UNIVERSIDADE DE SÃO PAULO

Escola de Artes, Ciências e Humanidades

Relatório Técnico PPgSI-001/2015

O córpus Stars2 de expressões de referência

Ivandré Paraboni, Michelle Reis Galindo, Douglas Iacovelli

Maio - 2015

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 córpus Stars2 de expressões de referência

Ivandré Paraboni[1], **Michelle Reis Galindo**[1], **Douglas Iacovelli**[1]

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

ivandre@usp.br

Resumo. *Este documento descreve o experimento de construção do córpus* Stars2 *de expressões de referência e seu conteúdo. O córpus foi construído para o estudo de diversas questões de pesquisa em geração de língua natural, como a geração de expressões de referência (GER) atômicas e relacionais envolvendo até três objetos por descrição, a questão da variação humana na seleção de conteúdo destas expressões, e a questão da superespecificação de objetos-alvo e pontos de referência com uso de propriedades atômicas e relacionais variadas. O conjunto de dados coletado foi anotado com suas propriedades semânticas e disponibilizado juntamente com as imagens de estímulo do experimento para fins de pesquisa em GER.*

1. Introdução

Sistemas de geração de língua natural (GLN) produzem descrições textuais a partir de uma entrada de dados tipicamente não linguística [Reiter e Dale 2000]. Sistemas deste tipo são utilizados quando é necessária uma maior variação linguística nos documentos gerados, ou para obter maior proximidade em relação ao desempenho humano. Aplicações de GLN incluem, por exemplo, sistemas de diálogo humano-computador, a produção de relatórios a partir de bases de dados (e.g., numéricos), a sumarização de documentos de texto da WEB etc.

A geração de expressões de referência (GER) [Krahmer e van Deemter 2012] é a subtarefa de GLN que produz descrições linguísticas de objetos do discurso, e pode também ser vista como a tarefa computacional 'simétrica' à interpretação de expressões de referência [Paraboni 1997, Cuevas e Paraboni 2008]. 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 (e.g., Pereira e Paraboni [2007, 2008], de Novais e Paraboni [2012]). Este trabalho trata da questão da seleção de conteúdo no microplanejamento.

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].

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, dos Santos Silva e Paraboni 2015]. Neste segundo caso pode-se tirar proveito de conjuntos de dados - ou córpus de GER - produzidos sob condições controladas.

Um córpus de GER tipicamente consiste 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 córpus *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 córpus *Stars2* de descrições sub e superespecificadas - e seus resultados preliminares.

2. Trabalho Relacionado

Um dos primeiros exemplos de conjunto de dados de uso público para GER é o córpus *Drawer* em Viethen e Dale [2006]. Este córpus apresenta uma única cena - um armário com 16 gavetas de cores variadas - e contém 140 descrições geradas por 20 participantes. O objetivo do estudo foi examinar a questão da variação humana na produção de expressões de referência.

Um exemplo mais representativo de córpus de GER é o córpus *TUNA* em Gatt et al. [2007]. TUNA contém expressões coletadas para o estudo de fenômenos e algoritmos de GER em uma série de competições [Gatt e Belz 2007, Gatt et al. 2008, 2009]. O córpus *TUNA* descreve situações de referência a peças de Mobília (Furniture) e fotos de Pessoas (People) e contem um total de 2280 expressões geradas por 50 participantes em língua inglesa. Nos experimentos *TUNA* o propósito da referência era sempre a identificação do objeto-alvo, e todas descrições são do tipo atômico.

Os córpus *GRE3D3* e *GRE3D7* em Dale e Viethen [2009], Viethen e Dale [2011] tratam do uso de relações espaciais em um contexto visual tridimensional simplificado. O córpus contém 4480 descrições de objetos geométricos produzidas por 294 participantes em uma série de experimentos on-line.

Experimentos deste tipo são também úteis à construção de córpus de diálogos [Eugenio et al. 2000, Guhe 2009]. O córpus *iMap* [Guhe 2009] de instruções de rota em mapas consiste de 256 diálogos nos quais uma dupla de participantes revezava-se nos papeis de instrutor e receptor das instruções. A tarefa do instrutor consistia em descrever um caminho de modo que o receptor pudesse desenhá-lo em seu próprio mapa. O conjunto de dados formado pelo córpus *iMap* não é entretanto disponíveis publicamente por razões de confidencialidade.

Finalmente, o córpus *GIVE-2* [Gargett et al. 2010] de instruções de navegação em mundos virtuais interativos tridimensionais foi construído como parte do projeto GIVE [Byron et al. 2007] e de uma série de competições de sistemas deste tipo [Byron et al. 2009, Koller et al. 2010, Striegnitz et al. 2011]. O córpus foi criado por meio de experimentos envolvendo 36 duplas de participantes de língua inglesa e alemã, que se alternaram nos papeis de instrutor e jogador. O conjunto de dados resultante (instruções, passos de navegação etc.) pode ser visualizado na forma de animação com uso da ferramenta *Replay* em Gargett et al. [2010].

3. Trabalho atual

Um experimento-piloto anterior a este trabalho resultou no córpus *Stars* de expressões de referência detalhado em Paraboni et al. [2014]. Apesar da semelhança de nomes com o presente córpus Stars2, os dois córpus são distintos em termos de objetivos, condições experimentais, sujeitos, materiais empregados e metodologia. Outras diferenças entre estes dois córpus são apresentadas na Seção 4.

O foco do presente experimento *Stars2* é a coleta de expressões de referência atômicas e relacionais envolvendo até três referentes em um domínio visual simples. As condições experimentais consideradas privilegiam 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.

A Fig.1 ilustra as condições 01..04 do experimento em que o uso de uma descrição relacional se faz necessário.

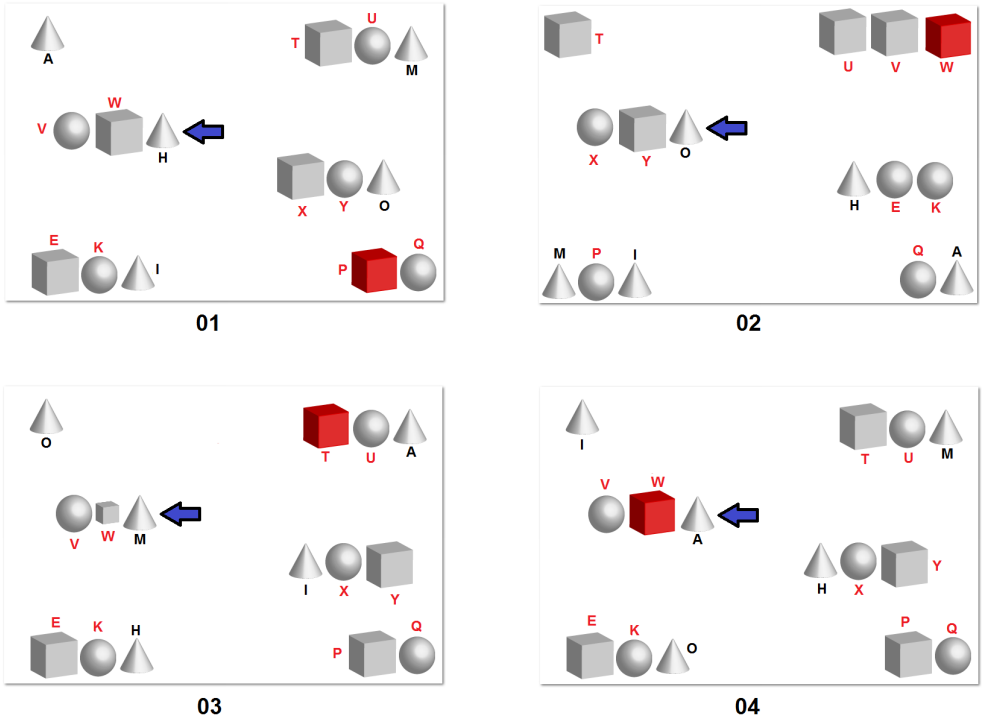

Figura 1. Situações de referência com uso de descrições relacionais

Nas condições 01..04, o objeto-alvo (no exemplo, um cone) compartilha suas principais propriedades atômicas (e.g., tipo, cor e tamanho) com diversos distraidores presentes no contexto. Por este motivo, e partindo-se do pressuposto de que não é possível fazer uso de atributos espaciais (e.g., referenciando-se posições da tela) como estipulado no experimento (cf. Seção 3.2), faz-se necessário incluir na descrição deste objeto uma relação com o ponto de referência mais próximo (nos exemplos 01..04, um cubo). Entretanto, como o objeto-alvo neste exemplo é o único cone ao lado de um cubo na cena inteira, uma descrição relacional mutuamente exclusiva [Teixeira et al. 2014] do tipo 'o cone ao

lado do cubo' sempre é possível, muito embora seus componentes individuais (i.e., as porções 'cone' e 'cubo') sejam, se consideradas de forma isolada, ambíguas.

Descrições relacionais mínimas como a deste exemplo representam a saída típica de diversos algoritmos de GER capazes de tratar propriedades relacionais (e.g., Dale e Haddock [1991], Krahmer et al. [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.

Ao superespecificar a descrição de um ponto de referência, diversas propriedades podem ser selecionadas. Por este motivo, 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, o ponto de referência (i.e., o cubo mais próximo do alvo) possui uma propriedade altamente discriminatória que, embora desnecessária para desambiguação, pode ser selecionada caso o locutor deseje superespecificá-lo. 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. A Fig.2 ilustra estas condições.

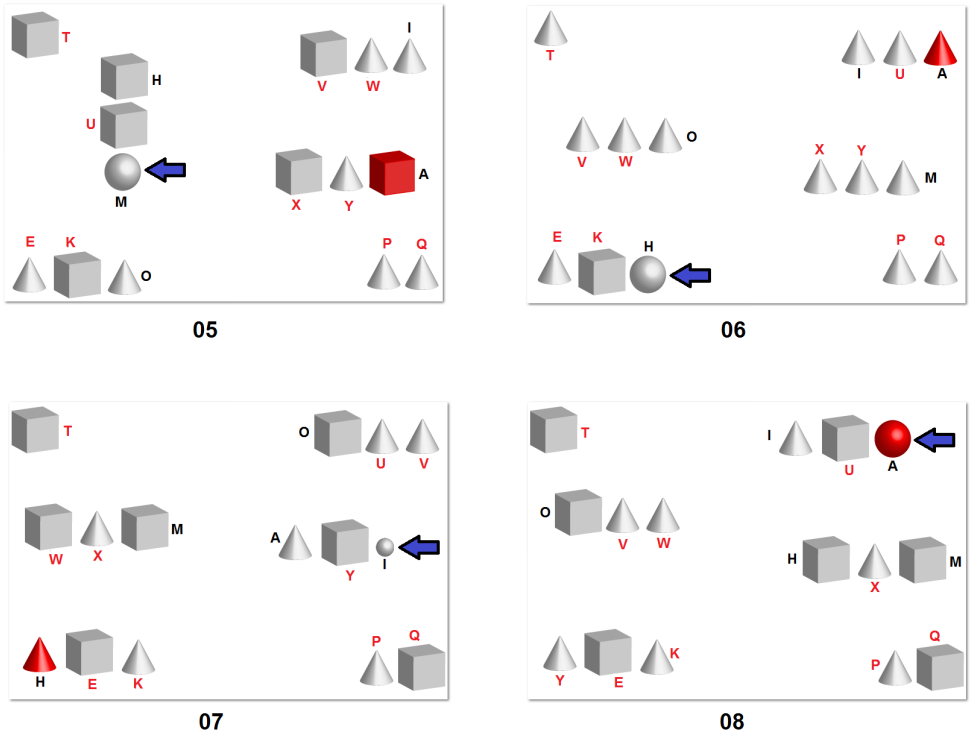

Figura 2. Situações de referência com uso de descrições atômicas

Nas condições 05..08, o objeto-alvo (no exemplo, uma esfera) é o único objeto do seu tipo na cena. Por este motivo, nestas condições é sempre possível identificar o alvo com

uso de uma descrição atômica mínima formada apenas pelo tipo do objeto, como em 'a esfera'. Entretanto, em cada uma destas condições, o objeto-alvo possui uma propriedade altamente discriminatória que, embora desnecessária para desambiguação, pode ser selecionada caso o locutor deseje superespecificá-lo. Esta propriedade corresponde a uma relação espacial única como *below-cube* na condição 05, a uma relação espacial de proximidade com um ponto de referência único como *near-cube* em 06, à propriedade atômica representando tamanho como *size-small* em 07, e à propriedade atômica representando cor como *colour-red* em 08.

As oito condições do experimento foram projetadas para a investigação de diversas questões relacionadas à superespecificação de descrições de objetos-alvo e pontos de referência com uso de diferentes tipos de propriedades. A comparação entre estas condições está fora do escopo do presente trabalho e será tratada à parte.

Para fins de coleta de dados, consideramos neste trabalho apenas a questão do uso de propriedades redundantes na descrição de pontos de referência e objetos-alvo representadas por duas hipóteses básicas, as quais são detalhadas como segue. Em primeiro lugar, consideramos que o uso de propriedades superespecificadas na descrição de pontos de referência é uma estratégia de reparo de descrições de difícil identificação (cf. Paraboni e van Deemter [2014]), e é possível assim que esta forma de superespecificação seja altamente frequente:

> *h1*: *Descrições relacionais superespecificadas são mais frequentes do que descrições mínimas.*

A hipótese $h1$ será testada comparando-se o número médio de descrições relacionais mínimas com o número médio de descrições relacionais superespecificadas nas condições 01..04.

Em segundo lugar, consideramos que o uso de propriedades superespecificadas na descrição de objetos-alvo é menos crítico, e que por isso esta forma de superespecificação seja menos frequente:

> *h2*: *Descrições atômicas superespecificadas são menos frequentes do que descrições mínimas.*

A hipótese $h2$ será testada comparando-se o número médio de descrições atômicas mínimas com o número médio de descrições atômicas superespecificadas nas condições 05..08.

3.1. Sujeitos

56 voluntários, estudantes de Sistemas de Informação da USP-EACH, que responderam a um convite enviado por email e redes sociais. Os participantes tinham em média 20 anos de idade e eram predominantemente do sexo masculino (91%). Todos participantes eram brasileiros nativos e tinham visão normal ou corrigida. O convite deixou explícito que os participantes trabalhariam em duplas. A maioria dos participantes escolheu o próprio parceiro, mas um pequeno número de indivíduos sem dupla teve seu parceiro designado no momento da realização do experimento. Um pequeno prêmio - um livro de Inteligência Artificial - foi sorteado entre os participantes ao final da série de experimentos. Alguns participantes não manifestaram interesse no sorteio e atuaram apenas como voluntários.

3.2. Procedimento

O experimento 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.

Antes do experimento propriamente dito, cada dupla recebeu instruções sobre a tarefa a ser desempenhada e um período de prática. Esta prática consistia da apresentação de uma série de imagens de estímulo semelhantes às usadas no experimento, que eram então descritas pelos participantes simulando a tarefa a ser realizada no experimento. Uma vez que ambos os sujeitos estavam familiarizados com o ambiente do experimento, eram informados de que o experimento real seria iniciado.

Todos os sujeitos participaram do experimento duas vezes, alternando os papeis de locutor e ouvinte. O papel inicial era designado de forma aleatória. Ao término de uma sessão, os papeis eram invertidos e um segundo experimento era iniciados imediatamente. Como diversos aspectos da apresentação dos estímulos são definidos de forma aleatória ao longo a execução do experimento (cf. Seção 3.3), 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. 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. 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 de cada objeto candidato eram visíveis apenas na tela do ouvinte. A Fig.3 ilustra estas duas telas.

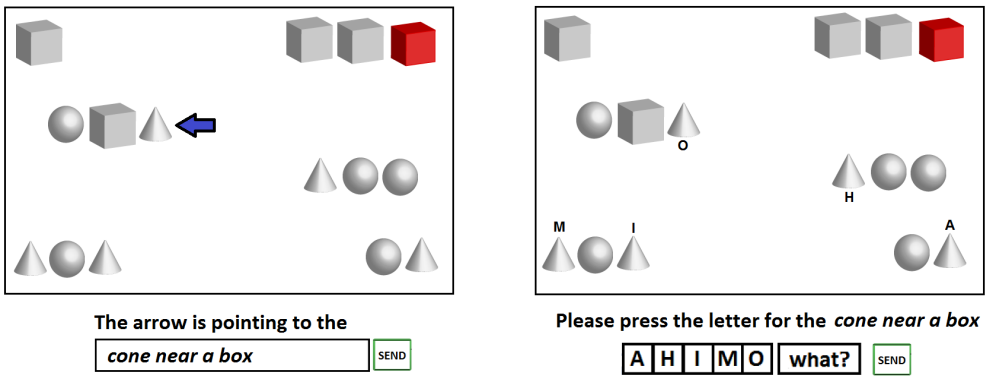

Figura 3. Telas do locutor (esquerda) e ouvinte (direita).

A única ação possível para o locutor era enviar descrições textuais ao ouvinte. Os locutores eram instruídos a descrever cada objeto-alvo em função de suas propriedades e

das de outros objetos da cena, mas evitando referência à posições da tela (e.g., 'o cone no canto esquerdo') sob a justificativa de que a disposição dos objetos na tela do ouvinte poderia ser diferente (embora isso não tenha ocorrido). Certas palavras ilegais eram filtradas automaticamente pela interface do experimento (e.g., 'canto', 'tela' etc.) e uma mensagem de erro era exibida ao locutor solicitando uma nova tentativa.

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, uma mensagem era enviada ao locutor.

As condições do experimento que produziram respostas incorretas ou pedidos de esclarecimento permaneciam na fila para reapresentação posterior. Conforme será discutido na Seção 3.3, isso geralmente não envolvia reapresentar a mesma imagem, embora isso possa ter ocorrido em algumas poucas ocasiões (i.e., se a seleção aleatória escolher exatamente o mesmo modo de apresentação, o mesmo padrão de objetos e a mesma orientação duas vezes para aquela mesma condição que teve de ser repetida).

3.3. Materiais

O experimento consistia de oito imagens (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 monotonia da tarefa.

Cada imagem exibia exatamente 15 objetos (cones, cubos e esferas). Os objetos eram organizados em seis grupos: um objeto isolado, dois pares e quatro triplas. Os grupos eram distribuídos aproximadamente em três camadas conforme ilustrado nos exemplos anteriores. Como forma de ocultar o papel do atributo cor nas condições 04 e 08, todas imagens exibiam exatamente um objeto vermelho. Todos os demais objetos eram de cor cinza.

Entre uma condição e outra (e também em uma mesma condição no caso de repetição ou erro de interpretação), 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, a ser discutida a seguir. Não houve necessidade de rotular todos os 15 objetos da cena porque a maioria deles era de tipos diferentes do objeto-alvo, o que tornava sua seleção pouco provável.

Conforme ilustrado na Fig.1 anterior, as condições 01..04 faziam uso de exatamente cinco objetos de cada tipo. O alvo era apresentado sempre na camada do meio. Entretanto, uma vez que a imagem podia ser exibida em orientação normal ou reversa, o objeto-alvo podia aparecer em lados diferentes da tela. Na tela do ouvinte, os rótulos eram atribuídos aleatoriamente ao alvo e aos quatro distraidores do mesmo tipo.

Conforme ilustrado na Fig.2 anterior, as condições 05..08 apresentavam o objeto-alvo como único objeto daquele tipo, ao lado de um ponto de referência de um segundo tipo, e ainda outro ponto de referência do terceiro tipo. A exceção era a condição 06, em que o ponto de referência também era de um tipo único e assim todos os outros treze objetos na cena eram do terceiro tipo. Esta condição foi utilizada como *filler* em experimentos realizados a partir destes dados. O objeto-alvo era exibido em camadas diferentes da tela para cada condição (não ilustrado nos exemplos). Na tela do ouvinte, os rótulos eram atribuídos aleatoriamente ao alvo e quatro distraidores também escolhidos de forma aleatória.

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 tal qual ilustrado na Fig.4. Embora este exemplo apresente duas cenas praticamente idênticas (exceto pela sobreposição), lembramos que no experimento real a forma de apresentação seria modificada de forma aleatória conforme discutido anteriormente.

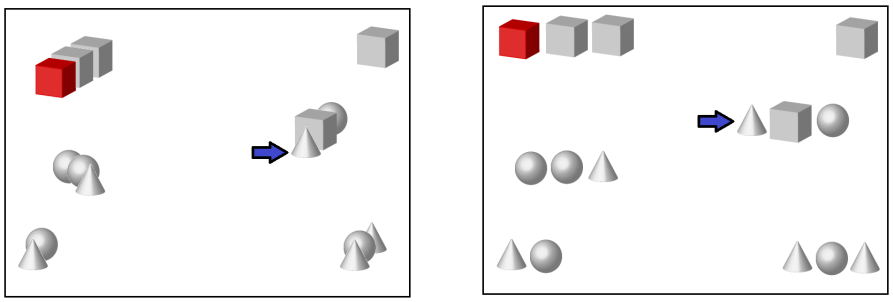

Figura 4. Uma condição experimental com e sem sobreposição de objetos

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.

3.4. Coleta e transcrição

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.

A anotação das expressões de referência coletadas cobriu apenas o grupo dos três objetos composto pelo objeto-alvo e seus dois pontos de referência. Nos raros casos em

que uma expressão relaciona o objeto-alvo e o segundo ponto de referência diretamente (e.g., 'a esfera ao lado do cubo', omitindo o fato de que havia um cone entre eles) os atributos do ponto de referência foram anotados como pertencentes ao segundo ponto de referência, ou seja, como se houvesse um primeiro ponto de referência oculto representado por uma descrição vazia. Referências a objetos distantes do grupo de interesse (e.g., 'a esfera do lado oposto ao grupo de cones') foram infrequentes, e não foram anotadas de modo a manter a simplicidade do esquema de anotação. Referências espaciais implícitas (e.g., 'cone esfera') foram anotadas como contendo uma relação de proximidade (ou seja, foram interpretadas como sendo 'o cone próximo à esfera'). Esta medida teve o objetivo de facilitar a manipulação dos dados por algoritmos de GER.

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	{cube, ball, cone}
colour	{red, grey}
size	{small, regular}
near	(id de objeto)
left	(id de objeto)
right	(id de objeto)
below	(id de objeto)
in-front-of	(id de objeto)

Assim como no córpus Stars [Paraboni et al. 2014], estes nomes de atributos são usados como especificado na tabela apenas para o caso do objeto-alvo. Para pontos de referência, estes nomes são antecedidos pelos identificadores 'LANDMARK-' ou 'SECOND-LANDMARK' conforme pertinente. Por exemplo, o atributo 'type' do segundo ponto de referência da descrição é designado pelo atributo 'SECOND-LAND MARK-TYPE'. Esta diferenciação de nomenclatura objetiva facilitar o cálculo automático de coeficientes Dice [Dice 1945] e outras tarefas de GER.

Dada a simplicidade do domínio, a transcrição não exigiu concordância entre juízes [Landis e Koch 1977]. Tanto imagens como descrições foram representados em formato XML, em uma versão simplificada do formato utilizado no córpus TUNA [Gatt et al. 2007].

Na representação atual, cada locutor do experimento é representado por um nó do tipo TRIAL contendo um identificador sequencial, um código identificador do sujeito, informações de faixa etária e gênero, um identificador da instância do experimento realizado, e um código identificador do ouvinte. O campo TURN indica se os dados foram coletados no primeiro (1) ou segundo (2) turno de cada dupla, ou seja, os dados para os quais TURN=1 são sempre os que foram coletados na primeira execução (i.e., antes da inversão de papeis).

Cada TRIAL é composto de 16 nós CONTEXT, cada qual representando uma condição doe experimento. O atributo ID representa a identificação completa da cena de estímulo na

forma NNM-tPD, onde NN é o número da condição (01..08), M é o modo de apresentação (f=flat e o=overlap, cf. Fig.4), P é o padrão de distribuição de objetos utilizado (1 para o padrão cone-cubo-esfera, e 2 para o padrão cubo-cone-esfera), e D é a direção da imagem (n=normal e r=reversa). As informações sobre a condição (COND), sobreposição (OVERLAP) e direção (DIR) aparecem também de forma redundante como atributos deste nó, bem como a ordem sequencial em que o estímulo foi exibido (SEQ) e o nome do arquivo de imagem utilizado (IMAGE).

Objetos e expressões de referência são representados por um nó ATTRIBUTE-SET contendo uma lista de atributos identificados por nome (NAME) e valor (VALUE). O nó ATTRIBUTE-SET apresenta ainda o identificador do objeto-alvo (TARGET) e ponto de referência mais próximo (LANDMARK), a expressão textual coletada (STRING), o seu tamanho em número de propriedades anotadas (LENGTH) e o número de propriedades relacionais anotadas (REL-COUNT). Estes dois últimos dados podem também ser inferidos a partir da própria lista de atributos subordinada a este nó.

O exemplo a seguir ilustra a representação de um objeto do tipo cone em um fragmento de TRIAL exibindo apenas um dos 16 contextos existentes.

```
<TRIAL ID="1"  SPEAKER="832" AGE="19" GENDER="m" EXP="g2k"
               HEARER="455" TURN="1" >

  <CONTEXT     ID="04f-t2r" COND="4" OVERLAP="f" DIR="r"
               SEQ="7" IMAGE="type2/04f-t2r">

    <ATTRIBUTE-SET TARGET="a" LANDMARK="w"
               STRING="the cube next to the red cone"
               LENGTH="4" REL-COUNT="1">

      <ATTRIBUTE NAME="type" VALUE="cube" />
      <ATTRIBUTE NAME="near" VALUE="w" />
      <ATTRIBUTE NAME="landmark-type" VALUE="cone" />
      <ATTRIBUTE NAME="landmark-colour" VALUE="red" />

    </ATTRIBUTE-SET>

  </CONTEXT>

...

</TRIAL>
```

A representação dos contextos é similar, ou seja, cada contexto é formado por uma lista de nós ATTRIBUTE-SET correspondendo aos objetos presentes em cada cena. Os 64 contextos possíveis são relacionados em um arquivo único (Stars2-context.xml).

4. Resultados

4.1. Visão geral

A Tabela 2 apresenta estatísticas gerais do córpus coletado e sua comparação com córpus de GER existentes. No caso do córpus GRE3D3, foi considerada a divisão do atributo *pos* em dois (i.e., *hpos* e *vpos*), resultando assim em 9 atributos possíveis, e não 8 como originalmente anotado em Dale e Viethen [2009].

Tabela 2. Comparação com córpus de GER existentes

	Domínio		Descrições					
córpus	Atrib.	Rel.	Tam.	Uso	Atôm.	1-Rel	2-Rel	Sup.
TUNA-Furniture(sing.)	4	0	3.1	0.8	1.00	0.00	0.00	0.88
TUNA-People(sing.)	10	0	3.1	0.3	1.00	0.00	0.00	0.95
GRE3D3	9	1	3.4	0.3	0.64	0.36	0.00	0.80
GRE3D7	6	1	3.0	0.4	0.87	0.13	0.00	0.46
Stars	8	2	4.4	0.4	0.07	0.68	0.25	0.58
Stars2	9	2	3.3	0.3	0.38	0.32	0.31	0.54

Os dados apresentados na Tabela 2 representam o número de atributos atômicos e propriedades relacionais existentes em cada domínio, e informações sobre o tipo de descrição encontrada: o tamanho médio da descrição em número de propriedades anotadas, a proporção de uso das propriedades possíveis em cada descrição (onde valores menores indicam maior complexidade, ou seja, mais opções de seleção de atributos), a proporção de descrições contendo 0 (i.e., atômicas), 1 ou 2 relações, e a proporção de superespecificação da descrição do objeto-alvo (i.e., sem considerar-se as descrições de pontos de referências a ele relacionados).

Observa-se por estes dados que o domínio Stars2 possui maior número de propriedades possíveis (mesmo considerando-se o desdobramento do atributo *pos* no córpus GRE3D3), uma proporção de uso destas propriedades baixa (semelhante aos domínios TUNA-People e GRE3D3), e um equilíbrio entre descrições contendo 0, 1 e 2 relações. A comparação mostra também que o domínio TUNA-Furniture possui uma proporção de uso de atributos muito alta (0,8), o que indica que a tarefa de GER é consideravelmente mais simples.

4.2. Superespecificação

A Tabela 3 mostra os resultados para a hipótese $h1$.

Analisando-se as descrições de pontos de referências nas condições 01..04 (onde o uso de uma propriedade relacional era sempre necessário para desambiguação), observa-se

Tabela 3. Descrições de pontos de referência superespecificadas vs. mínimas

Condição	01		02		03		04		Média	
	sup.	mín.	sup.	mín.	sup.	mín.	sup.	mín.	sup.	min.
média	1.20	0.70	0.95	1.02	1.93	0.05	1.82	0.18	1.47	0.49
desv.	0.86	0.85	0.90	0.90	0.26	0.23	0.51	0.51	0.45	0.47

Tabela 4. Descrições atômicas superespecificadas vs. mínimas

Condição	05		06		07		08		Média	
	sup.	mín.	sup.	mín.	sup.	mín.	sup.	mín.	sup.	min.
média	0.84	1.14	0.52	1.46	1.00	1.00	1.38	0.61	0.93	1.05
desv.	0.85	0.84	0.81	0.81	0.93	0.93	0.84	0.85	0.65	0.65

que o número de descrições superespecificadas é, na média geral, superior ao número de descrições mínimas. A diferença é altamente significativa (F(1,55)=65.63, MSE=0.4153, p<0.0001). Isso confirma a hipótese $h1$.

A Tabela 4 mostra os resultados para a hipótese $h2$.

Analisando-se as descrições atômicas nas condições 05..08 (onde o uso de uma propriedade relacional não era necessário para desambiguação), observa-se que a diferença entre o número de descrições superespecificadas e mínimas não é significativa (F(1,55)=0.48, MSE=0.8489, p=0.491338). A hipótese $h2$ não foi assim confirmada.

4.3. Geração de expressões de referência

Finalmente, foi analisado também o comportamento de um algoritmo de GER padrão na geração das descrições do córpus *Stars2*. Para este fim, foi utilizada uma versão relacional do algoritmo Incremental [Dale e Reiter 1995] com controle de referência circular, e levando em conta uma lista preferencial P ordenada com base nas frequências observadas no córpus:

P = <*type*, *colour*, *size*, *near*, *in-front-of*, *right*, *left*, *below*, *above*, *behind* >

A versão padrão do algoritmo - aqui denominada *Increm* - será utilizada como sistema de *baseline*. Esta opção seleciona propriedades discriminatórias em P até obter uma descrição livre de ambiguidade. Um segundo sistema de *baseline* aqui denominado *Random* segue a mesma estratégia, porém as propriedades em P são examinadas em ordem aleatória. O objetivo deste *baseline* é simplesmente o de ilustrar um limite inferior para a tarefa.

Além destas dos dois sistemas de *baseline*, foi considerada ainda uma variação - aqui denominada *Overspec* - em que uma propriedade superespecificada é adicionada à descrição do ponto de referência caso esta não esteja completamente especificada. A propriedade a ser incluída é a de maior poder discriminatório dentre as ainda disponíveis para seleção em P (ou seja, dentre as que ainda não foram incluídas na descrição). No caso de descrições atômicas, nenhuma superespecificação é realizada.

Os resultados dos algoritmos básicos e da versão superespecificada são apresentados na Tabela 5 considerando-se o número de acertos global (Accuracy) e os coeficientes de Dice [Dice 1945] e MASI [Passonneau 2006][1].

[1]Estas métricas são as mais comuns para avaliação da tarefa de seleção de conteúdo. Por outro lado, a realização superficial seria tipicamente avaliada com uso das métricas BLEU [Papineni et al. 2002] e NIST [NIST 2002].

Tabela 5. Geração de expressões do córpus Stars2

	Accuracy		Dice		MASI	
Algoritmo	mean	sd	mean	sd	mean	sd
Random	0.28	0.45	0.68	0.25	0.43	0.38
Increm	0.31	0.46	0.70	0.25	0.46	0.39
Overspec	0.36	0.48	0.73	0.25	0.49	0.38

Ressaltamos que os resultados desta análise são meramente ilustrativos, e podem servir de referência para a futura avaliação de soluções completas de GER que façam uso dos dados do córpus *Stars2*.

5. Conclusão

Este documento descreveu o experimento de construção do córpus *Stars2* de expressões de referência e sua estrutura. A disponibilização do córpus *Stars2* objetiva o reaproveitamento destes dados para o estudo e desenvolvimento de soluções de GER para situações complexas de referência, especialmente no que diz respeito ao uso de expressões relacionando até três objetos e com diversas estratégias de superespecificação.

O presente conjunto de dados foi parcialmente utilizado em Ferreira e Paraboni [2014b]. O estudo enfocou a geração de expressões de referência considerando a variação humana (i.e., envolvendo características dependentes do locutor, como sua lista de preferência individual, informações de gênero e faixa etária etc.). Outros estudos envolvendo as questões da variação humana e de estratégias de superespecificação encontram-se em andamento.

Agradecimentos

Este trabalho contou com apoio FAPESP e da Universidade de São Paulo. Os autores são gratos também aos voluntários que participaram do experimento.

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.

dos Santos Silva, D. e Paraboni, I. (2015). Generating spatial referring expressions in interactive 3D worlds. *Spatial Cognition and Computation*.

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.