UNIVERSIDADE DE SÃO PAULO

Escola de Artes, Ciências e Humanidades

Relatório Técnico PPgSI-002/2014

O corpus Stars de expressões de referência

Ivandré Paraboni

Novembro - 2014

Série de Relatórios Técnicos

PPgSI-EACH-USP. Rua Arlindo Béttio, 1000 - Ermelino Matarazzo - 03828-000
São Paulo, SP.
TEL: (11) 3091-8197 http://www.each.usp.br/ppgsi

O corpus Stars de expressões de referência

Ivandré Paraboni[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo
São Paulo – SP, Brazil

ivandre@usp.br

Resumo. *Este documento descreve a construção do corpus* Stars *de expressões de referência e sua estrutura. O corpus foi construído como um experimento-piloto para o estudo de diversas questões de interesse para a geração de língua natural, como a geração de expressões de referência relacionais, a variação humana na seleção de conteúdo e a questão da subespecificação de pontos de referência. As expressões coletadas foram anotadas com suas propriedades semânticas e disponibilizadas juntamente com as imagens de estímulo para fins de pesquisa.*

1. Introdução

Sistemas de geração de língua natural (GLN) - que produzem descrições textuais a partir de uma entrada de dados geralmente não linguística [Reiter e Dale 2000] - são empregados quando o uso de texto predefinido não é suficiente, ou seja, quando é necessária uma maior variação linguística nos documentos gerados e/ou maior proximidade em relação ao desempenho humano. Aplicações de GLN incluem sistemas de diálogo humano-computador em língua natural, geração de relatórios a partir de bases de dados, sumarização de documentos WEB, e muitas outras.

Dentre as diversas subtarefas desempenhadas por um sistema de GLN está a geração de expressões de referência (GER) [Krahmer e van Deemter 2012], que consiste em produzir descrições linguísticas de objetos de que trata o discurso. A tarefa de GER pode ser desempenhada com uso de algoritmos de propósito específico [Dale e Haddock 1991, Dale e Reiter 1995, Krahmer e Theune 2002, Krahmer et al. 2003, de Lucena et al. 2010] ou, mais recentemente, com uso de técnicas de aprendizagem de máquina [Viethen e Dale 2011, Ferreira e Paraboni 2014a]. Neste segundo caso pode-se tirar proveito de conjuntos de dados - ou corpus de GER - produzidos sob condições controladas.

Um corpus de GER tipicamente consistem de coleções de imagens de estímulo contendo um objeto-alvo e um certo número de distraidores, e descrições linguísticas do objeto-alvo produzidas por um grupo de sujeitos humanos. Exemplos de recursos deste tipo incluem o corpus *Coconut* [Eugenio et al. 2000], *Drawer* [Viethen e Dale 2006], *TUNA* [Gatt et al. 2007], *GRE3D3/7* [Dale e Viethen 2009, Viethen e Dale 2011] e outros. Neste trabalho descrevemos a construção de um novo recurso deste tipo - o corpus *Stars* de descrições relacionais - e seus resultados preliminares.

2. Trabalho Relacionado

2.1. A Geração de Expressões de Referência (GER)

Objetos de um determinado contexto podem ser referenciados com uso de descrições definidas, indefinidas e diversas outras formas, como em 'o prédio da esquina' ou 'uma rua à esquerda, depois do supermercado'. Em sistemas de geração de língua natural,

denominamos Geração de Expressões de Referência (GER) a tarefa computacional de produzir descrições deste tipo a partir de dados não-linguísticos fornecidos como entrada[1].

GER é uma ativa linha de pesquisa em GLN, e pode estar refletida nas três camadas da arquitetura de GLN tradicional [Reiter e Dale 2000]: macroplanejamento [Paraboni e van Deemter 1999, 2002b], microplanejamento [Krahmer e van Deemter 2012] e realização superficial[2]. Neste trabalho enfocamos apenas a questão da seleção de conteúdo implementada como uma das tarefas de microplanejamento[3].

A seleção de conteúdo em GER - a tarefa de decidir quais propriedades semânticas de um objeto-alvo devem ser incluídas em uma descrição do mesmo - conta com inúmeros algoritmos de propósito geral [Dale e Reiter 1995, Paraboni 2000, Paraboni e van Deemter 2002a, Paraboni 2003, Paraboni et al. 2006, Paraboni e van Deemter 2014]. Algoritmos deste tipo tipicamente seguem critérios de preferência por determinados tipos de propriedades [Pechmann 1989], mas estudos mais recentes demonstrem que a tarefa de GER é na realidade mais complexa, e estas preferências podem ser efetivamente redefinidas [van Deemter et al. 2012, Tarenskeen et al. 2014, van Gompel et al. 2014].

Um dos algoritmos de seleção de conteúdo mais conhecidos na área, e que ajudou a definir o próprio problema computacional de geração de expressões de referência, é o algoritmo Incremental apresentado em Dale e Reiter [1995]. Este algoritmo recebe como entrada um contexto C formado por um grupo de objetos denominados distraidores, o objeto-alvo ou referente r que se deseja descrever, e suas propriedades semânticas na forma de pares (*atributo* - *valor*), como em (*cor* - *azul*). O objetivo do algoritmo é produzir como saída uma descrição L composta de uma série de pares atributo-valor que representem o objeto-alvo r e nenhum outro distraidor do contexto C.

2.2. Corpus para GER

Assim como em diversas outras linhas de pesquisa do processamento de línguas naturais (PLN), a pesquisa em GLN/GER envolve a observação empírica do uso da língua para treinamento e teste de artefatos computacionais. Em outras palavras, também no caso da seleção de conteúdo faz-se necessário o uso de corpus representativos do fenômeno investigado. Diferentemente de outras áreas do PLN, entretanto, o uso de corpus de propósito geral (e.g., uma coleção de artigos jornalísticos) em GLN é limitado pelo fato de que o conhecimento a partir do qual o texto foi produzido normalmente não se encontrar disponível. Em outras palavras, o texto que encontramos em um corpus de propósito geral é apenas o produto final de um processo de produção humana da língua, e que normalmente traz pouca informação sobre o processo em si. Além disso, aspectos relevantes do processo, mesmo que presentes, tendem a ocorrer misturados a diversos outros fenômenos, e permanecem assim fora do controle do pesquisador.

Na pesquisa em GLN e áreas afins como psicolinguística, ciências cognitivas etc., a forma usual de reproduzir condições de produção de língua natural é a realização de

[1]GER pode também ser vista como a tarefa computacional 'simétrica' à interpretação de expressões de referência [Paraboni 1997, Cuevas e Paraboni 2008].

[2]Exemplos de sistemas de realização superficial para o Português são apresentados em Pereira e Paraboni [2007, 2008], de Novais e Paraboni [2012]

[3]Seleção de conteúdo e realização superficial são efetivamente linhas de pesquisa distintas, contando inclusive com métodos de avaliação próprios, e.g., sistemas de realização superficial são avaliados com base em métricas como BLEU [Papineni et al. 2002] e NIST [NIST 2002].

experimentos controlados com uso de participantes humanos. Experimentos deste tipo fornecem um certo tipo de estímulo textual ou visual aos participantes - que podem assumir o papel de falante, ouvinte ou ambos - e registram suas reações, geralmente manifestas na forma oral, escrita, ou ainda por ações de interação com o ambiente computacional. Em outras palavras, para pesquisa e validação de vários tipos de sistemas e artefatos de GLN, faz-se necessário examinar não apenas o texto resultante do processo, mas também modelar as condições contextuais nas quais este texto for produzido. Dependendo do objetivo da pesquisa, isso pode envolver, por exemplo, a definição do conteúdo que estava em discussão no momento em que a sentença foi produzida, do objetivo da comunicação, do conhecimento prévio do falante, do seu estado de atenção etc.

Experimentos para coleta de corpus para GLN fornecem estímulo textual ou visual aos participantes e registram suas reações, manifestas na forma oral, escrita, ou por ações de interação com um ambiente computacional. De acordo com o papel assumido pelo participante humano no experimento, estes estudos podem fazer uso de experimentos orientados ao falante, experimentos orientados ao ouvinte, ou ambos. A seguir apresentamos um breve levantamento dos principais recursos produzidos que se encontram publicamente disponíveis para pesquisa na área.

Experimentos orientados ao falante são utilizados tanto para ganhar entendimento sobre um determinado aspecto da produção de língua (e.g., coletando-se dados de treinamento) como para validação de artefatos ou sistemas previamente existentes (e.g., coletando-se dados de teste). Um dos primeiros exemplos de experimento de GLN orientado ao falante com construção de um conjunto de dados de uso público é o caso do corpus *Drawer* de expressões de referência em Viethen e Dale [2006]. Este corpus é baseado em uma única cena representando um armário com 16 gavetas de cores variadas, contendo 140 descrições geradas por 20 participantes. O objetivo do estudo foi o de examinar a questão da variação humana de estratégias de produção de língua, a qual é de interesse para a presente proposta. O pequeno número de instâncias coletadas, entretanto, naturalmente limita o escopo de seu eventual reuso.

Um exemplo mais representativo de experimento orientado ao falante é o projeto de construção do corpus *TUNA* em Gatt et al. [2007], que contém expressões coletadas propositalmente para o estudo de fenômenos e algoritmos de geração de expressões de referência em uma série de competições de algoritmos de GER [Gatt e Belz 2007, Gatt et al. 2008, 2009]. O corpus *TUNA* descreve situações de referência em dois domínios distintos: peças de Mobília (*Furniture*), e fotos de Pessoas (*People*). O corpus *TUNA* contém 2280 expressões (780 singulares e 1500 plurais) e seus respectivos contextos. As descrições foram geradas a partir de experimentos controlados realizados com 50 participantes que usavam o Inglês como idioma nativo, ou possuíam fluência no mesmo. O propósito da descrição era apenas a identificação do objeto indicado. Descrições no domínio TUNA são no entanto do tipo atômico, ou seja, não contemplam situações mais complexas de uso de relações entre objetos.

Um exemplo recente de experimento orientado ao falante é a construção dos corpus *GRE3D3* e *GRE3D7* em Dale e Viethen [2009], Viethen e Dale [2011], que trata da questão do uso de relações espaciais em um contexto visual tridimensional simplificado. Nestes experimentos, 294 participantes foram instruídos a descrever objetos geométricos do tipo esfera e cubo. O procedimento resultou no corpus *GRE3D3* [Dale e Viethen 2009],

posteriormente ampliado para o corpus *GRE3D7* [Viethen e Dale 2011], contendo 4480 descrições destes tipos de objetos. Apesar do volume considerável de dados, o domínio *GRE3D7* possui ainda um número reduzido de atributos atômicos e relacionais possíveis.

Experimento orientados ao falante são também úteis para a construção de corpus de diálogos [Eugenio et al. 2000], como no caso do corpus *iMap* [Guhe 2009] de instruções de rota em mapas bidimensionais. O corpus *iMap* consiste de 256 diálogos construídos com base em mapas. Em cada diálogo, uma dupla de participantes revezava-se nos papeis 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 corpus *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 de psicologia), não há uso prático além do estudo apresentado em Guhe [2009].

Finalmente, o corpus *GIVE-2* [Gargett et al. 2010] de instruções em mundos virtuais foi construído como parte do projeto GIVE [Byron et al. 2007] e de uma série de competições de geração de instruções em mundos virtuais [Byron et al. 2009, Koller et al. 2010, Striegnitz et al. 2011]. O corpus foi criado por meio de experimentos envolvendo 36 pares de participantes de língua inglesa e alemã alternando-se nas tarefas de instrutor e jogador. Este corpus foi produzido em condições semelhantes às do modelo instrutor-receptor empregado no corpus *iMap*, 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 corpus *GIVE-2* também apresenta algumas dificuldades de reuso. A representação em formato de animação acompanhada de diálogos torna as tarefas de anotação e processamento significativamente complexas, e provavelmente não justificáveis face ao número reduzido de descrições de interesse (cerca de 992 descrições de objetos do tipo botão, que é o único elemento manipulável em mundos *GIVE*).

3. Trabalho realizado

Corpus de GER são normalmente construídos com um propósito altamente específico, o que tende a limitar o escopo das conclusões possíveis sobre outros tipos de fenômeno além daqueles originalmente contemplados em seus experimentos de origem, e já exaustivamente discutidos em seus respectivos projetos de pesquisa. Por estes motivos, e apesar da inegável contribuição de recursos deste tipo para a pesquisa na área, novas questões de pesquisa tendem a exigir recursos com características ainda não oferecidas. Nesta seção descrevemos o projeto do experimento de coleta do corpus de expressões de referência Stars e seus resultados.

3.1. Experimento

A coleta de dados que deu origem ao corpus Stars foi planejada de modo a obter um grande número de descrições relacionais contendo até três referentes - aqui denominados

objeto-alvo principal, ponto de referência e segundo ponto de referência. Diferentemente das descrições relacionais nos corpus GRE3D3/7 [Dale e Viethen 2009, Viethen e Dale 2011], no domínio Stars o uso de propriedades relacionais é frequentemente necessário para desambiguação do objeto-alvo.

O experimento objetivou investigar condições em que são produzidas descrições relacionais mínimas, ou com ponto de referência subespecificado (situação em que objeto-alvo e ponto de referência permitem desambiguação mútua). As hipóteses consideradas no experimento e seus resultados são discutidos em Paraboni et al. [2014].

Foram definidas seis condições críticas, sendo três do tipo *abs* (de absoluto) e três do tipo *proj* (projetiva). Estas condições objetivam comparar situações em que um ponto de referência (uma caixa) pode ser totalmente especificada com uso de um atributo absoluto (cor) ou apenas com uso de um atributo projetivo (uma relação espacial com um terceiro objeto do tipo círculo). A Fig. 1 exemplifica os dois tipos de condição.

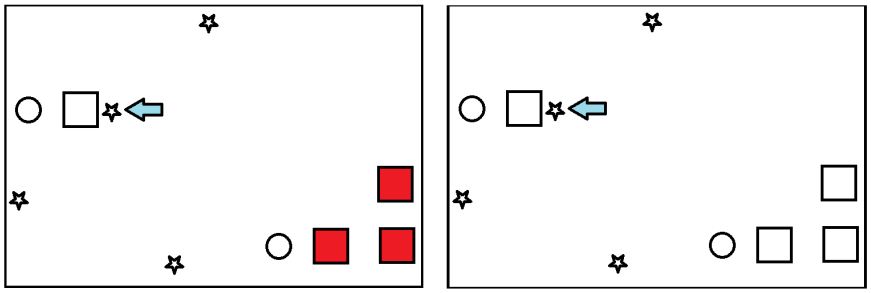

Figura 1. Duas situações de possível referência a um objeto-alvo (estrela) via ponto de referência (caixa). Adaptado de Paraboni et al. [2014]

A diferença entre os dois tipos de condição é o fato de que somente na cena da esquerda o ponto de referência (caixa) possui uma cor discriminatória (branca). Os dois tipos de condição foram apresentados em três tamanhos de contexto (1, 3 e 5) correspondendo ao número de distraidores do mesmo tipo que o ponto de referência (caixa). A Fig. 1 ilustra exemplos de condições *abs3* e *proj3*, respectivamente, ambas adaptadas a partir das imagens de estímulo usadas no experimento e descritas na seção Materiais a seguir.

Em todas as condições críticas, o objeto-alvo é sempre uma estrela, e é acompanhado por três distraidores idênticos, o que força o uso de informação adicional para fins de desambiguação (por exemplo, o uso de informação espacial como 'a estrela no canto esquerdo' ou uma propriedade relacional, como em 'a estrela ao lado da caixa'). O objeto mais próximo do alvo, e que tem maior probabilidade de ser selecionado como ponto de referência, é sempre do tipo caixa. Em todos os casos, é sempre possível produzir uma descrição relacional subespecificada como em 'a estrela ao lado da caixa'. Detalhes sobre as hipóteses do experimento Stars são discutidos em Paraboni et al. [2014].

3.2. Sujeitos

73 estudantes de Sistemas de Informação da USP-EACH, que responderam a um convite enviado por email, e concordaram em participar de forma voluntária. Os participantes tinham em média 20,9 anos de idade e eram predominantemente do sexo masculino (62, ou 84,9%). Todos participantes eram brasileiros nativos.

3.3. Procedimento

Os participantes foram solicitados a executar um pequeno aplicativo JAVA (.jar) anexo ao email de convite, e a seguir as instruções exibidas na tela. Ao executar o aplicativo, uma breve página de instruções informava ao participante que se tratava de uma participação voluntária e anônima em pesquisa em Ciência da Computação. O número de matrícula do estudante era solicitado como forma de computar a faixa etária média e gênero do grupo de participantes.

As instruções explicavam também que a tarefa a ser realizada consistia em completar a frase 'O objeto apontado pela seta azul é um/a...' em uma série de imagens, e que os participantes deveriam completar a frase da forma mais natural possível, como se estivessem conversando com um amigo sobre os objetos na cena sem ambiguidade. Para não influenciar o tipo de descrição produzida, nenhum exemplo foi fornecido.

As imagens eram exibidas uma-a-uma em ordem aleatória, juntamente com a frase a ser completada e um botão 'Próximo' para avançar para a próxima imagem de estímulo. Casos triviais de ambiguidade como 'a estrela' eram detectados automaticamente pelo aplicativo, que então exibia uma mensagem do tipo 'Eu não sei de qual estrela você está falando. Por favor seja mais específico' e solicitava uma nova descrição. Casos mais complexos de ambiguidade não eram verificados.

Ao fornecer uma descrição e pressionar o botão 'Próximo', uma nova imagem de estímulo era exibida. Ao final do experimento, um arquivo com as respostas criptografadas era gerado e o participante era solicitado a enviá-lo por email aos responsáveis pelo experimento.

3.4. Materiais

O experimento fez uso de software desenvolvido especificamente para este fim. O aplicativo em questão exibia as instruções gerais, as imagens de estímulo em ordem aleatória, coletava as respostas e as armazenava em um arquivo criptografado.

O conjunto de estímulo consistia de 11 imagens representando as seis condições críticas ilustradas na Fig. 2 e cinco *fillers* ilustrados na Fig. 3. Embora sem função no experimento original, as imagens do tipo os *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*.

3.5. Coleta e transcrição

Foi coletado um total de 803 descrições produzidas por 73 participantes, que levaram em média 5,5 minutos cada para completar a tarefa. 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

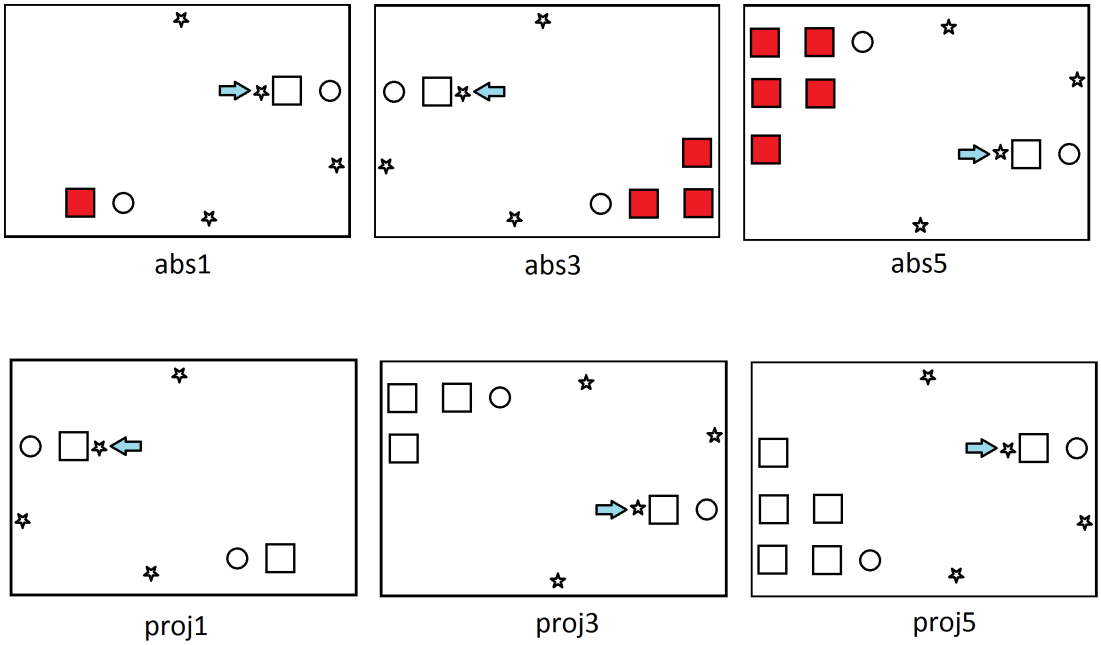

Figura 2. Condições críticas

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 corpus final foi constituido de 704 descrições produzidas por 64 participantes.

Dada a disposição dos objetos de interesse em triplas estrela-caixa-círculo, foram anotados somente os atributos referenciais destes três objetos. Neste esquema, o objeto-alvo principal é considerado sempre a estrela, o (primeiro) ponto de referência é sempre a caixa, e o segundo ponto de referência é sempre o circulo. Nos poucos casos em que uma expressão relaciona o objeto-alvo e o segundo ponto de referência diretamente (e.g., 'a estrela ao lado do círculo', omitindo o fato de que há uma caixa entre eles) os atributos da caixa foram anotados como pertencentes ao segundo ponto de referência, ou seja, como se houvesse um primeiro ponto de referência omitido. Referências a objetos distantes deste grupo (e.g., 'a estrela do lado oposto ao grupo de cubos vermelhos') foram infrequentes, e não foram anotadas de modo a manter a simplicidade do esquema de anotação.

Os atributos anotados e seus valores possíveis para cada objeto citado (i.e, em função de alvo-principal, primeiro e segundo pontos de referência) são resumidos na Tabela 1.

Tabela 1. Atributos anotados

Atributo	Valores possíveis
type	{star, cube, ball}
colour	{red, white}
hpos	{left, right} ou nulo
vpos	{centre} ou nulo
near	(id de objeto) ou nulo
left	(id de objeto) ou nulo
right	(id de objeto) ou nulo

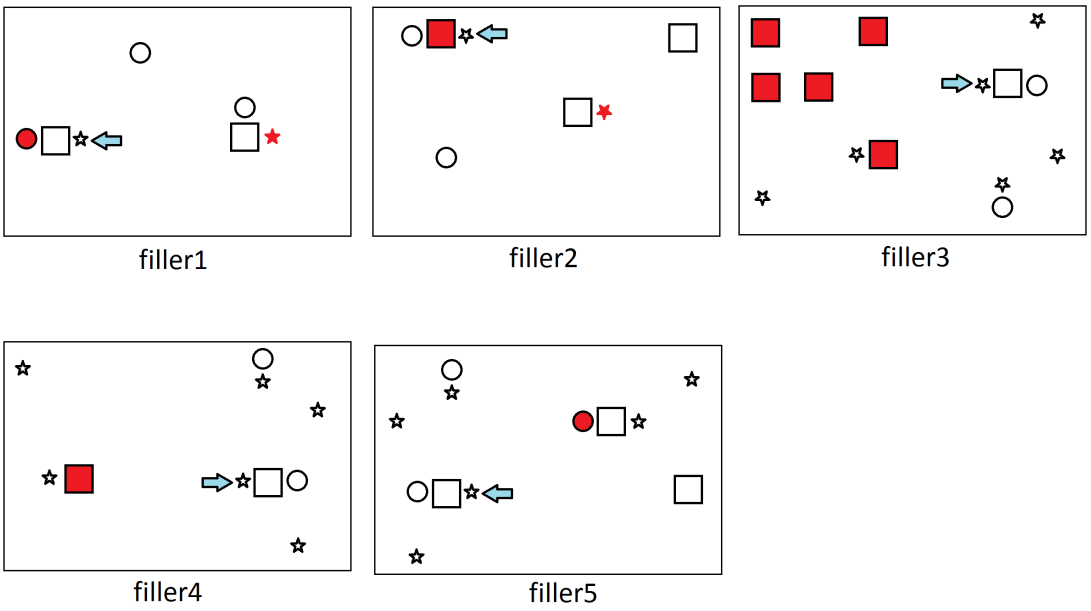

Figura 3. Fillers

Os nomes de atributos são usados como especificado apenas para o caso do objeto-alvo. No caso dos pontos de referência, estes nomes são antecedidos pelos identificadores 'LANDMARK-' ou 'SECOND-LANDMARK' conforme necessário. Por exemplo, o atributo 'tipo' do segundo ponto de referência da descrição é designado pelo atributo 'SECOND-LANDMARK-TYPE'. A diferenciação entre atributos do objeto-alvo, primeiro e segundo pontos de referência facilita a implementação de algoritmos de GER e o cálculo automático de coeficientes Dice [Dice 1945] e outros.

Ocorrências de outros atributo eram raras e não forma transcritas, embora possam ser inferidas a partir da forma superficial coletada. Este é o caso, por exemplo, de instâncias de referência ordinal como 'a segunda estrela da esquerda para direita'.

Três anotadores trabalharam de forma independente e a representação final foi consolidado por um quarto avaliador. Dada a simplicidade do domínio, a transcrição não exigiu concordância entre juízes [Landis e Koch 1977].

Tanto imagens como descrições foram transcritas em formato XML. O formato utilizado na representação do corpus é uma versão simplificada do formato utilizado no corpus TUNA [Gatt et al. 2007]. Nesta representação, tanto objetos de cada contexto quanto expressões de referência são representados por um nó ATTRIBUTE-SET contendo uma série de atributos identificador por nome (NAME) e valor (VALUE). O exemplo a seguir ilustra este tipo de representação para um objeto do tipo cubo.

```
<ATTRIBUTE-SET ID="q1">
  <ATTRIBUTE NAME="type" VALUE="cube" />
  <ATTRIBUTE NAME="colour" VALUE="red" />
  <ATTRIBUTE NAME="left" VALUE="c2" />
  <ATTRIBUTE NAME="hpos" VALUE="left" />
  <ATTRIBUTE NAME="vpos" VALUE="bottom" />
</ATTRIBUTE-SET>
```

No caso da representação dos contextos, o nó raiz CONTEXT armazena informações sobre o identificador de cada cena (e.g., *abs1*, *filler1* etc.), o arquivo de imagem utilizado e o identificador do objeto-alvo. Esta última informação pode também ser inferida a partir dos conjuntos de atributos, já que o nó ATTRIBUTE-SET pode inclui um campo FUNCTION indicando o papel do objeto em questão no experimento. Os valores possíveis para este atributo são TARGET, LANDMARK e SECOND-LANDMARK. As nove imagens são agrupadas em um único arquivo denominado context.xml. Um fragmento da representação XML de um contexto é ilustrado no Apêndice A.

No caso das descrições coletadas, estas foram agrupadas por sujeito, formando 64 arquivos individuais contendo 9 expressões cada, e tendo um nó TRIAL como raiz. Cada trial possui um identificador sequencial único, um identificador do sujeito (também único), e informações de idade (AGE) e gênero (GENDER) do mesmo. Cada conjunto de atributos que forma uma descrição é subordinado a um nó CONTEXT contendo o identificador da cena em questão. O nó ATTRIBUTE-SET da descrição inclui informação sobre o alvo principal (TARGET) e, se pertinente, ponto de referência (LANDMARK) e segundo ponto de referência (SECOND-LANDMARK). Além disso, é apresentada também a forma superficial produzida pelo sujeito (STRING), o número de atributos transcritos (LENGTH) e a quantidade de relações (REL-COUNT). Descrições atômicas são anotadas como REL-COUNT=0. Um exemplo de representação XML de uma expressão de referência é apresentado no Apêndice B.

4. Resultados

As descrições obtidas tinham em média 4.4. atributos cada, com desvio de 1,6 e faixa de 1 a 10 atributos. A Tabela 2 ilustra a distribuição de descriçoes por quantidade de objetos referenciados. Dadas as características do estímulo empregado, a alta incidência (92,3%) de descrições relacionais era esperada.

Tabela 2. Nro. de referentes

Referentes	Quant.	%
0	54	7.7%
1	477	67.7%
2	173	24.6%
total	704	100.0%

Para as 384 descrições produzidas nas condições críticas do experimento (i.e., excetuando-se as imagens do tipo *filler*), a Tabela 3 ilustra a distribuição por estratégia de especificação do primeiro ponto de referência. A especificação de valor zero representa descrições de ponto de referência subespecificadas (i.e., sem adição de nenhuma propriedade além do tipo básico do objeto). Especificações superiores a zero indicam que o número correspondente de propriedades foi adicionado à descrição para evitar desambiguação mútua entre objeto-alvo e ponto de referência. Observa-se nestes resultados a significativa (63,8%) preferência por descrições superespecificadas.

Tabela 3. Especificação de pontos de referência

Especificação	Quant.	%
0	139	36.2%
1	211	54.9%
2	34	8.9%
total	384	100.0%

Finalmente, analisamos o comportamento de um algoritmo de GER padrão na geração das descrições do corpus. Para este fim, utilizamos uma versão relacional do algoritmo Incremental [Dale e Reiter 1995] com controle na de referência circular e levando-se em conta uma lista preferencial P definida como segue:

P = <*type*, *colour*, *near*, *left*, *right*, *hpos*, *vpos* >

A versão padrão do algoritmo - a ser utilizada como sistema de *baseline* - inclui propriedades até obter uma descrição livre de ambiguidade. Além desta, foi avaliada também uma variação em que uma propriedade superespecificada é adicionada à descrição do ponto de referência caso esta não esteja completamente especificada. Este procedimento, no caso das situações de referência críticas do corpus Stars (i.e., *abs* e *proj*) tem o efeito prático de prevenir a desambiguação mútua entre objeto-alvo e ponto de referência. Os resultados do algoritmo básico e da versão superespecificada são apresentados na Tabela 4 considerando-se o número de acertos global (Acc) e os coeficientes de Dice [Dice 1945] e MASI [Passonneau 2006].

Tabela 4. Geração de expressões do corpus Stars

Algoritmo	Acc.	Dice	MASI
Baseline	0.08	0.63	0.30
Superespecificado	0.18	0.65	0.33

Embora estes resultados tenham caráter meramente ilustrativo (i.e., uma vez que a proposta de uma estratégia de GER para obtenção de resultados ótimos estaria fora do escopo do presente trabalho), observa-se que a superespecificação das descrições do ponto de referência apresenta ligeira vantagem sobre a estratégia padrão. Esta diferença se explica pelo fato do corpus Stars oferecer um grande número de oportunidades para desambiguação mútua entre objeto-alvo e ponto de referência. Esta estratégia é o comportamento padrão de muitos algoritmos de GER, e é reproduzida pelo algoritmo usado como *baseline*. Observando-se os dados do corpus, entretanto, e possível constatar que esta estratégia foi seguida com menor frequência pelos sujeitos do experimento, o que é refletido no melhor desempenho da estratégia superespecificada.

5. Conclusão

Este documento descreveu o experimento de construção do corpus *Stars* de expressões de referência e sua estrutura. O corpus consiste de um conjunto de imagens de estímulo e descrições linguística de determinados objetos produzidas por participantes de um experimento controlado. Tanto imagens como as descrições obtias foram anotadas semanticamente e representadas em formato XML para utilização na pesquisa em GER.

O corpus *Stars* foi utilizado em um estudo sobre a subespecificação de descrições relacionais em Teixeira et al. [2014], posteriormente expandido em Paraboni et al. [2014] com uma investigação sobre a produção de descrições relacionais mínimas. O corpus foi também utilizado para fins de treinamento e teste de modelos computacionais de GER baseados em aprendizagem de máquina em [Ferreira e Paraboni 2014b].

Agradecimentos

Este trabalho contou com apoio FAPESP e da Universidade de São Paulo.

Referências

Byron, D., Koller, A., Oberlander, J., Stoia, L., e Striegnitz, K. (2007). Generating instructions in virtual environments (GIVE): A challenge and evaluation testbed for NLG. Em *Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*.

Byron, D., Koller, A., Striegnitz, K., Cassell, J., Dale, R., Moore, J., e Oberlander, J. (2009). Report on the first NLG challenge on generating instructions in virtual environments (GIVE). Em *12th European Workshop on Natural Language Generation (ENLG)*, Athens.

Cuevas, R. R. M. e Paraboni, I. (2008). A machine learning approach to portuguese pronoun resolution. *Advances in Artificial Intelligence–IBERAMIA 2008*, LNAI 5290:262–271.

Dale, R. e Haddock, N. J. (1991). Content determination in the generation of referring expressions. *Computational Intelligence*, 7(4):252–265.

Dale, R. e Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Dale, R. e Viethen, J. (2009). Referring expression generation through attribute-based heuristics. Em *Proceedings of ENLG-2009*, páginas 58–65.

de Lucena, D. J., Paraboni, I., e Pereira, D. B. (2010). From semantic properties to surface text: The generation of domain object descriptions. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 14(45):48–58.

de Novais, E. M. e Paraboni, I. (2012). Portuguese text generation using factored language models. *Journal of the Brazilian Computer Society*, 19(2):135–146.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Eugenio, B. D., Jordan, P. W., Thomason, R. H., e Moore, J. D. (2000). The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human-Computer Studies*, 53(6):1017–1076.

Ferreira, T. C. e Paraboni, I. (2014a). Classification-based referring expression generation. *Lecture Notes in Computer Science*, 8403:481–491.

Ferreira, T. C. e Paraboni, I. (2014b). Referring expression generation: taking speakers' preferences into account. *Lecture Notes in Artificial Intelligence*, 8655:539–546.

Gargett, A., Garoufi, K., Koller, A., e Striegnitz, K. (2010). The GIVE-2 corpus of giving instructions in virtual environments. Em *Proceedings of LREC-2010*.

Gatt, A. e Belz, A. (2007). The attribute selection for GRE challenge: Overview and evaluation results. Em *UCNLG+MT: Language Generation and Machine Translation*.

Gatt, A., Belz, A., e Kow, E. (2008). The TUNA challenge 2008: Overview and evaluation results. Em *Fifth International Natural Language Generation Conference (INLG-2008)*, páginas 198–206.

Gatt, A., Belz, A., e Kow, E. (2009). The TUNA challenge 2009: Overview and evaluation results. Em *Proceedings of the 12nd European Workshop on Natural Language Generation*, páginas 174–182.

Gatt, A., van der Sluis, I., e van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. Em *Proceedings of ENLG-07*.

Guhe, M. (2009). Generating referring expressions with a cognitive model. Em *Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference*.

Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R., Moore, J., e Oberlander, J. (2010). Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). Em *6th International Natural Language Generation Conference (INLG)*, Dublin.

Krahmer, E. e Theune, M. (2002). Efficient context-sensitive generation of referring expressions. Em van Deemter, K. e Kibble, R., editores , *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, páginas 223–264. CSLI Publications, Stanford, CA.

Krahmer, E. e van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Krahmer, E., van Erk, S., e Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

Landis, J. R. e Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

NIST (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.

Papineni, S., Roukos, T., ward, W., e Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. Em *Proceddings of ACL-2002*, páginas 311–318.

Paraboni, I. (1997). Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa. Master's thesis, PUCRS, Porto Alegre.

Paraboni, I. (2000). An algorithm for generating document-deictic references. Em *Procs. of workshop Coherence in Generated Multimedia, associated with First Int. Conf. on Natural Language Generation (INLG-2000), Mitzpe Ramon*, páginas 27–31.

Paraboni, I. (2003). *Generating references in hierarchical domains: the case of Document Deixis*. Tese de Doutorado , University of Brighton.

Paraboni, I., Masthoff, J., e van Deemter, K. (2006). Overspecified reference in hierarchical domains: measuring the benefits for readers. Em *Proc. of INLG-2006*, páginas 55–62, Sydney.

Paraboni, I. e van Deemter, K. (1999). Issues for the generation of document deixis. Em *Procs. of workshop on Deixis, Demonstration and Deictic Belief in Multimedia Contexts, in association with the 11th European Summers School in Logic, Language and Information (esslli99)*, páginas 44–48.

Paraboni, I. e van Deemter, K. (2002a). Generating easy references: the case of document deixis. Em *INLG-2002, New York*, páginas 113–119.

Paraboni, I. e van Deemter, K. (2002b). Towards the generation of document-deictic references. Em *Information sharing: reference and presupposition in language generation and interpretation*, páginas 329–352. CSLI Publications.

Paraboni, I. e van Deemter, K. (2014). Reference and the facilitation of search in spatial domains. *Language, Cognition and Neuroscience*, 29(8):1002–1017.

Paraboni, I., Yamasaki, A. K., da Silva, A. S. R., e Teixeira, C. V. M. (2014). Generating underspecified descriptions of landmark objects. *Lecture Notes in Artificial Intelligence*, 8655:76–83.

Passonneau, R. (2006). Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation. Em *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27(1):98–110.

Pereira, D. B. e Paraboni, I. (2007). A language modelling tool for statistical NLP. Em *5th Workshop on Information and Human Language Technology (TIL-2007). Anais do XXVII Congresso da SBC*, páginas 1679–1688, Rio de Janeiro. Sociedade Brasileira de Computação.

Pereira, D. B. e Paraboni, I. (2008). Statistical surface realisation of portuguese referring expressions. *Advances in Natural Language Processing*, LNAI 5221:383–392.

Reiter, E. e Dale, R. (2000). *Building natural language generation systems*. Cambridge University Press, New York, NY, USA.

Striegnitz, K., Denis, A., Gargett, A., Garoufi, K., Koller, A., e Theune, M. (2011). Report on the second second challenge on generating instructions in virtual environments (GIVE-2.5). Em *Generation Challenges Session at the 13th European Workshop on Natural Language Generation (ENLG)*, páginas 270–279.

Tarenskeen, S., Broersma, M., e Geurts, B. (2014). Referential overspecification: Colour is not that special. Em *RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production*, Edinburgh, Scotland.

Teixeira, C. V. M., Paraboni, I., da Silva, A. S. R., e Yamasaki, A. K. (2014). Generating relational descriptions involving mutual disambiguation. *Lecture Notes in Computer Science*, 8403:492–502.

van Deemter, K., van Gompel, A. G. R., e Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4:166–183.

van Gompel, R., Gatt, A., Krahmer, E., e Deemter, K. V. (2014). Testing computational models of reference generation as models of human language production: The case of size contrast. Em *RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production*, Edinburgh, Scotland.

Viethen, J. e Dale, R. (2006). Algorithms for generating referring expressions: Do they do what people do? Em *Proceedings of INLG-2006*, páginas 63–70.

Viethen, J. e Dale, R. (2011). GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. Em *Proceedings of UCNLG+Eval-2011*, páginas 12–22.

Apêndice A. Representação XML de imagens de estímulo

O fragmento XML abaixo exemplifica a representação de parte da cena *abs1* ilustrada na Fig. 2.

```
<CONTEXT ID="abs1" IMAGE="abs1.png" MAIN-TARGET="e2">

  <ATTRIBUTE-SET ID="e1">
    <ATTRIBUTE NAME="type" VALUE="star" />
    <ATTRIBUTE NAME="colour" VALUE="white" />
    <ATTRIBUTE NAME="hpos" VALUE="centre" />
    <ATTRIBUTE NAME="vpos" VALUE="top" />
  </ATTRIBUTE-SET>

  <ATTRIBUTE-SET ID="e2" FUNCTION="target">
    <ATTRIBUTE NAME="type" VALUE="star" />
    <ATTRIBUTE NAME="colour" VALUE="white" />
    <ATTRIBUTE NAME="left" VALUE="q2" />
    <ATTRIBUTE NAME="near" VALUE="q2" />
    <ATTRIBUTE NAME="hpos" VALUE="right" />
    <ATTRIBUTE NAME="vpos" VALUE="centre" />
  </ATTRIBUTE-SET>

  <ATTRIBUTE-SET ID="q1">
    <ATTRIBUTE NAME="type" VALUE="cube" />
    <ATTRIBUTE NAME="colour" VALUE="red" />
    <ATTRIBUTE NAME="left" VALUE="c2" />
    <ATTRIBUTE NAME="near" VALUE="c2" />
    <ATTRIBUTE NAME="right" VALUE="e2" />
    <ATTRIBUTE NAME="near" VALUE="e2" />
    <ATTRIBUTE NAME="hpos" VALUE="left" />
    <ATTRIBUTE NAME="vpos" VALUE="bottom" />
  </ATTRIBUTE-SET>

  <ATTRIBUTE-SET ID="q2" FUNCTION="landmark">
    <ATTRIBUTE NAME="type" VALUE="cube" />
    <ATTRIBUTE NAME="colour" VALUE="white" />
    <ATTRIBUTE NAME="left" VALUE="c1" />
    <ATTRIBUTE NAME="near" VALUE="c1" />
    <ATTRIBUTE NAME="hpos" VALUE="right" />
    <ATTRIBUTE NAME="vpos" VALUE="centre" />
  </ATTRIBUTE-SET>

  <ATTRIBUTE-SET ID="c1" FUNCTION="second-landmark">
    <ATTRIBUTE NAME="type" VALUE="ball" />
    <ATTRIBUTE NAME="colour" VALUE="white" />
    <ATTRIBUTE NAME="right" VALUE="q2" />
    <ATTRIBUTE NAME="near" VALUE="q2" />
    <ATTRIBUTE NAME="hpos" VALUE="right" />
    <ATTRIBUTE NAME="vpos" VALUE="centre" />
  </ATTRIBUTE-SET>

  ...

</CONTEXT>
```

Apêndice B. Representação XML de expressões de referência

O fragmento XML abaixo exemplifica a representação de quatro expressões de referência produzidas pelo sujeitos nro. 18 do corpus.

```
<TRIAL ID="18" SPEAKER="18" AGE="20" GENDER="m">

  <CONTEXT ID="abs1">
    <ATTRIBUTE-SET TARGET="e2" LANDMARK="q2"
                   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="q2" />
    </ATTRIBUTE-SET>
  </CONTEXT>

  <CONTEXT ID="abs3">
    <ATTRIBUTE-SET TARGET="e2" LANDMARK="q1"
                   STRING=" a estrela à direita 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="right" VALUE="q1" />
    </ATTRIBUTE-SET>
  </CONTEXT>

  <CONTEXT ID="abs5">
    <ATTRIBUTE-SET TARGET="e3" 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>

  <CONTEXT ID="fill1">
    <ATTRIBUTE-SET TARGET="e1" LANDMARK="q1" SECOND-LANDMARK="c3"
                   STRING=" estrela do lado do quadrado vermelho "
                   LENGTH="4" REL-COUNT="1">
      <ATTRIBUTE NAME="type" VALUE="star" />
      <ATTRIBUTE NAME="landmark-type" VALUE="cube" />
      <ATTRIBUTE NAME="second-landmark-colour" VALUE="red" />
      <ATTRIBUTE NAME="near" VALUE="q1" />
    </ATTRIBUTE-SET>
  </CONTEXT>

  ...

</TRIAL>
```