v. 30, n. 1, p. 24–50, 2025
ISSN: 2238-3824 · DOI: 10.35699/2238-3824.2024.57565
Submetido em: 14/02/2025 · Aprovado em: 31/03/2025

caligrama
revista de estudos românicos

A gramática em datasets multimodais: um estudo de caso de legendas de imagens

Grammar in Multimodal Datasets: A Case-study of Image Captions

André V. Lopes Coneglian
Universidade Federal de Minas Gerais (UFMG) | Belo Horizonte | MG | BR
coneglian@ufmg.br
https://orcid.org/0000-0003-1726-8890

Adriana Pagano
Universidade Federal de Minas Gerais (UFMG) | Belo Horizonte | MG | BR
CNPq
FAPEMIG
apagano@ufmg.br
https://orcid.org/0000-0002-3150-3503

Resumo: Este artigo apresenta um estudo linguístico de uma amostra de 150 legendas de imagem, que compõem o dataset multimodal Framed Multi30k (Viridiano, 2024). O objetivo é propor uma análise das legendas de imagem segundo fatores sociocognitivos implicados no processo de verbalização da experiência (Chafe, 2002, 2005; Croft, 2007) de modo que seja possível explicar, pelo menos parcialmente, a variação observada na construção léxico-gramatical das legendas. Para tanto, o artigo discute uma metodologia computacional de anotação linguística das legendas, com base na qual são extraídas informações gramaticais. Discute-se, ainda, o caráter experimental de elicitação controlada da produção de legendas de imagem e as consequências linguística disso. O exame gramatical das legendas centra-se nos processos de verbalização, como seleção, categorização e orientação. Por fim, discute-se a natureza da legenda de imagem como uma peça textual descritiva, explicitando-se operações descritivas que se verificam na produção dessas peças.

Palavras-chave: Verbalização da experiência; teoria baseada no uso; construção gramatical; significação; multimodalidade.

Abstract: This article presents a linguistic study of a sample of 150 image captions from the multimodal dataset Framed Multi30k (Viridiano, 2024). The objective is to analyze image captions based on socio-cognitive factors involved in the process of verbalizing experience (Chafe, 2002, 2005; Croft, 2007), in order to partially explain the variation observed in the lexico-grammatical construction of captions. To this end, the article discusses a

computational methodology for linguistic annotation of captions, which serves as the basis for extracting grammatical information. Additionally, it examines the experimental nature of controlled elicitation in caption production and its linguistic implications. The grammatical analysis focuses on verbalization processes, such as selection, categorization, and orientation. Finally, the study explores the nature of image captions as descriptive textual units, highlighting descriptive operations observed in their production.

Key words: Verbalization of experience; usage-based theory; grammatical construction; meaning; multimodality.

1 Introdução: do processamento de língua natural à análise linguística

Legendas de imagens têm sido amplamente usadas para, dentre outras tarefas de processamento de língua natural (PLN), treinar e avaliar sistemas de descrição de imagens. Datasets de legendas de imagens se compõem, minimamente, da imagem e da legenda (ou, das legendas) que a descreve. Tais datasets têm sido produzidos para diversas línguas (veja-se Viridiano *et al*., 2024 para uma lista compreensiva) e, mais recentemente, têm incorporado informação semântica de diversos aspectos contextuais, chegando, assim, a uma verdadeira integração entre as modalidades escrita e visual (p.e., Belcavelo, 2023; Viridiano, 2024; Viridiano *et al*., 2024).

Enquanto datasets dessa natureza têm recebido ampla atenção da comunidade de PLN, um aspecto virtualmente ignorado na construção desses datasets é a linguagem que informantes (ou seja, usuários de língua natural)[1] mobilizam para compor legendas. É exatamente para esta direção que este artigo se encaminha.

No contexto experimental, legendas para imagens são produzidas por informantes, os quais podem assumir diferentes perspectivas ou enfocar diferentes regiões da imagem ao produzir as legendas. Não é surpreendente que seja assim, uma vez que a interpretação de um *input* não verbal e a construção verbal desse *input* são processos regulados por inúmeros fatores sociocognitivos. De todo modo e exatamente por isso mesmo, uma legenda não pode constituir uma descrição única da imagem ou, como se considera nos estudos de PLN, um único padrão ouro ou referência para essa imagem. Uma mesma imagem pode ter tantas descrições distintas quanto é o número de usuários que as produzem, assim como uma mesma experiência pode ser verbalizada, isto é, empacotada linguisticamente, de formas diferentes (Chafe, 1977a, 1998; Givón, 1995; Clark 2003).

De um modo geral, a atenção que se dirige à linguagem das legendas está, quase sempre, condicionada por tarefas de identificação de vieses e estereotipagem em datasets.

[1] O termo "usuário de língua natural" vem de Dik (1997).

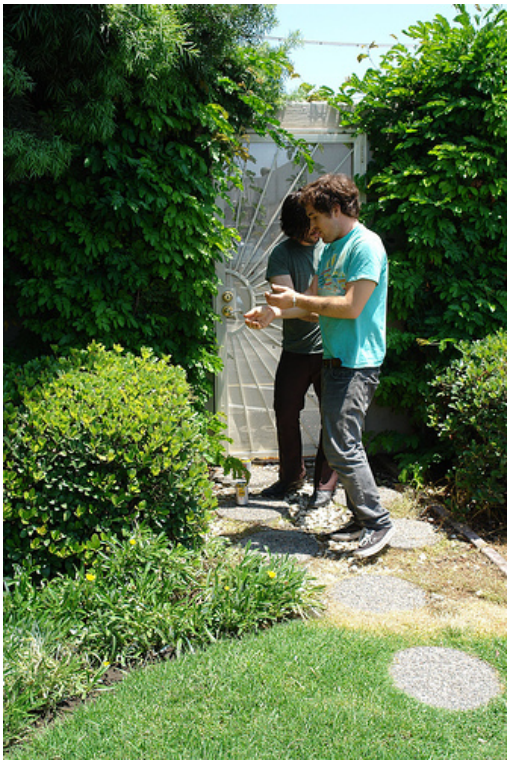
Por exemplo, examinando as mais de 30.000 imagens no dataset Flickr30K (Young *et al*., 2014), Miltenburg (2016) apresenta dois tipos de estereótipos presentes nesse dataset, o que o autor chama de vieses linguísticos (*linguistic biases*) e inferência infundada (*unwarrented inference*). São exemplos do primeiro caso descrições como *male nurse* ('masculino enfermeiro', literalmente) e *female surgeon* ('feminina cirurgiã', literalmente). Expressões desse tipo podem construir significados sexistas e racistas como já é bem documentado na literatura especializada. Inferências infundadas, por outro lado, caracterizam descrições que constroem significados mais específicos que aqueles que as imagens embasam, por exemplo, categorizar pessoas mais velhas com crianças como pais ou mães das crianças, sem qualquer evidência de que seja este o caso.

Análises dos datasets nessa orientação têm se multiplicado e o campo novo de perspectivismo em PLN (Basile *et al*., 2023) tem se consolidado, no qual integram-se aos datasets perspectivas e opiniões de indivíduos envolvidos na anotação.

Análises dessa natureza tendem a considerar a legenda como produto, isto é, como uma proposição fechada em si. No entanto, neste artigo, propõe-se considerar a legenda do ponto de vista da produção linguística, isto é, considerando-a como um enunciado descritivo que é construído por um indivíduo para fins de verbalização de uma experiência não linguística. Nesse modo de condução, independentemente do ponto de consideração das legendas com relação às imagens, é inquestionável que uma característica fundamental de datasets multimodais de imagens e suas legendas é a variação, tanto no que diz respeito a operações descritivas quanto à linguagem desses enunciados descritivos.

A título de ilustração, considere-se a Figura 1 e as cinco legendas produzidas por indivíduos diferentes.

Figura 1: Imagem 1000092795.jpg



Fonte: dataset *Framed Multi30k*.

(01) Legendas da imagem 1000092795.jpg[2]
#1 Dois homens conversam no jardim perto do portão.
#2 Dois jovens vestindo calças e camisetas encontram-se próximos a um portão cercados por uma área verde com muitas plantas e um gramado.
#3 Dois homens vestindo calça e camisa de malha estão em pé em frente a um portão que dá acesso a um jardim.
#4 Dois homens parados, em um ambiente aberto com folhagem verde, próximos a um portão de metal e vidro.
#5 Dois homens em frente a um portão cercado de plantas.

Note-se que, do ponto de vista da descrição, as legendas diferem quanto à inclusão ou não de informação a respeito da roupa dos homens na imagem, quanto ao que eles fazem (se conversam ou se estão parados) e quanto ao ambiente em que se encontram (se um jardim ou se é um lugar cercado de plantas). Do ponto de vista da composição construcional desses enunciados, dentre outras coisas, pode-se apontar, por exemplo, que apenas #1 e #2 contêm predicados verbais (cujos núcleos são *conversam* e *encontram-se*, respectivamente), #3 a #5 constroem-se como predicados não verbais adverbais. Quanto aos aspectos semânticos desses enunciados, note-se que #1 constrói um predicado dinâmico, ao passo que #2 a #5, predicados estativos.

Este artigo propõe uma análise de uma amostra de 150 legendas produzidas para 30 imagens do dataset Framed Multi30k (Viridiano, 2024). Mais especificamente, a análise que se

[2] As legendas são acompanhadas de #n, em que 'n' indica o número da legenda da imagem no dataset.

propõe aqui parte do reconhecimento de que as legendas de imagens são peças linguísticas, ou ainda, enunciados descritivos, que verbalizam um estímulo não verbal (a imagem) com base em diretrizes de verbalização (as instruções da tarefa de legendagem). Nesse contexto, este artigo examina a construção linguística das legendas, mais especificamente, em aspectos gramaticais da organização das legendas, valendo-se de uma metodologia de anotação automática e revisão manual dos textos segundo o modelo de anotação das Dependências Universais, UDs, (De Marneffe *et al*., 2021).

Nesse encaminhamento, o objetivo deste artigo é apresentar uma análise das legendas de imagem segundo fatores sociocognitivos implicados no processo de verbalização da experiência de modo que seja possível explicar, pelo menos parcialmente, a variação observada na construção léxico-gramatical das legendas, avaliando em que medida essa variação em forma corresponde a variação em significação. Com isso, busca-se mostrar que metodologias de PLN, como a anotação automática de textos, neste caso, podem servir de base para o mapeamento de dimensões de variação gramatical em enunciados linguísticos.

A proposta deste artigo justifica-se pelo fato de que, enquanto legendas de imagens são amplamente discutidas no contexto do PLN, pouco se estudam legendas de imagem do ponto de vista linguístico. Mais amplamente, espera-se que as análises linguísticas empreendidas neste artigo possam colaborar com e se somar às análises computacionais de modo a contribuir para o PLN.

Na seção 2, apresenta-se o dataset Framed Multi30k (Viridiano, 2024) e o recorte desse dataset que constitui a amostragem anotada segundo as UDs, neste estudo. Na seção 3, discute-se o enquadre experimental da produção de legendas de imagens, de modo que seja possível explicitar o cenário de produção linguística em que informantes criam essas peças textuais. Na seção 4, discutem-se aspectos linguísticos e cognitivos da produção de legendas de imagem pelos processos de verbalização da experiência, com base na extração de informações do dataset anotado com as UDs, com vistas ao mapeamento da variação da estrutura gramatical verificada na amostra de legendas. Na seção 5, discute-se a natureza descritiva das legendas de imagem a fim de apontar padrões regulares no uso da linguagem.

2 A construção de um dataset (amostral) de legendas com anotação linguística

Como se indicou na Introdução, este artigo apresenta a análise da variação gramatical em uma amostra de 150 legendas descritivas de 30 imagens, que, em conjunto, compõem o dataset multimodal Framed Multi30k[3] (Viridiano, 2024; Viridiano *et al*., 2024).

O Framed Multi30k é composto de legendas originais em inglês e suas traduções para o português, bem como de legendas originalmente produzidas em português. Neste artigo, a amostra de legendas analisadas provém de um recorte do grupo de legendas originalmente produzidas em português. Portanto, o que se discute a seguir é válido majoritariamente para as legendas originais em português. O dataset de legendas originais do português é com-

[3] O dataset completo pode ser acessado por este link: https://huggingface.co/datasets/FrameNetBrasil/FM30K. Acesso em: 12 de fevereiro de 2024.

posto de 158.915 legendas para 31.783 imagens. Cada imagem tem 5 legendas produzidas por informantes distintos.

Detalhes a respeito da composição do dataset Framed Multi30k podem ser encontrados em Viridiano (2024). É importante destacar que esse dataset multimodal apresenta anotação para frames semânticos, segundo as diretrizes da FrameNet Brasil, tanto para as imagens quanto para as legendas. No dataset original, as legendas não apresentam anotação morfossintática e sintática. Para o estudo que ora se apresenta, foi criado um dataset com uma amostra de 150 legendas, de 30 imagens diferentes (5 legendas para cada imagem), com anotação morfossintática e sintática segundo o modelo das Dependências Universais, UDs, (de Marneffe *et al*., 2021).

As 158.915 legendas foram extraídas do Framed Multi30k em arquivo txt codificação UTF8. Com base nesse arquivo, foram escolhidas 150 legendas aleatoriamente para se realizar a anotação piloto segundo o modelo das UDs, que consiste em 17 etiquetas para anotação de classes gramaticais, juntamente com suas características (*features*) e 37 etiquetas de relações sintáticas, além de sub-relações. A amostra de 150 legendas foi primeiramente anotada de forma automática por meio da ferramenta UDpipe (c), com um modelo de língua portuguesa multigênero que utiliza o Porttinari-UD-2.15 (Pardo et al., 2021). Os arquivos CONLLU foram importados na ferramenta de anotação Arborator-Grew-NILC,[4] uma versão expandida e aprimorada de Arborator-Grew (Guibon et al., 2020). A revisão da anotação automática foi realizada por 1 anotador (o primeiro autor deste artigo) familiarizado com as diretrizes das UDs. Após a revisão, o arquivo em formato CONLLU foi exportado da ferramenta Arborator-Grew-NILC e processado por script em linguagem Python para contagem das categorias anotadas.

Para este artigo, com base na anotação das UDs são feitas extrações de informações lexicais e gramaticais da amostra de 150 legendas de modo que seja possível proceder à análise linguística das legendas.[5]

De todo modo, com a divulgação desse dataset amostral com anotação de UDs, este artigo busca apresentar mais evidências para a tese de que, para além da realização de tarefas de PLN, datasets anotados podem servir, também, à análise linguística (Coneglian, 2023).

3 O enquadre experimental da produção de legendas de imagem

Esta seção discute a legenda de imagens como produto da produção linguística. Assim, propõe-se, aqui, que as legendas constituem, na verdade, enunciados descritivos, os quais são produzidos no contexto experimental de elicitação controlada. Antes de proceder ao desenvolvimento dessas duas teses, é necessário fazer os seguintes apontamentos sobre a natureza do dataset Framed Multi30k, com particular destaque às legendas:

1 as legendas constituem descrições conceituais (Jaimes e Chang, 1999) das imagens, na medida em que permitem a identificação do que é mostrado na imagem;

[4] Disponível em: https://arborator.icmc.usp. Acesso em: 12 fev.2024.

[5] Este artigo não elabora os pormenores da constituição do *dataset*, como diretrizes e dificuldades de anotação, o que é efeito em Pagano e Coneglian (em preparação). A expectativa é de que essa amostra inicial sirva de base para treinar modelos de anotação de legendas.

2 essas descrições não são objetivas, isto é, isentas de perspectiva, podendo legendas diferentes de uma mesma imagem diferir no modo pelo qual conceptualizam a cena apresentada pela imagem (como se mostrou com o exemplo em (01) anteriormente);

3 no contexto do PLN, as legendas são um meio para um fim – treinamento de modelos de visão computacional e constituição de datasets multimodais.

Os dois primeiros apontamentos valem para qualquer legenda, seja ela original, seja ela traduzida, e ligam-se, especificamente, à natureza linguística da legenda. O terceiro apontamento, entretanto, diz respeito apenas às legendas originais produzidas por falantes em uma determinada língua, e liga-se à natureza experimental da produção de legendas. Nessa condução, o restante desta seção dedica-se a explicitar a natureza descritiva das legendas de imagens e a enquadrá-las no contexto experimental. Inicie-se por este aspecto.

O procedimento de coleta de legendas originais para as imagens enquadra-se como uma tarefa experimental de elicitação controlada, como já dito. Obviamente, a tarefa de coleta de legendas para fins de consecução de tarefas de PLN não constitui trabalho de campo linguístico no sentido clássico do termo, porque a finalidade não é descrição e análise linguística, mas, sim, a realização de uma tarefa da qual a linguagem é uma parte componente. No entanto, enquadrar a elicitação de legendas no campo de tarefas de trabalho de campo linguístico pode iluminar alguns aspectos dessa tarefa, como se desenvolve nesta seção.

Do ponto de vista puramente linguístico, a tarefa de elicitação pode ser caracterizada por três fatores (Samarin, 1967, p. 106-107): (i) as sentenças obtidas numa tarefa de elicitação são, em geral, descontextualizadas e têm extensão de uma única sentença; (ii) a tarefa de elicitação é desenhada para a análise de um domínio linguístico específico; (iii) a tarefa de elicitação se define pela relação entre pesquisador e sujeito de pesquisa. Esses três fatores prontamente se verificam na tarefa de elicitação de legendas originais do Framed Multi30k.

Viridiano (2024, p. 70-73) apresenta detalhadamente o procedimento para a coleta de legendas originais em português para as imagens do Framed Multi30k. As legendas foram criadas por 148 estudantes universitários (graduandos e pós-graduandos), divididos em dois grupos, um grupo permanente, responsável pela criação de demais de 50% do total das legendas, e um grupo itinerante (fator (iii) de Samarin). Os indivíduos da pesquisa receberam as instruções em português para a criação das legendas. Várias legendas produzidas por diferentes indivíduos foram coletadas para uma mesma imagem, para garantir a "um número mínimo de variação na forma como cada uma das fotos foi descrita" (Viridiano, 2024, p. 71).

Considerem-se as instruções disponibilizadas aos anotadores:

> 1. Descrever cada uma das imagens usando apenas uma sentença – que, para os fins desta tarefa, equivale a uma sequência de palavras delimitada por um ponto final – com menos de 140 caracteres, tendo atenção ao uso padrão da gramática e ortografia;
> 2. Fornecer uma descrição precisa das atividades sendo executadas na imagem, das pessoas ou animais que executam a atividade e de quaisquer objetos envolvidos nela;

> 3. Quando possível, utilizar adjetivos e descrever também elementos relevantes que não estejam diretamente envolvidos na atividade (como elementos em segundo plano) (Viridiano, 2024, p. 72).

Note-se que a primeira instrução se liga ao primeiro fator de Samarin (1967), com a restrição de número de caracteres totais da sentença. As instruções 2 e 3 dizem respeito à dimensão semântica dessa tarefa de elicitação (fator (ii) de Samarin), a busca pela representação linguística de um estímulo não verbal (veja-se discussão na seção 4).

A coleta de legendas de imagens pode ser enquadrada como uma tarefa de elicitação baseada em estímulo imagético (Chelliah e De Reuse, 2011), um subtipo de elicitação controlada pela análise (Samarin 1967),[6] que, no trabalho de campo linguístico, em geral, é usada para elicitar itens lexicais isolados. No caso da tarefa do Framed Multi30k, é usada para elicitar sentenças descritivas.

Uma das vantagens desse tipo de experimentação está justamente no fato de que, se coletada para um número considerável de informantes, as descrições resultantes podem ser diferentes o suficiente para que seja possível mapear o espaço variação semântica. Ilustre-se essa questão com a Figura 2 e suas respectivas legendas em (02).

Figura 2: Imagem 106514190.jpg

Fonte: dataset Framed Multi30k.

(02) Legendas da imagem 106514190.jpg
#1 Uma pessoa caminhando em uma montanha de neve.

[6] Em inglês, analysis-controlled elicitation. Samarin (1967) – e posteriormente Chelliah e De Reuse (2011) – distinguem as tarefas de elicitação em dois tipos: (i) elicitação controlada por agenda e (ii) elicitação controlada pela análise. Do primeiro tipo, fazem parte técnicas ligadas a questionários estruturados. Do segundo tipo, fazem parte técnicas baseadas em algum tipo de *prompt*.

#2 Uma pessoa está no topo de uma montanha coberta por neve, durante o dia.
#3 Homem que escala uma montanha em meio a neve observa a paisagem.
#4 Um esquiador observa o horizonte do alto de uma montanha.
#5 Um esquiador em uma parte mais alta de um monte.

O que se observa nas cinco legendas da imagem na Figura 2 são diferenças quanto à categorização do ser humano na imagem, pelo uso dos substantivos *pessoa*, *homem*, *esquiador*, quanto à categorização do evento, pelo uso de expressões verbais como *caminhar, estar no topo de uma montanha*, *escalar uma montanha*, *observar a paisagem*, *observar o horizonte*. No entanto, o lugar é o mesmo, diferindo a topologia, se é o percurso que é construído ou um ponto específico do percurso.

Do ponto de vista linguístico e com o foco no exame da construção linguística das legendas, o interesse reside justamente no mapeamento do espaço de variação léxicogramatical nas legendas. Por exemplo, valendo-se das legendas em (02), verifica-se uma clara diferença na organização das predicações: #1 e #4 com predicados verbais simples, #3 com predicado verbal complexo, #2 e #5 com predicados não verbais, com cópula e sem cópula, respectivamente. Essa diversidade no acionamento básico das predicações pode resultar, pelo menos em parte, das escolhas que os informantes, como usuários de uma língua natural, fazem para verbalizar um conhecimento não verbal.

Tendo explicitado a natureza experimental com base na qual as legendas de imagens são produzidas, é possível passar para um exame geral das legendas. Como Viridiano (2024) e outros (*p.e.*, Young et al., 2014) já apontaram, uma legenda é boa para uma tarefa de PLN multimodal na medida em que apresenta uma descrição precsa e compreensiva de uma imagem. A seguir, discuste-se a natureza descritiva das legendas de imagens.

4 Da informação visual ao empacotamento linguístico da informação

A proposta deste estudo se constrói sobre os pilares de uma linguística baseada no uso (Bybee, 2023; c), para que seja possível articular texto, gramática e significado na análise de legendas de imagens (seção 5). O estudo aproveita, também, princípios desenvolvidos no âmbito da Linguística do Texto, particularmente no modelo de Adam (2018), para o tratamento das legendas de imagens como peças textuais descritivas (seção 4).

A vertente baseada no uso da linguística assume uma estreita relação entre estrutura linguística e instâncias de uso de linguagem (Barlow e Kemmer, 2000). De acordo com Bybee (2023, p. 9), "a teoria baseada no uso olha para a maneira pela qual a experiência com a linguagem impacta diretamente a representação cognitiva da linguagem".[7] Para este estudo, faz-se um recorte teórico muito específico da teoria baseada no uso pelo modelo da verbalização da experiência, inicialmente desenvolvido por Chafe (1975a, 1975b, 1977a, 1977b, 1998, 2002, 2005) e posteriormente elaborado por Croft (2007, 2010, 2020). O processo de **verbalização** diz respeito a um conjunto de processos pelos quais um indivíduo transforma infor-

[7] Texto original: "...Usage-Based Theory looks at the way experience with language directly impacts the cognitive representation of language".

mação cognitiva em linguagem. Mais especificamente, é o conjunto de processos pelos quais falantes empacotam linguisticamente conteúdo experiencial-cognitivo.

A análise das legendas de imagens pode em muito ser beneficiada por esse aparato teórico, uma vez que, dada sua natureza, a tarefa de produção de legendas de imagens é uma tarefa de verbalização, na qual um usuário (ou informante) empacota informação não linguística (visual) em enunciados linguísticos.

Chafe (idem) estabelece uma distinção entre pensamento e linguagem. Para ele, o pensamento é mais complexo, denso e multivariado do que a linguagem humana pode acomodar em seus meios de expressão. Logo, o autor propõe que uma série de ajustes (processos de verbalização) devem acontecer para que pensamento seja empacotado linguisticamente. Segundo o autor,

> ... pensamentos são mais ricos, mais extensos e mais complicados do que qualquer coisa que possa ser expressa na língua. A língua é simplesmente muito limitada para comportar tudo que possamos pensar, e, portanto, é necessário ser seletivo (Chafe, 2002, p. 400).[8]

Ao longo de inúmeros artigos sobre o assunto, Chafe apresenta cinco processos de ajuste entre pensamento e linguagem, são eles: **seleção**, **categorização**, **orientação**, **combinação** e **linearização**.

4.1 A gramática em ajustes de seleção

O ajuste de **seleção** decorre do fato de que o pensamento contém mais informação do que se pode empacotar linguisticamente. A seleção implica, portanto, o recorte da experiência que um indivíduo faz para verbalizá-la. No caso das legendas de imagens, por exemplo, não é incomum que a imagem tenha muito mais elementos representados do que a legenda pode acomodar linguisticamente.

Considere-se a imagem na Figura 3, a seguir, e suas respectivas legendas em (03).

[8] Tradução destes autores. Texto original: "... thoughts are richer, more extensive, and more complicated than anything that can be expressed in language. Language is simply too limited to accommodate everything we may be thinking, and so it is always necessary to be selective".

Figura 3: Imagem 1063866640.jpg

Fonte: dataset Framed Multi30k.

(03) Legendas da imagem 1063866640.jpg

#1 Silhueta de cinco homens conversando, sentados na beira do mar.

#2 Um grupo de pessoas conversam sentadas em troncos de árvore, em frente ao mar, durante o final da tarde.

#3 Pessoas conversam em uma rodinha sob a sombra de uma árvore perto de um corpo de água.

#4 Pessoas estão sentadas em pedras na beira de um rio, embaixo da sombra de uma árvore.

#5 Grupo de amigos sentados à beira de um lago conversam entre si.

Note-se que, enquanto as cinco legendas fazem alguma menção ao grupo de pessoas, ao mar e a alguma atividade que esse grupo esteja fazendo (no entanto há diferenças na categorização dessa atividade), nem todas as legendas fazem menção à arvore e nenhuma faz menção ao barco que passa no mar e às plantas atrás da árvore. Nessa discussão, não se pode ignorar o fato de que, para a composição das legendas do Framed Multi30k, os informantes tinham restrições quanto ao número de caracteres para cada legenda (seção 3). Essa restrição, ainda que artificial, não garante quais elementos visuais entram ou deixam de entrar nas legendas, do mesmo modo que restrições diversas no uso da linguagem em situações não controladas podem gerenciar o que um falante seleciona ou deixa de selecionar em termos de experiência não linguística na construção dos textos.

O fato linguístico relevante sobre o ajuste de seleção é especificamente o estabelecimento de **tópicos** (Chafe, 2005), informações focais sobre as quais o falante elabora seu enunciado. Novamente, no caso das legendas de imagens, que são, por natureza, descritivas, uma operação descritiva básica que se verifica na composição desses enunciados é a tematização por ancoragem referencial (Adam, 2018), isto é, todo o enunciado é construído com base em um referente. Esse ponto é elaborado e discutido na seção 5.

Analisando o processo de seleção do ponto de vista cognitivo, Croft (2007) propõe que a seleção integra, na verdade, o processo de **particularização**, juntamente com a **situação**. Assim, **seleção** e **situação** são dois processos simultâneos que compõem a **particularização**

(Croft, 2007). A situação diz respeito à ancoragem de uma entidade, de modo que o ouvinte, ou leitor da legenda, consiga fazer a identificação segundo algum modelo mental adequado.

No caso da **seleção**, particularmente de entidades nominais, é interessante identificar se as entidades selecionadas são construídas como indivíduos ou grupos. No primeiro caso, tem-se a distinção entre singular e plural e, no segundo caso, no uso de substantivos coletivos ou que indicam coletividade. No caso da **situação**, particularmente no domínio nominal, a contraparte gramatical é o acionamento de significados de definitude e indefinitude.

O dataset anotado segundo o modelo das UDs (seção 2) permite que sejam analisadas essas questões. A seguir, mostram-se os resultados especificamente sobre a distribuição de expressões (in)definidas na amostra de legendas.

Definitude e indefinitude são significados que, acima de tudo, são construídos na negociação do universo discursivo e decorrem de um contrato interacional entre falante e ouvinte quanto ao compartilhamento de informação (Clark, 1992). As línguas variam quanto aos expedientes gramaticais e construcionais que constroem significados de (in)definitude. No caso do português, a principal classe de palavras associada esses significados são os artigos, mas podem construir definitude pronomes demonstrativos (como *este*, *esse*, *isso*...) e pronomes possessivos (como *seu*, *meu*, *nosso*...). Diferentemente dos artigos, essas duas classes de pronome acrescentam alguma propriedade significativa à definitude própria do seu estatuto categorial; os demonstrativos fazem distinções dêiticas, ao passo que os pronomes possessivos sempre fazem referência bipessoal. Assim, a (in)definitude não é o único significado que esses expedientes constroem, como é o caso dos artigos.

Para a extração de informações sobre essas classes, no dataset anotado, foram isoladas três *features* de `PronType` (tipo de pronome): `PronType:Art` (artigo), `PronType:Dem` (demonstrativo), `PronType:Poss` (possessivo). No caso dos artigos, foram ainda isoladas as *features* `Definite:Def` (definido) e `Definite:Ind` (indefinido). O Quadro 1 apresenta os resultados quantitativos.

Quadro 1: Número de ocorrência de pronomes demonstrativos, pronomes possessivos e artigos no dataset deste estudo.

Expediente gramatical na anotação das UDs	Número de ocorrência
`PronType:Art` `Definite:Ind`	320
`PronType:Art` `Definite:Def`	134
`PronType:Poss`	06
`PronType:Dem`	01

Fonte: Elaboração própria.

Na amostra deste estudo, há apenas uma instância de pronome demonstrativo, que vem apresentada em (04). No caso desse enunciado em particular, o pronome demonstrativo é usado para fazer referência anafórica a um elemento já introduzido, *um banco*.

(04) #2 Um homem branco com cabelos pretos e sem camisa deitado em um banco sem encosto enquanto um bulldog francês branco e preto está amarrado pela coleira neste mesmo banco (imagem 1003163366.jpg).

Quanto aos pronomes possessivos, apareceram as formas *seu* e *sua*, num total de 06 ocorrências, como se vê em (05) a (10).

(05) #2 Um limpador de janelas usa um rodo para realizar seu trabalho no alto de uma escada apoiada na parede externa de um edifício de tijolos (imagem 1000344755.jpg).

(06) #3 Um guitarrista posa em frente a uma luminária de set de filmagem enquanto um produtor ajusta seu figurino de origem indígena (imagem 1000523639.jpg).

(07) #5 Homem em um armazém segura uma guitarra enquanto um segundo homem costura sua roupa (imagem 1000523639.jpg).

(08) #3 Um homem sem camisa está deitado sobre um banco em um deck de observação próximo a um rio, acompanhado de seu cachorro. (imagem 1003163366.jpg)

(09) #4 Um homem deitado sobre um banco de praça descansa enquanto segura a guia do seu cachorro (imagem 1003163366.jpg).

(10) #4 Homem deitado no meio da rua ao lado de seu carro enquanto o trânsito está parado (imagem 1067675215.jpg).

Com exceção da ocorrência (05) em que o pronome acompanha um substantivo abstrato (*trabalho*), o pronome possessivo acompanha um substantivo concreto. Os informantes que compuseram as legendas certamente estabeleceram alguma relação entre duas entidades e daí escolheram o pronome possessivo, por exemplo, em (06), o pronome possessivo estabelece uma relação entre o guitarrista e o figurino. Os sintagmas nominais com os pronomes possessivos são construídos como definidos, não porque sejam intrinsecamente definidos, mas a definitude nasce como uma inferência devido à referência bipessoal (Rinke, 2010). Uma outra questão a se considerar no caso dos pronomes possessivos, mas que não será discutida aqui, é a semântica do pronome possessivo, que não necessariamente indica 'posse' em sentido estrito (Neves, 2011).

Examine-se, então, a distribuição de artigos definidos e indefinidos na amostra de legendas deste estudo.

Como se vê pelos resultados do Quadro 1, dos expedientes de (in)definitude no português, os artigos apresentam o maior número de ocorrência, 454 ocorrências no total. Na amostra, nenhuma legenda iniciou-se com artigo definido, os quais ocorreram em contextos informacionais no qual a definitude já é pressuposta. Comparem-se o caso das duas primeiras legendas de (01), retomadas, aqui, como (11) e (12), respectivamente.

(11) #1 Dois homens conversam no jardim perto do portão.

(12) #2 Dois jovens vestindo calças e camisetas encontram-se próximos a um portão cercados por uma área verde com muitas plantas e um gramado.

Nessas duas legendas, que descrevem a mesma imagem (Figura 1), dois informantes diferentes optaram por duas estratégias diferentes no que diz respeito à instalação da (in) definitude. Em (11), o *portão* é construído como uma entidade definida, ao passo que em (12), é construído como indefinido.

Em geral, os informantes tendem a ser uniformes quanto à escolha do artigo definido ou indefinido, como se vê nas legendas em (13) e (14), que descrevem uma mesma imagem no dataset Framed Multi30k.

(13) #1 Um rapaz sentado em uma cadeira brinca com um leão de pelúcia.

(14) #2 Um jovem sentado em um banco e vestindo bermuda, tênis e camiseta segura um leão de pelúcia e mexe na boca do brinquedo.

No caso das legendas, a preferência pelo artigo indefinido nas descrições pode-se explicar pelo fato de que, do ponto de vista de uma semântica referencial, esses artigos apenas trazem à existência, no universo sentencial, as entidades que aparecem na imagem, assim, não fazem referência a tipos, mas fazem referência a classes denotadas pelos substantivos que determinam (cf. Neves, 2011). Nesse sentido, parece haver uma preferência, do ponto de vista informacional, pela referência a classes nas legendas. A preferência por descrições genéricas, de classe, fica ainda mais evidente se se considerar o fato de que se observam outras estratégias que constroem essa genericidade (cf. Abbott, 2010), como a modificação por numeral sem artigo definido, como em (15), o sintagma nominal nu, com núcleo singular, como em (16), ou com núcleo plural, como em (17). As três legendas a seguir descrevem a mesma imagem 1066252238.jpg do Framed Multi30k.

Figura 4: Imagem 1066252238.jpg

Fonte: dataset Framed Multi30k.

(15) #1 **Três** meninos andam de bicicleta em uma rua.

(16) #3 **Criança** de blusa vermelha anda de bicicleta por cima de monte de areia na rua.

(17) #4 **Jovens** saltam sobre um monte de terra de construção em uma rua residencial.

A preferência atestada com base na amostra bastante restrita deste artigo deve ser testada nas mais de 15 mil legendas do FramedMulti30k. Assim, duas hipóteses podem ser testadas para explicar esse fato: em primeiro lugar, tem de ser verificado se a ausência de um contexto interacional definido na tarefa de elicitação das legendas pode ter alguma influência na genericidade com que se descrevem os cenários; em segundo lugar, tem de ser testado se a genericidade decorre do fato de as legendas serem peças descritivas conceituais. Essas duas hipóteses não são mutuamente excludentes.

4.2 A gramática em ajustes de categorização

O ajuste de **categorização** tem a ver com o modo pelo qual os indivíduos categorizam as entidades (eventos e coisas) que compõem um todo experiencial. Na Introdução deste artigo, discutiram-se os casos de vieses de gênero e de raça em datasets multimodais. Tais vieses decorrem de escolhas de categorização feitas pelos informantes que produzem as legendas. Ainda no caso das legendas, é interessante notar que o ajuste de categorização é responsável pela granularidade com que um determinado referente ou evento é construído linguisticamente. Viridiano (2024, p. 30) compara as sentenças "O lateral esquerdo fez um gol de bicicleta" e "O jogador chutou a bola", como sendo sentenças que podem descrever uma mesma imagem com diferentes graus de detalhamento. A questão, em si, não é detalhamento propriamente, mas é a maneira pela qual as entidades (coisas e eventos) são categorizadas. A categorização é um processo altamente sensível a questões contextuais, discursivas e interacionais (Mauri, 2021).

No caso da categorização dos eventos, é notável a variação nas legendas. Por um lado, há casos em que, para uma mesma imagem, o evento é categorizado como sendo o mesmo pelos cinco informantes, mas há variação léxico-construcional no modo pelo qual o evento é

expresso. É o caso das legendas em (18), para a imagem 106691539.jpg do Framed Multi30k. Em quatro das cinco legendas, descreve-se uma equipe médica realizando uma operação. Nas legendas #1, #3 e #4, em vez de um verbo simples, como *operar*, os falantes verbalizam o evento por meio de uma construção de verbo suporte, formada do verbo *realizar* e do substantivo deverbal *operação*, ou *procedimento cirúrgico*. Novamente, precisam ser investigados fatores que regulam esse tipo de escolha. Um dos possíveis motivos é o fato de que, enquanto uma verbalização com *operar* requer uma construção transitiva, com dois argumentos, uma verbalização com verbo suporte requer apenas um argumento. Na imagem, não aparece o paciente, assim, a verbalização com *operar* pode ser despreferida uma vez que o argumento paciente necessariamente deve aparecer no predicado.

Figura 5: Imagem 106691539.jpg

Fonte: dataset Framed Multi30k.

(18) Legendas da imagem 106691539.jpg

#1 Uma equipe médica realiza uma operação.

#2 Três pessoas com vestes médicas, toucas e máscaras estão participando de uma operação, na mesa há muitos instrumentos médicos.

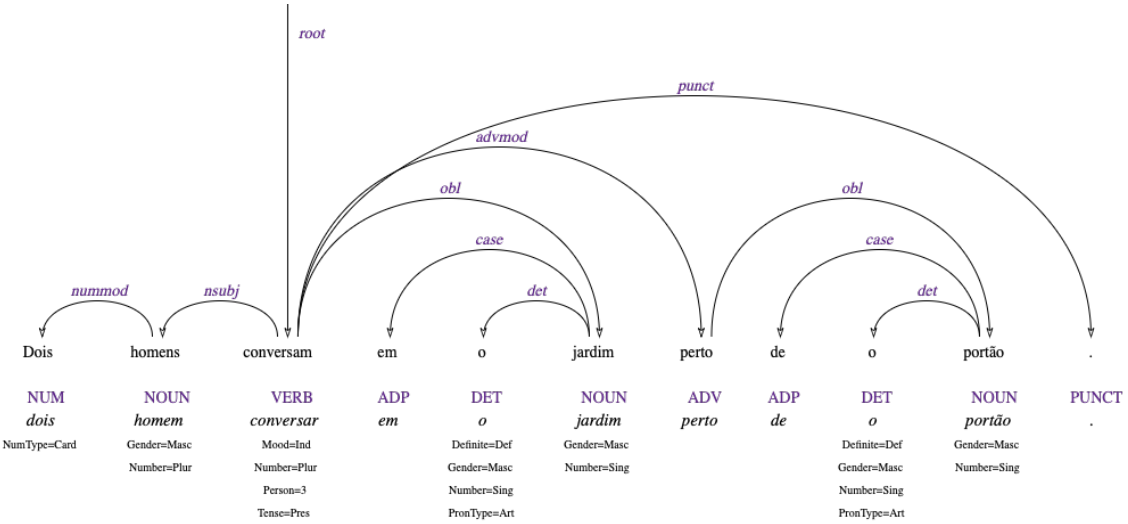
#3 Um médico realiza uma cirurgia com dois auxiliares, em uma sala de cirurgia.

#4 Um veterinário realiza um procedimento cirúrgico.

#5 Médicos manipulando materiais de cirurgia.

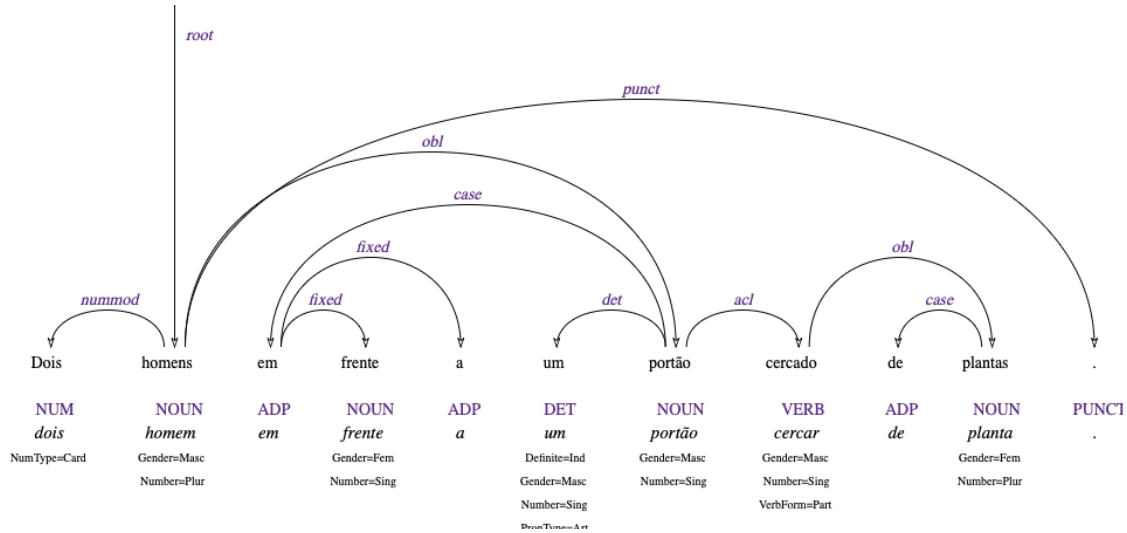
A análise dos meios de expressão de um evento nas legendas em (18) pode ser estendida para uma outra dimensão igualmente relevante, que diz respeito ao tipo de construção de predicado das legendas, como se ilustrou inicialmente na Introdução com as legendas da Figura 1. Os tipos de predicado são construções gramaticais que permitem aos usuários de uma língua fazer a categorização de tipos esquemáticos de evento. Acontece que cada construção gramatical apresenta sua própria organização quanto às relações de dependência interna entre as suas partes componentes. No caso de um predicado verbal, o núcleo (*root*) da oração é sempre o verbo, ao passo que num predicado não verbal adverbial, o núcleo (*root*) é o advérbio). As figuras 6 e 7, respectivamente, ilustram esses casos para as legendas #1 e #5, de (01), retomadas em (19) e (20).

Figura 6: Árvore com anotação das UDs para a legenda #1 da imagem 1000092795.jpg em (19).



Fonte: Elaboração própria.

Figura 7: Árvore com anotação das UDs para a legenda #5 da imagem 1000092795.jpg em (20).



Fonte: Elaboração própria.

(19) #1 Dois homens conversam no jardim perto do portão.

(20) #5 Dois homens em frente a um portão cercado de plantas.

Ao fim desta seção, é importante fazer uma conexão entre os ajustes de categorização e os vieses apontados no início deste artigo. O que os vieses revelam do ponto de vista linguístico não é outra coisa senão um modelo *folk* (ou de senso comum) implicado nos procedimentos de categorização das entidades retratadas em uma imagem.

4.3 A gramática em ajustes de orientação

O ajuste de **orientação** diz respeito às diversas ancoragens de uma ideia. A ancoragem pode ser temporal, espacial, epistemológica, emocional. As línguas diferem quanto aos recursos linguísticos disponíveis para fazer orientação: algumas línguas apresentam mais ou menos tempos verbais gramaticalizados, ou expressões e construções evidenciais, locativas etc. Interessa, nessa questão, a orientação que se observa nas legendas. Como elas são peças textuais descritivas (seção 5), quando uma legenda apresenta um verbo na forma finita, é muito mais provável que essa forma apareça no presente do que em qualquer outro tempo.

Novamente, nesse ponto, a anotação das UDs pode auxiliar na análise. Para a extração de informações sobre essas classes, no dataset anotado, foram isoladas as *feature* `VerbForm:Fin` (forma verbal finita), que pode apresentar os valores de `Tense` (tempo verbal), no `Mood:Ind` (modo indicativo)[9]. O Quadro 2 apresenta os resultados dos valores para `Tense`. Os resultados da anotação mostram que de 161 formas verbais finitas nas 150 legendas, 160 são no presente e 1 no pretérito.

Quadro 2: Número de ocorrência da feature Tense no dataset deste estudo.

Valores da feature `Tense`	Número de ocorrência
`Tense:Pres` (presente)	160
`Tense:Past` (passado)	01

Fonte: Elaboração própria.

A orientação não tem a ver apenas com a ancoragem temporal dos eventos. Mas é possível, também, orientar ideias pela epistemologia, segundo a qual eventos são avaliados como certos ou prováveis ou improváveis. Tal orientação pode ser construída linguisticamente por meio de expressões modais (Neves, 2006), por exemplo. No caso da amostra de legendas analisada neste trabalho, não se encontrou nenhuma ocorrência de expressão desse tipo. Por hipótese, pode-se esperar que, na totalidade de legendas do FramedMulti30k, não sejam encontradas expressões assim, uma vez que as descrições das legendas sempre se constroem como factuais.

4.4 A gramática em ajustes de combinação e linearização

A **combinação** e a **linearização** são ajustes que decorrem da natureza discursiva e sequencial da linguagem. A linearização é um aspecto fundamental das legendas, uma vez que diz respeito à disposição sequencial de informações numa sentença (seção 3). A combinação é um aspecto igualmente fundamental na medida em que diferentes unidades de ideias são combinadas em uma única sentença.

Nesse modo de condução, é necessário fazer uma ressalva. O modelo de verbalização que Chafe e Croft desenvolvem está ligado à experiência que já está armazenada e que

[9] Não houve ocorrência de nenhum outro modo verbal.

é recuperada para a construção de um discurso no cenário interacional. Nesse ponto, a produção das legendas é diferente, como se discute na seção 3. As legendas são produzidas em contextos controlados de elicitação, fora de um contexto interacional. E elas não resultam de experiência armazenada, mas de experiência (estímulo visual) no momento de sua produção. Essa configuração, no entanto, não invalida o modelo de verbalização da experiência tal como proposto pelos autores. O modelo de verbalização precisa dar conta da dimensão *on-line* da produção das legendas, de modo que seja possível mapear as consequências do fato de ser produção de linguagem simultaneamente à experimentação de um estímulo, para a produção linguística de legendas descritivas, como se discutiu na seção 3.

5 A natureza descritiva das legendas de imagem

Nesta seção, desenvolve-se a ideia de que as legendas são peças textuais, no sentido semântico do termo (Halliday e Hasan, 1976), caracterizadas por serem exclusivamente descritivas. Se, por um lado, a variação léxicogramatical é um aspecto escancarado na construção linguística das legendas, o acionamento de operações descritivas parece ser o mesmo na macroestrutura das legendas. Para desenvolver essa proposta, nesta seção, mobilizam-se pressupostos teóricos no âmbito da Linguística do Texto (Adam e Petit Jean, 1989; Adam, 2018).

Do ponto de vista da organização textual, a descrição é um tipo de estrutura sequencial que, diferentemente de outras sequências (como a narrativa, a dissertativa ou a expositiva), é menos rígida, não apresentando um agrupamento ordenado de proposições ou sequências (Adam, 2018). Isso não significa que a descrição não apresente algum encadeamento ou agrupamento lógico de ideias. Comparem-se as legendas em (21) e (22), ambas da imagem na Figura 8.

Figura 8: Imagem 1000344755.jpg

Fonte: dataset Framed Multi30k.

(21) #1 Um trabalhador limpa os vidros de um prédio em cima de uma escada.

(22) #3 Pessoa vestindo boné e camiseta azuis, em uma escada escorada em um edifício, limpando janelas de vidro com um acessório.

Claramente as duas legendas diferem no que diz respeito à sua constituição lexical e gramatical, na esteira da discussão conduzida até aqui. Mas elas também diferem num outro aspecto até então não discutido neste artigo, o qual diz respeito à distribuição da informação, ou linearização, na sentença. Em (21) e (22) a legenda se inicia ancorada no indivíduo, operação descritiva de tematização, como se discute a seguir, mas a ordem dos demais elementos, a ação de limpar a janela e estar em uma escada, varia. Em (1), a ordem das partes componentes responde a ordem de constituintes não marcada em português (c), SVO-Adposição, enquanto em (22), a ordem das partes constituintes já corresponde à ordem marcada, em que o sintagma preposicionado *em uma escada escorada em um edifício* aparece entre o sujeito (*pessoa vestindo boné e camiseta azuis*) e o sintagma verbal (*limpando janelas de vidro com um acessório*).

A diferença na organização da informação observada em (21) e (22) tem a ver com a linearização da informação na sentença, particulamente com a ordenação das partes componentes da sentença.

Um caso semelhante de diferença na organização da informação é o das legendas em (23) e (24), da Figura 9. No entanto, a diferença na organização da informação tem a ver com a organização de planos descritivos, isto é, a configuração de figura e fundo (Talmy, 2000), primeiro e segundo planos, respectivamente. Em (23), a ação de *polvilhar açúcar*, com a forma verbal finita, é construída linguisticamente em primeiro plano, ao passo que em (24), ela é construída linguisticamente em segundo plano, com a forma verbal no gerúndio.

Figura 9: Imagem 100207720.jpg

Fonte: dataset Framed Multi30k.

(23) #1 Mulher de óculos e blusa preta na cozinha polvilhando açúcar em um bolo de chocolate.

(24) #3 Uma mulher de cabelos castanhos decora um bolo de chocolate polvilhando açúcar de confeiteiro sobre ele.

A respeito da questão da ordem em sequências descritivas, Adam (2018, p. 85) diz que

> Na medida em que o protótipo da sequência descritiva não há nenhuma indicação de ordem [...], não comporta linearidade intrínseca que lhe permite estar (ou não) em sintonia com a linearidade própria da linguagem articulada, *as organizações períodicas, os planos de textos e suas marcas específicas têm uma importância decisiva para a legibilidade e para a interpretação de toda descrição* (Adam, 2018, p. 85, grifo próprio).

Aí está, pois, a relevância de se considerar as legendas de imagens do ponto de vista linguístico-textual. A ideia de que "organizações periódicas, planos de textos e suas marcas específicas" é fundamental para compreender a unidade no processo de criação das legendas. A defesa que se faz aqui é que tal unidade decorre do fato de que todas as legendas, pelo menos aquelas que compõem a mostra deste estudo, são construídas com base na **operação descritiva de tematização por ancoragem referencial** (Adam, 2018), na qual "a sequência descritiva assinala desde o início quem ou que vai estar em questão" (p. 85).

No caso específico da amostra de legendas deste estudo, a ancoragem referencial, em todos os casos, são entidades animadas, humana ou não humana. Isso se explica, pelo menos em partes, pela hierarquia de animacidade (Comrie, 1989; Croft, 2003),[10] em que entidades animadas tendem a ser mais focais e proeminentes do que entidades não animadas. Assim, é pouco provável que um falante construa uma legenda como (25'), uma vez que a legenda em (25) parece manter a proeminência nas entidades humanas.

(25) #1 Um grupo de pessoas em um equipamento de construção (10002456.jpg).

(25') Um equipamento de construção com um grupo de pessoas.

Na amostra de 150 legendas deste estudo, são apenas dois os casos em que não vêm tematizada uma entidade animada, como em (26) e (27), na qual, pela ordem linear, está tematizado o lugar.

(26) #2 Em frente a um prédio de cimento onde é possível ver o número 23 encontra-se duas jovens vestindo preto e cinza e preto e um jovem vestindo preto próximo a outra pessoa cuja apenas as mãos são visíveis (1001465944.jpg).

(27) #2 Perto do carro preto, três meninos andam de bicicleta na rua enquanto um deles, de capacete azul, está em cima de um monte de areia (1065323785.jpg).

Compare-se esse caso em (26) com as outras legendas, em (28), produzidas para a mesma imagem, na Figura 10. Note-se que nas quatro legendas em (28) a descrição é tematizada pelas entidades humanas da imagem.

[10] Escala de animacidade estendida, segundo Croft (2003, p. 130): pronomes de primeira/segunda pessoa < pronome de terceira pessoa < nomes próprios < substantivo comum humano < substantivo comum não humano animado < substantivo comum inanimado.

Figura 10: Imagem 1001465944.jpg

Fonte: dataset Framed Multi30k.

Figura 11: Imagem 1065323785.jpg do dataset

Fonte: dataset Framed Multi30k.

(28) Quatro legendas da imagem 1001465944.jpg
#1 Um rapaz de preto olha para uma mulher na rua.
#3 Uma mulher e um homem vestidos de roupas pretas na calçada enquanto uma mulher de cabelos curtos e roupa casual passa pela rua.
#4 Um homem e uma mulher trocam olhares na rua ao passarem um pelo outro em uma fila demarcada por gradis.
#5 Um grupo de jovens aguarda em uma fila atrás de uma grade de proteção.

Em todos esses casos, pode-se dizer que as entidades humanas são focais na imagem, ou seja, têm alguma proeminência. No entanto, considere-se o caso da legenda em (29), da imagem na Figura 12, como um exemplo de peça descritiva em que estão tematizadas entidades humanas, enquanto uma entidade não animada (*os prédios*) é construída como focal. As outras quatro legendas, em (30), tematizam as entidades humanas e as constroem como ponto focal da imagem.

Figura 7: Imagem 100197432.jpg

Fonte: dataset Framed Multi30k.

(29)#2 Através de um grupo de jovens mulheres brancas na rua vê-se prédios próximos com alturas, cores e arquiteturas distintas (100197432.jpg).

(30) Quatro legendas da imagem 100197432.jpg

#1 Pessoas passam por uma rua em frente a um ônibus e prédios.

#3 Duas mulheres loiras e uma asiática em frente a uma fachada de prédios enquanto passa um ônibus na rua.

#4 Um grupo de mulheres caminha por uma rua movimentada de uma cidade, passando diante de um ônibus.

#5 Um grupo de pessoas caminha em meio aos prédios de uma cidade.

Diante dos fatos apresentados nesta seção, abre-se uma possível nova frente de investigação sobre a configuração linguística das legendas de imagem no que diz respeito às restrições (ou preferências) semânticas no empacotamento linguístico da informação imagética. A operação descritiva de tematização por ancoragem referencial parece ser a operação preferida para a composição das legendas cujas imagens apresentam indivíduos animados, ainda que esses indivíduos não sejam, necessariamente, os elementos focais da imagem. De todo modo, essa operação parece estar com confirmidade com o que se sabe a respeito de preferências tipológicas de se por em proeminência, nos enunciados, entidades animadas, a chamada escala de animacidade.

Considerações finais

Este artigo buscou ancorar linguisticamente a discussão a respeito de legendas de imagens no contexto do PLN, ao mesmo tempo em que tentou mostrar de que maneira métodos de PLN, como modelos de anotação linguística automática, podem contribuir para a análise linguística. A discussão empreendida, aqui, equacionou gramática, cognição e texto para chegar a uma explicação da variabilidade gramatical da construção das legendas, bem como da regularidade de operações e processos cognitivo-textuais que estão na base dessas peças textuais.

Um dos argumentos pivotais desenvolvidos neste artigo é o de que legendas descritivas de imagens são peças textuais linguísticas que estão submetidas aos mesmos processos de verbalização da experiência aos quais estão a produção de linguagem em contextos interacionais (Chafe, 2002, 2005; Croft, 2007, 2010, 2020). A verbalização diz respeito ao processo pelo qual informação cognitiva, visual, no caso das legendas de imagens, é empacotada linguisticamente.

A premissa na base deste argumento pode ser sumarizada com a citação a seguir,

> Os enunciados não são registros verbais diretos de eventos. Um evento não pode ser plenamente representado na linguagem, pois a expressão linguística exige algum grau de esquematização. Cada enunciado representa uma seleção de características, cabendo ao ouvinte completar os detalhes com base no contexto em curso e no conhecimento prévio. Parte desse conhecimento prévio inclui a compreensão do que é obrigatório ou típico na língua em uso (Slobin, 2003, p. 159).[11]

Note-se que a proposição de Slobin (2003) tem na sua base a relação entre falante e ouvinte. Obviamente essa relação não está prevista diretamente na produção das legendas, devido ao contexto artificial de experimentação no qual essa produção se insere (seção 3). De todo modo, é exatamente a explicitação desse conhecimento linguístico prévio que as análises propostas aqui buscaram explicitar. Como se mostrou, diferentes expedientes construcionais relacionam-se a diferentes processos de ajustes entre uma ideia e um enunciado linguístico (seções 4 e 5).

As análises apresentadas neste estudo necessitam ser ampliadas para uma amostra mais abrangente de legendas, compreendendo as mais de 15.000 legendas originais e traduzidas que compõem o Framed Multi30k. Além disso, são requeridos experimentos psicolinguísticos que possam validar algumas das propostas de verbalização apresentadas aqui.

Referências

ABBOTT, B. *Reference*. Oxford: Oxford University Press, 2010.

ADAM, J. M.; PETIT JEAN, A. *Le texte descriptif*. Paris: Édition Nathan, 1989.

[11] Texto original: "Utterances are not verbal filmclips of events. An event cannot be fully represented in language: linguistic expression requires schematization of some sort. Every utterance represents a selection of characteristics, leav- ing it to the receiver to fill in details on the basis of ongoing context and background knowledge. Part of the background is a knowledge of what is obligatory or typical of the language being used".

ADAM, J. M. *Textos*: tipos e protótipos. Tradução de Monica Cavalcante. São Paulo: Editora Contexto, 2018.

BASILE, V. et al. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 39, n. 1, p. 1-17, 2023. Disponível em: h https://ojs.aaai.org/index.php/AAAI/article/view/25840. Acesso em: 14 abr. 2025.

BELCAVELLO, F. *FrameNet Annotation for Multimodal Corpora***:** Devising a Methodology for the Semantic Representation of Text-image Interactions in Audiovisual Productions. 2023. 134f. Tese (Doutorado em Estudos Linguísticos) – Universidade Federal de Juiz de Fora, 2023. Disponível em: https://repositorio.ufjf.br/jspui/handle/ufjf/15527. Acesso em: 14 abr. 2025.

BYBEE, J. What is Usage-based Linguistics? *In*: DÍAZ-CAMPOS, M.; BALASCH, S. (orgs.) *The Handbook of Usage-based Linguistics.* New York: Wilet Blackwell, 2023. p. 9-30.

CHAFE, W. Some Thoughts on Schemata. *In*: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings TNLAP'75*, p. 89-91, 1975a.

CHAFE, W. Creativity on Verbalization as Evidence for Analogic Knowledge. *In*: DEPARTAMENT OF LINGUISTICS. *Proceedings TNLAP'75*, p. 144-145, 1975a.

CHAFE, W. Creativity in Verbalization and its Implications for the Nature of Stored Knowledge. *In*: FREEDLE, R. O. (org.) *Discourse Production and Comprehension*. Norwood: Ablex, 1977a. p. 41-55.

CHAFE, W. The Recall and Verbalization of Past Experience. *In*: COLE, R. (org.) *Current Issues in Linguistic Theory*. London: Indiana Univeristy Press, 1977b. p. 215-246.

CHAFE, W. Things we can Learn from Repeated Tellings of the Same Experience. *Narrative Inquiry*, v. 8, n. 2, p. 269-285, 1998.

CHAFE, W. Putting Grammaticalization in its Place. *In*: WISCHER, Ilse; DIEWALD, Gabriele (org.) *New Reflections on Rrammaticalization*. Amsterdam: John Benjamins, 2002. p. 395-412.

CHAFE, W. The Relation of Grammar to Thought. *In*: BUTLER, C. S.; GÓMEZ-GONZÁLLEZ; M. de los Á.; DOVAL-SUÁREZ, S. (org.) *The Dynamics of Language Use*. Amsterdam: John Benjamins: 2005. p. 57-78.

CHELLIAH, S. L.; DE REUSE, W. J. *Handbook of Descriptive Linguistic Fieldwork.* New York: Springer, 2011.

CLARK, H. H. *Arenas of Language Use*. Cambridge, UK: Cambridge University Press, 1992.

CLARK, E. Languages and Representations. *In*: GENTNER, D.; GOLDIN-MEADOW, S. (orgs.) *Language in Mind*: Advances in the Study of Language and Thought. Cambridge, MA: The MIT Press, 2003. p. 13-27.

COMRIE, B. *Language Universals and Linguistic Typology*. 2. ed. Chicago: The University of Chicago Press, 1989.

CONEGLIAN, A. V. L. O modelo das dependências universais: assentando bases teóricas e revisando diretrizes metodológicas. *Revista da Abralin*, v. 23, n. 2, p. 187-214, 2023.

CROFT, W. Typology and Universals. 2. ed. Cambridge, UK: 2012.

CROFT, W. The Origins of Grammar in the Verbalization of Experience. *Cognitive Linguistics*, v. 18, n. 3, p. 339-382, 2007.

CROFT, W. The origins of grammaticalization in the verbalization of experience. *Linguistics*, v. 48, n. 1, p. 1-48, 2010.

CROFT, W. *Ten lectures on construction grammar and typology*. New York: Brill, 2020.

DE MARNEFFE, Marie-Catherine *et al*. Universal dependencies. *Computational Linguistics*, v. 47, n. 2, p. 255-308, 2021.

DIK, S. *The theory of functional grammar*. 2. ed. Berlin: Mouton de Gruyter, 1997.

FRAMED MULTI30K. Imagem 1000092795.jpg. *In*: FRAMED MULTI30K. Banco de dados. [*S. l.*]: [*S. n.*]. Disponível em:. Acesso em: https://github.com/FrameNetBrasil/framed-multi30k. Acesso em: 25 abr. 2025.

GIVÓN, T. *Functionalism and grammar*. Amsterdam: John Benjamins, 1995.

HALLIDAY, M.; HASAN, R. *Cohesion in English*. London: Longman, 1976.

KEMMER, S.; BARLOW, M. Introduction: a usage-based conception of language. *In*: BARLOW, M.; KEMMER, S. (orgs.) *Usage-based models of language*. Sanford: CSLI Publications, 2000. p. vii-xxviii.

JAIMES, A.; CHANG, S.-F. A Conceptual Framework for Indexing Visual Information at Multiple Levels. *IS&T/SPIE Internet Imaging*, v. 3964, s.p., 2000.

MAURI, C. Ad hoc categorization in linguistic interaction. *In*: MAURI, C. *et al*. (org.) *Building categories in interaction*: linguistic resources at work. Amsterdam: John Benjamins, 2021. p. 9-34.

MILTENVERG, E. Stereotyping and bias in the Flickr30K dataset. *In*: MULTIMODAL CORPORA: COMPUTER VISION AND LANGUAGE PROCESSING (MMC 2016). *Proceedings* [...]. [*S. l.*] p. 1-4, 2016. Disponível em: http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-MCC-2016-proceedings.pdf, Acesso em: 14 fev. 2024.

NEVES, M. H. M. *Gramática de usos do português*. 2. ed. São Paulo: Editora Unesp, 2011.

PARDO, T *et al*. Porttinari: a Large Multi-genre Treebank for Brazilian Portuguese. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DE INFORMAÇÃO E DA LINGUAGEM HUMANA. *Anais*... Porto Alegre: SBC, 2021. p. 1-10.

PEZZATI, E. *A ordem das palavras no português*. São Paulo: Parábola, 2014.

RINKE, E. A combinação de artigo definido e pronome possessivo na história do português. *Estudos de Linguística Galega*, v. 2, p. 121-139, 2010.

SAMARIN, W. J. *Field Linguistics*. New York: Holt, 1967.

SLOBIN, D. Language and Thought Online: Cognitive Consequences of Linguistic Relativity. *In*: GENTNER, D.; GOLDIN-MEADOW, S. (org.) *Language in Mind*: Advances in the Study of Language and Thought. Cambridge, MA: The MIT Press, 2003. p. 157-192.

STRAKA, M. *et al*. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. *In*: CALZOLARI, Nicoletta *et al*. (org.) *c* (LREC'16). European Language Resources Association, 2016. p. 4290-4297. Disponível em: https://aclanthology.org/L16-1680.pdf. Acesso em: 08 jan. 2025.

TALMY, L. *Toward cognitive semantics*. Cambridge, MA: The MIT Press, 2000. 2 v.

VIRIDIANO, M. *Framed Multi30k*: um dataset multimodal-multilíngue baseado em semântica de frames. 2024. 107f. Tese (Doutorado em Estudos Linguísticos) – Universidade Federal de Juiz de Fora, 2024. Disponível em: https://repositorio.ufjf.br/jspui/handle/ufjf/16854. Acesso em: 14 abr. 2025.

VIDIRIANO, M. *et al*. Framed Multi30k: a frame-based multimodal multilingual dataset. *In*: THE 2024 JOINT INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, LANGUAGE RESOURCES AND EVALUATION (LREC-COLING 2024). *Proceedings* [...]. Torino, Italia, 2024. p.7438-7449. Disponível em: https://aclanthology.org/2024.lrec-main.656.pdf. Acesso em: 14 abr. 2025.

YOUNG, P. *et al*. From image descriptions to visual descriptions: new similarity metrics for semantic inference over event descriptions. *Transactions for Computational Linguistics*, v. 2, p. 67-68, 2014.