UNIVERSIDADE DE SÃO PAULO

Escola de Artes, Ciências e Humanidades

Relatório Técnico PPgSI-001/2017

O córpus b5 de textos e inventários de personalidade (v. 1.0.)

Ricelli M. S. Ramos, Georges B. Stavracas Neto,
Bárbara B. C. da Silva, Danielle S. Monteiro,
Fabio B. F. Lopes, Vitor G. dos Santos, Ivandré Paraboni

Fevereiro - 2017

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 b5 de textos e inventários de personalidade (v. 1.0.)

Ricelli M. S. Ramos[1] **, Georges B. Stavracas Neto**[1] **, Bárbara B. C. da Silva**[1] **,**
Danielle S. Monteiro[1] **, Fabio B. F. Lopes**[1] **, Vitor G. dos Santos**[1] **,**
Ivandré Paraboni[1]

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

{ivandre}@usp.br

Resumo. *Este relatório descreve o experimento de coleta de dados e o conteúdo que constitui a versão 1.0. do córpus* $b5$ *de textos e inventários de personalidade de seus autores.*

1. Introdução

O córpus $b5$ é um conjunto de dados paralelo contendo textos e os inventários de personalidade de seus respectivos autores. O objetivo do córpus é oferecer suporte a diversos tipos de estudo de tratamento computacional da personalidade humana a partir de textos, tanto para reconhecimento de traços de personalidade a partir destes (uma tarefa típica de Processamento de Língua Natural - PLN) como para geração de texto com base em um conjunto de traços de personalidade de interesse (uma tarefa de Geração de Língua Natural - GLN). Todos os dados foram coletados com consentimento explícito dos participantes, e mediante aprovação da comissão da ética em pesquisa desta instituição.

Os inventários de personalidade considerados são baseados no modelo dos Cinco Grandes Fatores ou CGF [John et al. 2008], fazendo uso de um inventário validado para o Português Brasileiro [de Andrade 2008]. Os inventários foram respondidos de forma autônoma pelos próprios participantes, em um experimento presencial ou, majoritariamente, por meio de um aplicativo Facebook. O conjunto de inventários é representado por uma tabela *b5-subject* descrita na seção 4.

Os textos produzidos pelos participantes são divididos em duas categorias: texto livre, obtidos a partir das atualizações de status na rede social Facebook do participante, e texto controlado, obtido a partir de um experimento presencial a partir de estímulos pré-definidos extraídos das bases Face Place [Righi et al. 2012] e GAPED [Dan-Glauser e Scherer 2011]. Uma visão geral desta organização é ilustrada na Fig. 1.

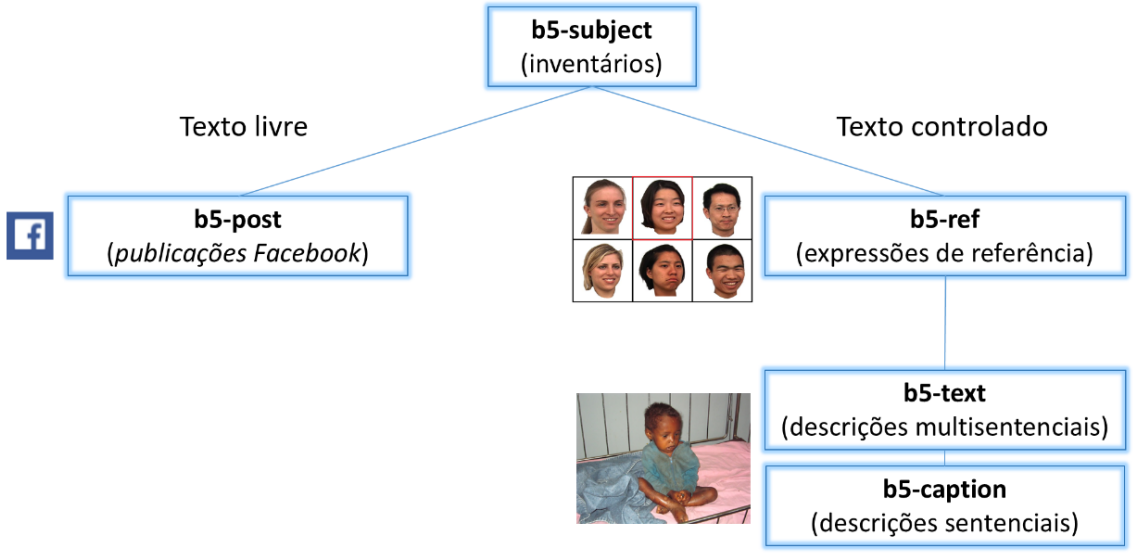

Figura 1. As cinco bases de dados do córpus b5.

Os textos livres constituem o subcórpus *b5-post* descrito na seção 5. Além deste, o experimento controlado foi dividido em três subtarefas, resultando assim em três subcórpus de texto controlado: o subcórpus *b5-ref* de expressões de referência detalhado na seção 6, e os subcórpus *b5-text b5-caption* de descrições multisentenciais e descrições sentenciais, ambos discutidos na seção 7. Diferentes participantes completaram tarefas distintas, de modo que cada subcórpus pode contemplar um subconjunto de sujeitos que não necessariamente aparecem nos outros conjuntos.

2. Coleta de dados

O córpus $b5$ foi coletado de forma descentralizada, em uma série de iniciativas paralelas de experimentos presenciais e divulgação do aplicativo Facebook para obtenção de um conjunto de inventários de personalidade com base no modelo *CGF* [John et al. 1991, 2008], e de textos produzidos em formato livre e controlado. Esta etapa contemplou quatro atividades principais: (1a) o desenvolvimento de ferramentas de coleta de dados; (1b) a aplicação do inventário de personalidade a um grupo de participantes; (1c) a coleta de textos produzidos por estes mesmos participantes de forma livre (i.e., a partir da rede social Facebook); (1d) a coleta de textos de forma controlada (como resposta a estímulos de um experimento presencial).

As ferramentas relacionadas à atividade (1a) consistem de um aplicativo para a rede social Facebook e uma interface de experimento presencial. O aplicativo Facebook permite a resposta ao inventário de personalidade e faz a coleta simultânea das publicações de cada sujeito (mediante autorização prévia). Este aplicativo, que segue uma metodologia semelhante à discutida em Schwartz et al. [2013], foi utilizado tanto na aplicação do inventário de personalidade (1b) como na coleta de textos de produção livre (atividade 1c). Para participantes que não eram usuários Facebook, foi disponibilizada também uma versão off-line do inventário, embora obviamente neste caso não houve coleta de textos provenientes da rede social. Ao invés disso, o inventário era seguido do experimento de coleta de textos de modo controlado (atividade 1d).

O inventário de personalidade utilizado foi o *IGFP-5*, que foi validado para o Português brasileiro em de Andrade [2008]. Assim como em vários dos trabalhos mais influentes desta área como Argamon et al. [2005], Oberlander e Nowson [2006], Mairesse et al. [2007], foi utilizado o método de autoavaliação da personalidade, ou seja, utilizando-se questionários respondidos pelo próprio sujeito avaliado. Embora estudos como Mairesse et al. [2007] indiquem que resultados mais precisos podem ser obtidos com o emprego direto de especialistas humanos (i.e., psicólogos) na tarefa, o custo de uma abordagem deste tipo seria excessivo e possivelmente não justificável face ao caráter exploratório deste projeto. Além disso, se resultados satisfatórios puderem ser obtidos a partir de dados de autoavaliação, é razoável supor que estes resultados sejam ainda melhores se/quando os modelos propostos puderem ser recriados com dados de avaliações produzidas por especialistas.

Assim como em Schwartz et al. [2013], considera-se o uso de texto livre proveniente da rede social Facebook. No entanto, observamos que no caso específico da pesquisa em GLN é necessário conhecer não apenas exemplos de produção da língua, mas também os estímulo iniciais que os motivaram. Por este motivo, a base textual construída contemplou também três tipos de texto produzido sob condições controladas. Estes textos foram produzidos por participantes de um experimento presencial em resposta a determinados estímulos visuais de interesse, e estão relacionados a três tarefas de produção de língua natural: a identificação de entidades visuais, a descrição livre e multi-sentencial de imagens; e a produção de legendas descritivas na forma mono-sentencial das mesmas imagens. Estas tarefas são descritas com mais detalhes nas respectivas seções que tratam de cada subcórpus.

O tempo estimado para resposta ao inventário de personalidade e coleta simultânea de publicações Facebook foi de cerca de 10 minutos. O experimento presencial tomou cerca de 40-60 minutos para ser completado, descontando-se o tempo do inventário. As permissões necessárias para realização dos dois tipos de coleta de dados foram concedidas pelo Comitê de Ética em Pesquisa desta instituição, e são acompanhadas de um termo de concordância aceito pelos sujeitos antes da participação, tanto em modo off-line como via Facebook. Os textos coletados - tanto em modo livre como controlado - passaram por correção ortográfica e diversos outros tipos de pré-processamento. Este procedimento é discutido em Paraboni [2016].

3. Visão geral

O córpus $b5$ contém 1082 inventários de personalidade do tipo *IGFP-5* [de Andrade 2008] preenchidos via Facebook ou de forma presencial. Para cada uma das bases textuais coletadas - *post*, *text*, *caption* e *ref* - a Tabela 1 ilustra o número de sujeitos, sentenças (ou atualizações de status, no caso de *post*), itens (palavras, símbolos de pontuação), tokens e tokens após normalização (e.g., tratamento de nomes próprios, gírias, estrangeirismos, palavras desconhecidas etc., com transformação em símbolos especiais como $name$ etc., cf. Paraboni [2016]). A diferença entre tokens e tokens normalizados só é expressiva no caso de texto livre proveniente da rede social Facebook (*post*) uma vez que nas modalidades de texto controlado a variação entre a forma original e normalizada tende a ser mínima, já que nesta modalidade praticamente não há ocorrências de palavras fora do vocabulário.

Tabela 1. Textos coletados

Base	Sujeitos	Sentenças	Itens	Tokens	Tokens(n)
post	1039	194400	2219622	866274	714158
text	151	1510	84463	37210	37005
caption	151	1510	4896	4121	4116
ref (strings)	152	4558	64518	18700	18666

Quanto à base *ref*, cabe ressaltar que os dados apresentados na Tabela 1 são baseados no conjunto de strings coletados, e não na semântica destas expressões, que é o verdadeiro foco da coleta destes dados específicos (cf. Seção 6). Assim, os dados são apresentados para fins meramente ilustrativos, e incluem não apenas os strings produzidos na tarefa de identificação em si, mas também os strings coletados na fase de prática da tarefa e os strings produzidos em resposta aos estímulos do tipo *filler*, que no presente caso eram formados a partir de Greebles [Gauthier e Tarr 1997].

4. A base de inventários de personalidade b5-subject

A base *b5-subject* contém os inventários de personalidade e informações adicionais sobre os sujeitos do córpus. Esta base é disponibilizada na forma de um arquivo único *subject.csv* em formato anonimizado, em que cada sujeito é representado por um identificador numérico. A Figura 2 ilustra a distribuição de personalidade do conjunto completo dos 1082 sujeitos do córpus.

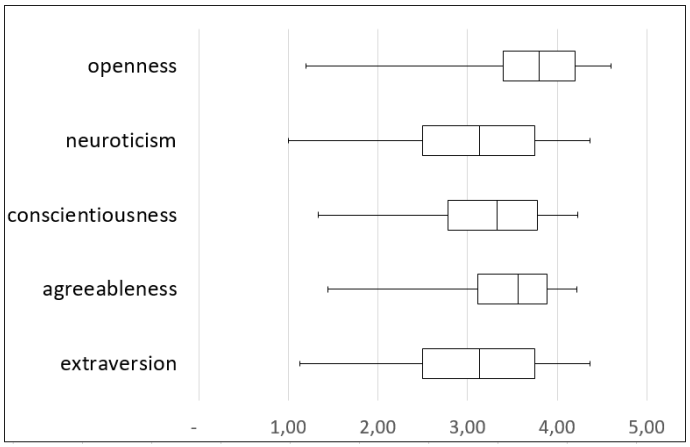

Figura 2. Distribuição de personalidade (1082 sujeitos).

É importante observar que os valores escalares das dimensões de personalidade não devem ser interpretados de forma absoluta, mas sim relativos ao grupo (ou subgrupo) de sujeitos analisados. Por exemplo, não é correto afirmar que um indivíduo com valor de Extroversão abaixo de x é necessariamente introvertido. O correto seria afirmar que, dado dois graus de Extroversão $x1$ e $x2$ tais que $x1 < x2$, um indivíduo de grau de Extroversão $x1$ é menos extrovertido do que um indivíduo de grau $x2$.

Além das informações de personalidade, a tabela *subject.csv* contém informações sobre o formato do inventário realizado (presencial ou via Facebook), e um identificador do experimento *exp* para uso interno da equipe de desenvolvimento. Estas informações

são ainda acrescidas de anotações parciais de informações de gênero, idade, um indicador binário representando se o sujeito tinha formação na área de TI, o grau de religiosidade (1-5) do sujeito e um indicador numérico do seu curso de formação (com códigos de 1 a 11 para os cursos mais frequentes, e zero para todos os demais).

A informação de gênero é conhecida para 1081 (99,9%) sujeitos. Destes, 597 (55,2%) são do sexo feminino. A informação de idade é conhecida para 810 (74,9%) sujeitos conforme distribuição da Figura 3.

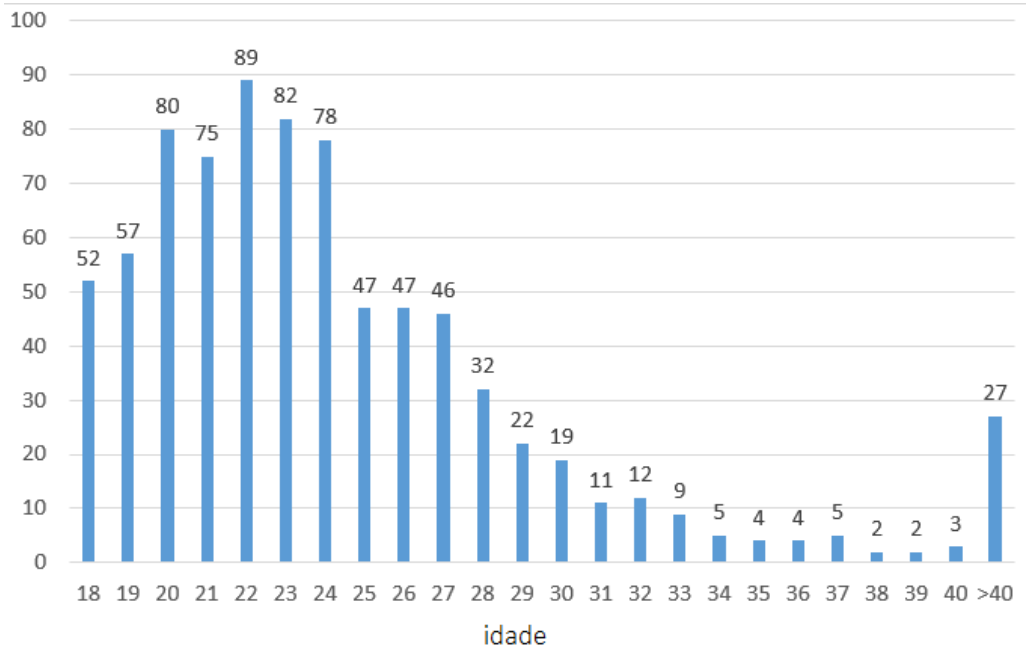

Figura 3. Distribuição de sujeitos por idade.

A informação sobre relação com a área de TI - aqui entendida tanto como formação básica como no sentido de atuação profissional, independente da formação - é conhecida para 871 (80,5%) sujeitos. Destes, 513 (58,9%) não são da área.

O grau de religiosidade dos sujeitos foi obtido apresentando-se a pergunta ilustrada na Figura 4.

Considerando seu grau de RELIGIOSIDADE, você se define como:

(1) Nada religioso
(2) Pouco religioso
(3) Mais ou menos religioso
(4) Religioso
(5) Muito religioso

Figura 4. Questão sobre o grau de religiosidade

No total, 481 sujeitos (44,5%) responderam à questão sobre religiosidade. A distribuição das respostas é ilustrada na Tabela 2.

Tabela 2. Grau de religiosidade (481 sujeitos)

Grau	Sujeitos	%
1	118	24,6%
2	119	24,7%
3	107	22,2%
4	115	23,9%
5	22	4,6%

Assim como no caso dos valores escalares de personalidade, cabe destacar que os graus de religiosidade também devem ser interpretados de forma relativa ao grupo (ou subgrupo) de indivíduos analisados. Além disso, um indivíduo de grau de religiosidade 4 não deve ser entendido como sendo "duas vezes mais religioso" do que um indivíduo de grau 2, mas simplesmente como sendo mais religioso. Dada a natureza subjetiva deste tipo de autoavaliação, sugere-se inclusive que a modelagem da informação de religiosidade para fins de aprendizagem de máquina seja feita de forma limitada, possivelmente na forma de uma classe binária.

A informação do tipo de curso de graduação é conhecida para 479 (44,3%) sujeitos, divididos em 11 cursos conforme distribuição da Figura 5.

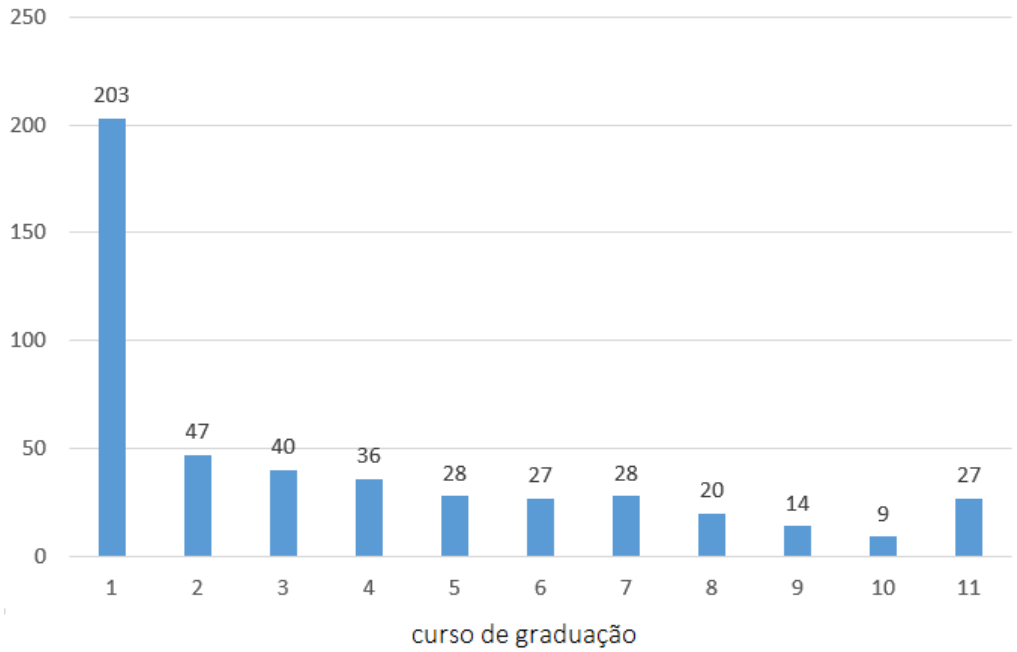

Figura 5. Distribuição de sujeitos por curso de graduação.

Finalmente, cabe destacar que os números apresentados referem-se ao conjunto completo do córpus, incluídos aqui tanto os sujeitos que eram usuários Facebook, os que realizaram apenas o experimento presencial, e os que atendem ambas condições. Assim, para aplicações de caracterização autoral (de gênero, idade, etc.) a partir de dados do Facebook, por exemplo, é preciso considerar que o número de instâncias real em cada classe será necessariamente menor do que o apresentado acima.

5. A base de atualizações de status Facebook b5-post

A base *b5-post* foi construída por meio da coleta de textos provenientes das atualizações de status na rede social Facebook após autorização concedida pelos sujeitos que preencheram o inventário de personalidade utilizando um aplicativo específico. 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 não controlados provenientes de redes sociais.

Para cada sujeito que preenchia o inventário de personalidade, eram coletadas até 1.000 atualizações de status. No total, 1039 sujeitos realizaram este procedimento, mas 11 destes não possuíam atualizações na rede social, ou utilizavam algum tipo de configuração que impediu a captura dos dados. A Tabela 3 apresenta um resumo da coleta realizada.

Tabela 3. Atualizações de status (posts) Facebook

Medida	Posts	Itens	Tokens	Tokens(n)
Mínimo	0	0	0	0
Máximo	901	22905	4710	3985
Média	187	2136	834	687

Por razões de confidencialidade, os dados da base *b5-post* não são disponibilizados em formato textual, mas apenas na forma de índices numéricos em uma pasta [indexed] (que pode ser o suficiente para aplicações de classificação de documentos, por exemplo), e resumidos na forma de uma tabela de estatísticas lexicais [lexical-features.csv] computada conforme descrito em Paraboni [2016].

6. A base de expressões de referência b5-ref

Para estudo do fenômeno de produção de expressões de referência, tanto do ponto de vista da seleção de conteúdo como da realização superficial destas expressões[1], uma porção presencial do experimento *b5* contemplou uma subtarefa de identificação de entidades visuais. Diferentemente de trabalhos prévios baseados em domínios simplificados (e.g., objetos geométricos), entretanto, o domínio considerado neste caso faz uso de imagens de estímulo com maior potencial de explicitar diferenças entre traços de personalidade. De forma mais específica, os contextos de referência a serem considerados fazem uso de imagens extraídas da base *Face Place* [Righi et al. 2012] de fotografias humanas validadas para diversos tipos de emoções e características físicas. Um exemplo de imagem de estímulo a ser utilizada nesta atividade é ilustrado na Fig. 6.

[1]Aqui entendido como sendo do ponto de vista de GLN - para questões de interpretação de expressões de referência, ver por exemplo Paraboni e de Lima [1998], Cuevas e Paraboni [2008].

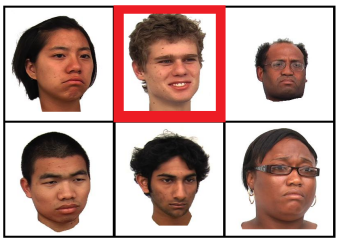

Figura 6. Imagem de estímulo com objeto-alvo em destaque, criada a partir da base *Face Place* [Righi et al. 2012].

Nesta tarefa, o participante é instruído a referenciar de forma única (i.e., sem ambiguidade) a pessoa ou entidade destacada de tal modo que outra pessoa possa identificá-la. No caso do presente exemplo, isso poderia ser feito, por exemplo, com uso de descrições como 'o rapaz loiro' ou 'o homem que está sorrindo, no centro superior da imagem'. O objetivo deste tipo de coleta de dados é modelar as decisões sobre o conteúdo - ou 'o que dizer', cf. Krahmer e van Deemter [2012] - de expressões produzidas por indivíduos com diferentes traços de personalidade.

Esta tarefa levou à construção do subcórpus *b5-ref* de expressões de referência anotadas com suas propriedades semânticas, e acompanhadas das informações de personalidade de seus autores (obtidas a partir dos inventários discutidos na Seção 4. Em sua versão atual, o córpus *b5-ref* possui 1810 descrições produzidas por 152 sujeitos. Destes, 86 (56,6%) são do sexo feminino. A faixa etária média era de 25,8 anos (mínimo 18 e máximo 59).

As expressões coletadas foram submetidas à correção ortográfica, e anotadas segundo um esquema composto de 27 atributos propostos com base nos tipos de informação mais frequentes observados nos dados. Estes atributos incluem características de aparência física dos objetos-alvo (rostos humanos) presentes em cada imagem de estímulo, como cor da pele ou comprimento do cabelo, e também características de teor mais subjetivo, como emoções. Diversos tipos de atributos que fugiam ao propósito de identificação (como em 'a pessoa que parece ter bom gosto para se vestir') não fizeram parte deste esquema, e estas informações não foram anotadas (embora ainda permaneçam na cadeia de caracteres original disponível no córpus).

Como forma de evitar a anotação de um número excessivo de atributos esparsos, diversos atributos foram combinados em classes mais gerais por afinidade. Por exemplo, toda e qualquer referência a pelos faciais como barba, bigode, cavanhaque e afins é representada por um único atributo *facial.hair* com valores possíveis *yes/no* indicando apenas que houve uma referência deste tipo na expressão, mas sem detalhar qual exatamente foi esta referência.

Além disso, atributos cujo valor não era passível de estimativa objetiva, como o formato mais arredondado de um determinado rosto, foram modelados como possuindo apenas o valor *others*. Isso indica que estes atributos, embora ocorram nas expressões do córpus, não possuem valor discriminatório (e.g., porque, dependendo do ponto de vista, qualquer das pessoas apresentadas como estímulo pode ter, de certa forma, um rosto arredondado etc.).

A subjetividade dos valores de certos atributos foi tratada de forma uniforme quando

possível considerando-se as informações fornecidas pela própria base *Face Place*. Isso inclui, por exemplo, propriedades relativas ao tipo étnico (negro, caucasiano etc.) e também à representação de emoções (alegre, triste etc.) já que as imagens provenientes da base *Face Place* são disponibilizadas com este tipo de informação. Todos atributos foram anotados com seu valor padrão quando possível ou, na falta deste, segundo o julgamento da maioria dos sujeitos para aquela situação independentemente do valor exato empregado na expressão de referência. Por exemplo, todas as descrições de uma mesma imagem que comprovadamente exibe uma pessoa de cabelo curto (segundo a anotação *Face Place*) são anotadas como *hair.length-short* mesmo que alguns sujeitos tenham descrito o cabelo como sendo longo. Em outras palavras, o valor anotado é indicativo apenas de que houve referência ao atributo *hair.length*, mas não necessariamente que houve referência ao valor (*short* ou *long*) específico.

A lista completa dos 27 atributos e seus valores possíveis é apresentada na Tabela 4. Cabe destacar ainda que alguns tipos de informação infrequente não fazem parte deste esquema, e portanto não foram anotados. Dentre estes, destacamos referências à 'único' (e.g., 'a única moça sorridente'), quantificadores (e.g., muito, pouco, ligeiramente, bastante), comparativos (e.g., maior / menor do que) e referências a uma segunda pessoa (e.g., 'perto do moço de óculos').

O córpus *b5-ref* consiste de dois componentes principais: uma pasta [descriptions-xml] contendo um conjunto de arquivos XML (aqui denominados TRIALS) representando as expressões produzidas pelos participantes, e um arquivo *b5-ref-contexts.xml* contendo a especificação completa de cada uma das 12 cenas de estímulo.

A notação XML utilizada no córpus é semelhante à utilizada em diversos outros projetos da área como em de Lucena et al. [2010], Teixeira et al. [2014], Paraboni et al. [2016]. Além da anotação semântica das expressões coletadas e das cenas de estímulo, o córpus contém ainda o conjunto completo de descrições em uma única planilha *descriptions.xlsx*. Nesta planilha as descrições são relacionadas uma abaixo da outra, e as colunas identificam o participante *subject* que as produziu (referente ao arquivo *subject.csv* da pasta principal do córpus, cf. Seção 4), a cena *id*, e a expressão original produzida *string* (em Português). As demais 27 colunas contêm a anotação semântica de cada descrição.

O córpus *b5-ref* objetiva propiciar estudos em geração de expressões de referência (GER) baseada em córpus [Ferreira e Paraboni 2014a]. Considerando-se a informação de personalidade disponível na tabela *subject.csv* do córpus principal (cf. Seção 4), torna-se possível projetar algoritmos de GER baseados em personalidade, que podem ser vistos como uma extensão de modelos de variação humana para esta tarefa [Ferreira e Paraboni 2014b, 2017]. Além disso, e embora não seja o foco original deste projeto, ambas as versões do córpus (na representação XML e na versão em planilha) incluem os strings produzidos em Português em formato próximo do original, salvo correções ortográficas mínimas. Estes strings podem ser aproveitados em futuros projetos de realização superficial [Pereira e Paraboni 2007, 2008] com uso de informações de personalidade.

7. As bases de textos e legendas b5-text e b5-caption

Para estudo de questões mais gerais de produção de texto, como o planejamento de documentos, geração texto-para-texto e sumarização, o experimento *b5* contempla também duas subtarefas de descrição de imagens: em versão detalhada (em texto livre e multi-

Tabela 4. Esquema de anotação do córpus b5-ref

Atributo	Valores	Descrição
mf	{*male*, *female*}	gênero explícito (homem, rapaz) ou implícito (japonesa, loira, o/a)
isyoung	{*yes*}	jovem, rapaz, moça, garota
race	{*asian*, *black*, *caucasian*}	*asian*: oriental, asiático, japonês, chinês; *black*: raça negra, afrodescendente (sem menção à pele) como em 'mulher negra'; *caucasian*: branco, ocidental, homem branco. Convenções: 'alemão' é anotado como cor de cabelo, não como raça) mas sem menção à pele; 'mestiço' é anotado como *race-asian* e *skin-dark*
skin	{*fair*,*dark*}	menção explícita à cor da pele: clara, escura, negra - sempre com menção à pele. Convenção: 'mestiço' é anotado como *race-asian* e *skin-dark*.
emotion	{*positive*, *negative*, *neutral*}	emoção positiva / negativa / neutra: semblante / expressão alegre, feliz, extrovertida (sem menção explícita ao sorriso). Ou triste, cansado, impaciente, rancoroso, com raiva, chateado, entediado, emburrado, fechado; ou normal, cara de paisagem, 'pokerface', calmo, tranquilo, sem fazer bico. Olhar distante.
smile	{*yes, no*}	com ou sem sorriso (sério)
mouth	{*shut, open*}	boca fechada, aberta; mostrando ou não mostrando os dentes
facial.hair	{*yes*}	toda referência à barba, barbicha, cavanhaque, pelos ou penugem no queixo ou face (independentemente de cor)
shape	{*other*}	referência ao formato do rosto (arredondado, fino) mas não bochecha. Obesidade só é marcada quando menciona face, caso contrário é *face-others*.
spots	{*yes*}	toda referência a espinhas/marcas/sinais/sardas no rosto, testa ou pescoço
lips	{*other*}	toda referência a lábios, exceto quando cerrado/aberto, que é marcado como *mouth-shut* ou *mouth-open*.
nose	{*other*}	toda referência a nariz (tamanho, formato, aparência etc.)
eyebrows	{*other*}	toda referência a sobrancelhas (grossa, fina, formato, curvatura, cor)
face	{*other*}	qualquer outro aspecto da face (e.g., magro, obeso, com sobrepeso) ou pele (macia - mas não quando referente a cor), queixo ou bochechas (e.g., grandes), covinhas tamanho da boca (grande, pequena); pescoço.
hair.colour	{*dark, blonde*}	*dark*: cabelo escuro / preto / castanho / castanho escuro; *blonde*: cabelo claro / loiro / castanho / castanho claro. Convenção: 'moreno' geralmente é anotado como *hair.colour-dark*, a menos que haja referência explícita à pele morena. No entanto, para as imagens 9 e 11 (mostrando pessoas negras), 'moreno' foi entendido como *race-black*.
hair.length	{*short, long*}	cabelo curto / raspado / na altura do pescoço ou ombros / do queixo para cima, ou cabelo longo / comprido
hair.style	{*straight, curly*}	liso ou crespo / ondulado / encaracolado
unkempt	{*yes*}	cabelo desarrumado, despenteado, assanhado
ponytail	{*yes, no*}	rabo-de-cavalo, cabelo preso
fringe	{*yes*}	toda referência à franja, cabelos cobrindo a face, caindo sobre a testa, mecha de cabelo na testa, cabelos presos
hair	{*other*}	qualquer outro aspecto especial do cabelo (grande, volumoso, espesso)
narrow.eyed	{*yes, no*}	para orientais: olhos fechados, quase fechados, estreitos, profundos, 'como os de um japonês', olhos orientais / asiáticos. Para outras raças, 'olhos fechados' é marcado como *eyes-other*.
eye.colour	{*light, dark*}	olhos claros / azuis / verdes ou escuros / pretos
glasses	{*yes, no*}	com ou sem óculos
eyes	{*other*}	qualquer outro aspecto especial dos olhos (e.g., pequenos, grandes, estrábicos, com olheiras, desenhados com lápis); olhar para a frente;
ears	{*other*}	toda referência a orelhas (e.g., pequenas, grandes, visíveis ou não - mas não quando falando do comprimento do cabelo, anotado em hair.length)
earring	{*yes, no*}	com ou sem brincos, com ou sem furo na orelha (mas orelhas furadas são marcadas como *ears-other*)

sentencial) e em versão resumida (na forma de uma sentença única). Os estímulos visuais neste caso serão provenientes da base *GAPED* [Dan-Glauser e Scherer 2011] de imagens classificadas por valência e significância normativa designadas de modo a despertar diferentes graus de emoção. Um exemplo de imagem disponibilizada pela base *GAPED* e ilustrado na Fig. 7.

Figura 7. Imagem de estímulo da base *GAPED* [Dan-Glauser e Scherer 2011].

Diferentemente da tarefa de identificação discutida na seção anterior, o objetivo neste caso não é observar a estratégia referencial do sujeito do experimento, mas sim a estratégia utilizada para determinar os elementos mais importantes da imagem, a ordem e estruturação da descrição, e suas escolhas lexicais e sintáticas. Esta coleta de dados foi realizada em duas versões - detalhada e resumida - como forma de exercer um maior grau de controle sobre o texto produzido, sem no entanto influenciá-lo.

Estas duas versões do texto permitirão o estudo de questões ligadas à geração do tipo texto-para-texto tanto em nível sentencial quanto discursivo, contemplando diversas subtarefas relacionadas ao planejamento sentencial [Paraboni e van Deemter 2002, Stent e Molina 2009] como questões de ordenação sentencial, agregação e inserção de marcadores de discurso e relacionadas ao próprio planejamento discursivo, além da própria realização da forma superficial [de Novais e Paraboni 2012].

Os textos b5-text e b5-caption são disponibilizados em três formatos: *original*, em 10 arquivos representando os conjuntos de descrições de cada imagem de estímulo e contendo o identificador de cada participante (conforme a tabela b5-subject descrita na Seção 4); *per-speaker*, de arquivos individuais contendo todo texto produzido por cada um dos participantes, e *parsed*, representando os mesmos 10 arquivos originais com informação de análise sintática fornecida pela versão online da ferramenta PALAVRAS [Bick 2000]. De modo geral, os formatos *original* e *parsed* são mais úteis para pesquisa de geração de língua natural, pois indicam o que cada participante escreveu em resposta a cada estímulo, enquanto que *per-speaker* é mais útil para pesquisa de interpretação de língua natural (por exemplo, para caracterização autoral ou classificação de documentos).

As 10 imagens de estímulo constantes desta base foram parcialmente anotadas de forma semi-automática com rótulos representando os principais (i.e., mais frequentes) conceitos que cada imagem representa. Estes conceitos foram extraídos a partir dos substantivos identificados na análise sintática, removendo-se repetições e agregados em sinônimos mediante anotação manual realizada por dois anotadores e revisadas por um terceiro. Observa-se entretanto que não foi realizada a correspondência entre elementos (ou regiões) da imagem e seus conceitos, ou seja, não se trata de um córpus de imagens do tipo utilizado em geração de texto a partir de imagens Gilbert et al. [2015].

8. Disponibilização

A versão 1.0. do córpus está disponível para fins de pesquisa mediante solicitação aos seus autores.

Agradecimentos

Este trabalho conta com apoio FAPESP nro. 2016/14223-0. Os autores são também gratos aos participantes da coleta de dados, e a todos que de alguma forma nela colaboraram.

Referências

Argamon, S., Dhawle, S., Koppel, M., e Pennebaker, J. W. (2005). Lexical predictors of personality type. Em *The joint annual meeting of the interface and the classification society of North America*.

Bick, E. (2000). *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework*. Tese de Doutorado , Aarhus University.

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.

Dan-Glauser, E. S. e Scherer, K. R. (2011). The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, 43(2):468–477.

de Andrade, J. M. (2008). *Evidências de validade do inventário dos cinco grandes fatores de personalidade para o Brasil*. Tese de Doutorado , Universidade de Brasília.

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.

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.

Ferreira, T. C. e Paraboni, I. (2017). Generating natural language descriptions using speaker-dependent information. *Natural Language Engineering (to appear)*.

Gauthier, I. e Tarr, M. J. (1997). Becoming a greeble expert: Exploring mechanisms for face recognition. *Vision Research*, 37(12):1673–1682.

Gilbert, A., Piras, L., Wang, J., Yan, F., Dellandrea, E., Gaizauskas, R., Villegas, M., e Mikolajczyk, K. (2015). Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. Em *CLEF2015 Working Notes*, CEUR Workshop Proceedings, Toulouse, France. CEUR-WS.org.

John, O. P., Donahue, E., e Kentle, R. (1991). The Big Five inventory - versions 4a and 54. Technical report, Inst. Personality Social Research, University of California, Berkeley, CA, USA.

John, O. P., Naumann, L. P., e Soto, C. J. (2008). *Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues*, páginas 114–158. Guilford Press, New York, NY.

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

Mairesse, F., Walker, M., Mehl, M., e Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)*, 30:457–500.

Oberlander, J. e Nowson, S. (2006). Whose thumb is it anyway? classifying author personality from weblog text. Em *COLING/ACL 2006 Poster Sessions*, páginas 627–634, Sydney, Australia.

Paraboni, I. (2016). Tratamento lexical para computação de personalidade a partir de textos. Technical Report PPgSI-000/2016, USP / EACH.

Paraboni, I. e de Lima, V. L. S. (1998). Possessive pronominal anaphor resolution in portuguese written texts. Em *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, páginas 1010–1014. Association for Computational Linguistics.

Paraboni, I., Galindo, M., e Iacovelli, D. (2016). Stars2: a corpus of object descriptions in a visual domain. *Language Resources and Evaluation*.

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

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.

Righi, G., Peissig, J. J., e Tarr, M. J. (2012). Recognizing disguised faces. *Visual Cognition*, 20(2):143–169.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M., e Ungar, L. H. (2013). Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One*, 8(9):e73791.

Stent, A. e Molina, M. (2009). Evaluating automatic extraction of rules for sentence plan construction. Em *SIGDIAL '09 Proceedings*, páginas 290–297.

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.