE="q6" />
  </ATTRIBUTE-SET>
</CONTEXT>
```

Fonte: Ivandré Paraboni, 2017.

Os resultados da investigação referente às hipóteses do experimento são detalhados em Paraboni et al. (2014). Brevemente, o experimento complementa os resultados do trabalho prévio (PARABONI; DEEMTER, 2014) demonstrando que (*h1*) em contextos referenciais mais simples - como imagens estáticas bidimensionais - os efeitos da subespecificação do ponto de referência são minimizados, a ponto de boa parte dos locutores adotá-la como estratégia de referência. Por outro lado, este tipo de subespecificação (*h2*) permanece ainda um fenômeno relativamente raro

se comparado à produção de expressões superespecificadas, um resultado que é consistente com diversos estudos prévios sobre este fenômeno (PECHMANN, 1989; ARTS et al., 2011).

O *cópus Stars* foi apresentado em Paraboni et al. (2014) com a investigação original da questão de produção de descrições relacionais mínimas, e utilizado em um experimento prático de GER enfocando a questão da subespecificação de descrições relacionais em Teixeira et al. (2014). O *cópus* foi também utilizado para fins de treinamento e teste de modelos computacionais de GER baseados em aprendizagem de máquina em Ferreira e Paraboni (2014b), e é um dos domínios considerados no estudo de variação humana na tarefa de GER em Ferreira e Paraboni (2017a). O *cópus* encontra-se publicamente disponível para reuso em pesquisa na página do projeto ao qual está vinculado¹.

4.2.3 Contribuições

O experimento Stars foi uma primeira iniciativa de construção de *cópus* próprio (e em português brasileiro) para pesquisa em GLN sob a perspectiva do locutor. Como tal, é uma iniciativa relativamente modesta, especialmente pelo número reduzido de condições experimentais consideradas, e pelo método de execução (via WEB), que é naturalmente sujeito a imprecisões. Estas e outras lições aprendidas foram no entanto levadas em conta no projeto subsequente, que originou o *cópus Stars2* discutido a seguir.

4.3 O *cópus Stars2*

Esta seção descreve o experimento de construção do *cópus Stars2* detalhado em Paraboni, Galindo e Iacovelli (2017). A ferramenta computacional utilizada neste experimento foi apresentada em Iacovelli, Galindo e Paraboni (2014), e resultados preliminares de um experimento de GER com base nos dados coletados foram discutidos em Paraboni, Galindo e Iacovelli (2015). Apesar da semelhança de nome com o experimento anterior, Stars2 não é uma extensão do *cópus Stars*. Os dois experimentos diferem em objetivos, projeto, metodologia, participantes e materiais utilizados.

Sob vários aspectos, o *cópus Stars2* constitui um conjunto de dados mais sofisticado do que seu antecessor Stars, ou outros recursos amplamente reconhecidos na área. Em especial, houve considerável investimento de tempo e de recursos humanos e computacionais para desenvolvimento e implementação de um experimento envolvendo duplas de participantes humanos

¹<http://ivandreparaboni.wix.com/research#!visconde/cwwa>

em situações de diálogo, em um paradigma experimental do tipo *director-matcher* (KRAUSS; WEINHEIMER, 1966). Deste modo, as expressões de referência coletadas são possivelmente mais precisas (do ponto de vista da identificação do objeto-alvo) e mais realistas (do ponto de vista da tarefa de comunicação) do que em experimentos do tipo monólogo baseados na WEB, incluídos aqui não apenas o *cópus Stars* mas também os *cópus TUNA* (GATT; SLUIS; DEEMTER, 2007) e *GRE3D3/7* (DALE; VIETHEN, 2009; VIETHEN; DALE, 2011).

O objetivo do experimento *Stars2* foi a construção de um *cópus* de GER com situações de referência projetadas de modo a proporcionar diversos desafios a algoritmos tradicionais da área, como o algoritmo Greedy (DALE, 2002), Incremental (DALE; REITER, 1995) e outros. Em especial, procurou-se criar situações de referência que pudessem questionar princípios estabelecidos como a preferência por cor Pechmann (1989), por propriedades absolutas (GOMPEL et al., 2014) e por descrições mais breves (e.g., atômicas vs. relacionais) (DALE; HADDOCK, 1991; KELLEHER; COSTELLO, 2009) conforme discutido a seguir.

O projeto do experimento *Stars2* considerou quatro requisitos principais: (i) emprego de cenas com um número de objetos distraidores superior ao de outros *cópus* para GER; (ii) disponibilização de múltiplas relações espaciais entre objetos (de modo a apresentar oportunidades de produção de descrições relacionais variadas); (iii) maior escolha de alternativas de superespecificação; e (iv) disponibilização de uma propriedade altamente *saliente*, porém estritamente desnecessária (i.e., redundante) para fins de identificação do objeto-alvo em cada contexto.

O domínio *Stars2* segue a tradição de outros *cópus* de GER como *GRE3D* (DALE; VIETHEN, 2009; VIETHEN; DALE, 2011) ao fazer uso de cenas contendo objetos geométricos simples, e de um conjunto básico de atributos compatível (e.g., cor, tamanho etc.). Como forma de estimular a superespecificação referencial, entretanto, foi definido - conforme requisito (i) acima - um número mais elevado de objetos distraidores em cada cena. Considerando-se que os domínios *TUNA* e *GRE3D* contêm entre três e sete objetos por cena, as cenas no domínio *Stars2* foram definidas de modo a exibir 15 objetos cada.

A definição de contextos contendo um grande número de objetos pode entretanto ser pouco efetiva a menos que um número igualmente maior de atributos referenciais seja considerado. Isso, por consequência, pode exigir a coleta de um número muito maior de expressões de referência de modo a garantir que todos os atributos definidos ocorram em números significativos (i.e., de modo a evitar que o conjunto de dados seja esparso). Como alternativa à coleta de um conjunto de dados de grandes proporções deste tipo, optou-se, conforme requisito (ii), por modelar contextos em que um número potencialmente grande de relações espaciais poderia ser utilizado para fins

de referência. Assim, objetos nas cenas Stars2 foram organizados em grupos de dois ou três elementos cada, de modo que não apenas o objeto-alvo pudesse ser referenciado via objeto mais próximo (que funciona como ponto de referência ou *landmark*) mas também o próprio ponto de referência podia ser referenciado via um terceiro objeto (ou segundo ponto de referência), em uma configuração semelhante à já adotada (de forma mais simplificada) no *córpus Stars* (TEIXEIRA et al., 2014) discutido na seção anterior.

Um exemplo de contexto neste domínio é ilustrado pela figura 15, em que o objeto-alvo marcado pela seta pode ser descrito como, e.g., ‘o cone próximo ao cubo’, ‘o cone cinza próximo ao cubo, ao lado de uma bola’ etc.

Figura 15: Exemplo de contexto referencial no *córpus Stars2*.

Fonte: Ivandré Paraboni, 2017.

O uso de cenas relativamente complexas como neste exemplo é um motivador para a superespecificação de objeto-alvo e pontos de referência conforme exigido pelo requisito (iii).

Com relação ao requisito (iv), observa-se que a saliência relativa dos atributos atômicos como cor e tamanho é facilmente manipulável (e.g., podemos modelar objetos fazendo uso de uma cor especialmente brilhante ou de um tamanho incomum), mas o mesmo não pode ser dito da saliência das relações espaciais existentes. Modelar relações espaciais com maior ou menor grau de saliência exigiria a construção de cenas visuais excessivamente complexas. Além disso, embora propriedades altamente salientes sejam facilmente perceptíveis por leitores humanos, esta informação não é normalmente disponível a um modelo computacional.

Com base nestas observações, optou-se assim por modelar propriedades altamente salientes na forma de propriedades com alto poder discriminatório, aqui entendido como o número de distraidores que seria ‘eliminado’ pela seleção da propriedade em questão (DALE; REITER, 1995). Em outras palavras, como forma de garantir que certas propriedades de interesse em cada contexto sejam percebidas como altamente salientes, optou-se por defini-las como sendo

altamente *discriminatórias*. Esta decisão de projeto permite a observação dos efeitos da saliência sobre o grau de especificação referencial e, ao mesmo tempo, faz do conjunto Stars2 uma base adequada para desenvolvimento de algoritmos práticos de GER uma vez que a computação do poder discriminatório de atributos referenciais é uma tarefa trivial.

As condições experimentais do experimento Stars2 foram definidas de modo a privilegiar situações de referência de dois tipos: um grupo de quatro condições 01..04 em que a descrição do objeto-alvo exige o uso de uma propriedade relacional, e um segundo grupo de quatro condições 05..08 em que o alvo pode ser identificado de forma única com uso de uma descrição atômica. Dentro de cada grupo, as condições eram idênticas exceto pelo tipo de propriedade definida como altamente saliente, e que era disponibilizada para superespecificação do ponto de referência (nas condições 01..04) ou objeto-alvo (condições 05..08). Esta propriedade saliente poderia ser cor, tamanho, uma relação espacial projetiva (e.g., à esquerda de) ou topológica (e.g., próximo a).

A figura 16 ilustra as condições 01..04 do experimento em que o uso de uma descrição relacional é altamente provável.

Figura 16: Situações de referência com uso de descrições relacionais (rótulos adicionados para fins ilustrativos).

Fonte: Adaptado de Paraboni, Galindo e Iacovelli (2017).

Nas condições 01..04, o objeto-alvo (nos quatro casos, um cone) compartilha suas principais

propriedades atômicas (e.g., seu tipo, cor e tamanho) com diversos distraidores presentes no contexto. Por este motivo, e considerando-se que os sujeitos foram instruídos a não fazer uso de atributos espaciais (e.g., referenciando-se posições da tela), faz-se necessário incluir na descrição deste objeto uma relação com o ponto de referência mais próximo, ou seja, um cubo. Entretanto, como o objeto-alvo neste exemplo é o único cone ao lado de um cubo na cena inteira, este ponto de referência pode ser subespecificado sem prejuízo à identificação do objeto-alvo, ou seja, em todas as quatro cenas é sempre possível produzir descrições do tipo ‘o cone *ao lado do cubo*’ muito embora seus componentes individuais (i.e., as porções ‘cone’ e ‘cubo’) sejam, se consideradas de forma isolada, ambíguas.

Descrições relacionais deste tipo representam a saída típica de diversos algoritmos de GER capazes de tratar propriedades relacionais (DALE; HADDOCK, 1991; KRAHMER; ERK; VERLEG, 2003). Entretanto, locutores humanos frequentemente preferem superespecificar estas expressões com o acréscimo de informação redundante à descrição do ponto de referência, como em ‘o cone ao lado do cubo *vermelho*’ no caso 04 do exemplo. O projeto do domínio Stars2 procurou assim explorar diferentes estratégias de superespecificação possíveis oferecendo oportunidades baseadas em diferentes atributos, ou seja, a diferença entre as condições 01..04 é justamente o tipo de propriedade que se apresenta como oportunidade para superespecificação.

Em cada uma destas condições 01..04, o ponto de referência (i.e., o cubo mais próximo do alvo) possui uma propriedade altamente saliente (e altamente discriminatória) que, embora desnecessária para desambiguação, pode ser selecionada caso o locutor deseje superespecificá-la. Esta propriedade corresponde a uma relação espacial de direção como *right-ball* na condição 01, a uma relação espacial de proximidade como *near-ball* em 02, à propriedade atômica representando tamanho como *size-small* em 03, e à propriedade atômica representando cor como *colour-red* em 04.

Além das condições relacionais 01..04, foram definidas também quatro condições 05..08 em que o uso de uma descrição atômica seria suficiente para fins de desambiguação. Este conjunto de oito condições experimentais, e as questões de pesquisa que as motivaram, são centrados na ideia de que a propriedade mais saliente - seja ela o atributo cor, tamanho, uma relação espacial topológica ou projetiva, conforme cada caso - constitui a estratégia preferencial de superespecificação em todas as situações consideradas. Este pressuposto, que traz consequências não-triviais para o projeto de algoritmos de GER, é discutido em detalhes em Paraboni, Galindo e Iacovelli (2017).

A definição destas condições experimentais oferece suporte a múltiplas questões de pesquisa. Em especial, a ocorrência de um atributo saliente distinto em cada tipo de situação dificulta o

uso de métodos de GER baseados em estimativas de frequência de atributos - como uma possível implementação do algoritmo Incremental (DALE; REITER, 1995) - já que o número de situações que privilegia cada tipo de atributo é balanceado. Além disso, preferências por atributos como cor (PECHMANN, 1989) tendem a ser válidas apenas para certas condições (como 04 no exemplo anterior), e não podem mais ser usadas de modo indiscriminado, e também o uso de relações espaciais (DALE; VIETHEN, 2009) passa a ser favorecido em casos específicos (possivelmente do tipo topológica no caso 01 anterior, ou projetiva no caso 02, mas não nos casos 03 e 04).

4.3.1 Materiais e método

O experimento Stars2 contou com 56 participantes que trabalharam em duplas. O experimento foi realizado de forma presencial, com uso de dois computadores. O procedimento consistiu na apresentação de uma série de imagens de estímulo a pares de sujeitos desempenhando papéis de locutor e ouvinte. O papel do ouvinte era limitado à validação das descrições produzidas pelo locutor, garantindo que o objeto-alvo de cada expressão fosse de fato identificável. Todos os sujeitos participaram do experimento duas vezes, alternando os papéis de locutor e ouvinte, e os papéis iniciais eram designados de forma aleatória. Ao término de uma sessão, os papéis eram invertidos e um segundo experimento era iniciado imediatamente. Como diversos aspectos da apresentação dos estímulos eram definidos de forma aleatória ao longo a execução do experimento, conforme discutido a seguir, as imagens apresentadas em cada experimento consistem majoritariamente de material inédito.

Locutor e ouvinte trabalharam em computadores individuais, e a comunicação entre eles era restrita a certos tipos de mensagem eletrônica. Um ambiente computacional descrito em Iacovelli, Galindo e Paraboni (2014) foi desenvolvido especificamente para a realização de experimentos online em duplas locutor-ouvinte deste tipo. Neste ambiente, a tarefa do locutor era preencher uma caixa de texto com uma descrição do objeto apontado por uma seta na sua própria tela, e então pressionar um botão ‘Send’. Ao proceder desta forma, a descrição fornecida era exibida na tela do ouvinte como parte de uma instrução na forma ‘Por favor selecione a letra correspondente a X’, em que X era a descrição fornecida pelo locutor. A tarefa do ouvinte consistia assim em interpretar a descrição recebida e selecionar o objeto correspondente em sua própria tela. Cada imagem de estímulo era apresentada simultaneamente aos dois participantes, e ambas imagens representavam a mesma situação de referência exceto por detalhes exigidos para a manipulação solicitada a cada participante: a seta apontando para o objeto-alvo era visível apenas na tela do locutor, e os rótulos identificadores (letras) de cada objeto candidato eram visíveis apenas na tela do ouvinte. A figura 17 ilustra estas duas telas.

Figura 17: Telas do locutor (esquerda) e ouvinte (direita).

Fonte: Iacovelli, Galindo e Paraboni (2014).

A única ação possível para o locutor era enviar descrições textuais ao ouvinte, e a única ação possível para o ouvinte era selecionar uma letra correspondente a um dos objetos exibidos em sua tela, ou pressionar um botão ‘*What?*’ caso a instrução não estivesse suficientemente clara. Após cada seleção correta, o locutor era solicitado a descrever um novo objeto em uma nova cena, até o término da série de estímulos. No caso de uma seleção incorreta por parte do ouvinte, ambos os participantes recebiam uma mensagem de alerta. Quando o ouvinte sinalizava que não havia entendido a descrição, ou quando este selecionava o objeto errado, uma mensagem era enviada ao locutor e a condição experimental era inserida em uma fila para futura reexibição.

O experimento consistia de oito imagens básicas (quatro representando as condições de referência relacional 01..04 e quatro representando condições de referência atômica 05..08). As imagens foram misturadas e apresentadas em ordem aleatória, uma a uma. Uma vez que as condições relacionais e atômicas representam situações que exigem estratégias de referência consideravelmente distintas, a combinação aleatória contribui para quebrar a monotonicidade da tarefa.

Entre uma condição e outra (e também em uma mesma condição no caso de repetição ou erro de identificação), diversos aspectos da apresentação eram modificados. Metade das imagens era exibida com orientação normal como visto nos exemplos anteriores, e a outra metade em reverso (i.e., como se refletida em um espelho). A orientação - normal ou reversa - era definida aleatoriamente. Além disso, metade das imagens apresentava objetos em um padrão cone-cubo-esfera, e a outra metade em um padrão cubo-cone-esfera. A escolha de padrão era também definida de forma aleatória.

As imagens apresentadas ao locutor exibiam o objeto-alvo apontado por uma seta e sem rótulos. As imagens apresentadas ao ouvinte apresentavam rótulos (A,H,I,O,M) associados a cinco objetos (o objeto-alvo e quatro distraidores) conforme a condição de teste. Tendo em

vista o objetivo de coletar um conjunto de expressões de referência de grandes proporções, cada condição foi testada duas vezes usando modos de apresentação ligeiramente diferentes, com grupos de objetos sobrepostos e não sobrepostos.

Cada participante foi solicitado a descrever, no papel de locutor, 8 imagens em 2 modos de apresentação, totalizando assim $8 * 2 = 16$ imagens. Além disso, cada participante também atuou como ouvinte na identificação de outras 16 imagens. Havendo 16 imagens distintas $* 2$ orientações possíveis $* 2$ padrões de grupos de objetos, cada conjunto de 16 imagens presente em um experimento era selecionado a partir de 64 possibilidades.

4.3.2 Resultados

O experimento produziu um total de 884 expressões de referência corretamente identificadas pelos 56 ouvintes, sendo que 61 instâncias foram repetidas em virtude do uso do botão ‘What?’. Após verificação manual, 12 descrições foram descartadas por serem malformadas (e.g., o locutor produziu ‘o cone’ ao descrever uma esfera, e mesmo assim o ouvinte selecionou, provavelmente por acaso, a esfera correta). Cada material de estímulo específico foi utilizado entre 0,8% a 2,6% dos testes. As imagens e descrições foram anotadas por dois anotadores. Ambos os conjuntos de anotações foram representados no mesmo formato XML adotado no *cópus* Stars descrito na seção anterior.

Os dados coletados oferecem suporte a uma série de questões de pesquisa discutidas em Paraboni, Galindo e Iacovelli (2017), comprovando a hipótese de que é a saliência do atributo - e não necessariamente o seu tipo - que determina a escolha feita na superespecificação referencial. Conforme discutido em Paraboni, Galindo e Iacovelli (2017), este resultado é uma contribuição significativa para o projeto de algoritmos de GER, que via de regra fazem uso de heurísticas mais simples (e.g., preferência pelo atributo cor ou pelo atributo mais frequente no domínio em questão) para modelar o problema.

O experimento Stars2 e o conjunto de dados propriamente dito foram apresentados em detalhes em Paraboni, Galindo e Iacovelli (2017). O *cópus* está disponível para download na página do projeto que lhe deu origem², e a ferramenta de implementação do experimento em modo online e colaborativo é descrita em Iacovelli, Galindo e Paraboni (2014).

O *cópus* Stars2 serviu de apoio a uma série de estudos desenvolvidos pelos seus autores e colaboradores. O *cópus* foi parcialmente utilizado em Ferreira e Paraboni (2014b), trabalho que enfocou a questão da variação humana em GER usando os modelos computacionais discutidos

²<http://ivandreparaboni.wix.com/research#!visconde/cwwa>

em Ferreira e Paraboni (2014a). Posteriormente, o *córpus* foi empregado na sua totalidade na versão final destes mesmos estudos em Ferreira e Paraboni (2017a), a ser discutido na seção 6.2. Finalmente, o *córpus* foi também utilizado em um estudo sobre o fenômeno de superespecificação referencial apresentado em Paraboni, Galindo e Iacovelli (2015).

4.3.3 Contribuições

O *córpus Stars2* é um recurso linguístico-computacional relativamente sofisticado, tendo sido o primeiro do gênero a ser desenvolvido em situações de comunicação envolvendo locutor e ouvinte. Além disso, as situações de referência contempladas oferecem diversos desafios inéditos para o projeto de algoritmos de GER, contribuindo para o melhor entendimento do papel da noção de saliência na superespecificação referencial, o que em última análise pode facilitar o desenvolvimento de algoritmos de GER mais plausíveis.

4.4 O *córpus Zoom*

Experimentos altamente controlados como *Stars* e *Stars2*, discutidos nas seções anteriores, permitem a investigação de aspectos específicos da produção da língua natural por locutores humanos. Esta necessidade de controle, entretanto, frequentemente leva ao projeto de domínios simplificados - muitas vezes baseados em objetos geométricos ou outras formas simples - e em certo sentidos mais distantes de aplicações computacionais mais realistas.

Como forma de preencher esta lacuna entre estudos como *Stars* e *Stars2* e aplicações práticas de GLN foi construído um *córpus* de expressões em formato ‘livre’, isso é, com pouca manipulação do contexto de referência, denominado *Zoom*. Diferentemente dos estudos apresentados nas seções anteriores, o *córpus Zoom* não foi desenvolvido com o intuito de responder questões de pesquisa específicas de GER. Ao invés, disso, trata-se de um domínio mais aberto, e com maior ênfase à questão de realização superficial (que não foi abordada em *Stars* e *Stars2*, por exemplo), e de caráter mais linguístico do que o dos estudos das seções anteriores. O objetivo desta iniciativa foi o de observar o uso de expressões referenciais em um domínio razoavelmente próximo de uma aplicação computacional hipotética, porém mantendo-se ainda um certo grau de controle que permitisse, por exemplo, a anotação semântica de forma sistemática para desenvolvimento de modelos computacionais de GER nos moldes dos estudos anteriores.

O *córpus Zoom* é um recurso bilíngue que faz uso de imagens de mapas de cidades reais.

O trabalho de construção e uso do *cópus Zoom* foi descrito em Altamirano et al. (2015), e é discutido em linhas gerais nesta seção.

4.4.1 Materiais e método

As situações de referência consideradas no domínio *Zoom* fazem uso de mapas com diferentes graus de detalhamento (representados por diferentes níveis de zoom), abordando descrições de objetos simples e conjuntos. O experimento fez uso de uma interface especialmente projetada para este fim, conforme ilustrado na figura 18, e um conjunto de imagens de mapas obtidas do projeto *OpenStreetMap*³.

Figura 18: Interface do experimento *Zoom*.

Fonte: Altamirano et al. (2015).

Assim como em outros experimentos deste tipo, o contexto visual apresentava diversos objetos distraidores (pontos de interesse como restaurante, bares etc.), com um ou dois objeto(s) apontado(s) por uma seta. Os participantes eram então solicitados a descrever a(s) localidade(s) indicadas como se estivessem fazendo uma recomendação a um amigo completando uma frase na forma 'Seria interessante visitar...' com a descrição do local. Ao fornecer a descrição e pressionar um botão 'próximo', outro estímulo era exibido. O procedimento era repetido até o fim do conjunto de imagens.

³openstreetmap.org

Os mapas utilizados na versão em português do experimento eram de localidades da cidade de Lisboa, e os utilizados na versão em espanhol eram de localidades da cidade de Madrid. Para cada cidade, foram utilizados 10 mapas. As localidades a serem descritas eram exibidas em duas versões, com mais ou menos zoom, totalizando assim 20 imagens em cada idioma. Em ambos níveis de zoom o objeto-alvo era o mesmo, mas o contexto era exibido com diferente grau de detalhe (e.g., afetando a quantidade e tipo de distraidores, a exibição de nomes de ruas etc.) dependendo da representação. Metade das imagens exibia um objeto único apontado pela seta, exigindo uma descrição singular como em ‘o restaurante A na rua X’. A outra metade das imagens exibia dois objetos apontados por setas, exigindo uma descrição do conjunto como em ‘os restaurantes na rua X’ ou ‘o restaurante A e o restaurante B’.

Voluntários foram recrutados via e-mail e redes sociais. O experimento foi realizado via WEB, e por este motivo a qualidade dos dados coletados apresentava grande variação.

4.4.2 Resultados

A porção em português (aqui denominada Zoom-Pt) dos dados teve 93 participantes, e a porção em espanhol (Zoom-Sp) teve 85 participantes. Mediante verificação manual, 602 descrições mal formadas em português e 366 em espanhol foram descartadas. Assim, a porção em português do *córpus* (aqui denominada Zoom-Pt) contém 1358 descrições, e a porção em espanhol (Zoom-Sp) contém 1234 descrições.

Dada a complexidade do domínio, o esquema de anotação do *córpus Zoom* é consideravelmente mais elaborado do que outros estudos do gênero. Cada expressão de referência foi representada como contendo a descrição do alvo principal e, opcionalmente, até quatro pontos de referência relacionados. Cada objeto-alvo foi anotado com três atributos atômicos (tipo, nome e outros) e sete relações relacionais possíveis (e.g., acima, abaixo, na esquina com, etc.). Além disso, cada um dos quatro pontos de referência possíveis era anotado com os mesmos três atributos atômicos e seus identificadores, resultando assim em um esquema de anotação composto de 26 atributos para cada descrição de objeto singular. No caso de descrições de conjuntos, o número de atributos a serem anotados era dobrado.

As descrições da porção Zoom-Pt do *córpus* foram anotadas por dois anotadores independentes. Ao final da tarefa, um terceiro anotador assumiu o papel de juiz e encarregou-se de produzir a anotação final. Uma vez que o esquema de anotação era razoavelmente direto (em especial pelo uso do campo *others* para qualquer tipo de informação fora do esquema) a concordância Kappa (LANDIS; KOCH, 1977) entre juízes no nível do atributo foi de 84%. Assim como no caso dos *córpus Stars* e *Stars2*, esta porção dos dados foi também representada no formato XML

já discutido nas seções anteriores.

4.4.3 Contribuições

O *cópus Zoom* foi apresentado em Altamirano et al. (2015), juntamente com um experimento para geração destas descrições utilizando um modelo de aprendizagem baseado em máquinas de vetores suporte. O *cópus Zoom* ainda está sendo utilizado em uma pesquisa em nível de mestrado sob orientação deste autor, tratando da comunicação humano-computador (do tipo utilizada na construção deste experimento) e sua comparação com a modalidade de comunicação entre humanos (do tipo utilizado no experimento *Stars2*, da seção anterior), e ainda não foi publicamente disponibilizado para reuso.

4.5 Considerações finais

Neste capítulo foi discutido o processo de construção de *cópus* de GER, seus desafios e contribuições por meio de três exemplos de experimentos de coleta de dados para pesquisa de GLN sob a perspectiva do locutor (cf. capítulo 3). Partindo-se do princípio de que a língua gerada é plausível quando coincidente com a forma produzida por um locutor humano, estes conjuntos de dados oferecem subsídios variados para a investigação de diversas questões da área de GER (por exemplo, ligados à superespecificação referencial, à questão da escolha entre descrições atômicas e relacionais etc.). Além disso, todos os três *cópus* aqui apresentados se prestam ao estudo da questão da variação humana em GLN (a ser discutida no capítulo 6).

Os *cópus Stars* e *Stars2* seguiram a metodologia tradicional para construção de experimentos altamente controlados de GER, nos quais o contexto de referência é explicitamente manipulado de modo a estimular determinados tipos de comportamento humano a serem investigados. O experimento *Zoom*, por outro lado, foi criado com o propósito específico de obter exemplos linguísticos mais próximos dos que seriam observados em uma aplicação computacional real, fazendo assim uso de contextos de referência mais realistas e com manipulação mínima. O *cópus Zoom* é também o único destes três exemplos que fornece subsídios para pesquisa em realização superficial (neste caso bilíngue), embora todos os três conjuntos aqui discutidos incluam as formas superficiais originalmente produzidas. Estas diferenças não devem entretanto ser interpretadas como vantagem ou desvantagem de nenhum dos três *cópus*. Os domínios *Stars*, *Stars2* e *Zoom* são de natureza complementar, e cada *cópus* apresenta sua própria contribuição científica.

Como forma de melhor contextualizar estes estudos, a tabela 7 apresenta uma breve comparação entre os três corpú aqui discutidos e alguns dos principais recursos deste tipo que são publicamente disponíveis para pesquisa na área de GER, incluindo também o corpú *b5-ref* (PARABONI; MONTEIRO; LAN, 2017) que será discutido na seção 6.3.2. Nesta comparação, as informações relativas aos corpú TUNA e Zoom contemplam apenas a porção de descrições no singular destes conjuntos de dados, e o número de descrições do corpú Zoom-Sp é aproximado. Detalhes adicionais são discutidos a seguir.

Tabela 7: Corpú de GER.

Corpú	Descrições	Sujeitos	Atributos	Pt.ref.	Tamanho	Uso
TUNA-Furniture	1200	60	4	0	3,1	0,8
TUNA-People	1080	60	10	0	3,1	0,3
GRE3D3	630	63	9	1	3,4	0,3
GRE3D7	4480	287	6	1	3,0	0,4
Stars	704	64	8	2	4,4	0,4
Stars2	884	56	9	2	3,3	0,3
Zoom-Pt	669	100	19	4	6,7	0,3
Zoom-Sp	617	80	19	4	7,2	0,3
b5-ref	1810	152	27	0	4,0	0,2

Fonte: Ivandré Paraboni, 2017.

Na tabela 7, as colunas ‘Descrições’ e ‘Sujeitos’ apresentam estatísticas gerais de cada corpú. A coluna ‘Atributos’ indica o número de atributos atômicos possíveis em cada domínio (e é assim um indicativo da sua variedade), e a coluna ‘Pt.ref’ indica o número máximo de objetos de ponto de referência (ou ‘landmarks’) que pode ser relacionado a um objeto-alvo nas descrições de cada corpú (servindo como indicativo da sua densidade em número de objetos presentes em cada cena de estímulo). Nota-se que o corpú TUNA só possui descrições atômicas (com zero ponto de referência), enquanto que no corpú Zoom (cf. seção 4.4) um objeto-alvo pode ser relacionado a até quatro pontos de referência simultaneamente (e.g., ‘a farmácia entre as ruas X e Y, perto do teatro e do parque’).

A coluna ‘Tamanho’ apresenta o tamanho médio das descrições obtidas em cada corpú, em número de propriedades semânticas anotadas. Finalmente, a coluna ‘Uso’ indica a proporção média de propriedades possíveis que aparece em cada descrição. Do ponto de vista da tarefa computacional de GER, descrições mais longas e uso médio menor sugerem situações de referência mais complexas.

A investigação das principais questões de pesquisa que motivaram a construção dos corpú Stars, Stars2 e Zoom encontra-se concluída, e os resultados destes estudos foram publicados na sua totalidade. Assim, estes corpú são agora utilizados predominantemente como conjuntos de treinamento e teste de novos modelos de GER, mas não podem oferecer subsídios para a

investigação de questões de pesquisa realmente inéditas. Apesar dos esforços consideráveis investidos no desenvolvimento destes recursos - não por acaso suportados por um projeto de pesquisa de dois anos de duração e com apoio de uma instituição estrangeira - a observação feita no início deste capítulo permanece válida: novas questões de pesquisa tendem a exigir novos recursos. De modo geral, a busca pela plausibilidade da língua natural sob a perspectiva do locutor tende a exigir recursos de custo elevado deste tipo, e estes recursos tendem a se esgotar uma vez que as questões de pesquisa que os motivaram tenham sido investigadas. Uma alternativa de pesquisa em GLN de caráter complementar a esta perspectiva será o foco do próximo capítulo.

Capítulo 5

A PERSPECTIVA DO OUVINTE

5.1 Visão geral

No desenvolvimento de sistemas de GLN, a perspectiva do locutor discutida no capítulo 3 parte do princípio de que uma forma linguística plausível é pura e simplesmente aquela produzida por locutores humanos. No entanto, considerando-se que uma forma linguística - seja proveniente de um locutor humano ou de um modelo computacional - é necessariamente destinada a um agente receptor, e que procura exercer certos efeitos sobre este, é evidente que esta visão não é a única forma possível de desenvolver ou avaliar sistemas de GLN. A contrapartida da perspectiva do locutor - aqui denominada a perspectiva do ouvinte - é o foco deste capítulo.

Sob a perspectiva do ouvinte, a decisão sobre qual forma linguística gerar não é baseada na semelhança com exemplos de como um locutor humano produz língua natural. Ao invés disso, esta decisão é guiada pelos efeitos de diferentes estratégias de GLN sobre o comportamento do ouvinte, ou seja, avaliando-se a forma como o texto gerado automaticamente é recebido e interpretado. Esta diferença não é entretanto apenas uma questão de como interpretamos o conceito de plausibilidade da língua: as perspectivas do locutor e ouvinte guardam também diferenças metodológicas importantes. De modo geral, estudos de GLN baseados na perspectiva do ouvinte substituem a necessidade de um *cópus* de exemplos linguísticos pela medição de variáveis que representam o comportamento do ouvinte em resposta às formas linguísticas de interesse (e.g., provenientes de um sistema de GLN previamente existente). Exemplos de medição deste tipo incluem questionários de julgamento humano, tempos de leitura, erros de interpretação textual, ações de navegação em estruturas de *hyperlinks*, tempos ou distâncias de navegação em mundos virtuais, movimentos ou tempos de fixação do olhar, potenciais relacionados a eventos (ERP) e muitos outros.

Assim como em estudos sob a perspectiva do locutor, estudos de GLN sob a perspectiva do ouvinte frequentemente fazem uso de experimentos controlados envolvendo sujeitos humanos, e que podem ter alto custo de implementação. Além disso, a medição das variáveis de interesse pode exigir um investimento inicial alto em recursos de hardware ou software especializados, como sistemas que implementam mundos virtuais navegáveis, equipamento de rastreamento de olhar, câmeras e outros. Entretanto, cabe destacar que estes recursos, de modo geral, possuem ampla capacidade de reúso. Experimentos que medem o comportamento do ouvinte podem ser facilmente modificados, estendidos ou adaptado a outros estudos, muitas vezes com a simples substituição das formas linguísticas sob avaliação.

Como forma de ilustrar a perspectiva do ouvinte em GLN, neste capítulo serão discutidos dois exemplos deste tipo de pesquisa: um estudo sobre o impacto de diferentes estratégias de geração de instruções textuais sobre o tempo e distância de navegação de usuários em mundos virtuais (PARABONI; DEEMTER, 2014), e os efeitos da superespecificação referencial sobre o esforço exigido para a identificação de objetos em cenas visuais com uso de técnicas de rastreamento de olhar (PARABONI et al., 2017). O primeiro destes estudos é o resultado do projeto de pós-doutorado de curta duração (agosto-2011 a fevereiro 2012) desenvolvido junto à University of Aberdeen (Escócia) intitulado *‘Extensão de Algoritmos de Geração de Expressões de Referência para Domínios Físicos’* (CAPES nro. 6415/10-5), e o segundo é fruto de colaborações com alunos de graduação e um pesquisador da área de rastreamento de olhar no contexto do projeto de pesquisa regular intitulado *‘Geração de língua natural para descrição de objetos em contextos visuais’* (FAPESP nro. 2013/23169-0).

5.2 Efeitos da estratégia referencial e memória do ouvinte sobre a busca em domínios espaciais

A popularização da plataforma GIVE (BYRON et al., 2007) para desenvolvimento de sistemas de geração de instruções em mundos virtuais e a série de desafios computacionais a ela associados (BYRON et al., 2009; KOLLER et al., 2010; STRIEGNITZ et al., 2011) viabilizou a investigação de diversas questões de pesquisa em domínios deste tipo, como o estudo sobre a geração de relações espaciais apresentado na seção 3.4 anterior. Além deste, a plataforma GIVE foi utilizada também no estudo apresentado em Paraboni e Deemter (2014), que investiga o impacto de diferentes estratégias de GER e da memória na busca por referentes com diferentes graus de saliência perceptual sobre o comportamento de navegação de usuários GIVE. Este estudo é discutido a seguir.

O trabalho em Paraboni e Deemter (2014) é um exemplo típico da visão de pesquisa de GLN aqui denominada ‘perspectiva do ouvinte’. Neste trabalho, o ambiente GIVE - originalmente projetado para teste de sistemas de geração de instruções de navegação, cf. Byron et al. (2007) - foi adaptado de modo específico para a avaliação de diferentes estratégias de GER (e.g., contemplando diferentes graus de especificação referencial), e estendido para cômputo de diversas métricas de tempo e distância de navegação indicativas do esforço de busca pelo objeto-alvo da expressão gerada de forma automática. Cabe assim ressaltar que, diferentemente do trabalho em Silva e Paraboni (2015), o presente estudo não procura resolver nenhum problema específico de sistemas de navegação em mundos virtuais, mas sim utilizar a infraestrutura GIVE existente para um estudo de produção de expressões de referência que podiam envolver o uso de relações espaciais ou não.

O ambiente GIVE modificado foi utilizado em um experimento controlado em que sujeitos humanos eram solicitados a seguir instruções contendo expressões geradas por uma série de algoritmos de GER de interesse, e que eram produzidas em contextos em que certos objetos menos salientes (botões interativos) poderiam ser parcial ou totalmente ocultados por outros que eram mais salientes (e.g., *landmarks* como plantas ou peças de mobília). Um exemplo deste ambiente é ilustrado pela figura 19.

Figura 19: Interface do experimento, incluindo quatro pontos de referência de alto grau de saliência (plantas e cadeiras) e objetos do tipo botão de baixa saliência (ocultos).

Fonte: Paraboni e Deemter (2014).

Uma visão esquemática do mesmo contexto ilustrado na figura 19 é apresentada na figura 20, onde são evidenciados os quatro objetos ocultos ao final de corredores atrás de cada *landmark*. Com base em uma série de contextos deste tipo, o experimento realizado consistiu em gerar instruções de acordo com várias estratégias de GER sob avaliação como ‘pressione o botão azul

atrás da planta’ e monitorar os passos de navegação do ouvinte do momento em que a instrução era fornecida até a identificação do objeto-alvo (sinalizada pela execução da tarefa solicitada, ou seja, pressionando-se o botão em questão).

Figura 20: Mapa esquemático representando o contexto da figura 19, com quatro botões ocultos (três azuis e um verde). O ouvinte está posicionado em [start] e o alvo está localizado em [target].

Fonte: Adaptado de Paraboni e Deemter (2012).

O estudo tratou de questões relacionadas ao impacto de diferentes estratégias de GER e da memória (ou conhecimento prévio do ouvinte) sobre a identificação do objeto-alvo em domínios espaciais tridimensionais, em contraste aos domínios bidimensionais estáticos normalmente considerados na pesquisa da área. De forma mais específica, procurou-se demonstrar que expressões breves - do tipo que seria preferido de acordo com a máxima de quantidade de Grice (GRICE, 1975) - podem causar dificuldades de identificação do objeto-alvo em domínios espaciais deste tipo (em que objetos são percebidos como tendo graus variados de saliência) mesmo quando são perfeitamente corretas e livres de ambiguidade.

Considere-se por exemplo o objetivo de referenciar o objeto-alvo definido como *target* no contexto das figuras 19 e 20, em que tanto o objeto-alvo como dois distraidores compartilham o mesmo tipo (botão) e cor (azul). Uma vez que nenhum destes distraidores é inicialmente visível do ponto de vista do ouvinte em *start*, o sistema de GLN precisa assim gerar uma instrução fazendo referência ao ponto de referência (planta ou cadeira) mais próximo. Neste caso, um exemplo de instrução possível - e que de fato seria gerada pela grande maioria dos algoritmos tradicionais de GER - seria na forma ‘o botão azul *atrás da planta*’. Expressões deste tipo

são livres de ambiguidade, mas acarretam grande esforço de busca (i.e., levando o ouvinte a examinar regiões desnecessárias do contexto) pelo fato de que os objetos distraidores (botões) possuem baixa saliência relativa. Além disso - e de forma especialmente relevante para o estudo realizado - há o agravante de que o contexto apresenta outro objeto do tipo planta com maior saliência (por estar mais próximo do ouvinte), e que a expressão assim tende a produzir uma interpretação errônea que levaria o ouvinte a examinar a região onde se encontra o botão verde, e onde obviamente o objeto-alvo não se encontra. Este tipo de situação exigiria não apenas maior esforço de busca, mas suscitaria dúvidas sobre a própria identidade do objeto-alvo e sobre a correção da instrução fornecida e, uma vez instauradas, dúvidas deste tipo podem persistir mesmo após a localização do objeto correto (PARABONI; DEEMTER, 2014)

Problemas de identificação como o deste exemplo sugerem um limite inferior para o volume de informação a ser incluída na descrição do objeto-alvo. Este limite, que está relacionado ao grau de saliência dos diversos obstáculos presentes no contexto, é o foco de uma teoria de busca proposta em Paraboni e Deemter (2014), e que não é discutida em mais detalhes na presente monografia. Como forma de evitar diversos problemas de busca e identificação como este e outros, discutidos em Paraboni e Deemter (2014), uma alternativa mais adequada neste caso seria superespecificar a descrição, produzindo algo como ‘Pressione o botão azul atrás da planta, à direita’. No entanto, mesmo sendo uma estratégia à primeira vista simples, não é óbvio quando este tipo de superespecificação seria de fato recomendado, já que há bons motivos para evitar a produção de expressões mais longas do que o estritamente necessário (GRICE, 1975). Em outras palavras, há também um limite superior para o volume de informação a ser incluída na expressão, e este limite também está relacionado ao grau de saliência dos elementos contextuais.

Além de limites inferior e superior para o grau de especificação referencial, o experimento realizado procurou ainda demonstrar que as dificuldades relacionadas à interpretação de expressões com diferentes graus de especificação podem ser amenizadas - ou mesmo suprimidas - nos casos em que há conhecimento prévio do domínio. Em especial, um ouvinte com memória ‘perfeita’, ou com amplo conhecimento do contexto em questão (por exemplo, porque possui um mapa do domínio como o da figura 20 anterior), poderia identificar o objeto-alvo corretamente mesmo quando instruído de forma imperfeita, como em ‘Pressione o botão verde’. Com base nestas observações, o experimento investigou problemas de busca tanto em situações em que o contexto era visitado pela primeira vez como em situações em que o participante tinha desempenhado uma tarefa de reconhecimento prévio do ambiente (e que portanto tinha alguma memória da sua estrutura e da distribuição dos objetos).

5.2.1 Materiais e método

O projeto do experimento levou às adaptações do sistema GIVE discutidas na seção anterior para implementação das métricas de tempo e distância de navegação. Além disso, como forma de implementar as condições experimentais de forma controlada (e.g., evitando possível ruído decorrente de certas decisões de apresentação, como os tipos de objetos exibidos, cores, posicionamento à esquerda ou direita etc.) foi desenvolvido um procedimento para geração de mundos virtuais parametrizáveis em tempo de execução. Em outras palavras, diversos aspectos da exibição do contexto eram customizados automaticamente e de forma aleatória para cada participante e para cada condição experimental, constituindo uma ferramenta completa para realização de experimentos psicolinguísticos deste tipo para a plataforma GIVE.

O experimento contou com 50 voluntários que receberam uma pequena remuneração pela participação. Os experimentos foram realizados individualmente, em sessões de 30-60 minutos cada, incluído aqui o tempo de prática para familiarização com a interface GIVE.

Conforme já mencionado, o procedimento propriamente dito consistiu em fornecer instruções geradas automaticamente segundo as diversas estratégias de GER sob avaliação, e em medir os tempos e distâncias de busca por elas exigidos. Os algoritmos em questão eram assim avaliados com base nestas métricas, e estas eram portanto tomadas como sendo indicativas da qualidade da solução computacional sob a perspectiva do ouvinte (i.e., o usuário do sistema).

5.2.2 Resultados

Como resultado do experimento, foi coletada uma grande massa de dados contendo informações sobre o comportamento de cada participante na tarefa de identificação de cada objeto-alvo. Estes dados incluem o registro dos tempos e distâncias de navegação necessários para identificar o objeto-alvo em cada condição experimental (e.g., descrições de diferentes graus de especificação, e com ou sem conhecimento prévio do domínio), e também o registro completo de cada trajetória e dos tempos de inspeção de todos os objetos encontrados ao longo do percurso.

Os resultados apresentados em Paraboni e Deemter (2014) comprovam as hipóteses para os limites inferior e superior de superespecificação referencial com base na saliência relativa dos objetos contextuais, e oferece subsídio para a proposta de uma nova estratégia de seleção de conteúdo para domínios espaciais - em que a identificação do objeto-alvo pode ter que considerar obstáculos naturais (i.e., *landmarks* de alto grau de saliência, como a planta do exemplo anterior) - que generaliza a noção de busca e produz descrições que minimizam o esforço de localização do objeto-alvo.

5.2.3 Contribuições

O estudo em Paraboni e Deemter (2014) apresentou contribuições potencialmente úteis em diversas frentes. Do ponto de vista científico, o trabalho realizado apresenta contribuições teóricas para a área de GER ao demonstrar o efeito da memória combinada a diferentes tipos de superespecificação referencial sobre a tarefa de busca em domínios espaciais. Este estudo difere de outros trabalhos da área de GER também por investigar os efeitos de interesse de forma contínua, ou seja, com base no tempo e distância de navegação no ambiente virtual, e não com base em medidas discretas de acurácia.

Do ponto de vista metodológico, o trabalho propôs uma transformação do sistema GIVE (BYRON et al., 2007) em um ambiente de experimentação em GER com a adição de vários recursos inéditos, como a geração dinâmica de contextos e a implementação de novas métricas de avaliação. Neste sentido, o método utilizado pode ser visto como um novo paradigma de avaliação de algoritmos de GER, e foi de fato um dos primeiros estudo do gênero a utilizar o ambiente GIVE de forma extensiva para um propósito de pesquisa diferente do originalmente previsto¹.

Finalmente, cabe destacar que o estudo realizado foi um primeiro exemplo de experimento psicolinguístico relativamente complexo, cuja metodologia seria uma influência para estudos posteriores realizados por este autor e seus colaboradores. Estes incluem as iniciativas de construção de *córpus* detalhadas no capítulo anterior (TEIXEIRA et al., 2014; PARABONI; GALINDO; IACOVELLI, 2017; ALTAMIRANO et al., 2015) e o experimento em Paraboni et al. (2017). Além disso, algumas das experiências adquiridas com o estudo em Paraboni e Deemter (2014) seriam posteriormente aproveitados no projeto apresentado na seção 3.4, que também fez uso do sistema GIVE, porém neste caso desenvolvendo um estudo de GLN sob a perspectiva do locutor.

5.3 Esforço cognitivo e superespecificação

Novas tecnologias propiciam novas oportunidades de investigação do impacto das decisões de GLN segundo a perspectiva do ouvinte. De interesse específico para a presente discussão, consideramos aqui o uso de técnicas de rastreamento de olhar (ou *gaze tracking*) (DUCHOWSKI, 2002; MORIMOTO; MIMICA, 2005; HANSEN; JI, 2010). Técnicas deste tipo são amplamente utilizadas na pesquisa em psicolinguística, interação humano-computador e muitas outras. No caso específico da pesquisa em GLN/PLN, o rastreamento de olhar pode viabilizar a observação

¹Estudos como Koller et al. (2012) também podem ser categorizados desta forma.

direta de certos fenômenos que não seria possível, por exemplo, com uso de análise de corpus ou outras técnicas mais tradicionais (e.g., segundo a perspectiva do locutor).

O estudo em Paraboni et al. (2017) utiliza rastreamento de olhar para medir as reações de sujeitos humanos expostos a alternativas textuais produzidas por diferentes algoritmos de GER de interesse. Neste caso, o tempo de fixação do olhar do ouvinte sobre objetos em um contexto visual é tomado como indicativo do esforço cognitivo exigido para a interpretação da forma linguística e, por extensão, como indicativo do desempenho do algoritmo que a produziu. Trata-se assim de um exemplo típico de estudo de GLN sob a perspectiva do ouvinte no sentido aqui discutido.

O estudo foi motivado pela observação fundamental de que, apesar do grande corpo de trabalho existente na área de superespecificação referencial, seu efeito sobre a interpretação destas expressões ainda é pouco conhecido, e que diferentes estudos oferecem explicações aparentemente contraditórias a este respeito. De um lado, trabalhos como em Arts et al. (2011) e Paraboni e Deemter (2014) demonstram que a superespecificação facilita a identificação do objeto-alvo por oferecer mais informação a respeito dele. Por exemplo, em uma cena representando uma multidão de pessoas, dentre as quais há apenas uma mulher, pode ser mais fácil identificar esta pessoa com base em uma expressão superespecificada como ‘a mulher *de vermelho*’ do que com base em uma expressão mínima como ‘a mulher’. De outro lado, trabalhos como Engelhardt, Demiral e Ferreira (2011) e Engelhardt, Baileyand e Ferreira (2006) sugerem o efeito exatamente oposto, ou seja, que a superespecificação referencial exige mais esforço cognitivo por parte do ouvinte pelo fato de que ele ou ela precisa processar um volume maior de informação. Por exemplo, em um contexto em que há apenas um objeto do tipo quadrado, é postulado que o acréscimo de informação redundante - como a referência ao atributo cor em ‘o quadrado *vermelho*’ - acarreta maior tempo de reação por parte do ouvinte do que uma descrição mínima como ‘o quadrado’.

Diante destas visões aparentemente conflitantes sobre o fenômeno de superespecificação referencial, o objetivo do trabalho em Paraboni et al. (2017) foi o de propor uma teoria capaz de conciliá-las. De forma mais específica, o estudo realizado investiga a simples hipótese de que a superespecificação referencial pode tanto facilitar como prejudicar a identificação de um objeto *dependendo do tipo de informação que for acrescentada*.

Como forma de testar esta hipótese, foram realizados dois experimentos envolvendo sujeitos humanos e que utilizam técnicas de rastreamento de olhar para medir o tempo de inspeção exigido por expressões de referência produzidas segundo diferentes estratégias de GER com e sem superespecificação referencial. O primeiro experimento trata da questão da superespecificação

utilizando propriedades *relacionais*, e o segundo complementa os resultados obtidos abordando a questão da superespecificação utilizando propriedades *atômicas*. Estes dois experimentos são discutidos brevemente a seguir.

5.3.1 Experimento 1: Propriedades relacionais

O objetivo do primeiro experimento realizado em Paraboni et al. (2017) foi o de verificar as circunstâncias em que a superespecificação de propriedades relacionais podia facilitar ou dificultar a identificação do objetos-alvo. A figura 21 apresenta um exemplo de estímulo utilizado para este fim, composto de uma série de objetos do tipo estrela de sete ou oito pontas contendo rótulos alfanuméricos no centro. Os objetos são distribuídos a intervalos regulares ao longo da imagem como forma de facilitar o uso da técnica de rastreamento de olhar.

Figura 21: Exemplo de contexto com estrelas de 7/8 pontas contendo letras ou números (objeto-alvo destacado para fins ilustrativos).

Fonte: Adaptado de Paraboni et al. (2017).

Em um contexto deste tipo, partiu-se do pressuposto de que decidir se uma determinada estrela possui 7 ou 8 pontas é *mais difícil* do que decidir se um rótulo representa uma letra ou número. Mais especificamente, partiu-se do pressuposto de que um sujeito humano precisaria de *mais tempo* examinando um determinado objeto para determinar o seu formato do que para determinar o seu tipo de rótulo².

O experimento realizado fez uso de séries de imagens de estímulo como a deste exemplo, representando condições em que uma descrição poderia ser mínima ou superespecificada, e contendo propriedades de reconhecimento mais difícil (e.g., fazendo referência ao formato da estrela) ou mais fácil (e.g., fazendo referência ao tipo de rótulo). As condições experimentais - detalhadas em Paraboni et al. (2017) - foram então testadas com auxílio de sujeitos humanos em

²Embora plausível neste contexto, este pressuposto também foi validado como parte do estudo realizado.

uma tarefa de identificação monitorada por rastreamento de olhar. Partindo-se de uma instrução linguística do tipo ‘Olhe para o objeto *X*’, na qual *X* é uma descrição definida não ambígua produzida por um algoritmo de interesse, o dispositivo de rastreamento de olhar media o tempo gasto pelos participantes examinando cada objeto da cena, de modo a estabelecer o impacto de estratégias variadas de superespecificação na resolução destas expressões.

De especial relevância para esta metodologia é a questão do tipo de medição realizada. Em testes preliminares realizados para refinamento dos materiais e métodos do experimento, foi observado que o tempo de reconhecimento do objeto-alvo tende a apresentar forte ruído, e que este tipo de medição não se prestaria aos objetivos do estudo proposto. Uma possível explicação para esta discrepância é o fato de que a identificação do objeto-alvo inclui a decisão consciente do sujeito de concluir a tarefa ou não (e, uma vez identificado o alvo, pressionar a barra de espaço para prosseguir). Por este motivo, o experimento considerou o tempo médio de inspeção dos objetos *distraidores*, e não o tempo de inspeção do objeto-alvo, como à primeira vista talvez parecesse mais intuitivo. Neste experimento, o cômputo do tempo de inspeção de objetos *distraidores* garantiu não apenas maior acurácia (já que os participantes estavam concentrados na tarefa de identificar o objeto-alvo) como também proporcionou a coleta de um volume maior de dados (já que as cenas exibem sete objetos *distraidores* para cada alvo).

O procedimento adotado consistiu em exibir cada um dos estímulos ao participante em ordem aleatória, e rastrear seu olhar do momento em que a expressão de referência era exibida no centro da tela, até a fixação sobre o objeto-alvo em questão. Ao pressionar a tecla de espaço, o participante indicava a conclusão da identificação do objeto, e o estímulo seguinte era exibido. Nenhuma menção era feita ao papel dos objetos *distraidores* (que eram o verdadeiro interesse do presente experimento). A participação no experimento foi realizada de forma individual em horário pré-agendado com um total de 25 sujeitos. Após a calibração do dispositivo de rastreamento de olhar, os participantes recebiam instruções sobre como proceder e tinham algum tempo para se familiarizar com a tarefa utilizando imagens de prática semelhantes aos estímulos reais do experimento. Após a sessão prática, seguia-se o experimento propriamente dito.

Os dados coletados - inicialmente na forma de um fluxo de coordenadas da tela indicando o foco de atenção visual de cada indivíduo - foram processados para extrair a sequência de objetos examinados pelo participante e o tempo de inspeção de cada um, começando pelo primeiro objeto visualizado na tela, e terminando no momento em que o objeto-alvo é visualizado pela primeira vez. Os resultados obtidos neste experimento oferecem suporte à hipótese de que o impacto da superespecificação referencial sobre a identificação depende do tipo de propriedade considerada, ou seja: acrescentar propriedades de difícil reconhecimento (como o formato da

estrela no exemplo) tornam a tarefa de identificação do objeto-alvo mais difícil do que seria caso não fosse acrescentada nenhuma informação, e acrescentar propriedades de fácil reconhecimento (como o rótulo da estrela no exemplo) tornam a tarefa mais fácil.

5.3.2 Experimento 2: Propriedades atômicas

O experimento, descrito na seção anterior fez uso de exemplos extremos de propriedades referenciais ditas ‘fáceis’ e ‘difíceis’ para investigar a hipótese de que o esforço cognitivo envolvido no reconhecimento destas propriedades é um fator determinante do tipo de impacto que a superespecificação referencial terá sobre a identificação do objeto-alvo. No experimento seguinte, procurou-se assim demonstrar que estes princípios permanecem válidos mesmo quando a diferença entre graus de esforço cognitivo é menos evidente. Além disso, como forma de ampliar o escopo dos resultados obtidos, considerou-se neste segundo experimento o caso de superespecificação com uso de propriedades atômicas, e não relacionais. Este experimento é discutido brevemente a seguir.

O experimento realizado fez uso de um domínio representando propriedades mais gerais ou mais específicas, as quais foram representadas por cenas contendo letras (vogais e consoantes) e números (pares e ímpares). Um exemplo de estímulo utilizado neste experimento é ilustrado na figura 22.

Figura 22: Exemplo de contexto baseado em letras e números (objeto-alvo destacado para fins ilustrativos).

Fonte: Adaptado de Paraboni et al. (2017).

De forma análoga ao experimento da seção anterior, o pressuposto neste caso é o de que o uso de informação mais específica (e.g., ‘Olhe para o número seguido de uma *consoante*’) exigiria tempo de reconhecimento maior do que o uso de informação mais geral (e.g., ‘Olhe para o número seguido de *uma letra*’). Para teste desta hipótese, foi adotado o mesmo procedimento

do experimento anterior, e utilizando-se a mesma infraestrutura para rastreamento de olhar com o auxílio de um novo grupo de 25 participantes voluntários.

Os resultados obtidos neste experimento confirmam a hipótese de que o uso de uma propriedade mais geral (e.g., ‘letra’) está relacionado a um tempo menor de inspeção de objetos em relação à expressão mínima, e que no uso de uma propriedade mais específica (e.g., ‘consoante’) é observado um efeito no sentido oposto. Em outras palavras, a superespecificação utilizando propriedades atômicas mais gerais facilita a identificação do objeto-alvo, enquanto que o uso de propriedades mais específicas dificulta esta tarefa. Deste modo, além de confirmar os resultados do experimento da seção anterior, o presente experimento generaliza estes resultados para o caso de referência utilizando propriedades atômicas.

5.3.3 Contribuições

Ambos os experimentos aqui relatados - tratando do uso de propriedades relacionais e atômicas, respectivamente - comprovam a hipótese de que a superespecificação referencial pode ter impacto positivo ou negativo sobre a identificação. Propriedades que demandam baixo esforço cognitivo facilitam a identificação (como em Arts et al. (2011) e Paraboni e Deemter (2014)) e propriedades que demandam maior esforço têm o efeito oposto (como em Engelhardt, Demiral e Ferreira (2011) e Engelhardt, Baileyand e Ferreira (2006)).

Apesar de à primeira vista simples, o trabalho realizado representa uma contribuição potencialmente importante para a área de GER, ao propor pela primeira vez - ainda que em caráter preliminar - uma teoria que procura unificar visões divergentes sobre este fenômeno. Essencialmente, os dois experimentos realizados sugerem que o acréscimo de informação redundante não é inerentemente positivo nem negativo para a identificação do objeto-alvo. A identificação se torna mais difícil do que seria de outra forma necessário (e.g., usando uma expressão mínima) quando a expressão inclui propriedades de difícil reconhecimento, e a identificação se torna mais fácil do que seria de outra forma possível (e.g., também com base em uma expressão mínima) quando a expressão inclui informação de fácil reconhecimento.

5.4 Considerações finais

Neste capítulo foi discutida a questão do desenvolvimento de sistemas ou artefatos de GLN sob a perspectiva do ouvinte. A discussão foi ilustrada com dois exemplos de pesquisa em GLN em que foram abordados aspectos específicos da questão da superespecificação referencial.

O estudo em Paraboni e Deemter (2014) fez uso da infraestrutura de planejamento e suporte à navegação em mundos virtuais disponibilizada pela plataforma GIVE (BYRON et al., 2007), estendida de modo a propiciar a avaliação de diferentes estratégias de GER com base em diversas métricas de desempenho da tarefa de navegação (e.g., baseadas em tempo, distância, complexidade do percurso etc.). Assim, são consideradas mais plausíveis - sob a ótica do ouvinte ou receptor das instruções geradas - as expressões que facilitam a navegação do mundo virtual e a execução de tarefas associadas à aplicação. Os resultados deste estudo identificam certos tipos de situações em que a saliência das propriedades do objeto-alvo pode não ser suficiente para sua identificação por meio de descrições mínimas, e motivam a produção de descrições superespecificadas como forma de minimizar erros ou outras dificuldades de navegação. O estudo propõe assim uma estratégia algorítmica de uso judicioso da informação redundante como forma de prevenir estas dificuldades, e que possui aplicações imediatas para sistemas de GLN em domínios espaciais interativos.

O estudo em Paraboni et al. (2017) fez uso de hardware e software específico de rastreamento de olhar para medir o esforço cognitivo exigido por expressões de referência produzidas segundo diferentes estratégias de superespecificação. Neste caso são consideradas mais plausíveis - novamente sob a ótica do ouvinte - as expressões que minimizam o tempo de reconhecimento do objeto-alvo em questão. Os resultados deste estudo representam uma contribuição teórica potencialmente relevante para a área ao conciliar visões conflitantes provenientes de trabalhos prévios sobre o papel deste tipo de conhecimento na tarefa de GER, e abrem caminho para o futuro projeto de algoritmos orientados ao ouvinte com o objetivo de minimizar o esforço cognitivo exigido para a identificação.

Tanto no caso do uso do ambiente virtual interativo como no uso de rastreamento de olhar, observa-se que houve grande esforço inicial de desenvolvimento (de métricas de navegação, rotinas de captura e tratamento do foco de atenção visual, cômputo de tempos de fixação etc.) para criação do ambiente de experimentação exigido. Uma vez realizados os estudos originalmente previstos, entretanto, estes recursos permanecem válidos e com alto potencial de reúso em novas pesquisas, de modo geral exigindo apenas a substituição dos estímulos utilizados para que a investigação de novas questões de pesquisa seja viabilizada.

Em ambos os estudos aqui discutidos, observa-se também que a plausibilidade da língua gerada foi medida não apenas para fins de avaliação de algoritmos de GER pré-existentes, mas principalmente com o objetivo de ganhar um maior entendimento sobre fenômenos linguísticos de interesse. Não por acaso os exemplos aqui apresentados possuem um caráter psicolinguístico mais pronunciado do que os estudos sob a perspectiva do locutor discutidos no capítulo 3, de

caráter claramente computacional e com forte ênfase na avaliação de artefatos e sistemas. É importante ressaltar entretanto que esta relação é incidental, e que ambas as perspectivas se prestam indistintamente a estudos de caráter computacional e psicolinguístico. Por exemplo, estudos sob a perspectiva do locutor como os de construção de corpus de GLN discutidos no capítulo 4 possuem uma contribuição evidentemente psicolinguística, na qual a possível aplicação computacional é de certa forma secundária. Da mesma forma, observa-se que os métodos orientados ao ouvinte - como o monitoramento de passos de navegação ou rastreamento de olhar aqui discutidos - são técnicas comuns de avaliação de sistemas computacionais em diversas áreas.

Capítulo 6

PLAUSÍVEL PARA QUEM?

6.1 Visão geral

Nos capítulos anteriores foram discutidas duas visões da plausibilidade da língua natural a ser gerada automaticamente: a perspectiva do locutor e a do ouvinte. Apesar de serem em certo sentido opostas, estas duas visões partem do pressuposto comum de que a plausibilidade da língua seja determinada a partir do *comportamento médio de um grupo* de indivíduos, sejam estes locutores (e.g., que produzem córpus para GLN), ou ouvintes (e.g., que respondem a formas linguísticas produzidos por artefatos de GLN). Em outras palavras, os estudos discutidos até aqui partem do princípio de que os sistemas a serem desenvolvidos serão baseados em estratégias fixas de GLN, e que estas estratégias seriam, por definição, suficientemente plausíveis sob a ótica de todos os locutores ou ouvintes que elas pretendem representar.

Modelos de GLN baseados no comportamento médio de um grupo de locutores ou ouvintes são entretanto uma simplificação da realidade, pois não há uma forma única ou ideal de expressar conteúdo em língua natural. Pessoas diferentes podem se expressar de forma distinta, e um mesmo indivíduo pode se expressar de forma distinta em diferentes ocasiões. Esta variação humana é inerente à produção de língua natural, e está refletida em todas as suas etapas, incluindo-se a seleção de conteúdo e a realização superficial. Um sistema de GLN que se limite a reproduzir o comportamento médio de um grupo de locutores ou ouvintes, como os exemplos discutidos nos capítulos anteriores, será assim considerado mais ou menos plausível para cada indivíduo que utilizá-lo.

Estas observações nos levam a questionar *para quem* um sistema de GLN deve de fato ser plausível. Por um lado, se decidirmos que o sistema deva contemplar um usuário médio dentro de um grupo-alvo de interesse, então as abordagens dos capítulos anteriores serão em

princípio satisfatórias. Por outro lado, se decidirmos que o texto a ser gerado deva ser de alguma forma customizado para cada indivíduo de interesse - de modo que a língua gerada seja o mais plausível possível - então torna-se necessário incorporar a noção de *variação humana* ao modelo pretendido.

Este capítulo discute duas interpretações possíveis da questão da variação humana em GLN: o uso de métodos de aprendizagem de máquina para construção de modelos de GER dependentes do locutor, ou seja, que levam em conta preferências individuais para geração de língua natural customizada para cada locutor de interesse (FERREIRA; PARABONI, 2017a); e modelos de interpretação e geração de língua natural baseados em uma nova visão de variação humana que faz uso do conceito de *traços de personalidade* propostos na área de Psicologia (SANTOS; PARABONI; SILVA, 2017; PARABONI; MONTEIRO; LAN, 2017). O primeiro conjunto de trabalhos é parte integrante do já citado projeto de pesquisa regular ‘*Geração de língua natural para descrição de objetos em contextos visuais*’ (FAPESP nro. 2013/23169-0), e o segundo é um resultado inicial do recente projeto de pesquisa regular ‘*Tratamento Computacional da Personalidade Humana para Aplicações de Processamento de Língua Natural*’ (FAPESP nro. 2016/14223-0). Em ambos os casos, os resultados aqui apresentados são fruto de orientações em nível de mestrado e contam com a colaboração de alunos de graduação.

6.2 Modelos de locutores individualizados

Em anos recentes, a pesquisa em GER produziu diversos conjuntos de expressões de referência anotados semanticamente - ou *cópus* para GER - como TUNA (GATT; SLUIS; DEEMTER, 2007), GRE3D3/7 (DALE; VIETHEN, 2009; VIETHEN; DALE, 2011), e recursos similares desenvolvidos por este autor e seus colaboradores, como os já citados *cópus* Stars (TEIXEIRA et al., 2014), Stars2 (PARABONI; GALINDO; IACOVELLI, 2017) e Zoom (ALTAMIRANO et al., 2015). A disponibilização de recursos deste tipo, aliada ao fato de que nestes recursos é normalmente possível distinguir o locutor que produziu cada descrição disponibilizada, levou a um estudo em nível de mestrado enfocando a questão da variação humana na tarefa de seleção de conteúdo para GER.

O estudo realizado propõe novos modelos de GER capazes de gerar expressões de referências customizadas para cada locutor-alvo de interesse, e com acurácia superior à de modelos baseados no comportamento médio de um grupo de locutores. Para este fim, o trabalho realizado envolveu uma investigação sobre o uso de métodos de aprendizagem de máquina na tarefa de GER (FERREIRA; PARABONI, 2014a), sobre a questão da variação humana propriamente dita

(FERREIRA; PARABONI, 2014b) e da clusterização de locutores (FERREIRA; PARABONI, 2017b). Os resultados finais deste estudo foram apresentados em Ferreira e Paraboni (2017a) e são discutidos resumidamente a seguir.

6.2.1 Modelos propostos

O objetivo do estudo em Ferreira e Paraboni (2017a) foi demonstrar que, havendo disponibilidade de dados de treinamento em volume suficiente, (a) o uso de conhecimento sobre o locutor (ou sobre seu comportamento linguístico) aumenta a acurácia de algoritmos de GER; e (b) métodos de aprendizagem de máquina podem apresentar desempenho superior ao uso do algoritmo Incremental (DALE; REITER, 1995) nesta tarefa. Para investigar estas questões, foram propostos diversos modelos de GER baseados em métodos de aprendizagem de máquina supervisionada - em especial, fazendo uso de máquinas de vetores suporte (SVM) e árvores de decisão - com e sem uso de informação sobre o locutor que produziu cada descrição. Assim, a pesquisa realizada adota mais uma vez a perspectiva do locutor (cf. capítulo 3), fazendo uso extensivo de *cópus* para fins de treinamento e teste de modelos computacionais de GER.

A abordagem de GER proposta em Ferreira e Paraboni (2017a) é baseada em uma série de classificadores binários que determinam se cada atributo referencial do objeto-alvo em questão deve ou não ser selecionado para compor a expressão de referência de saída. Assim como em modelos de GER tradicionais, a seleção de conteúdo é baseada em informações contextuais como o poder discriminatório de cada atributo referencial e a frequência relativa com que cada atributo aparece nas descrições disponibilizadas na porção de treinamento dos dados. No presente caso, estas informações são modeladas na forma de características de aprendizagem para uso nas diversas tarefas de classificação a serem realizadas.

Diferentemente do trabalho prévio sobre este tema, entretanto, os modelos propostos exploram também a questão da variação humana em GER, ou seja, incluem características adicionais relativas a cada locutor e a cada descrição por ele/a gerada. Estas características - que compõem o modelo de locutor proposto - são consideradas em diversas combinações, incluindo desde o uso de um identificador único do locutor como em Viethen e Dale (2010), até informações específicas sobre a estratégia referencial de cada indivíduo (e.g., os atributos mais frequentemente utilizados, o tamanho médio das expressões produzidas etc.).

Em linhas gerais, a abordagem proposta em Ferreira e Paraboni (2017a) funciona da seguinte forma. Assim como em outros estudos da área, dado um objeto-alvo r que se deseja descrever, e um contexto C formado pelos objetos distraidores de interesse, o objetivo é produzir um conjunto de propriedades L que permitam distinguir r dos demais objetos em C . O primeiro passo do

algoritmo proposto é assim obter as estimativas fornecidas pelo conjunto de classificadores (construídos a partir da porção de treinamento dos dados) associados a cada um dos atributos possíveis de r , e a estimativa de uma possível propriedade relacional que, se utilizada, permitiria complementar a descrição de r com uma referência a um segundo objeto (ou *landmark*) lm .

Com base nestas estimativas, a seleção de conteúdo consiste em avaliar os atributos do objeto-alvo r um a um, na ordem de preferência específica para cada locutor (informação também obtida a partir da porção de treinamento dos dados), e selecionar todos os atributos que sejam recomendados pelo respectivo classificador. É importante destacar que, ao contrário da abordagem fundamental de GER discutida na seção 2.1.2, a proposta não faz nenhum teste explícito com base no poder discriminatório dos atributos, ou seja, todos atributos recomendados pelo classificador são selecionados independentemente de contribuir para a desambiguação do objeto-alvo ou não¹. Isso no entanto não significa que o poder discriminatório seja desconsiderado, já que esta informação está implícita na definição dos classificadores considerados.

Este procedimento é repetido até que todos atributos de r tenham sido examinados. Se a descrição L resultante for do tipo atômica, o algoritmo é encerrado. Por outro lado, se L incluir uma propriedade relacional referenciando um segundo objeto lm , o algoritmo é chamado recursivamente para descrever lm utilizando-se o mesmo procedimento.

Além da proposta de implementação de GER como uma tarefa de classificação e do uso de modelos de locutor, o trabalho apresentado em Ferreira e Paraboni (2017a) também compara diversos modelos de GER - com acesso a diferentes graus de informação sobre o locutor - com uma extensão do algoritmo Incremental (DALE; REITER, 1995) adaptada para manipular propriedades relacionais (e não apenas atômicas), já que estas ocorrem com frequência em vários dos corpúsculos de teste utilizados na avaliação da proposta. Embora este algoritmo - denominado *Extended Incremental Algorithm* (EIA) - desempenhe apenas o papel de *baseline* na proposta principal em Ferreira e Paraboni (2017a), seu projeto agrega decisões importantes sobre como estas relações são combinadas com propriedades atômicas, e representa uma contribuição adicional independente do trabalho principal realizado.

No total, três abordagens de GER são avaliadas em Ferreira e Paraboni (2017a): a proposta principal baseada em máquinas de vetor suporte (SVM) e uma variação desta fazendo uso de árvores de decisão (DT), e o sistema de *baseline* EIA. Estas três abordagens são avaliadas em duas versões, com (+) e sem (−) acesso ao modelo do locutor, resultando assim em seis implementações distintas (SVM+, SVM−, DT+, DT−, EIA+ e EIA−).

¹A exceção é o caso de propriedade relacional, já que o algoritmo proposto prevê a inclusão de apenas uma propriedade deste tipo, evitando assim a produção de expressões como ‘o objeto que está ao lado de X e acima de Y’, que eram pouco comuns nos domínios considerados.

Considerando estas seis alternativas, a investigação das duas questões de pesquisa propostas - relacionadas ao uso do modelo de locutor e dos métodos de aprendizagem de máquina - foi realizada por meio de uma comparação extensiva dos resultados de cada alternativa com base em diversos conjuntos de dados. Esta avaliação fez uso dos corpúscos para GER desenvolvidos por terceiros TUNA-Furniture e TUNA-People (GATT; SLUIS; DEEMTER, 2007) e GRE3D3/7 (DALE; VIETHEN, 2009; VIETHEN; DALE, 2011), ambos discutidos na seção 2.2, e também dos corpúscos Stars (TEIXEIRA et al., 2014) e Stars2 (PARABONI; GALINDO; IACOVELLI, 2017) desenvolvidos pela equipe local, e discutidos no capítulo 4.

6.2.2 Resultados

Os resultados em Ferreira e Paraboni (2017a) foram significativamente superiores aos do estado-da-arte identificado na literatura para quase todos os conjuntos de dados utilizados na avaliação das abordagens propostas, e de modo especial para os domínios TUNA e GRE3D3/7. No caso do domínio TUNA, a proposta apresentada é superior ao algoritmo IS-FBN, vencedor do desafio TUNA em Gatt e Belz (2007) tanto nos domínios *Furniture* como *People*. No caso do domínio GRE3D3/7, a proposta abre uma vantagem de 29 pontos de acurácia no caso do corpúscos GRE3D3, e de 10 pontos de vantagem no caso do corpúscos GRE3D7, em ambos os casos considerando-se o trabalho prévio dos próprios autores destes corpúscos (VIETHEN; DALE, 2010; VIETHEN; MITCHELL; KRAHMER, 2013).

6.2.3 Contribuições

Além de resultados superiores aos trabalhos prévios da área, o estudo em Ferreira e Paraboni (2017a) apresenta contribuições científicas potencialmente relevantes ao formalizar a tarefa de GER como um problema de classificação (especialmente no que diz respeito ao uso de SVMs nesta tarefa) e pela experimentação com múltiplos modelos de locutor. Além disso, ao incorporar o modelo de locutor a algoritmos práticos de GER para gerar descrições reais, o estudo representa uma evolução em relação a estudos prévios sobre variação humana como Viethen (2011), que são limitados à proposta de modelos de locutor sem no entanto chegar a utilizar estes modelos na tarefa de GER propriamente dita.

6.3 Modelos baseados em personalidade humana

Do ponto de vista da plausibilidade da língua gerada de forma automática, a modelagem de locutores individuais discutida na seção anterior representa uma melhoria significativa em relação aos modelos baseados no comportamento médio de um grupo de indivíduos. Pode-se assim afirmar que o uso de modelos de locutores individualizados constitui uma estratégia possivelmente ideal de geração de língua natural de forma plausível. Este nível de detalhamento permite, por exemplo, que aplicações de GLN não apenas produzam língua com alto grau de realismo, mas proporcionem uma experiência de comunicação humano-computador mais completa, na qual o sistema pode não apenas expressar-se de forma a atender os requisitos da aplicação, mas que esta forma seja perceptivelmente distinta da forma de expressão adotada por outros locutores, sejam estes artificiais ou humanos.

Apesar destes atrativos, entretanto, a modelagem de locutores individuais esbarra na dificuldade prática de obtenção de dados de treinamento. Para a construção de modelos deste tipo, faz-se necessário coletar exemplos de formas linguísticas produzidas em todas as situações de interesse *para todos os indivíduos* que se deseja modelar, ou seja, faz-se necessário construir um *cópus* de GLN contemplando todos os locutores-alvo da aplicação. A depender de quais e quantos locutores devam ser modelados, entretanto, esta abordagem pode não ser factível, seja por razões de custo ou simplesmente porque os indivíduos de interesse não são totalmente conhecidos ou acessíveis.

Uma solução intermediária entre os modelos de GLN baseados no comportamento médio de um grupo e os modelos de locutores individualizados seria o agrupamento de indivíduos que apresentem *escolhas linguísticas semelhantes em perfis* específicos. Por exemplo, podemos considerar perfis de locutores que tendem a se comunicar de forma breve, ou que tendem a preferir o uso de palavras de maior carga afetiva, dentre inúmeras outras possibilidades. Desde que a computação de perfis deste tipo possa ser implementada de alguma forma simples (i.e., com custo menor do que a modelagem de cada um dos locutores individualizados), a geração de língua natural baseada em perfis pode em princípio agregar alguns dos benefícios da modelagem individualizada (e.g., como o maior realismo na reprodução de suas escolhas linguísticas) sem algumas das suas principais desvantagens (e.g., a necessidade de obter exemplos linguísticos produzidos por todos os locutores de interesse).

Dentre diversas possibilidades de construção de perfis de locutores de baixo custo, consideramos no restante deste capítulo a questão da relação entre produção linguística e *personalidade humana* tal qual representada pelo modelo dos Cinco Grandes Fatores (CGF) (GOLDBERG, 1990).

Conforme já discutido na seção 2.5, o modelo CGF constitui, em virtude da sua fundamentação linguística, uma base teórica possivelmente sólida para uma ampla gama de estudos de PLN/GLN. Entretanto, sendo este tema parte de um projeto de pesquisa recente (iniciado em março de 2017), a discussão que se segue terá um caráter menos conclusivo - ou mais especulativo - do que o dos capítulos anteriores, e será em grande parte limitada a questões de construção de um recurso linguístico-computacional básico na forma de um *cópus* de textos e inventários de personalidade, e a alguns estudos preliminares de tratamento computacional de personalidade humana baseados neste *cópus*.

O *cópus* coletado, que é apresentado em Santos, Paraboni e Silva (2017), objetiva oferecer suporte à investigação de relações entre traços de personalidade e a produção de língua em português brasileiro, e também para aplicações computacionais práticas como o reconhecimento de personalidade a partir de texto proveniente de redes sociais. Uma vez tratada a construção deste recurso, a questão da plausibilidade da língua em GLN será então retomada com a proposta - de caráter preliminar - de um novo modelo de superespecificação referencial para GER baseado em traços de personalidade. Este modelo, que é apresentando em Paraboni, Monteiro e Lan (2017), pode ser visto como uma generalização do trabalho em Ferreira e Paraboni (2017a) discutido na seção anterior, e que era baseado em modelos de locutores individualizados.

6.3.1 O *cópus b5* de textos e inventários de personalidade

Com o objetivo de oferecer suporte a diversos tipos de estudo de tratamento computacional da personalidade humana a partir de texto - tanto para reconhecimento de traços de personalidade a partir destes (uma tarefa típica de PLN) como para geração de texto com base em um conjunto de traços de personalidade de interesse (uma tarefa de GLN) - foi construído um conjunto de dados contendo textos e os inventários de personalidade de seus respectivos autores, denominado *cópus b5*.

O *cópus b5* possui 1082 conjuntos de textos acompanhados dos inventários de personalidade de seus respectivos autores e, até onde temos conhecimento, é o maior conjunto de dados do gênero já produzido para o português Brasileiro. Uma versão preliminar do *cópus* foi utilizada em um estudo sobre o reconhecimento de personalidade a partir de textos controlados e provenientes da rede social Facebook em Santos, Paraboni e Silva (2017). A construção e validação deste *cópus* é o foco desta seção.

6.3.1.1 Visão geral

O *cópus b5* consiste de uma coleção de inventários de personalidade representada pela base *b5-subject*, e quatro bases textuais (ou subcórpus) *b5-post*, *b5-ref*, *b5-text* e *b5-caption* discutidas a seguir. Uma visão geral desta organização é ilustrada na figura 23.

Figura 23: Componentes do *cópus b5*.

Fonte: Adaptado de Ramos et al. (2017).

A base *b5-subject* contém 1082 inventários de personalidade do tipo *IGFP-5* (ANDRADE, 2008), preenchidos de modo *online* ou de forma presencial pelos próprios sujeitos, e outras informações de caracterização autoral. Os inventários foram respondidos majoritariamente por meio do aplicativo Facebook, e em alguns casos por meio de um experimento presencial. Além das informações de personalidade, *b5-subject* contém também anotações parciais de gênero e idade dos participantes, informações sobre sua área de atuação profissional, grau de religiosidade e outras. Estas informações oferecem suporte a diversos projetos correlacionados, atualmente em desenvolvimento, nas áreas de caracterização e identificação autoral a partir de textos.

Os textos integrantes do *cópus b5* são de dois tipos: texto livre obtido a partir da rede social Facebook do participante, e texto controlado, obtido por meio de experimentos presenciais de coleta de dados a partir de estímulos pré-definidos envolvendo um subconjunto dos participantes. A porção de textos livres constitui o subcórpus *b5-post*, e a porção de textos controlados é composta de três subcórpus: o subcórpus *b5-ref* de expressões de referência, e os subcórpus *b5-text* e *b5-caption* de descrições multisentenciais e descrições sentenciais, respectivamente.

O subcórpus *b5-post* foi construído por meio da coleta de textos provenientes das atualizações de status na rede social Facebook de usuários que preencheram o inventário de personalidade utilizando um aplicativo específico, e com o consentimento prévio destes. Estes dados são destinados à pesquisa de modelos computacionais de reconhecimento automático de personalidade e caracterização autoral (e.g., de gênero, faixa etária etc.) a partir de textos em formato livre provenientes de redes sociais. Para cada sujeito que preenchia o inventário de personalidade, foram coletadas até 1.000 atualizações de status Facebook. No total, 1039 sujeitos realizaram este procedimento.

Os subcórpus *b5-text* e *b5-caption* foram construídos para estudo de questões mais gerais de produção de texto com base em personalidade, como o planejamento de documentos, geração texto-para-texto e sumarização. Para este fim, foi realizada uma série de experimentos presenciais de coleta de dados contemplando duas subtarefas de descrição de imagens: em versão detalhada (em texto multi-sentencial) e em versão resumida (na forma de uma sentença única). Os estímulos visuais empregados nestas tarefas são provenientes da base *GAPED* (DAN-GLAUSER; SCHERER, 2011) de imagens classificadas por valência e significância normativa, e designadas de modo a despertar diferentes graus e tipos de emoção. Um exemplo deste tipo de imagem é ilustrado na figura 24.

Figura 24: Exemplo de estímulo usado na coleta dos corpúscos *b5-text* e *b5-caption*.

Fonte: Extraído da base *GAPED* em Dan-Glauser e Scherer (2011).

O objetivo desta tarefa descritiva foi o de observar a estratégia utilizada pelos participantes para determinar os elementos mais importantes da imagem, a ordem e estruturação da descrição, e suas escolhas lexicais e sintáticas. A coleta de um mesmo conjunto de descrições em duas versões - detalhada e resumida - foi realizada com o objetivo de viabilizar o estudo de questões ligadas à geração do tipo texto-para-texto tanto em nível sentencial quanto discursivo. Estas questões incluem tarefas de GLN relacionadas ao planejamento sentencial (STENT; MOLINA, 2009) como a ordenação sentencial, agregação e inserção de marcadores de discurso, tarefas relacionadas ao planejamento discursivo e à realização de formas superficiais.

Finalmente, e de especial interesse para o tema da presente monografia, o *córpus b5* inclui ainda um *córpus* para GER no sentido tradicional do termo (cf. capítulo 4) denominado *b5-ref*. Assim como TUNA (GATT; SLUIS; DEEMTER, 2007), GRE3D3/7 (DALE; VIETHEN, 2009; VIETHEN; DALE, 2011), Stars (TEIXEIRA et al., 2014), Stars2 (PARABONI; GALINDO; IACOVELLI, 2017) e muitos outros, este *córpus* também consiste de uma coleção de descrições definidas produzidas por sujeitos humanos em uma tarefa de identificação de entidades visuais. Diferentemente dos anteriores, entretanto, o *córpus b5-ref* é o único recurso do gênero a incluir informações sobre a *personalidade* de cada locutor. Assim, este *córpus* destina-se principalmente ao estudo do fenômeno de produção de expressões de referência baseada em personalidade - aqui entendido tanto do ponto de vista da seleção de conteúdo como da realização superficial - embora naturalmente a disponibilização de informação de personalidade não represente nenhum obstáculo ao seu uso como *córpus* de GER de propósito geral.

Tendo em vista este objetivo, o projeto do *córpus b5-ref* privilegiou o uso de imagens de estímulo com maior potencial de explicitar diferenças entre traços de personalidade, e não apenas baseadas em domínios simplificados (e.g., objetos geométricos etc.) usados em outros recursos do tipo. De forma mais específica, os contextos de referência considerados fazem uso de imagens extraídas da base *Face Place* (RIGHI; PEISSIG; TARR, 2012) de fotografias humanas validadas para diversos tipos de emoções e características físicas. Estímulos deste tipo foram intercalados com contextos referenciais contendo objetos do tipo Greeble (GAUTHIER; TARR, 1997), amplamente utilizadas na pesquisa em reconhecimento facial². As imagens representando Greebles são essencialmente *fillers* empregados para reduzir a repetição de padrões linguísticos durante a coleta de dados, mas também constituem uma base textual própria investigada em Santos, Paraboni e Silva (2017). Exemplos dos dois tipos de imagem de estímulo utilizados são ilustrados pela figura 25.

Figura 25: Exemplos de estímulo usado na coleta do *córpus b5-ref*.

Fonte: Santos, Paraboni e Silva (2017).

Assim como em outros experimentos de coleta de dados para GER, também neste caso os

²As imagens Face Place e Greebles são uma cortesia de Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University.

participantes eram instruídos a referenciar de forma única (i.e., sem ambiguidade) a pessoa ou entidade destacada, de tal modo que uma segunda pessoa (hipotética) pudesse identificá-la com base na descrição produzida. No presente exemplo, isso poderia ser feito com uso de descrições como, e.g., ‘a moça loira’ ou ‘a mulher que não está sorrindo, no canto esquerdo da imagem’ etc.

O subcórpus *b5-ref* possui 1810 descrições produzidas por 152 participantes. Dado o propósito de utilizar os dados coletados para estudos de seleção de conteúdo referencial baseada em personalidade, o subcórpus *b5-ref* difere das demais porções - puramente textuais - do córpus *b5* no sentido de que estas expressões de referência foram acrescidas de informação semântica na forma de propriedades (ou pares atributo-valor) como em *gender-female*. Após correção ortográfica, estas expressões foram assim rotuladas segundo um esquema composto de 27 atributos referenciais propostos com base nos tipos de informação mais frequentes observados nos dados coletados. Os valores destes atributos são em parte provenientes da base *Face Place* (RIGHI; PEISSIG; TARR, 2012), e em parte provenientes de anotação manual, e incluem tanto características físicas dos objetos-alvo presentes em cada estímulo (e.g., cor da pele, comprimento do cabelo etc.) como características de teor mais subjetivo, como emoções (e.g., alegre, com expressão neutra etc.).

6.3.1.2 Reconhecimento de personalidade no córpus *b5*

Como forma de verificar a validade dos dados coletados para o desenvolvimento de modelos de GLN baseados em personalidade (a serem discutidos na próxima seção), o córpus *b5* foi inicialmente empregado na tarefa de reconhecimento de personalidade a partir de texto livre. Posteriormente, esta mesma tarefa foi reavaliada em (SANTOS; PARABONI; SILVA, 2017) considerando-se também os gêneros textuais obtidos de forma controlada. Estes estudos - que são resultados de colaborações com alunos de mestrado e de graduação - são discutidos brevemente a seguir como evidência preliminar de que a base textual coletada reflete uma gama de personalidades suficientemente ampla para suporte a futuros estudos de tratamento computacional da personalidade.

A primeira etapa da investigação realizada consistiu de um estudo piloto sobre o problema computacional de reconhecimento de personalidade a partir de texto proveniente de redes sociais em português Brasileiro. O objetivo deste estudo foi investigar interpretações alternativas do problema de reconhecimento de personalidade a partir de texto como um problema de classificação, e de possíveis fontes de conhecimento linguístico aplicáveis a esta tarefa. Para este fim, foi realizada a coleta inicial do córpus de atualizações de status Facebook *b5-post*, e de grande parte dos inventários de personalidade integrantes da base *b5-subject*, ambos descritos na

seção anterior.

O estudo realizado fez uso de métodos supervisionados de aprendizagem de máquina aplicados à tarefa de reconhecimento das cinco dimensões de personalidade do modelo CGF. Os modelos propostos tomam como entrada um conjunto de 164 atributos de aprendizagem divididos em três categorias, a saber: 30 atributos obtidos a partir do pré-processamento dos textos, como número de erros ortográficos, abreviações típicas de redes sociais etc., obtidos com auxílio das bases disponibilizadas em Duran e Nunes (2015) e Duran, Avanco e Nunes (2015); 70 atributos obtidos do dicionário Unitex (MUNIZ, 2004), representando o número de ocorrências de cada característica representada no dicionário; e 64 atributos representando a frequência das categoria psicolinguísticas LIWC em português Brasileiro (FILHO; ALUÍSIO; PARDO, 2013).

Com base nestes atributos, foram treinados diversos modelos de classificação binária de personalidade a partir das atualizações de status Facebook implementados com o uso de SVMs. Resultados preliminares sugerem que, assim como já demonstrado em outros idiomas, também no caso do português Brasileiro o texto proveniente de redes sociais permite o reconhecimento de certas dimensões de personalidade. Também foi estabelecido que, para o volume de dados em questão, o uso de classificação binária seria o mais adequado, e que o conhecimento proveniente do dicionário LIWC é o que possui maior peso na solução do problema. Estes resultados constituem uma primeira evidência de que a base textual coletada reflete certos traços de personalidade de seus autores, e fornecem subsídio para estudos mais aprofundados de interpretação e geração de língua natural baseadas em personalidade.

A partir destes resultados iniciais, o trabalho em Santos, Paraboni e Silva (2017) apresenta uma investigação sobre o reconhecimento de personalidade a partir de textos de diferentes gêneros disponibilizados pelos corpús *b5*. O objetivo deste estudo foi identificar tarefas de produção textual em português mais fortemente correlacionadas às dimensões CGF, considerando diferenças entre texto livre *b5-post* e texto controlado dos tipos *b5-ref*, *b5-text* e *b5-caption*.

A comparação entre estes subcorpús foi baseada em um subconjunto de 154 locutores do corpús *b5* que produziram todos os quatro tipos de texto considerados. Para cada uma das cinco dimensões CGF e para cada subcorpús, foram construídos modelos supervisionados de reconhecimento de personalidade. Assim como no estudo piloto, este problema foi modelado como uma tarefa de classificação binária, atribuindo-se rótulos (positivos e negativos) a cada locutor relativos ao valor médio de cada dimensão de personalidade no conjunto de dados.

Os resultados em Santos, Paraboni e Silva (2017) sugerem que certas dimensões de personalidades são mais observáveis em alguns tipos de texto do que em outros, e que de modo geral o

uso de texto livre (Facebook) apresenta resultados superiores ao uso de texto controlado. Além disso, foi constatado que certas dimensões de personalidade são observáveis mesmo em textos muito curtos, como as legendas em *b5-caption*. De especial interesse para o presente trabalho, os resultados demonstram também que o reconhecimento de personalidade a partir de expressões de referências *b5-ref* apresentou resultados superiores ao uso de texto descritivo nesta tarefa, o que pode ser tomado como evidência de que este domínio pode ser utilizado em um estudo de GER baseada em personalidade, tal qual será discutido na próxima seção.

6.3.2 GER com base em traços de personalidade

Os estudos apresentados nas seções anteriores serviram ao propósito de construção do *cópus b5* de textos e inventários de personalidade, e de validação do conjunto de dados por meio da identificação de relações entre traços de personalidade CGF e textos em português brasileiro. Nesta seção, o tema central em discussão na presente monografia - a geração de língua natural de forma plausível - é finalmente retomado com a proposta de novos modelos de GER baseada em personalidade humana desenvolvidos com base no *cópus* coletado.

O estudo em Paraboni, Monteiro e Lan (2017) apresenta um experimento de GER com base em traços de personalidade utilizando o sub*cópus b5-ref* de expressões de referência discutido nas seções anteriores. Este estudo pode ser visto como uma extensão do trabalho em Ferreira e Paraboni (2017a) - que tratava da geração de expressões com base em modelos de locutores individualizados - fazendo neste caso uso de perfis de grupos de locutores agrupados de acordo com seus traços de personalidade.

Partindo-se da observação prática de que as características mais salientes de um objeto-alvo tendem a ser selecionadas pela maioria dos locutores, e que portanto a seleção destes atributos não revelaria diferenças significativas entre eles, o experimento realizado concentra-se assim na questão da *superespecificação referencial* (cf. seção 2.1.1). Em outras palavras, o experimento busca identificar diferenças entre locutores (que podem ou não ser decorrentes de suas diferenças de personalidade) examinando a informação redundante que cada locutor acrescenta à expressão de referência além do que seria estritamente necessário para desambiguação (já que esta presumidamente apresenta menor variação entre indivíduos). Para ilustrar esta questão, considere-se o objetivo de identificar de forma única a pessoa destacada em vermelho no contexto da figura 26, extraído do *cópus b5-ref* descrito anteriormente.

Neste exemplo, observa-se que o objeto-alvo de interesse é a única pessoa do sexo feminino na cena, o que torna a propriedade *gender-female* altamente saliente. Possivelmente por este motivo, a informação de gênero neste contexto é de fato selecionada pela maior parte dos

Figura 26: Um contexto de referência do *b5-ref* corpus.

Fonte: Construído a partir de imagens Face Place em Righi, Peissig e Tarr (2012).

locutores do *cópus*, a tal ponto de não oferecer subsídios para diferenciar um indivíduo de outro.

O mesmo não necessariamente ocorre, entretanto, com as informações adicionais (i.e., acrescentadas de forma superespecificada) à expressão básica. Por exemplo, alguns indivíduos podem optar por acrescentar informação sobre a expressão facial do referente (e.g., ‘a menina *emburrada*’) ou sua aparência física (e.g., ‘a mulher *de cabelos loiros sobre a testa*’), dentre muitas outras possibilidades. Uma vez que a informação redundante não é estritamente necessária para a identificação do objeto-alvo, a inclusão deste tipo de informação tende a variar amplamente de um locutor para outro, representando preferências linguísticas específicas de cada indivíduo. Estas preferências, segundo a hipótese de pesquisa em Paraboni, Monteiro e Lan (2017), seriam ao menos parcialmente relacionadas a diferenças de personalidade.

Com base nestas observações, o estudo em Paraboni, Monteiro e Lan (2017) apresenta um algoritmo de GER em que a tarefa de superespecificação referencial é guiada pelos traços de personalidade do locutor-alvo. Este algoritmo é dividido em duas partes: primeiramente, uma expressão livre de ambiguidade é produzida por um método de GER tradicional - no caso, usando o algoritmo Incremental em Dale e Reiter (1995) - e então propriedades adicionais são inseridas com base no modelo de personalidade de cada locutor-alvo em questão.

Assim como em Ferreira e Paraboni (2017a), o algoritmo proposto segue uma organização baseada em uma série de classificadores binários individuais (i.e., um para cada atributo sob consideração) obtidos por indução de árvores de decisão. No presente caso, entretanto, estes classificadores são aplicados especificamente à tarefa de superespecificação referencial, e não para a construção da expressão básica propriamente dita.

O conjunto de classificadores considerado é treinado com base nas expressões de referência do *cópus b5-ref* tomando-se como atributos de aprendizagem características do contexto (valores de poder discriminatório, frequências relativas etc.) e os valores de personalidade CGF de cada

locutor. Este procedimento é ilustrada pelo Algoritmo 2, extraído de Paraboni, Monteiro e Lan (2017).

Algoritmo 2: Superespecificação referencial com base em personalidade.

```

1 describe( $r, C, P, \text{Class}[], \text{length}$ )
2    $L \leftarrow \text{MakeReferringExpression}(r, C, P)$ 
3   for  $A_i \in P$  do
4     if  $|L| \leq \text{length}$  then
5       if  $\text{Class}[A_i] == \text{true}$  then
6          $v \leftarrow \langle A_i, \text{value}(r, A_i) \rangle$ 
7         if  $\langle A_i, v \rangle \notin L$  then
8            $L \leftarrow L \cup \langle A_i, v \rangle$ 
9         end
10      end
11    end
12  end
13  return  $L$ 

```

O algoritmo proposto recebe como entrada o objeto-alvo r a ser descrito, o contexto C contendo os objetos distraidores, uma lista de atributos preferenciais P , o conjunto de classificadores class contendo as predições para cada atributo possível e o limite length de propriedades a serem superespecificadas para este domínio. O objetivo do algoritmo é produzir uma descrição de saída L de tamanho máximo length contendo propriedades superespecificadas que são selecionadas de acordo com o modelo de personalidade do locutor.

Inicialmente, o algoritmo Incremental (DALE; REITER, 1995) é invocado (linha 2) para obtenção da expressão básica L . A superespecificação referencial propriamente dita (linhas 3-12) consiste em examinar os atributos possíveis na ordem de preferência P (3) e, estando dentro do limite length (4) e sendo previstas pelo respectivo classificador (5), acrescentá-las à descrição de saída L (7-8). O procedimento é repetido até que o tamanho limite length seja atingido, ou até que não haja mais propriedades a considerar.

O algoritmo proposto foi comparado com uma versão reduzida que leva em conta apenas atributos contextuais (ou seja, sem informação de personalidade), e dois sistemas de *baseline*: os algoritmos Incremental (DALE; REITER, 1995) e Greedy (DALE, 2002). Os resultados em Paraboni, Monteiro e Lan (2017) demonstram que o uso de informação de personalidade na tarefa de GER resulta em expressões mais próximas das que são produzidas por locutores humanos do que as que são geradas por todas as alternativas consideradas.

Apesar de já haver um certo corpo de estudos de GLN enfocando a questão da personalidade humana (MAIRESSE; WALKER, 2011; WALKER et al., 2014), o estudo em Paraboni, Monteiro e Lan (2017) é possivelmente o primeiro do gênero a explorar o modelo CGF especificamente nesta tarefa computacional. Estes resultados iniciais são consideravelmente positivos, embora

novos estudos - possivelmente cobrindo um espectro mais amplo de atributos referenciais - ainda sejam necessários para estabelecimento de conclusões mais concretas sobre a relação entre personalidade e estratégias de seleção de conteúdo.

6.4 Considerações finais

Este capítulo abordou duas alternativas à modelagem do comportamento médio de indivíduos em GLN, e que exploram a noção de variação humana na produção de língua natural: o estudo em Ferreira e Paraboni (2017a) e trabalhos relacionados, que tratam de modelos de locutores individualizados para a tarefa de GER e, em caráter mais exploratório, o estudo em Paraboni, Monteiro e Lan (2017) (antecedido pelas investigações de interpretação de língua natural em Santos, Paraboni e Silva (2017)) que propõe uma nova visão do conceito de variação humana baseada no modelo de personalidade CGF (ALLPORT; ALLPORT, 1921).

Modelos de locutores individualizados representam uma forma possivelmente ideal de gerar língua de modo mais plausível. Entretanto, dada a dificuldade de coleta de exemplos linguísticos produzidos por todos os indivíduos de interesse, esta abordagem pode ser inviável. O uso de informação de personalidade representa assim uma possível alternativa ao uso de modelos individualizados, e pode em princípio oferecer soluções mais gerais para o problema de GLN, e com custo de treinamento mais baixo, já que a computação de traços de personalidade (e.g., a partir de texto em redes sociais, como em Santos, Paraboni e Silva (2017)) requer em princípio menor esforço do que a coleta de exemplos linguísticos para cada domínio e cada locutor de interesse.

Embora os estudos discutidos neste capítulo - tanto a proposta de modelos de locutores individualizados em Ferreira e Paraboni (2017a) como de modelos de personalidade humana em Paraboni, Monteiro e Lan (2017) - tenham sido considerados exclusivamente à luz da tarefa computacional de GER, cabe destacar que não há em princípio motivos para esta limitação, já que ambas as técnicas podem ser adaptadas a diversas outras tarefas de GLN. Em especial, observamos que a tarefa de realização superficial pode ser potencialmente beneficiada pelo uso de modelos de variação humana baseados em personalidade, já que a relação entre o modelo CGF e lexicalização é amplamente reconhecida (JOHN; NAUMANN; SOTO, 2008). Estudos desta natureza, enfocando tanto a questão da lexicalização de expressões de referência proveniente do subcorpus *b5-ref* como a realização de texto livre dos subcorpus *b5-text* *b5-caption* e aqui discutidos encontram-se atualmente em andamento.

Capítulo 7

CONCLUSÕES

Esta monografia discutiu o problema computacional da geração de língua natural dita plausível sob as perspectivas do locutor e do ouvinte, apresentando exemplos de estudos que trataram de questões de seleção de conteúdo, realização superficial e outras. Em linhas gerais, foram apresentados exemplos de estudos que seguem uma definição mais convencional (ou talvez mais intuitiva) de plausibilidade baseada na perspectiva do locutor, e foi destacado que estudos deste tipo tendem a apresentar necessidade constante de novos conjuntos de dados para que novas questões de pesquisa possam ser investigadas. Como alternativa a esta abordagem, foram então apresentados exemplos de estudos que consideram a perspectiva do ouvinte, os quais tendem a apresentar amplo potencial de reúso dos recursos de pesquisa desenvolvidos, e podem trazer contribuições inéditas ou de outra forma difíceis de obter com base puramente na perspectiva do locutor. Finalmente, a discussão sobre a plausibilidade na tarefa computacional de GLN foi ampliada de modo a considerar a questão da variação humana na produção da língua, com a proposta de dois modelos deste tipo: modelos baseados no comportamento de locutores individuais, e modelos baseados em perfis de personalidade humana.

A discussão aqui apresentada - envolvendo as perspectivas do locutor, do ouvinte e suas possíveis generalizações - permite algumas considerações que podem ser sumarizadas pelos dois pontos principais a seguir:

- *Locutor versus ouvinte.* De modo geral, as contribuições científicas aqui relatadas sugerem que a perspectiva do ouvinte seja de natureza complementar à perspectiva do locutor, e que um mesmo fenômeno de produção da língua pode em princípio ser investigado de ambas as formas. No entanto, cabe observar que o uso de medidas de reação ou impacto da forma linguística sobre o ouvinte (como as medidas de tempo de navegação ou de reconhecimento de objetos visuais discutidas) nos permitem investigar não apenas como

gerar língua natural mas, o que é igualmente importante, investigar também como *não* gerá-la. Em outras palavras, um sistema de GLN avaliado apenas com base em um *cópus* do tipo *gold standard* pode apresentar alta fidelidade nos resultados, e ainda assim cometer erros graves diante de entradas inesperadas. Por este motivo, acreditamos que estudos orientados ao ouvinte sejam os que melhor refletem a natureza interdisciplinar da pesquisa em GLN, e talvez representem algumas de suas contribuições científicas mais expressivas.

- *Variação humana*. Como alternativa a modelos de comportamento médio de um grupo de locutores, os estudos de variação humana em GLN aqui relatados consideraram, em um primeiro momento, características individuais destes locutores. Apesar do desempenho superior ao estado-da-arte em GER, entretanto, esta abordagem depende da disponibilidade de exemplos de uso da língua produzidos pelos locutores de interesse, exigência esta que pode inviabilizar seu uso em aplicações práticas ou de caráter mais geral. Por este motivo, foi considerada a proposta de geração de língua natural com uso de informações de personalidade dos locutores de interesse. Estas informações, que podem ser obtidas de forma menos intrusiva e com baixo custo (e.g., mediante análise de publicações em redes sociais), podem representar uma solução intermediária entre sistemas de GLN que produzem sempre a mesma saída fixa para uma dada entrada, e aqueles que são customizados para cada locutor de interesse.

Feitas estas considerações, cabe destacar que as iniciativas de pesquisa aqui discutidas deixam diversas oportunidades para melhorias. De modo especial, enfatizamos a questão do tratamento computacional da personalidade humana e suas implicações para a tarefa computacional de GLN. Este tema, que é o foco principal de um recém iniciado projeto de pesquisa, sugere múltiplas questões a serem investigadas a partir do *cópus b5* aqui discutido, incluindo estudos sobre planejamento de documentos, lexicalização e reescrita sentencial com base em traços de personalidade. Além disso, esta linha de investigação possui também diversas ramificações na área de interpretação de língua natural, incluindo o reconhecimento de traços de personalidade a partir de textos, aplicações diversas de caracterização e identificação autoral, e outras. Projetos de pesquisa em nível de mestrado enfocando alguns destes tópicos, e que fazem uso tanto do atual *cópus* em português como de bases textuais similares disponibilizadas para outros idiomas, encontram-se atualmente em desenvolvimento.

Em resumo, esta monografia procurou demonstrar que os desafios da pesquisa em GLN são muitos, e que os exemplos aqui discutidos representam apenas uma breve introdução à questão da plausibilidade da língua natural. Espera-se assim que este texto introdutório possa servir de motivação para a leitura completa das publicações que fundamentaram esta discussão, e

que a partir deste estudo novos pesquisadores e estudantes (e.g., de Ciência da Computação, Psicolinguística e áreas relacionadas) sintam-se incentivados a abordar o problema computacional da geração automática de língua natural de modo plausível - seja com base na perspectiva do ouvinte, do locutor, ou perfis destes - e a desenvolver suas próprias metodologias, auxiliando assim no crescimento e consolidação da linha de pesquisa em GLN de modo geral, e em especial para o caso do português brasileiro.

REFERÊNCIAS

- ALLPORT, F. H.; ALLPORT, G. W. Personality traits: Their classification and measurement. *Journal of Abnormal And Social Psychology*, v. 16, p. 6–40, 1921.
- ALTAMIRANO, R. et al. Zoom: a corpus of natural language descriptions of map locations. In: *Proceedings of ACL-2015*. Beijing, China: Association for Computational Linguistics, 2015. p. 69–75.
- ANDRADE, J. M. de. *Evidências de validade do inventário dos cinco grandes fatores de personalidade para o Brasil*. Tese (Doutorado) — Universidade de Brasília, 2008.
- ARAUJO, R. P. A. de et al. SINotas: the evaluation of a NLG application. In: *7th International Conference on Language Resources and Evaluation (LREC-2010)*. Valletta, Malta: ELRA, 2010. p. 2388–2391.
- ARTS, A. et al. Overspecification facilitates object identification. *Journal of Pragmatics*, v. 43, p. 361–374, 2011.
- AZIZ, W. F.; PARDO, T. A. S.; PARABONI, I. Construção de um corpus paralelo alinhado para a tradução automática estatística. In: *VI Encontro de Linguística de Corpus*. São Paulo: [s.n.], 2007.
- AZIZ, W. F.; PARDO, T. A. S.; PARABONI, I. Building a Spanish-Portuguese parallel corpus for statistical machine translation. In: *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*. Vila Velha, Brazil: ACM, 2008. p. 369–371.
- AZIZ, W. F.; PARDO, T. A. S.; PARABONI, I. An experiment in Spanish-Portuguese statistical machine translation. In: *SBIA-2008, Lecture Notes in Artificial Intelligence 5249*. Salvador, Brazil: Springer-Verlag, 2008. p. 248–257.
- AZIZ, W. F.; PARDO, T. A. S.; PARABONI, I. Fine-tuning in Portuguese-English statistical machine translation. In: *Seventh Brazilian Symposium in Information and Human Language Technology (STIL)-2009*. São Carlos, Brazil: IEEE, 2009. p. 158–162.
- AZIZ, W. F.; PARDO, T. A. S.; PARABONI, I. Statistical phrase-based machine translation: Experiments with Brazilian Portuguese. In: *Anais do VII Encontro Nacional de Inteligência Artificial*. Bento Gonçalves, Brazil: [s.n.], 2009. p. 769–778.
- BARCLAY, M. *Reference Object Choice in Spatial Language: Machine and Human Models*. Tese (PhD thesis) — University of Exeter, 2010.

- BARTIE, P. et al. The REAL corpus: a crowd-sourced corpus of human generated and evaluated spatial references to real-world urban scenes. In: *Proceedings of LREC-2016*. Portoroz, Slovenia: ELRA, 2016.
- BELZ, A. Statistical generation: Three methods compared and evaluated. In: *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG-2005)*. Aberdeen, UK: Association for Computational Linguistics, 2005.
- BELZ, A. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, v. 14, n. 4, p. 431–455, 2008.
- BILMES, J.; KIRCHHOFF, K. Factored language models and generalized parallel backoff. In: *Proceedings of HLT-NAACL-2003*. Edmonton, Canada: Association for Computational Linguistics, 2003.
- BYRON, D. et al. Generating instructions in virtual environments (GIVE): A challenge and evaluation testbed for NLG. In: *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*. Arlington, VA, USA: [s.n.], 2007.
- BYRON, D. et al. Report on the first NLG challenge on generating instructions in virtual environments (GIVE). In: *12th European Workshop on Natural Language Generation (ENLG-2009)*. Athens, Greece: The Association for Computer Linguistics, 2009.
- CARENINI, G.; MOORE, J. D. Generating and evaluating evaluative arguments. *Artificial Intelligence*, v. 170, p. 925–952, 2006.
- CASIMIRO, C. R.; PARABONI, I. Temporal aspects of content recommendation on a microblog corpus. In: *Computational Processing of the Portuguese Language (PROPOR-2014), Lecture Notes in Computer Science 8775*. São Carlos, Brazil: Springer, 2014. p. 189–194.
- CELLI, F. *Adaptive Personality Recognition from Text*. Tese (Doutorado) — University of Trento, 2012.
- CLARK, H.; WILKES-GIBBS, D. Referring as a collaborative process. *Cognition*, v. 22, p. 1–39, 1986.
- CLARKE, A. D. F.; ELSNER, M.; ROHDE, H. Where's Wally: The influence of visual salience on referring expression generation. *Frontiers in Psychology*, v. 4, n. 329, 2013.
- CUEVAS, R. R. M. et al. Portuguese pronoun resolution: resources and evaluation. In: *Computational Linguistics and Intelligent Text Processing (CICLing-2008), Lecture Notes in Computer Science 4919*. Haifa, Israel: Springer-Verlag, 2008. p. 344–350.
- CUEVAS, R. R. M.; PARABONI, I. A machine learning approach to Portuguese pronoun resolution. In: *IBERAMIA-2008, Lecture Notes in Artificial Intelligence 5290*. Lisboa, Portugal: Springer-Verlag, 2008. p. 262–271.
- CUEVAS, R. R. M.; PARABONI, I. Using low-cost learning features for pronoun resolution. In: *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation (PACLIC22)*. Cebu City, Philippines: De La Salle University (DLSU), 2008. p. 377–383.
- DALE, R. Cooking up referring expressions. In: *Proceedings of ACL-2002*. Philadelphia, PA, USA: Association for Computational Linguistics, 2002. p. 68–75.

- DALE, R.; HADDOCK, N. J. Content determination in the generation of referring expressions. *Computational Intelligence*, v. 7, p. 252–265, 1991.
- DALE, R.; REITER, E. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, v. 19, 1995.
- DALE, R.; VIETHEN, J. Referring expression generation through attribute-based heuristics. In: *12th European Workshop on Natural Language Generation*. Athens, Greece: Association for Computational Linguistics, 2009. (ENLG '09), p. 58–65.
- DAMERAU, F. J. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, v. 7, n. 3, p. 171–176, 1964.
- DAN-GLAUSER, E. S.; SCHERER, K. R. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, v. 43, n. 2, p. 468–477, 2011.
- DEEMTER, K. van. *Computational Models of Referring. A Study in Cognitive Science*. Cambridge, Massachusetts, USA: MIT Press, 2016. ISBN 9780262034555.
- DEEMTER, K. van. Designing algorithms for referring with proper names. In: *Proceedings of the 9th International Natural Language Generation conference*. Edinburgh, UK: Association for Computational Linguistics, 2016. p. 31–35.
- DICE, L. R. Measures of the amount of ecologic association between species. *Ecology*, v. 26, n. 3, p. 297–302, 1945.
- DODDINGTON, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: *Second International Conference on Human Language Technology Research*. San Diego, CA, USA: Morgan Kaufmann, 2002. p. 138–145.
- DUCHOWSKI, A. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, v. 34, n. 4, p. 455–470, 2002.
- DURAN, M. S.; AVANCO, L.; NUNES, M. G. V. A normalizer for UGC in Brazilian Portuguese. In: *Proceedings of the Workshop on Noisy User-generated Text (WNUT) - 53rd Annual Meeting of the ACL*. Beijing, China: Association for Computational Linguistics, 2015. p. 38–47.
- DURAN, M. S.; NUNES, M. G. V. A importância dos falsos homógrafos para a correção automática de erros ortográficos em Português. In: *STIL-2015 IV Jornada de Descrição do Português -*. Natal: [s.n.], 2015.
- DUSEK, O.; JURCICEK, F. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In: *Proceedings of ACL-2016*. Berlin, Germany: Association for Computational Linguistics, 2016. p. 45–51.
- ENGELHARDT, P. E.; BAILEYAND, K.; FERREIRA, F. Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language*, v. 54, n. 4, p. 554–573, 2006.
- 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.

- EUGENIO, B. D.; GLASS, M.; TROLIO, M. The DIAG experiments: Natural language generation for tutoring systems. In: *Proceedings of the Second International Conference on Natural Language Generation (INLG-2002)*. Harriman, NY: Association for Computational Linguistics, 2002. p. 120–127.
- EUGENIO, B. di et al. PatientNarr: Towards generating patient-centric summaries of hospital stays. In: *Proceedings of the 8th International Natural Language Generation Conference (INLG-2014)*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2014. p. 6–10.
- FERREIRA, T. C.; PARABONI, I. Classification-based referring expression generation. In: *Computational Linguistics and Intelligent Text Processing (CICLing-2014), Lecture Notes in Computer Science 8403*. Kathmandu, Nepal: Springer, 2014. p. 481–491.
- FERREIRA, T. C.; PARABONI, I. Referring expression generation: taking speakers' preferences into account. In: *Text, Speech and Dialogue (TSD-2014), Lecture Notes in Artificial Intelligence 8655*. Brno, Czech Republic: Springer, 2014. p. 539–546.
- FERREIRA, T. C.; PARABONI, I. Generating natural language descriptions using speaker-dependent information. *Natural Language Engineering*, Cambridge University press, v. 23, n. 6, p. 813–834, 2017.
- FERREIRA, T. C.; PARABONI, I. Improving the generation of personalised descriptions. In: *10th International Conference on Natural Language Generation (INLG-2017)*. Santiago de Compostela, Spain: [s.n.], 2017.
- FILHO, P. P. B.; ALUÍSIO, S. M.; PARDO, T. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: *9th Brazilian Symposium in Information and Human Language Technology - STIL*. Fortaleza, Brazil: [s.n.], 2013. p. 215–219.
- FITZGERALD, N.; ARTZI, Y.; ZETTLEMOYER, L. Learning distributions over logical forms for referring expression generation. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, 2013. p. 1914–1925.
- FLESCHE, R. *The Art of Readable Writing*. New York: Harper Brothers, 1949.
- FUKUMURA, K.; GOMPEL, R. P. van. Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language*, v. 62, n. 1, p. 52–66, 2010.
- GARGETT, A. et al. The GIVE-2 corpus of giving instructions in virtual environments. In: *Proceedings of LREC-2010*. Valletta, Malta: ELRA, 2010.
- GATT, A.; BELZ, A. The attribute selection for GRE challenge: Overview and evaluation results. In: *UCNLG+MT: Language Generation and Machine Translation*. Copenhagen, Denmark: [s.n.], 2007.
- GATT, A.; BELZ, A.; KOW, E. The TUNA challenge 2008: Overview and evaluation results. In: *Fifth International Natural Language Generation Conference (INLG-2008)*. Salt Fork, Ohio, USA: The Association for Computer Linguistics, 2008. p. 198–206.

- GATT, A.; BELZ, A.; KOW, E. The TUNA challenge 2009: Overview and evaluation results. In: *12nd European Workshop on Natural Language Generation*. Athens, Greece: The Association for Computer Linguistics, 2009. p. 174–182.
- GATT, A.; KRAHMER, E. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *arXiv preprint arXiv:1703.09902*, 2017.
- GATT, A.; REITER, E. SimpleNLG: A realization engine for practical applications. In: *European Natural Language Generation workshop (ENLG-2009)*. Athens, Greece: Association for Computational Linguistics, 2009.
- GATT, A.; SLUIS, I. van der; DEEMTER, K. van. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In: *Proceedings of ENLG-07*. Schloss Dagstuhl, Germany: Association for Computational Linguistics, 2007.
- GAUTHIER, I.; TARR, M. J. Becoming a greeble expert: Exploring mechanisms for face recognition. *Vision Research*, v. 37, n. 12, p. 1673–1682, 1997.
- GHODSI, A.; DENERO, J. An analysis of the ability of statistical language models to capture the structural properties of language. In: *Proceedings of the 9th International Natural Language Generation conference*. Edinburgh, UK: Association for Computational Linguistics, 2016. p. 227–231.
- GOLDBERG, L. R. An alternative description of personality: The Big-Five factor structure. *Journal of Personality and Social Psychology*, v. 59, p. 1216–1229, 1990.
- GOMPEL, R. van et al. Testing computational models of reference generation as models of human language production: The case of size contrast. In: *RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production*. Edinburgh, UK: [s.n.], 2014.
- GRICE, H. P. Logic and conversation. In: COLE, P.; MORGAN, J. L. (Ed.). *Syntax and semantics*. New York, USA: Academic Press, 1975. v. 3.
- GROSZ, B. J.; WEINSTEIN, S.; JOSHI, A. K. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, v. 21, n. 2, p. 203–225, 1995.
- GUHE, M. Generating referring expressions with a cognitive model. In: *Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference*. Amsterdam, The Netherlands: [s.n.], 2009.
- HALLIDAY, M. A. K.; HASAN, R. *Cohesion in English*. London: Longman, 1976.
- HANSEN, D. W.; JI, Q. In the eye of the beholder: A survey of models for eyes and gaze. *TPAMI*, v. 32, n. 3, p. 478–500, 2010.
- HOCKENMAIER, J.; STEEDMAN, M. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, v. 33, n. 3, p. 355–396, 2007.
- HUNTER, J. et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial intelligence in medicine*, Elsevier, v. 56, n. 3, p. 157–172, 2012.

- IACOBELLI, F. et al. Large scale personality classification of bloggers. In: D'MELLO, S. K. et al. (Ed.). *ACII (2)*. Memphis, TN, USA: Springer, 2011. (Lecture Notes in Computer Science, v. 6975), p. 568–577. ISBN 978-3-642-24570-1.
- IACOVELLI, D.; GALINDO, M. R.; PARABONI, I. Lausanne: a framework for collaborative online NLP experiments. In: *Computational Processing of the Portuguese Language (PROPOR-2014)*, *Lecture Notes in Computer Science 8775*. São Carlos, Brazil: Springer, 2014. p. 280–285.
- JOHN, O. P.; NAUMANN, L. P.; SOTO, C. J. Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. In: _____. *Handbook of personality: Theory and research*. New York, NY: Guilford Press, 2008. p. 114–158.
- JONES, K. S.; GALLIERS, J. R. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Berlin: Springer Verlag, 1995.
- JORDAN, P. et al. TBI-Doc: Generating patient & clinician reports from brain imaging data. In: *Proceedings of the 8th International Natural Language Generation Conference (INLG)-2014*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2014. p. 143–146.
- KARPATHY, A.; FEI-FEI, L. Deep visual-semantic alignments for generating image descriptions. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, 2015. p. 3128–3137.
- 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: Association for Computational Linguistics, 2014. p. 787–798.
- KELLEHER, J. D.; COSTELLO, F. J. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, v. 35, n. 2, p. 271–306, 2009.
- KEYSAR, B.; LIN, S.; BARR, D. J. Limits on theory of mind use in adults. *Cognition*, v. 89, p. 25–41, 2003.
- KING, D. L.; WHITE, M. Enhancing PTB universal dependencies for grammar-based surface realization. In: *Proceedings of the 9th International Natural Language Generation conference*. Edinburgh, UK: Association for Computational Linguistics, 2016. p. 232–236.
- KOEHN, P. et al. Moses: Open source toolkit for statistical machine translation. In: *Proceedings of ACL-2007*. Prague, Czech Republic: Association for Computational Linguistics, 2007.
- KOEHN, P.; OCH, F. J.; MARCU, D. Statistical phrase-based translation. In: *Proceedings of HLT-NAACL-2003*. Edmonton, Alberta, Canada: Association for Computational Linguistics, 2003. p. 48–54.
- KOLLER, A. et al. Enhancing referential success by tracking hearer gaze. In: *Proceedings of SIGDIAL-2012*. Seoul, South Korea: Association for Computational Linguistics, 2012. p. 30–39.
- KOLLER, A. et al. Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). In: *6th International Natural Language Generation Conference (INLG-2010)*. Dublin, Ireland: The Association for Computer Linguistics, 2010.

- KOOLEN, R. et al. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, Elsevier, v. 43, p. 3231–3250, 2011.
- KOSINSKI, M. et al. Facebook as a social science research tool: Opportunities, challenges, ethical considerations and practical guidelines. *American Psychologist*, v. 70, n. 6, p. 543–556, 2015.
- KRAHMER, E.; DEEMTER, K. van. Computational generation of referring expressions: A survey. *Computational Linguistics*, v. 38, n. 1, p. 173–218, 2012.
- KRAHMER, E.; ERK, S. van; VERLEG, A. Graph-based generation of referring expressions. *Computational Linguistics*, v. 29, n. 1, p. 53–72, 2003.
- 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.
- LAN, A. G. J.; PARABONI, I.; COUTINHO, F. L. Avaliação de algoritmos de seleção de conteúdo utilizando rastreamento de olhar. In: *STIL-2015 TILic workshop*. Natal, Brazil: [s.n.], 2015.
- LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*, v. 33, p. 159–174, 1977.
- LANGKILDE-GEARY, I. An empirical verification of coverage and correctness for a general-purpose sentence generator. In: *Proceedings of INLG-2002*. Harriman, New York, USA: Association for Computational Linguistics, 2002.
- LANGKILDE, I. Generation that exploits corpus-based statistical knowledge. In: *Proceedings of COLING-ACL-1998*. Montreal, Canada: Morgan Kaufmann Publishers / ACL, 1998. p. 704–710.
- LANGKILDE, I. Forest-based statistical sentence generation. In: *Proceedings of ANLP-2000*. Seattle, Washington, USA: Association for Computational Linguistics, 2000. p. 170–177.
- LUCENA, D. J. de; PARABONI, I. Combining frequent and discriminating attributes in the generation of definite descriptions. In: *IBERAMIA-2008, Lecture Notes in Artificial Intelligence 5290*. Lisboa, Portugal: Springer-Verlag, 2008. p. 252–261.
- LUCENA, D. J. de; PARABONI, I. USP-EACH frequency-based greedy attribute selection for referring expressions generation. In: *Proceedings of the Fifth International Natural Language Generation Conference*. Salt Fork, Ohio, USA: The Association for Computer Linguistics, 2008. p. 219–220.
- LUCENA, D. J. de; PARABONI, I. The design of an experiment in anaphora resolution for referring expressions generation. In: *Recent Advances in Natural Language Processing (RANLP-2009)*. Borovets, Bulgaria: RANLP 2009 Organising Committee / ACL, 2009. p. 14–16.
- LUCENA, D. J. de; PARABONI, I. USP-EACH: Improved frequency-based greedy attribute selection. In: *Proceedings of the 12th European Workshop on Natural Language Generation*. Athens, Greece: The Association for Computer Linguistics, 2009. p. 189–190.

- LUCENA, D. J. de; PARABONI, I. The anaphor-antecedent match: Issues for referring expressions generation. In: *International Conference on Computational Processing of Portuguese Language (PROPOR-2010)*. Porto Alegre, Brazil: [s.n.], 2010.
- LUCENA, D. J. de; PARABONI, I.; PEREIRA, D. B. From semantic properties to surface text: The generation of domain object descriptions. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, Asociacion Espanhola para la Inteligencia Artificial, v. 14, n. 45, p. 48–58, 2010.
- MAIRESSE, F. et al. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)*, v. 30, p. 457–500, 2007.
- MAIRESSE, F.; WALKER, M. A. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Model. User-Adapt. Interaction*, v. 20, n. 3, p. 227–278, 2010.
- MAIRESSE, F.; WALKER, M. A. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, v. 37, n. 3, p. 455–488, 2011.
- MANN, W. C.; THOMPSON, S. A. Rhetorical structure theory: Toward a functional theory of text organization. *text*, v. 8, n. 3, p. 243–281, 1988.
- MARCINIAK, T.; STRUBE, M. Classification-based generation using TAG. In: *3rd International Conference on Natural Language Generation (INLG-2004)*, LNCS 3123. Brockenhurst, UK: Springer Verlag, 2004. p. 100–109.
- MARCINIAK, T.; STRUBE, M. Using an annotated corpus as a knowledge source for language generation. In: *Corpus Linguistics Workshop Using Corpora for NLG (UCNLG-2005)*. Birmingham, UK: Association for Computational Linguistics, 2005. p. 19–24.
- MCROY, S.; CHANNARUKUL, S.; ALI, S. S. An augmented template-based approach to text realization. *Natural Language Engineering*, v. 9, n. 4, p. 381–420, 2003.
- MEI, H.; BANSAL, M.; WALTER, M. R. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. In: *Proceedings of NAACL-HLT-2016*. San Diego, CA, USA: Association for Computational Linguistics, 2016. p. 720–730.
- MIKOLOV, T. et al. Extensions of recurrent neural network language model. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Prague, Czech Republic: IEEE, 2011. p. 5528–5531.
- MITCHELL, M.; DEEMTER, K. van; REITER, E. Natural reference to objects in a visual domain. In: *Proceedings of INLG-2010*. Dublin, Ireland: The Association for Computer Linguistics, 2010.
- MORIMOTO, C. H.; MIMICA, M. R. M. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, v. 98, p. 4–24, 2005.
- 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.

- NAJIB, F.; CHEEMA, W. A.; ADEELNAWAB, R. M. Author's traits prediction on Twitter data using content based approach. In: CAPPELLATO, L. et al. (Ed.). *CLEF 2015 Working Notes*. Toulouse, France: CEUR-WS.org, 2015. (CEUR Workshop Proceedings, v. 1391).
- NOVAIS, E. M. de et al. A testbed for Portuguese natural language generation. In: *Seventh Brazilian Symposium in Information and Human Language Technology (STIL)-2009*. São Carlos, Brazil: IEEE, 2009. p. 154–157.
- NOVAIS, E. M. de; PARABONI, I. Portuguese text generation using factored language models. *Journal of the Brazilian Computer Society*, Springer-Verlag, v. 19, n. 2, p. 135–146, 2012.
- NOVAIS, E. M. de; PARABONI, I.; FERREIRA, D. T. Highly-inflected language generation using factored language models. In: *Computational Linguistics and Intelligent Text Processing (CICLing-2011)*, *Lecture Notes in Computer Science 6608*. Tokyo, Japan: Springer-Verlag, 2011. p. 429–438.
- NOVAIS, E. M. de; PARABONI, I.; Silva Junior, D. F. P. da. Portuguese text generation from large corpora. In: *8th International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey: ERLA, 2012. p. 4010–4014.
- NOVAIS, E. M. de; TADEU, T. D.; PARABONI, I. Improved text generation using n-gram statistics. In: *IBERAMIA-2010, Lecture Notes in Artificial Intelligence 6433*. Bahia Blanca, Argentina: Springer-Verlag, 2010. p. 316–325.
- NOVAIS, E. M. de; TADEU, T. D.; PARABONI, I. Text generation for Brazilian Portuguese: the surface realization task. In: *Proceedings of the NAACL-HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*. Los Angeles, USA: Association for Computational Linguistics, 2010. p. 125–131.
- NOVAIS, E. M. de; TADEU, T. D.; PARABONI, I. Text-to-text surface realisation using dependency-tree replacement. In: *IBERAMIA-2010, Lecture Notes in Artificial Intelligence 6433*. Bahia Blanca, Argentina: Springer-Verlag, 2010. p. 326–335.
- NOVIKOVA, J.; LEMON, O.; RIESER, V. Crowd-sourcing NLG data: Pictures elicit better data. In: *Proceedings of the 9th International Natural Language Generation conference*. Edinburgh, UK: Association for Computational Linguistics, 2016. p. 265–273.
- NOWSON, S.; GILL, A. J. Look! who's talking?: Projection of extraversion across different social contexts. In: *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*. Orlando, FL, USA: ACM, 2014. p. 23–26.
- NUNES, M. das G. V. et al. A construção de um léxico para o Português do Brasil: lições aprendidas e perspectivas. In: *Anais do II PROPOR*. Curitiba, Brazil: CEFET-PR, 1996. p. 61–70.
- OBERLANDER, J. Do the right thing... but expect the unexpected. *Computational Linguistics*, v. 24, p. 501–507, 1998.
- OBERLANDER, J.; NOWSON, S. Whose thumb is it anyway? classifying author personality from weblog text. In: *COLING/ACL-2006 Poster Sessions*. Sydney, Australia: Association for Computational Linguistics, 2006. p. 627–634.

- OLIVEIRA, R. L. de et al. A classification-driven approach to document planning. In: *Recent Advances in Natural Language Processing (RANLP-2009)*. Borovets, Bulgaria: RANLP 2009 Organising Committee / ACL, 2009. p. 324–329.
- OLSON, D. R. Language and thought: aspects of a cognitive theory of semantics. *Psychological Review*, v. 77, n. 4, p. 257–273, 1970.
- PAIVA, D. S.; EVANS, R. A framework for stylistically controlled generation. In: BELZ, A.; EVANS, R.; PIWEK, P. (Ed.). *Natural Language Generation*. [S.l.]: Springer Berlin Heidelberg, 2004, (Lecture Notes in Computer Science, v. 3123). p. 120–129.
- PAN, S.; SHAW, J. SEGUE: A hybrid case-based surface natural language generator. In: *3rd International Conference on Natural Language Generation (INLG-2004)*, LNCS 3123. Brockenhurst, UK: Springer-Verlag, 2004.
- PAPINENI, S. et al. Bleu: a method for automatic evaluation of machine translation. In: *Proceddings of ACL-2002*. Philadelphia, PA, USA: Association for Computational Linguistics, 2002. p. 311–318.
- PARABONI, I.; DEEMTER, K. van. Using overspecified descriptions to help people find their way in spatial domains. In: *Symposium on Influencing People with Information (SIPI-2012)*. Aberdeen, UK: [s.n.], 2012.
- PARABONI, I.; DEEMTER, K. van. Reference and the facilitation of search in spatial domains. *Language and Cognition Processes / Language, Cognition and Neuroscience*, v. 29, n. 8, p. 1002–1017, 2014.
- 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.
- 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.
- PARABONI, I. et al. Effects of cognitive effort on the resolution of overspecified descriptions. *Computational Linguistics*, v. 43, n. 2, p. 451–459, 2017.
- 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.
- PARABONI, I. et al. Generating underspecified descriptions of landmark objects. In: *Text, Speech and Dialogue (TSD-2014), Lecture Notes in Artificial Intelligence 8655*. Brno, Czech Republic: Springer, 2014. p. 76–83.
- PASSONNEAU, R. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In: *International Conference on Language Resources and Evaluation (LREC)*. Genoa, Italy: ELRA, 2006.
- PECHMANN, T. Incremental speech production and referential overspecification. *Linguistics*, v. 27, n. 1, p. 98–110, 1989.

- PEREIRA, D. B.; PARABONI, I. A language modelling tool for statistical NLP. In: *5th Workshop on Information and Human Language Technology (TIL-2007). Anais do XXVII Congresso da SBC*. Rio de Janeiro: Sociedade Brasileira de Computação, 2007. p. 1679–1688.
- PEREIRA, D. B.; PARABONI, I. O pré-processamento de córpus para construção de modelos estatísticos de língua. In: *VI Encontro de Linguística de Corpus*. São Paulo: [s.n.], 2007.
- PEREIRA, D. B.; PARABONI, I. From TUNA attribute sets to Portuguese text: a first report. In: *Proceedings of the Fifth International Natural Language Generation Conference*. Salt Fork, Ohio, USA: The Association for Computer Linguistics, 2008. p. 232–233.
- PEREIRA, D. B.; PARABONI, I. Statistical surface realisation of Portuguese referring expressions. In: *Gotal-2008, Lecture Notes in Artificial Intelligence 5221*. Gothenburg, Sweden: Springer-Verlag, 2008. p. 383–392.
- PEREIRA, H. V. L. et al. Corpus-based referring expressions generation. In: *8th International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey: ERLA, 2012. p. 4004–4009.
- PLANK, B.; HOVY, D. Personality traits on Twitter - or - how to get 1,500 personality tests in a week. In: *Proc. of WASSA-2015*. Lisbon, Portugal: Association for Computational Linguistics, 2015. p. 92–98.
- PORTET, F. et al. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, v. 173, p. 789–816, 2009.
- RAJKUMAR, R.; WHITE, M. Better surface realization through psycholinguistics. *Language and Linguistics Compass*, v. 8, n. 10, p. 428–448, 2014.
- RAMOS, R. M. S. et al. *O córpus b5 de textos e inventários de personalidade (v. 1.0.). Relatório técnico PPgSI nr.0001/2017*. São Paulo, Brazil, 2017.
- RANGEL, F. et al. Overview of the 3rd Author Profiling Task at PAN 2015. In: *CLEF 2015 Evaluation Labs and Workshop*. Toulouse, France: CEUR-WS.org, 2015.
- REEVES, B.; NASS, C. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York, NY, USA: Cambridge University Press, 1996. ISBN 1-57586-052-X.
- REITER, E.; BELZ, A. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, v. 35, n. 4, p. 529–558, 2009.
- REITER, E.; DALE, R. *Building natural language generation systems*. New York, NY, USA: Cambridge University Press, 2000. ISBN 0-521-62036-8.
- REITER, E.; ROBERTSON, R.; OSMAN, L. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, v. 144, p. 41–58, 2003.
- REITER, E. et al. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, v. 167, p. 137–169, 2005.

- REITER, E.; WILLIAMS, S. Generating texts in different styles. In: ARGAMON, S.; BURNS, K.; DUBNOV, S. (Ed.). *The Structure of Style*. [S.l.]: Springer Berlin Heidelberg, 2010. p. 59–75. ISBN 978-3-642-12336-8.
- RIGHI, G.; PEISSIG, J. J.; TARR, M. J. Recognizing disguised faces. *Visual Cognition*, v. 20, n. 2, p. 143–169, 2012.
- SANDRIM, D. C.; FREIRE, F.; PARABONI, I. Geração de instruções em mundos virtuais: primeiros passos. In: *9th Brazilian Symposium in Information and Human Language Technology (STIL-2013)*. Fortaleza, Brazil: Association for Computational Linguistics, 2013. p. 189–193.
- SANTOS, F. M. V. dos; PEREIRA, D. B.; PARABONI, I. Rule-based vs. probabilistic surface realisation of definite description. In: *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*. Vila Velha, Brazil: ACM, 2008. p. 372–374.
- SANTOS, V. G. dos; PARABONI, I.; SILVA, B. B. C. Big five personality recognition from multiple text genres. In: *Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415*. Prague, Czech Republic: Springer-Verlag, 2017. p. 29–37.
- SCHNEIDER, A. H. et al. Using technology to enhance rural resilience in pre-hospital emergencies. *Scottish Geographical Journal*, Taylor & Francis, v. 131, n. 3-4, p. 194–200, 2015.
- SCOTT, D.; MOORE, J. An NLG evaluation competition? eight reasons to be cautious. In: *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*. Arlington, VA, USA: [s.n.], 2007. p. 22–23.
- SILVA, D. dos S.; PARABONI, I. Geração de expressões de referência usando relações espaciais. In: *9th Brazilian Symposium in Information and Human Language Technology (STIL-2013)*. Fortaleza, Brazil: Association for Computational Linguistics, 2013. p. 88–97.
- SILVA, D. dos S.; PARABONI, I. Geração de expressões de referência em ambientes virtuais interativos. *Linguamática*, v. 6, n. 1, p. 15–27, 2014.
- SILVA, D. dos S.; PARABONI, I. Generating spatial referring expressions in interactive 3D worlds. *Spatial Cognition & Computation*, v. 15, n. 03, p. 186–225, 2015.
- Silva Junior, D. F. P. da; NOVAIS, E. M. de; PARABONI, I. Realização superficial baseada em regras. In: *9th Brazilian Symposium in Information and Human Language Technology (STIL-2013)*. Fortaleza, Brazil: Association for Computational Linguistics, 2013. p. 148–152.
- Silva Junior, D. F. P. da; PARABONI, I.; NOVAIS, E. M. de. Um sistema de realização superficial para geração de texto em português. *Revista de Informática Teórica e Aplicada*, v. 20, n. 3, p. 31–48, 2013.
- SRIPADA, S. et al. A case study: NLG meeting weather industry demand for quality and quantity of textual weather forecasts. In: *Proceedings of the 8th International Natural Language Generation Conference (INLG)-2014*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2014. p. 1–5.
- STENT, A.; MOLINA, M. Evaluating automatic extraction of rules for sentence plan construction. In: *SIGDIAL'09 Proceedings*. London, UK: Association for Computational Linguistics, 2009. p. 290–297.

STRIEGNITZ, K. et al. Report on the second second challenge on generating instructions in virtual environments (GIVE-2.5). In: *Generation Challenges Session at the 13th European Workshop on Natural Language Generation (ENLG-2011)*. Nancy, France: Association for Computational Linguistics, 2011. p. 270–279.

TADEU, T. D.; NOVAIS, E. M. de; PARABONI, I. Extracting surface realisation templates from corpora. In: *7th International Conference on Language Resources and Evaluation (LREC-2010)*. Valletta, Malta: ELRA, 2010. p. 2392–2395.

TEIXEIRA, C. V. M. et al. Generating relational descriptions involving mutual disambiguation. In: *Computational Linguistics and Intelligent Text Processing (CICLing-2014), Lecture Notes in Computer Science 8403*. Kathmandu, Nepal: Springer, 2014. p. 492–502.

TENBRINK, T. Identifying objects on the basis of spatial contrast: An empirical study. In: FREKSA, C. et al. (Ed.). *Spatial Cognition IV: Reasoning, Action, Interaction. International Conference Spatial Cognition 2004*. Frauenchiemsee, Germany: Springer Verlag, 2005. p. 124–146.

TINTAREV, N. et al. Personal storytelling: Using natural language generation for children with complex communication needs, in the wild... *International Journal of Human-Computer Studies*, v. 92–93, p. 1–16, 2016.

VARGES, S. Overgeneration and ranking for spoken dialogue systems. In: *Proceedings of INLG-2006*. Sydney, Australia: Association for Computational Linguistics, 2006. p. 20–22.

VIETHEN, H. A. E. *The Generation of Natural Descriptions: Corpus-based Investigations of Referring Expressions in Visual Domains*. Tese (Doutorado) — Macquarie University, 2011.

VIETHEN, J.; DALE, R. Speaker-dependent variation in content selection for referring expression generation. In: *Proceedings of the Australasian Language Technology Association Workshop 2010*. Melbourne, Australia: Australasian Language Technology Association, 2010. p. 81–89.

VIETHEN, J.; DALE, R. GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In: *UCNLG+Eval: Language Generation and Evaluation Workshop*. Edinburgh, UK: Association for Computational Linguistics, 2011. p. 12–22.

VIETHEN, J.; MITCHELL, M.; KRAHMER, E. Graphs and spatial relations in the generation of referring expressions. In: *14th European Workshop on Natural Language Generation*. Sofia, Bulgaria: Association for Computational Linguistics, 2013. p. 72–81.

WALKER, M. A. et al. Perceived or not perceived: Film character models for expressive NLG. In: SI, M. et al. (Ed.). *Lecture Notes in Computer Science 7069*. Vancouver, Canada: Springer Verlag, 2011. p. 109–121.

WALKER, M. A. et al. Murder in the arboretum: Comparing character models to personality models. In: *Intelligent Narrative Technologies IV, Papers from the 2011 AIIDE Workshop*. Palo Alto, California, USA: The AAAI Press, 2011.

WALKER, M. A. et al. Does personality matter? expressive generation for dialogue interaction. *Natural Interaction with Robots, Knowbots and Smartphones*, Springer, p. 285, 2014.

WEN, T. et al. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In: *Proceedings of the SIGDIAL 2015 Conference*. Prague, Czech Republic: Association for Computational Linguistics, 2015. p. 275–284.

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.

WILLIAMS, S.; REITER, E. Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, v. 14, p. 495–535, 2008.

XU, K. et al. Show, attend and tell: Neural image caption generation with visual attention. In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*. Lille, France: [s.n.], 2015. p. 2048–2057.

ZHANG, X.; LAPATA, M. Chinese poetry generation with recurrent neural networks. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 670–680.

ZHONG, H.; STENT, A. J. Building surface realizers automatically from corpora. In: *Proceedings of the Workshop on Using Corpora for Natural Language Generation (UCNLG-2005)*. Birmingham, UK: Association for Computational Linguistics, 2005. p. 49–54.