UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

Marcelo Barros Custódio

Framed Multi30K: Um dataset multimodal-multilíngue baseado em semântica de frames

Juiz de Fora
2024

Marcelo Barros Custódio

Framed Multi30K: Um dataset multimodal-multilíngue baseado em semântica de frames

Tese de doutoramento apresentada ao programa de Pós-Graduação em Linguística da Faculdade de Letras da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Doutor em Linguística. Área de concentração: Linguística.

Orientador: Professor Doutor Tiago Timponi Torrent

Juiz de Fora

2024

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Custódio, Marcelo Barros.
Framed Multi30K: Um dataset multimodal-multilíngue baseado em semântica de frames / Marcelo Barros Custódio. – 2024.
106 f. : il.

Orientador: Tiago Timponi Torrent
Tese (Doutorado) – Universidade Federal de Juiz de Fora, Faculdade de Letras. Programa de Pós-Graduação em Linguística, 2024.

1. Semântica de Frames. 2. Dataset Multimodal. 3. Representação Semântica Multimodal. I. Torrent, Tiago Timponi, orient. II. Título.

Marcelo Barros Custodio

Framed Multi30K: Um dataset multimodal-multilíngue baseado em semântica de frames

Tese apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de doutor em linguística. Área de concentração: linguística.

Aprovada em 16 de julho de 2024.

BANCA EXAMINADORA

Prof. Dr. Tiago Timponi Torrent - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Ely Edison da Silva Matos
Universidade Federal de Juiz de Fora

Profa. Dra. Aline Alves Fonseca
Universidade Federal de Juiz de Fora

Profa. Dra. Adriana Silvina Pagano
Universidade Federal de Minas Gerais

Profa. Dra. Helena de Medeiros Caseli
Universidade Federal de São Carlos

Juiz de Fora, 02/07/2024.

Documento assinado eletronicamente por **Tiago Timponi Torrent**, **Coordenador(a)**, em 16/07/2024, às 11:55, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Aline Alves Fonseca**, **Professor(a)**, em 16/07/2024, às 11:56, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ely Edison da Silva Matos**, **Técnico Administrativo em Educação**, em 16/07/2024, às 11:58, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Adriana Silvina Pagano**, **Usuário Externo**, em 16/07/2024, às 12:04, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Helena de Medeiros Caseli**, **Usuário Externo**, em 16/07/2024, às 12:04, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **1844879** e o código CRC **668F1669**.

Dedicado a todos os *framenetters* — passados, presentes e futuros — que através do seu legado acadêmico impulsionam não apenas os limites da ciência, mas também a vida de pesquisadores em todo o mundo (incluindo a minha).

AGRADECIMENTOS

Embora apenas meu nome apareça na capa desta tese, eu jamais teria concluído este trabalho sem o apoio das pessoas que estiveram ao meu lado durante estes anos.

Quero agradecer, primeiramente, ao meu orientador, Tiago Torrent, por seu incentivo e apoio constantes. Tiago foi um mentor paciente e generoso, que me ensinou não apenas sobre ética profissional e o propósito da pesquisa acadêmica, mas também sobre a importância de oferecer oportunidades extraordinárias a pessoas comuns (como eu).

Obrigado a todos os *framenetters* pelas discussões e colaborações, e a todos os coautores que publicaram comigo ao longo desses anos, e especialmente a Frederico Belcavello, Ely Matos e Arthur Lorenzi. Agradeço também a Natália Sigiliano pelo empréstimo do seu "caderno mágico", sem o qual eu sequer teria tido a confiança para começar a trilhar este caminho. Trabalhar ao lado de vocês fez de mim um pesquisador melhor e, o que é mais importante, uma pessoa melhor.

Obrigado também à FAPEMIG, por financiar as iniciativas de pesquisa do laboratório FrameNet Brasil (bolsas RED00106/21, CNPq 408269/2021-9 e 420945/2022-9) que permitiram as anotações do dados desta tese, e à CAPES, por financiar não apenas os anos de estudo que deram origem a esta tese (bolsa número 88887.816219/2023-00), mas também meu período como pesquisador visitante na Universidade de Leipzig, na Alemanha (bolsa número 88887.628830/2021-00). Sou também grato a Nicolás Hinrichs, Maryam Foradi, Felix Hoberg, e demais colegas da Universidade de Leipzig, por me acolherem e me fazerem sentir bem-vindo, e especialmente grato a Oliver Czulo, meu orientador estrangeiro, pelas oportunidades acadêmicas e, acima de tudo, por me ajudar a não ser deportado da Alemanha.

Obrigado aos membros da minha banca de qualificação e banca examinadora, por dedicarem seu tempo para ler e comentar esta tese.

Obrigado aos amigos – em especial a Julio Cassé, Mairon Samagaio, Ivanovna Marinho e Emma Gallardo Richards – por estarem sempre presentes.

Finalmente, obrigado à minha mãe, pelos milhares de pequenos gestos de carinho e atenção durante toda a vida, e à Waleska, por seu amor e cuidado incondicionais, hoje e para sempre.

"A few words can change our minds, change our marital status, or change our religion. Words affect who we are. As a species, language is our most powerful and pervasive tool. With language, we can communicate what we think and who we are. Without language, we would be isolated. We would have no fiction, no history, and no science. To understand how meaning works, then, is to understand part of what it is to be human."

– Benjamin Bergen, "Louder Than Words: The New Science of How the Mind Makes Meaning.'

RESUMO

A combinação de diferentes modalidades de comunicação é uma das características definidoras da expressão humana, no entanto, muitas pesquisas voltam seus esforços para a análise da semântica textual e imagética de forma isolada. Nas últimas décadas, sistemas que processam dados de imagem e texto de forma correlacionada vêm sendo aplicados em tarefas computacionais como recuperação de dados, tradução automática e criação de legendas. Neste trabalho, partimos da premissa de que o desenvolvimento de tais aplicações computacionais pode se beneficiar de um melhor entendimento dos significados que se estabelecem a partir de combinação de informações textuais e visuais. Em particular, esta tese contribui com um *dataset* inovador que agrega a uma família de datesets padrão ouro para o PLN multimodal e multilíngue – Flickr30k, Multi30k e Flickr30k Entities – informações semânticas estruturadas em termos de frames, conforme modelados na FrameNet Brasil. O *dataset* resultante, denominado Framed Multi30k (FM30K), contribui os seguintes dados inovadores: (i) 150.000 descrições conceituais originalmente redigidas em português brasileiro para cada uma das 30.000 imagens no *dataset* Flickr30k; (ii) 30.000 traduções para o português brasileiro de uma das descrições originalmente escritas em inglês para cada uma das imagens no Flickr30k; (iii) anotações automáticas para frames de todas as descrições conceituais constantes do *dataset* para o português brasileiro e para o inglês, totalizando 330.000 descrições anotadas semanticamente; (iv) anotações manuais para cada uma das *bounding boxes* provenientes do *dataset* Flickr30k Entities em três condições de anotação distintas: anotação de entidades com presença de descrição, anotação de entidades sem presença de descrição e anotação de eventos com presença de descrição. O *dataset* resultante foi analisado para aspectos formais das descrições criadas em português brasileiro e para a similaridade de cosseno entre as representações semânticas derivadas das anotações automáticas e manuais realizadas para as descrições e imagens, respectivamente. Foram realizadas adicionalmente análises qualitativas acerca das distintas perspectivas codificadas nas representações semânticas geradas para as imagens em cada uma das condições de anotação. As análises corroboram a hipótese de que diferentes condições de anotação no que concerne à interação entre modalidades levam a distintas representações semânticas para as imagens, o que reforça o argumento em favor da adoção de uma abordagem perspectivista para a curadoria humana de *datasets*.

Palavras-chave: 1. Semântica de Frames. 2. Dataset Multimodal. 3. Representação Semântica Multimodal.

ABSTRACT

The combination of different communication modalities is one of the defining characteristics of human expression. However, much research has focused on analyzing the semantics of text and images separately. In recent decades, systems that process image and text data in a correlated way have been applied to computational tasks such as data retrieval, machine translation, and caption generation. In this work, we start from the premise that the development of such computational applications can benefit from a better understanding of the meanings that are established from the combination of textual and visual information. In particular, this dissertation contributes an innovative dataset that adds to a family of gold standard datesets for multimodal and multilingual NLP - Flickr30k, Multi30k and Flickr30k Entities - semantic information structured in terms of frames, as modeled in FrameNet Brasil. The resulting dataset, called Framed Multi30k (FM30K), contributes the following innovative data: (i) 150,000 conceptual descriptions originally written in Brazilian Portuguese for each of the 30,000 images in the Flickr30k dataset; (ii) 30,000 translations into Brazilian Portuguese of one of the descriptions originally written in English for each of the images in the Flickr30K; (iii) automatic annotations for frames of all the conceptual descriptions in the dataset into Brazilian Portuguese and English, totaling 330,000 semantically annotated descriptions; (iv) manual annotations for each of the bounding boxes from the dataset Flickr30k Entities in three different annotation conditions: annotation of entities with presence of description, annotation of entities without presence of description and annotation of events with presence of description. The resulting dataset was analized for formal aspects of the descriptions created in Brazilian Portuguese and for the cosine similarity between the semantic representations derived from the automatic and manual annotations carried out for the descriptions and images, respectively. Qualitative analyses were also carried out on the different perspectives encoded in the semantic representations generated for the images in each of the annotation conditions. The analyses corroborate the hypothesis that different annotation conditions regarding the interaction between modalities lead to different semantic representations for the images, which reinforces the argument in favor of adopting a perspectivist approach to human curation of datasets.

Keywords: Frame Semantics. Multimodal Dataset. Multimodal Semantic Representation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DET	Traduções em alemão das descrições em inglês (*Deutsch Translations*)
EF	Elemento de Frame
ENO	Descrições originais em inglês (*English Originals*)
FM30K	Framed Multi30K
IcD	Imagens na presença de sua descrição
IcD-EV	Imagem na presença de sua descrição, para frame de Evento
IsD	Imagens na ausência de sua descrição
PLN	Processamento de Língua Natural
PTO	Descrições originais em português (*Portuguese Originals*)
PTT	Traduções em português das descrições em inglês (*Portuguese Translations*)
UL	Unidade Lexical

SUMÁRIO

1 INTRODUÇÃO

Grande parte das pesquisas atuais em áreas como Processamento de Língua Natural (PLN) e Inteligência Artificial (IA) se baseiam, em alguma medida, na utilização de algoritmos de aprendizado de máquina. Todos os modelos de dados – desde os baseados em regras, passando por modelos estatísticos, redes neurais convolucionais (CNNs) e recorrentes (RNNs) e, mais recentemente, redes de *transformers* – têm sido desenvolvidos a partir de dados produzidos por humanos, seja através do uso de conjuntos de dados coletados com a colaboração ativa de anotadores, como no caso das tarefas de *crowdsourcing*, ou extraídos automaticamente da web, via "*web scraping*".[1]

Metodologias como estas, que funcionaram bem no passado, começam, no entanto, a mostrar seus limites. Pesquisas recentes (NAVEED et al., 2023; HOFFMANN et al., 2022) têm apontado para o fato de que as abordagens matemáticas aplicadas sobre dados crus, isto é, sem quaisquer metadados a eles vinculados, têm chegado ao seu limite, não sendo possível melhorar a performance dos sistemas de PLN apenas pelo uso de mais dados. Nesse contexto, pesquisadores da área têm sugerido a necessidade de que grandes volumes de dados crus sejam combinados a dados curados por humanos para a melhoria de desempenho dos sistemas em diversas tarefas (ANTHIS et al., 2024; MØLLER et al., 2024; ROGERS, 2021; BENDER et al., 2021).

A curadoria humana sobre os dados envolve, segundo ROGERS (2021), fazer escolhas sobre o que deve ser incluído em um *dataset*, selecionando elementos com base em fatores como padrões linguísticos, características socioculturais e atributos demográficos, garantindo que diferentes perspectivas e valores sejam representados na composição do conjunto de dados e buscando evitar que os modelos treinados a partir desses dados aprendam padrões ou vieses indesejáveis. Por ser inevitável – considerando que todas as escolhas que fazemos, explicitamente ou implicitamente, serão refletidas na composição desses dados - a questão que se impõe é a de quanto esforço deve ser investido nesse processo de curadoria.

No contexto em que esta tese se desenvolveu – o do Laboratório FrameNet Brasil de Linguística Computacional – a curadoria humana sobre os dados se manifesta, majoritariamente, na forma de anotações semânticas realizadas sobre eles. Desde 2020, o modelo da FrameNet foi expandido para outros modos comunicativos que não sequências de caracteres em textos (BELCAVELLO et al., 2020; TORRENT et al., 2020, 2022; DÁNNELS et al., 2022). O argumento principal por trás dessa expansão é o de que, assim como elementos de língua verbal podem evocar cenas estruturadas chamadas frames, outros modos comunicativos, como imagens, por exemplo, também o fazem.

[1] Em tempo: muito recentemente, o campo do PLN tem debatido os efeitos das IAs gerativas na proliferação de conteúdo gerado por máquina na web. Essa discussão, entretanto, foge ao escopo desta tese.

Isso posto, e visando a contribuir para o desenvolvimento de sistemas de PLN mais eficazes, esta tese teve por objetivo a construção de um *dataset* curado por humanos que agregue metadados semânticos estruturados como frames a pareamentos de imagens e sentenças descrevendo essas imagens. A escolha de imagens como modalidade auxiliar para o desenvolvimento dessa pesquisa é motivada pela ideia de que "a forma como os humanos processam informações é inerentemente multimodal e (...) qualquer sistema computacional que busque obter resultados de processamento de língua semelhante ao humano precisa, necessariamente, processar dados de forma multimodal" (SANABRIA, 2018, p. 1). De um lado mais implementacional, trabalhos em PLN multimodal argumentam que as imagens são, de um modo geral, representações um pouco menos vagas de conceitos que nos cercam, o que as torna candidatas naturais para resolver diversas ambiguidades linguísticas (ELLIOTT et al., 2016, p. 4) como, por exemplo, a inferência sobre informações de gênero para a tradução de uma língua que tenha gênero neutro para uma que tenha gênero gramatical, ou a desambiguação de sentido de substantivos que tenham vários significados nas línguas envolvidas em uma tradução.[2]

Em específico, apresentamos o Framed Multi30k (FM30K), um *dataset* inovador a ser adicionado à família de *datasets* do Flickr30K (YOUNG et al., 2014), ou seja, conjuntos de dados multimodais e multilíngues que são expansões do Flickr30K. Dentre eles, destacam-se o Multi30K (ELLIOTT et al., 2016), *dataset* que acrescenta traduções em alemão para as descrições originais em inglês e novas descrições originais em alemão, e o Flickr30K Entities (PLUMMER et al., 2015), que, além de ampliar o Flickr30K pela criação de correlações imagem-texto – criando *bounding boxes*[3] que relacionam entidades na imagem a seus respectivos descritores nas sentenças, – estabelece cadeias de correferência que relacionam uma mesma entidade com os diferentes sintagmas nominais que a descreve em cada uma das cinco descrições associadas a cada imagem.

O Framed Multi30K, por incorporar metadados semânticos baseados em uma implementação computacional da Semântica de Frames (FILLMORE, 1982), adota o que BASILE et al. (2021) chamam de uma abordagem perspectivista fraca, ou seja, aquela que não se contenta com a busca por um *dataset gold standard* criado a partir da coleta de apenas uma única anotação para cada objeto ou entidade, mas, sim, através da adoção de uma metodologia que integra, de forma mais abrangente e inclusiva, as opiniões e perspectivas dos sujeitos envolvidos na etapa de representação do conhecimento, coletando um maior número de anotações a partir de um grupo variado de anotadores. (BASILE et al., 2021, p. 3)

[2] Contra-exemplos ao caráter menos vago das imagens seriam abundantes. Alguns deles serão discutidos no capítulo 2.

[3] Uma *bounding box* – em português, uma caixa delimitadora – é um elemento visual retangular, definido a partir de quatro coordenadas que representam os pontos extremos do retângulo, utilizado para delimitar uma região de interesse associada a um objeto em uma imagem.

BASILE et al. (2021) compartilham algumas recomendações para adotar uma postura perspectivista na criação de conjuntos de dados, como:

1. Desenvolver tarefas de anotação que permitam aos anotadores associar várias *tags* – rótulos ou categorias – a uma mesma entidade ou objeto, bem como um categoria que possa ser utilizada para marcar exemplos que não se encaixem em nenhum dos casos – para dar conta de perspectivas não previstas, mas que sejam reconhecidas e propostas pelos anotadores, – permitindo também que os anotadores possam sinalizar a inadequação das etiquetas ou categorias disponíveis;

2. Recrutar um número de anotadores que permita a extração de maiorias estatisticamente significativas – por exemplo, pelo menos 12 anotadores para tarefas dicotômicas;

3. Recrutar um grupo heterogêneo de anotadores, tanto no que diz respeito à origem e cultura quanto ao conhecimento e habilidades.

Nesta tese, seguimos tais recomendações e desenvolvemos protocolos de anotação que visam a explorar múltiplas perspectivas na construção das representações semânticas associadas às imagens e descrições textuais, sendo o primeiro deles a própria adoção da FrameNet como repositório de categorias para a anotação. Frames são sistemas de conceitos que, necessariamente, incoporam um ponto de vista sobre a cena que representam. Assim, ao enriquecer a família Flickr30k com anotações de frames e seus elementos, tanto para imagens quanto para as descrições que as acompanham, convertemos um *dataset* considerado padrão ouro em uma abordagem agregacionista em um *dataset* perspectivizado.

A adoção de uma visão perspectivista para a constituição de um *dataset* anotado para frames permite a investigação da medida em que as modalidades comunicativas – imagem e texto – interagem nos processos de produção da significação (MARTINEC & SALWAY, 2005). A hipótese que perseguimos é a de que **diferentes desenhos experimentais para anotação de imagens, no que concerne à interação entre modalidades – ou seja, se a anotação é feita com ou sem presença das descrições – levam a distintas representações semânticas para as imagens**.

Os resultados alcançados corroboram a hipótese, indicando que a presunção de que imagens representam meios de desambiguação incontestáveis de textos em língua verbal, considerada padrão nas abordagens não-perspectivizadas para o PLN multimodal, precisa ser revista, ainda que parcialmente. Como se pontuará no capítulo 5, a anotação de imagens – com ou sem a presença de descrições – para as categorias da Semântica de Frames releva distintas possibilidades de perspectivas sobre os dados, contribuindo para a mitigação de enviesamentos danosos no *dataset* multimodal resultante.

A produção do novo *dataset* tem, ainda, o potencial de impactar diversas pesquisas de PLN que exploram a intercessão entre imagem e texto, investigando questões relacionadas a análise semântica e geração automática de conteúdo, e usam a família Flickr30k para tarefas computacionais, como *Visual Question Answering (VQA), Visual Commonsense Reasoning (VCR), Image Captioning (IC), Video Captioning (VC), Multimodal Machine Translation (MMT)*, dentre outras (UPPAL et al., 2022).

Para além dos aspectos diretamente ligados ao aprimoramento de sistemas de PLN, cabe aqui destacar ao menos duas vantagens em vincular o processamento de línguas naturais ao processamento de cenas visuais (KEVITT, 2003, p. 2): 1. Implementações computacionais que integram percepção visual e processamento de língua natural podem beneficiar pesquisas – por exemplo, no campo da linguística cognitiva – sobre a natureza da cognição humana e; 2. A combinação dessas áreas pode dar origem a metodologias e aplicações que ajudem a solucionar problemas como o da produção automática de textos a partir de imagens, de geração automática de imagens a partir de textos e de interpretação automática de imagens a partir de texto.

Além desse capítulo de introdução e do capítulo de conclusões, esta tese é composta por outros quatro capítulos. No capítulo 2, apresentamos o conceito de *datasets* multimodais e a família do Flickr30k, base para o desenvolvimento do novo conjunto de dados apresentado nesta tese. Discutimos ainda o problema da objetividade nas descrições conceituais. Já no 3, introduzimos o modelo da Semântica de Frames e sua implementação computacional na FrameNet Brasil, relacionando tal aparato teórico-metodológico com os fundamentos das gramáticas multimodais. No capítulo 4, apresentamos os materiais e métodos utilizados para a constituição e análise descritiva do Framed Multi30k, enquanto o capítulo 5 se debruça sobre a análise do *dataset* resultante desta tese.

2 DATASETS MULTIMODAIS

Ao longo da última década, o crescimento no número de *datasets* multimodais vem atraindo atenção de pesquisadores do campo da linguística computacional, que vêm trabalhando na criação e expansão de modelos voltados para o desenvolvimento de tarefas de PLN e Visão Computacional como *Visual Question Answering*, *Visual Commonsense Reasoning*, *Image and Video Captioning*, e *Multimodal Machine Translation*, dentre outras (GARG et al., 2022). No escopo deste trabalho, usaremos o termo multimodalidade para nos referirmos à capacidade de um sistema ou modelo de processar dados obtidos simultaneamente a partir de diferentes modalidades comunicativas – como texto, imagem, som ou vídeo, – ou seja, para nos referirmos à integração dos múltiplos modos de comunicação ou informação utilizados por estes sistemas e modelos para interpretar e analisar dados. Nesse sentido, chamamos de *datasets* multimodais os conjuntos de dados que combinam duas ou mais destas modalidades comunicativas. Muitos destes conjuntos de dados multimodais multilíngues surgem a partir da expansão de *datasets* originalmente criados em inglês – como o Flickr8k (HODOSH et al., 2013), Flickr30k (YOUNG et al., 2014) e MS-COCO (LIN et al., 2014). Por exemplo, temos as expansões do Flickr8k para o chinês (LI et al., 2016) e do Flickr30K para o alemão (ELLIOTT et al., 2016), francês (ELLIOTT et al., 2017), holandês (VAN MILTENBURG, 2017) e tcheco (BARRAULT et al., 2018).

A seguir, apresentaremos os *datasets* multimodais multilíngues de referência tomados como base para o desenvolvimento da nossa pesquisa e, no capítulo 5, discutiremos as contribuições propostas pelo novo dataset apresentado nesta tese, o Framed Multi30K (FM30K).

2.1 FLICKR30K E SUAS EXPANSÕES

Para o desenvolvimento desta tese, tomamos como ponto de partida três *datasets* multimodais. O primeiro deles, o Flickr30k, é composto por 31.014 imagens – fotografias de atividades, eventos e cenas cotidianas, extraídas do site de compartilhamento de imagens Flickr, representando uma ampla variedade de cenários e situações – cada uma acompanhada por um conjunto de cinco descrições em inglês. As 158.915 descrições – cinco para cada fotografia – foram obtidas através de uma tarefa de criação de descrições elaborada em uma plataforma de *crowdsourcing* disponibilizada pela empresa de tecnologia Amazon, chamada *Amazon Mechanical Turk*. Nessa tarefa, participantes não familiarizados com as entidades e circunstâncias específicas apresentadas em cada uma das cenas retratadas foram instruídos a descrever as pessoas, objetos, cenas e atividades mostradas em cada fotografia sem ter acesso a nenhuma informação adicional sobre o contexto em que as imagens foram originalmente produzidas. Estas descrições, feitas por diferentes anotadores, garantem variações consideráveis na maneira como cada imagem é descrita, fornecendo ao

conjunto de dados diferentes níveis de especificidade para cada cena, o que torna possível fazer inferências sobre semelhanças entre sentenças que não são normalmente relacionadas por regras de reescrita sintática (YOUNG et al., 2014, p. 69). Um exemplo de imagem que compõe o Flickr30k, acompanhado de suas descrições (1), pode ser visto na Figura 1.

Figura 1 - Exemplo de imagem que compõe o dataset Flickr30K.

Fonte: Imagem 126594141.jpg - Flickr 30K (YOUNG et al., 2014)

Para esta imagem, as cinco descrições em inglês criadas por anotadores independentes foram as constantes em (1a-e)[1].

(1) a. A female athlete ties the laces of one of her cleats on the field.
Uma atleta feminina amarra os cadarços de uma de suas chuteiras em um campo.

b. A girl with a ponytail is tying her shoes with a bent knee while on a grassy field.
Uma garota com rabo de cavalo está amarrando os sapatos com o joelho dobrado em um campo gramado.

c. A female soccer player crouches to put on her shoes.
Uma jogadora de futebol se agacha para calçar os sapatos.

d. A girl tying her shoe in a large sports field.
Uma garota amarrando seu sapato em um grande campo de esportes.

e. Soccer player kneeling down to tie her shoe.
Jogadora de futebol ajoelhando-se para amarrar seu sapato.

[1] As traduções em português foram produzidas pela equipe de anotação que trabalhou na criação do FM30K

Esse exemplo nos mostra que, ao coletar várias descrições para uma mesma imagem, é possível ter uma mesma entidade, evento ou situação descritos de maneiras diferentes – por exemplo, *female athlete*, na primeira descrição, e *girl* nas seguintes, – e, mesmo que todas façam referência à pessoa presente na foto, nem todas mencionam o campo de futebol – *field*, *grassy field*, e *sports field*.

Dentre as expansões do Flickr30, o *dataset* Flickr30K Entities (PLUMMER et al., 2015) destaca-se por ter sido não apenas o primeiro conjunto de dados a estabelecer, através da criação de *bounding boxes* – retângulos que delimitam na imagem uma entidade ou área de interesse, – as correspondências entre entidades presentes nas imagens e os itens lexicais com as quais estas se relacionam nas descrições – em inglês, *"region-to-phrase correspondences"*, – mas também por correlacionar, entre as cinco descrições, os diferentes sintagmas que se referem a uma mesma entidade ou conjunto de entidades presentes na imagem, o que os autores chamam de cadeias de correferência – em inglês, *"coreference chains"*. O resultado dessa expansão é um novo conjunto de dados que adiciona 244.035 cadeias de correferência – correlacionando 513.644 entidades ou cenas identificadas, com uma média de 3,2 entidades por descrição, – e 275.775 *bounding boxes* – uma média de 8,7 por imagem – aos dados originais do Flickr30k. Um exemplo de imagem representando a estrutura de dados adicionais agregados por essa expansão pode ser visto na Figura 2. A cadeia de correlação entre sintagmas que fazem referência a uma mesma entidade pode ser vista nas sentenças do exemplo (2), onde as cores do texto que destacam cada sintagma em uma das cinco descrições faz referência à *bounding box* de mesma cor, mostrando qual parte da imagem está relacionada àquele sintagma.

Figura 2 - Exemplo de dados que compõem o dataset Flickr30K Entites.

Fonte: Imagem 126594141.jpg - Flickr 30K Entities (PLUMMER et al., 2015)

(2) a. A female athlete ties the laces of one of her cleats on the field.
b. A girl with a ponytail is tying her shoes with a bent knee while on a grassy field.
c. A female soccer player crouches to put on her shoes.
d. A girl tying her shoe in a large sports field.
e. Soccer player kneeling down to tie her shoe.

Nas expansões voltadas para pesquisas multimodais multilíngiues, tomamos como referência para esse trabalho o *dataset* Multi30K (ELLIOTT et al., 2016), um conjunto de dados que incorpora ao Flickr30k 155.070 novas descrições originais em alemão – coletadas através de tarefas de *crowdsourcing* realizadas por falantes nativos de alemão – e 31.014 traduções para o alemão – criadas por tradutores profissionais falantes nativos da língua alemã a partir de uma das cinco descrições originais em inglês para cada imagem. As sentenças em (3) exemplificam as novas descrições criadas originalmente em alemão a partir da imagem na Figura 1. Já a sentença em (4) exemplifica a tarefa de tradução para o alemão de uma das legendas produzidas originalmente em inglês para a mesma imagem.

(3) a. Eine Frau Schnürt sich die Schuhe auf Einem Sportfeld.
b. Eine Frau in einer Sportdress bindet sich das Schuhband.
c. Eine Sportlerin zieht sich andere Sportschuhe an.
d. Auf einem Rasensportplatz mit roten, gelben und weißen Spielfeldmarkierungen kniet eine Frau auf einem Bein, sie bindet die Schnürsenkel des Schuhs am Fuß des anderen Beines, neben ihr sind Schuhe, eine Beuteltasche und einige textile Gegenstände, daneben ein Tragnetzgeführt, das Spielfeld ist im Hintergrund mit gitterartigen Gestellen begrenz, davor sind Personen, jenseits der Abgrenzung Bäume und ein Berg.
e. Eine Sportlerin im blau-schwarzen T-Shirt mit blsuen Stutzen bindet sich am Sportrasen hockernlnd den Schuh zu

(4) a. Eine[a.SING.NOM.FEM] Sportlerin[female-athlete.SING.NOM] bindet[tie.SING.PERS3] auf[on] dem[the.SING.DAT] Feld[field.SING.DAT] die[the.PLUR.ACC] Schnürsenkel[laces.PLUR.ACC] eines[one.SING.GEN] ihrer[her.PLUR.GEN] Sportschuhe[sports-shoe.PLUR.GEN].
A female athlete ties the laces of one of her cleats on the field.
Uma atleta feminina amarra os cadarços de uma de suas chuteiras em campo.

Como a comparação entre a sentença originalmente produzida em inglês – doravante, ENO, de *Engilsh Original* – e sua tradução para Alemão – DET, de *Deutsch Translation* – demonstra, descrições de imagens não podem ser consideradas como absolutamente objetivas. Ainda que as instruções passadas aos anotadores indicassem a busca pela objetividade, fatores internos a cada língua influenciam a configuração final das descrições

e sua comparatibilidade.

Entretanto, ainda que a comparação seja feita dentro de uma mesma língua, considerando apenas imagem e texto, conforme demonstram as sentenças em (1), a objetividade das descrições é uma ilusão. É sobre essa questão que se debruça a próxima seção.

2.2 O PROBLEMA DA OBJETIVIDADE NAS DESCRIÇÕES DE IMAGENS

Todos esses *datasets* – *benchmarks* para tarefas de PLN e Visão Computacional (VC) relacionadas à criação multilíngue de legendas e tradução automática multimodal – têm como premissa a proposta de HODOSH et al. (2013) de que, quando anotadores são instruídos a descrever as pessoas, objetos, cenas e atividades que são mostradas em uma foto sem receber nenhuma informação sobre o contexto em que aquela imagem foi produzida, o resultado são "descrições conceituais que se concentram apenas nas informações que podem ser obtidas a partir da imagem."(HODOSH et al., 2013, p. 859).[2] Os autores definem descrições conceituais como sendo aquelas que, embora possam conter inferências sobre o contexto da cena retratada, buscam descrever de forma concreta os elementos e entidades presentes em uma imagem, seus atributos e relações, e os eventos em que estas entidades estão envolvidas. Esse tipo de descrição é apresentada em oposição ao que eles definem como descrições não-visuais, aquelas que fornecem informações adicionais que não podem ser obtidas apenas a partir dos elementos presentes na imagem – por exemplo, o local onde aquela fotografia foi tirada, ou o nome das pessoas fotografadas – e que, por isso, são menos relevantes para tarefas PLN que envolvem visão computacional por fazerem referência a elementos visuais que não podem ser identificados na imagem.

Entretanto, autores como VAN MILTENBURG (2017) afirmam que essa premissa de neutralidade na descrição das imagens – em outras palavras, de que é possível criar descrições objetivas baseadas apenas nos elementos visuais presentes em uma imagem – ignora os processos de interpretação e recontextualização inerentes à forma como humanos interpretam uma imagem, e serve apenas como uma "simplificação útil" à criação de *datasets* usados no desenvolvimento de modelos que se beneficiam de um mapeamento direto entre os elementos visuais e suas descrições em uma imagem. Segundo o autor, mesmo pessoas claramente instruídas a fornecer uma descrição simples, porém completa e objetiva, das entidades proeminentes em uma imagem, sem fazer suposições sobre o que está ocorrendo em uma determinada cena, frequentemente especulam ao descrever as imagens – por exemplo, inferindo, sem nenhum fundamento, o grau de parentesco ou status de relacionamento das pessoas que aparecem em uma fotografia. Observe-se que, na Figura 3, dentre as cinco sentenças em (5) que descrevem a imagem, apenas a primeira é composta

[2] "(...) conceptual descriptions that focus only on the information that can be obtained from the image alone." [Tradução nossa]

por informações que podem ser obtidas exclusivamente a partir da imagem. Nas outras quatro, vemos que as partes em destaque fazem referência a atividades desempenhadas por alguns dos participantes – "voluntários" e "de férias" – e a nacionalidade e condição social de outros – "moradores de um país estrangeiro" que vivem em "um gueto num país pobre".

Figura 3 - Exemplo de inferências nas descrições do Flickr30k.

Fonte: Imagem 4969473643.jpg - Flickr 30K (YOUNG et al., 2014)

(5) a. A group of people in an alley looking at the camera.
Um grupo de pessoas em um beco olhando para a câmera.

b. A dark-haired man in blue jeans and a light green polo is standing next to a man with khaki cargo shorts and a blue polo as they pose for a picture **with locals in a foreign land.**
Um homem de cabelos escuros vestindo jeans azul e uma polo verde clara está parado ao lado de um homem com bermuda cargo cáqui e uma polo azul enquanto eles posam para uma foto ***com moradores de um país estrangeiro.***

c. Photo of **a ghetto in a poor country.**
Foto de ***um gueto em um país pobre.***

d. **Volunteers** meeting with the locals.
Voluntários *se encontram com os locais.*

e. A man in the blue shirt is taking a photo with **a family he met while on vacation .**
Um homem de camisa azul está tirando uma foto com ***uma família que conheceu durante as férias.***

Segundo VAN MILTENBURG (2017), o fato de que descrições em *datasets* como o

Flickr30K são inerentemente subjetivas não é necessariamente algo ruim, pois mesmo ao fazer inferências injustificadas – aquelas que resultam da especulação sobre a cena representada na imagem, quando o anotador faz uso de seus conhecimentos e expectativas sobre o mundo para fornecer uma descrição, – elas ainda são relevantes na medida em que apontam quais são, para um anotador humano, os elementos mais importantes em uma imagem – por exemplo, na Figura 3, fica claro que as pessoas são elementos relevantes, mas as árvores ao fundo, não. No entanto, essas inferências podem, também, ter como resultado descrições baseadas em ideias pré-concebidas sobre as características de um grupo de pessoas e sobre como esse grupo geralmente se comporta – neste exemplo, sugerindo que o homem branco é um turista de férias em um país pobre. Nestes casos, segundo o autor, o problema não é necessariamente a inferência baseada em estereótipos, mas a possibilidade de que, quando os indivíduos são descritos e pré-julgados com base em estereótipos negativos em vez de informações individuais disponíveis, essas descrições introduzam vieses nos dados que, ao serem usados para treinar modelos computacionais, poderão propagar estereótipos prejudiciais, promovendo como a ideia de que famílias de determinado perfil étnico ou social são necessariamente moradoras de guetos, ou originárias de países pobres.

Neste ponto, é importante ressaltar que a manifestação de crenças estereotipadas sobre as entidades e eventos presentes em uma imagem – refletidas nas escolhas lexicais feitas pelos criadores de descrições – não são, necessariamente, indício de discriminação e preconceito. Ao demarcar que membros de uma categoria compartilham características subjacentes que fazem com que sejam fundamentalmente semelhantes uns aos outros, esses vieses linguísticos – descritos por BEUKEBOOM et al. (2014) como uma "assimetria sistemática na escolha de palavras em função da categoria social à qual alguém pertence"[3] – são, em muitos casos, altamente funcionais do ponto de vista linguístico, na medida em que facilitam a transmissão de conceitos essenciais sobre categorias sociais. Segundo os autores, esses vieses linguísticos podem ser percebidos através da análise dos termos usados para descrever entidades em uma determinada categoria – por exemplo, a tendência de usar uma linguagem mais específica ao descrever uma pessoa que não atende às expectativas. Neste sentido, BEUKEBOOM et al. (2014) listam uma série de recursos linguísticos utilizados para marcar indivíduos que, na perspectiva dos descritores de imagens, se desviam da norma.

Um dos casos descrito pelos autores é o uso de palavras adicionais – neste caso, adjetivos – para marcar que uma entidade se desvia do esterótipo esperado para uma determinada categoria. Um exemplo de como este "uso de rótulos mais restritivos para indivíduos que não se encaixam nas expectativas gerais de uma categoria social"[4] (BEU-

[3] "(...) a systematic asymmetry in word choice as a function of the social category to which the target belongs." [Tradução Nossa]

[4] "(...) use of more narrow labels for individuals who do not fit with general social category

KEBOOM et al., 2014, p. 3) pode variar entre as descrições criadas por pessoas com diferentes concepções de mundo pode ser visto nas sentenças em (6).

(6) a. A female soccer player crouches to put on her shoes.
b. Soccer player kneeling down to tie her shoe.

Nestes dois exemplos é possível perceber que, apesar de termos descrições muito semelhantes sobre as atividades desempenhadas pela pessoa retratada na imagem – ambas fazem referência a uma pessoa que se ajoelha para colocar seus sapatos, – um dos autores julgou importante ressaltar o gênero da atleta – em inglês, *"A female soccer player"* – enquanto o outro apenas se referiu a atividade esportiva praticada – *"Soccer player"*. Segundo BEUKEBOOM et al. (2014), casos como o da descrição que faz referência ao gênero podem ser reflexo das crenças e expectativas do autor das descrições sobre o mundo – neste caso, a de que jogadores de futebol geralmente são homens – mas não são, necessariamente, uma indicação de preconceito, podendo apenas sinalizar uma tentativa de fornecer uma descrição mais detalhada da cena. Esta percepção encontra suporte na teoria proposta por LAKOFF (1987), que sugere que os humanos "organizam seu conhecimento por meio de estruturas denominadas modelos cognitivos idealizados, ou ICMs, e que estruturas categoriais e efeitos de protótipo são um resultado dessa organização.[5]". Esta perspectiva sobre modelos cognitivos, desenvolvida no escopo da linguística cognitiva, tem como uma de suas referências a teoria da Semântica de Frames (FILLMORE, 1982), que nos fornece um exemplo do que ocorre quando um elemento pertencente a um determinado modelo cognitivo não encontra correspondente direto na categoria conceitual esperada.

A análise das descrições no Flickr30K também torna evidente a ocorrência de diversos casos em que os anotadores não seguiram a diretriz que os orienta a não fazer suposições que extrapolam aquilo que é possível extrair apenas a partir da observação da imagem – o que Van Miltenburg chama de inferências infundadas ou, em inglês, *unwarranted inferences* (VAN MILTENBURG, 2019, p. 35). Segundo o autor, a análise de alguns exemplos de sentenças que contêm este tipo de ocorrência permite agrupá-las em algumas categorias mais proeminentes.

Muitos anotadores parecem interessados em explicar ou justificar os motivos subjacentes aos eventos retratados em uma cena – em outras palavras, o porquê da ocorrência de determinada situação. Na Figura 4, temos como exemplo a descrição para uma imagem em que um praticante de escalada esportiva é mostrado amarrando seu arnês de escalada, e que foi descrita como "Um homem se prende a uma corda de escalada **para se divertir.**" – no original, *"A man is hooking himself up to the tether line* ***in order to have some***

expectations."[Tradução Nossa]

[5] "(...) we organize our knowledge by means of structures called idealized cognitive models, or ICMs, and that category structures and prototype effects are by-products of that organization." (LAKOFF, 1987, p.68) [Tradução nossa]

fun.". Note-se que o propósito da escalada – segundo o autor da legenda, "diversão" – não pode ser comprovado a partir da imagem.

Figura 4 - Inferência sobre o objetivo.

Fonte: Imagem 3963038375.jpg - Flickr30K (YOUNG et al., 2014)

Outro exemplo comum de inferência infundada ocorre quando os anotadores supõem a atividade decorrente de uma cena, ou mesmo as relações de poder estabelecidas entre os participantes da cena. Isso ocorre na Figura 5, que mostra uma mulher vestindo uniforme, de pé, em frente a um homem com os braços cruzados, e que é descrita como "Um gerente **fala com uma funcionária sobre o desempenho no trabalho.**" - no original, *"A manager* ***talks to an employee about job performance.****"*.

Figura 5 - Inferência sobre a atividade.

Fonte: Imagem 80630071.jpg - Flickr 30K (YOUNG et al., 2014)

Comum também são as inferências sobre o que é visto para além do enquadramento – ou seja, para além das margens da fotografia – quando as fotos mostram locais geralmente

associados a um tipo específico de evento. Como exemplo, temos a Figura 6, que mostra apenas pessoas sentadas atrás de um alambrado, com uma arquibancada ao fundo, e que foi descrita como "Espectadores **em um jogo de beisebol.**" - no oirignal, *"Spectators **at a baseball game.**"*

Figura 6 - Inferência sobre o evento.

Fonte: Imagem 208053776.jpg - Flickr30K (YOUNG et al., 2014)

Outro exemplo de descrição que não pode ser inferida a partir das imagens, mas que aparece com frequência, são sobre tipo de relacionamento ou parentesco existente entre as pessoas retratadas em uma cena. No exemplo da Figura 7 não há nada que possa indicar o grau de parentesco entre os indivíduos retratados, no entanto, uma das descrições fornecidas para esta imagem foi "**Avó e sua neta** abrindo presentes no Natal." – no original, "***Grandmother and her granddaughter*** *opening presents at Christmas time.*".

Figura 7 - Exemplo de inferência sobre as relações de parentesco.

Fonte: Imagem 6575692515.jpg - Flickr30K (YOUNG et al., 2014)

VAN MILTENBURG (2017) destaca também a frequência com que são inferidas as profissões ou ocupações dos indivíduos que aparecem nas imagens. Como exemplo, temos a descrição "**Um grupo de universitários** se reúne para jogar pôquer Texas Hold em - no original, *"**A group of college students** gathers to play texas hold em poker."* – fornecida para a imagem na Figura 8, que não contém nenhum elemento visual que permita inferir tal ocupação – a não ser o fato de que todos os indivíduos retratados são jovens com idade para cursar o Ensino Superior.

Figura 8 - Exemplo de inferência sobre a ocupação.

Fonte: Imagem 36979.jpg - Flickr30K (YOUNG et al., 2014)

Finalmente, cabe também destacar os exemplos recorrentes de inferências sobre etnia e nacionalidade onde, por exemplo, pessoas com traços faciais asiáticos são frequentemente descritas como sendo de origem chinesa ou japonesa. Para a Figura 9, uma das descrições fornecidas foi "Um **homem chinês** segurando uma **garotinha chinesa**."– em inglês, *"A **Chinese man** holding a **Chinese little girl**."*.

Exemplos como estes – apesar de não exaustivos quanto a todas as possíveis categorias de inferências presentes neste conjunto de dados – apontam para o fato de que humanos, mesmo quando trabalhando em uma tarefa de anotação altamente parametrizada e sendo especificamente instruídos a propor descrições conceituais, nem sempre são capazes de criar descrições objetivas de uma imagem, nos permitindo teorizar sobre as motivações para o surgimento destes tipos de descrições especulativas e sobre como se dá o processo de descrição da imagem feito por humanos.

Uma forma de explicar o comportamento dos anotadores deste *dataset* é reconhecer a artificialidade da metodologia de descrição de imagens proposta por HODOSH et al. (2013). Tal artificialidade, acreditamos, decorre do fato de que tarefas dessa natureza tomam a descrição de imagens como um simples mapeamento entre as entidades presentes na modalidade visual e seus correspondentes lexicais na modalidade textual, desconsiderando o fato de que, de um lado, uma mesma entidade ou cena pode dar origem a várias inter-

Figura 9 - Exemplo de inferência sobre a nacionalidade.

Fonte: Imagem 3483640715.jpg - Flickr30K (YOUNG et al., 2014)

pretações válidas, enquanto, de outro, cada escolha lexical codifica uma perspectiva sobre o universo representado. Quando se consideram, por exemplo, fatores como conhecimento de mundo, contexto e perspectiva – todos eles absolutamente centrais para a cognição linguística humana, conforme discutiremos no capítulo sobre Semântica de Frames – a noção de que descrições conceituais são livres de inferências e perspectivas pessoais fica seriamente comprometida.

Na mesma direção, autores como AROYO & WELTY (2015) questionam a premissa de que conjuntos de dados coletados em tarefas de *crowdworking* – em que vários humanos fornecem o mesmo tipo de anotação para os mesmos exemplos – podem ser tomados como uma "verdade absoluta" – em inglês, *"ground truth"* –, apresentando experimentos que mostram como múltiplas perspectivas humanas, refletidas nas anotações, possibilitam identificar uma série de problemas em tarefas que envolvem interpretação semântica, revelando o que eles chamam de "falácia da verdade única e universalmente constante." (AROYO & WELTY , 2015, p. 16)

Para os autores, o uso desses conjuntos de dados de referência – chamados em inglês de *gold standard datasets*[6]. Em tarefas de PLN, a concordância entre anotadores tornou-se

[6] *Gold standard datasets* – ou, conjuntos de dados "padrão ouro" – são conjuntos de dados utilizados no treinamento, validação e teste de algoritmos computacionais usados em tarefas de aprendizado de máquina – em inglês, *"Machine Learning"*. A criação destes conjuntos de dados se dá a partir de tarefas onde anotadores humanos analisam e validam pequenas quantidades de dados de exemplo que servem como um "valor de verdade" – ou *"ground truth"* – e expressam qual o resultado esperado, ou seja, verdadeiro, para cada instância em um conjunto de dados. Quando realizadas por diferentes anotadores, a qualidade das anotações nestes *datasets* é estabelecida medindo-se a concordância entre os anotadores, ou seja, a probabilidade média de que duas pessoas concordem sobre uma determinada informação – no caso do Flickr30K, o grau de similaridade entre as descrições para uma mesma imagem (AROYO & WELTY , 2015, p. 16)

tão difundida que seu caráter de *"ground truth"* não é questionado mesmo nos casos que envolvem tarefas altamente subjetivas – como, por exemplo, na interpretação e criação de descrições para imagens. Em tarefas desse tipo, os pesquisadores apontam vários exemplos que contrariam a premissa de que cada instância anotada possui apenas uma interpretação correta, destacando casos frequentes em que tarefas de anotação idênticas atribuídas a dois anotadores distintos apresentam algum grau de discordância, ressaltando que, nestes casos, a discordância entre anotadores não deveria ser considerada uma medida de má qualidade na execução da tarefa – resultado de um problema na metodologia da tarefa ou de anotadores mal treinados – mas, sim, um sinal de que as tarefas em questão apresentam algum grau de subjetividade.

Outro problema comum percebido durante a análise das metodologias utilizadas na criação desses *datasets* é a suposição de que a baixa concordância entre os anotadores decorre da falta de diretrizes de anotação claras e consistentes (AROYO & WELTY , 2015, p. 17). Segundo os autores, a preocupação com a elaboração de diretrizes precisas e detalhadas – que buscam garantir uma maior similaridade entre os dados criados por diferentes anotadores, forçando-os a fazer escolhas mais restritas, que não refletem necessariamente suas percepções – não resulta numa melhor qualidade dos dados coletados, mas, sim, num conjunto de dados com uma concordância criada de forma artificial, fruto da diminuição no número de dados que indicariam as ambiguidades e as diferenças de perspectivas inerentes à tarefa e adotadas por diferentes anotadores. Para AROYO & WELTY (2015), isso fica evidente quando são observados os resultados em tarefas que utilizam métricas como BLEU (PAPINENI et al., 2002), Meteor (BANERJEE & LAVIE, 2015) e CIDEr (VENDATAM et al., 2015), que analisam o desempenho de algoritmos calculando a similaridade textual entre descrições geradas automaticamente e descrições de um conjunto de referência gerado por humanos. Métricas como essas – em especial a BLEU – são frequentemente criticadas em função da sua baixa correlação com avaliações humanas (ELLIOTT & KELLER, 2014; KILICKAYA et al., 2017; REITER, 2018).

Outros autores, como TROTT et al. (2020), reforçam essa perspectiva ao basear-se em evidências psicolinguísticas e pesquisas recentes em PLN para propor que a expressão humana se dá através da utilização de diferentes estratégias de modulação semântica, sugerindo que os processos cognitivos envolvidos na criação de sentido durante tarefas como a de descrição de uma imagem passam, necessariamente, por etapas de construção do significado que envolvem aspectos como:

1. A **perspectiva** – ou a escolha de um ponto de vista (*"vantage point"*), – que se reflete não apenas em escolhas relacionadas ao domínio espacial (como dizer que um objeto está a direita ou a esquerda de algum outro objeto, dependendo de qual perspectiva é privilegiada), mas também no domínio do movimento (por exemplo, quando um mesmo movimento pode ser descrito em relação a diferentes centros

dêiticos), e até na escolha do tempo verbal (quando, por exemplo, a escolha por um tempo verbal no passado é utilizada para indicar o ponto de vista reflexivo do falante);

2. A **proeminência** – ou seja, quais elementos são mais salientes – ressaltando as diferentes estratégias e recursos utilizados para definir tanto o foco da atenção em uma determinada cena quanto o perfilamento entre entidades envolvidas (por exemplo, quando decidimos se vamos dizer que "O gato está sobre a almofada" ou "A almofada está debaixo do gato");

3. A **resolução**, que engloba tanto o aspecto da **especificidade** – por exemplo, nas relações estabelecidas entre "dálmata < cachorro < animal" – quanto a **granularidade** – quando, por exemplo, falamos de uma floresta a partir das folhas de uma árvore, de seus galhos, ou das árvores. Tais aspectos se manifestam na diferença percebida entre sentenças como "O lateral esquerdo fez um gol de bicicleta" ou "O jogador chutou a bola", que podem descrever a mesma cena com diferentes graus de detalhamento e;

4. A **configuração**, que se refere às "propriedades estruturais internas de entidades, grupos de entidades e eventos, indicando suas 'formas' e 'texturas' esquemáticas: multiplicidade (ou plexidade), homogeneidade, limitação, relações parte-todo, etc.[7]" (TROTT et al., 2020, p. 5173). A Figura 10 ilustra diferentes possibilidades de configuração manifestas em legendas. Nota-se que uma mesma imagem de uma pessoa pedalando em meio ao tráfego de veículos em uma cidade é descrita por um dos anotadores como "Uma mulher usando um suéter rosa anda de bicicleta ao lado dos **carros.**"[8] e, por outro, como "Uma garota de camisa rosa andando de bicicleta no **trânsito**"[9]. A escolha pelo uso do plural "carros" foca na natureza individual dos múltiplos veículos, enquanto o uso do singular "trânsito" sugere um conjunto homogêneo de veículos.

Para mitigar estes e outros problemas, abordagens mais recentes (BASILE et al., 2021) defendem um novo paradigma, chamado de **perspectivismo de dados** – em inglês, *data perspectivism*, – que se distancia da difundida ideia de *datasets gold standard*, propondo a adoção de metodologias que incorporem as opiniões e perspectivas dos anotadores humanos envolvidos na etapa de representação do conhecimento presente na criação de novos *datasets*. Segundo os autores, a necessidade dessa nova abordagem decorre do fato de que muitas das atuais metodologias utilizadas em tarefas de anotação

[7] (...) internal-structural properties of entities, groups of entities, and events, indicating their schematic "shape" and "texture": multiplicity (or plexity), homogeneity, boundedness, part-whole relations, etc."[Tradução nossa]

[8] "A woman in a pink sweater rides her bike alongside **cars.**" [Tradução nossa]

[9] "A girl in a pink shirt is riding a bicycle in traffic." [Tradução nossa]

Figura 10 - Exemplo de dimensões do construal.

Fonte: Imagem 1691573772.jpg - Flickr30K (YOUNG et al., 2014)

desenvolvidas para pesquisas em linguística ainda se baseiam em uma série de práticas desenvolvidas para a anotação linguística pura, ou seja, a anotação de traços linguísticos menos subjetivos onde, uma vez estabelecido o referencial teórico que fundamenta a anotação, apenas uma condição de verdade seja possível – ou seja, que haja apenas uma resposta correta para cada instância a ser anotada. Esta premissa faz sentido se considerarmos, por exemplo, anotações que buscam atribuir classes gramaticais às palavras em uma sentença, já que estas, em uma visão mais restrita, não poderiam desempenhar simultaneamente o papel de um substantivo e um verbo[10]. Entretanto, os problemas de tais metodologias tornam-se evidentes quando o foco da anotação se volta para fenômenos mais subjetivos e pragmáticos das línguas naturais como, por exemplo, o julgamento sobre o uso de linguagem abusiva e ofensiva em descrições de imagens – fenômenos que não podem ser analisados a partir das mesmas estruturas metodológicas utilizadas em anotações linguísticas formais tradicionais, já que que tais metodologias não contemplam nuances presentes na intenção comunicativa de tais expressões, que possam ter sido identificadas pelos anotadores. Sob o olhar desse novo paradigma, que preconiza uma abordagem perspectivizada dos dados, nos casos em que uma mesma palavra ou sentença pode ser percebida como abusiva por um anotador e não abusiva por outro, ambas as perspectivas estão corretas e, portanto, ambas as anotações devem ser consideradas verdadeiras e incorporadas ao conjunto de dados.

[10] Os autores reconhecem que, mesmo nestes casos, discordâncias entre anotadores são possíveis na medida em que estes podem ter opiniões diferentes, ou se equivocar durante a realização da tarefa. No entanto, divergências deste tipo tenderiam a ser corrigidas ou removidas, já que não seriam resultado da adoção de diferentes perspectivas sobre a classe gramatical de um item lexical, mas de problemas como, por exemplo, a falta de clareza sobre a metodologia de anotação. Num viés cognitivista, entretanto, mesmo a rotulação de classes de palavras estaria sujeita a distintas perspectivas.

O perspectivismo de dados é também apontado como um caminho para abordar a questão dos enviesamentos danosos em conjuntos de dados multimodais. Um número crescente de pesquisas recentes no campo da Visão Computacional e no do Processamento de Língua Natural que envolvem o uso de *datasets* criados por humanos – sejam eles coletados via de tarefas *crowdsourcing* ou extraídos automaticamente da web (*"web scraping"*) – têm apontado a presença de enviesamentos potencialmente danosos nesses conjuntos de dados, como o uso de linguagem abusiva e ofensiva, elementos de discurso de ódio, e propagação de estereótipos. O problema é agravado porque tais *datasets* adotam uma abordagem não perspectivizada, que igualará tais enviesamentos a uma *ground truth*. BURNS et al. (2018) mostram que anotadores recrutados para tarefas de criação de *datasets* multimodais frequentemente recorrem a pistas contextuais, geralmente baseadas em estereótipos, para rotular entidades em imagens – por exemplo, atribuindo descritores de gênero como "homem / rapaz" ou "mulher / moça" em casos onde as imagens não contêm elementos que permitam essa identificação. Um exemplo desse tipo de viés de gênero pode ser observado na Figura 11, descrita por um dos anotadores como "**Um rapaz** fazendo um movimento de snowboard em uma ladeira." – em inglês, *"**A guy** making a snowboarding move on a slope."* – sem que, no entanto, haja na imagem uma clara indicação do gênero do praticante de *snowboard*.

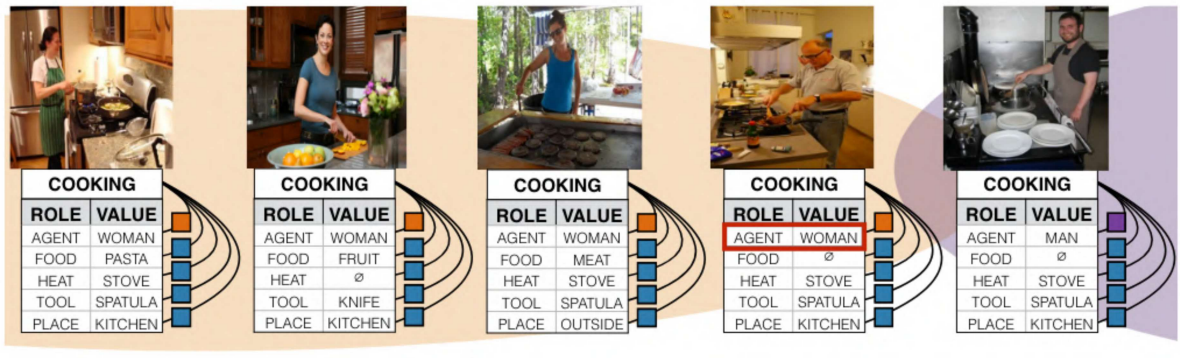
Figura 11 - Exemplo de viés de gênero em descrições de imagens.

Fonte: Imagem 3407584080.jpg - Flickr30K (YOUNG et al., 2014)

Um dos problemas, nesses casos, decorre do fato de que essas descrições de imagens são utilizadas, por exemplo, para o treinamento de algoritmos de rotulagem semântica – que, eventualmente, serão avaliados pela similaridade entre os rótulos gerados automaticamente e aqueles fornecidos pelos anotadores humanos –, reforçando os estereótipos e preconceitos presentes nos dados e, consequentemente, atribuindo mais rótulos incorretos. ZHAO et

al. (2017) abordam a questão da amplificação destes vieses apresentando exemplos de tarefas de rotulagem de papel semântico em imagens – em inglês, *Visual Semantic Role Labeling* (vSRL) – em que um modelo treinado a partir de um *dataset* multimodal contendo estereótipos de gênero não apenas perpetua estes vieses, mas os amplifica. No exemplo da Figura 12 – (ZHAO et al., 2017, p. 2) – um conjunto de dados criado a partir do estabelecimento de relações entre o verbo evocado por uma cena – no exemplo, *cozinhar.v* – o papel semântico do sujeito da ação – agente, – e os respectivos nomes que preenchem este papel – por exemplo, *mulher.n* – apresenta imagens de cenas de culinária em que mulheres desempenham o papel de agente em 66% dos casos e homens em 33% dos casos. Entretanto, segundo os autores, o modelo de rotulagem semântica treinado neste *dataset* atribui automaticamente o rótulo *mulher.n* ao papel de agente em 84% das imagens – uma ocorrência maior do que os 66% presentes no *dataset* original.

Figura 12 - Exemplo da amplificação do viés de gênero em tarefas de vSLR.



Fonte: Figura 1 - Zhao et al., 2017

Questões relacionadas aos riscos das aplicações destes *datasets* também são discutidos por autores como BIRHANE et al. (2021), que afirmam que os danos causados por modelos de IA treinados nestes conjuntos de dados podem tornar-se ainda mais graves com aumento recente no número de *datasets* elaborados a partir da extração automática de quantidades gigantescas de imagens disponíveis na Web. Segundo os autores, *datasets* coletados dessa maneira – como, por exemplo, o CommonCrowl[11], usado como base para o treinamento de grandes modelos de língua (em inglês, *Large Language Models*, ou LLMs) e de redes neurais para tarefas de aprendizado multimodal, como o CLIP (RADFORD et al., 2021) – apresentam graves problemas de curadoria, fomentando a difusão de imagens que propagam preconceitos de gênero, raciais e geográficos, imagens voyeurísticas não consensuais, com conteúdo inapropriado e rótulos ofensivos, dentre outros (BIRHANE et al., 2021, p. 3).

Como exemplo, a autora analisa os experimentos feitos com o *dadaset* LAION-400M (SCHUHMANN et al., 2021), que teve parte da curadoria dos dados feita através do uso

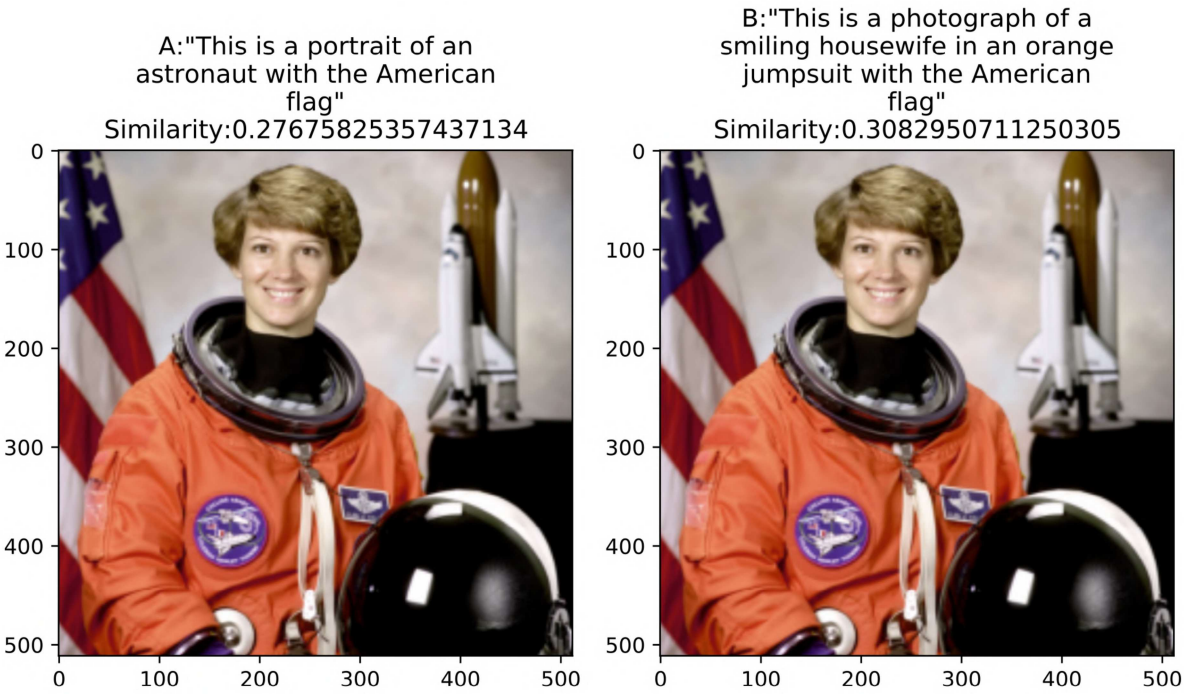
[11] https://commoncrawl.org

de um filtro algorítmico de imagens que buscou incluir apenas aquelas com alto nível de similaridade entre o conteúdo semântico da imagem e sua descrição textual. Isso foi feito calculando a similaridade de cosseno entre as *embeddings* da descrição textual e da imagem obtidas através do modelo CLIP, descartando aquelas com uma similaridade de cosseno abaixo de 0,3.

Para mostrar como a suposição de que um limite de 0,3 na similaridade de cosseno pode gerar problemas de enviesamento, BIRHANE et al. (2021) utiliza uma foto da astronauta norte-americana Eileen Collins, primeira mulher a pilotar um ônibus espacial, em 1995 (Figura 13). A fotografia, que mostra a astronauta em um traje espacial, vem acompanhada de duas descrições: (A) "Este é um retrato de uma astronauta com a bandeira americana." – no original, *"This is a portrait of an astronaut with the American flag.",* – e a segunda (B) "Esta é uma fotografia de uma dona de casa sorridente em um macacão laranja com a bandeira americana." – no original, *"This is a photograph of a smiling housewife in an orange jumpsuit with the American flag."*). Para os dois casos, o CLIP produziu, respectivamente, as seguintes similaridades de cosseno: 0,28 e 0,31. Com base nesses resultados, a autora propõe que imaginemos um cenário em que o filtro algorítmico – treinado para filtrar similaridades do cosseno inferiores a 0,3 – se depara com essas duas ocorrências de pareamento imagem-texto e, devido aos vieses de gênero incorporados no CLIP, trata como semanticamente mais relevante a segunda descrição, carregada com viés de gênero.

Considerados os fatos de que tanto a curadoria humana como a automática não são capazes de tornar os datasets multimodais livres de enviesamentos e perspectivas – danosos ou não – e dado que o conceito de "descrição conceitual" proposto por HODOSH et al. (2013), como demonstrado, não corresponde ao que os anotadores humanos de fato fazem ao criar tais descrições, uma questão que se impõe é: como conciliar tais limitações com a necessidade de se produzirem e ampliarem datasets multimodais? Esta tese busca na Semântica de Frames uma resposta para essa pergunta.

Figura 13 - Resultados dos experimentos com a imagem da astronauta Eileen Collins.



Fonte: Figura 1 em (BIRHANE et al., 2021)

3 REPRESENTAÇÕES SEMÂNTICAS PARA A MULTIMODALIDADE

Este capítulo aborda os conceitos fundamentais que embasam a adoção da Semântica de Frames como modelo para o tratamento de *datasets* multimodais. A primeira seção apresenta a FrameNet Brasil (TORRENT et al., 2022) como implementação computacional do modelo fillmoreano, expandido para o domínio multimodal. Já a segunda discute em que medida tal expansão dialoga com o proposto pelos estudos em multimodalidade acerca das relações entre imagem e texto.

3.1 SEMÂNTICA DE FRAMES E A FRAMENET

A Semântica de Frames (FILLMORE, 1982) é a teoria segundo a qual os significados das palavras devem ser compreendidos a partir das cenas que elas evocam. Essas cenas – representações esquemáticas do conhecimento humano – são chamadas de frames, que nada mais são do que estruturas conceituais que, partindo das experiências humanas, formam uma base de conhecimento que nos permite atribuir sentido a situações e eventos. Nas palavras de Fillmore:

> Com o termo 'frame', tenho em mente qualquer sistema de conceitos relacionados de tal maneira que, para entender qualquer um deles, você precisa entender toda a estrutura em que ele se encaixa; quando um dos elementos de uma dada estrutura é introduzido em um texto ou em uma conversa, todos os outros são automaticamente disponibilizados.[1]. (FILLMORE, 1982, p. 111)

Como exemplo, o autor apresenta pares de palavras similares, de uso cotidiano, que mostram como nossa atribuição de significado deriva do nosso reconhecimento sobre as diferentes maneiras pelas quais as palavras esquematizam o mundo (FILLMORE, 1982, p. 121). Ao considerarmos palavras como 'terra' e 'solo', vemos que uma das maneiras de diferenciar seus significados seria dizer que 'terra' designa a superfície seca do planeta, distinta do mar, enquanto o 'solo' se refere a superfície que se opõe ao ar acima dele. Para Fillmore, palavras como 'terra' e 'solo', então, diferem não apenas no que podem ser usadas para identificar, mas também em como codificam essa identificação em um contexto mais amplo, já que é pelo reconhecimento desse contraste de contexto que podemos entender, por exemplo, que um pássaro que "passa sua vida na terra" está sendo descrito como um animal que não passa nenhum tempo na água, e um pássaro que "passa sua vida no solo" está sendo descrito como um animal que não voa.

[1] By the term 'frame' I have in mind any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits; when one of the things in such a structure is introduced into a text, or into a conversation, all of the others are automatically made available

Desde 1997, a teoria da Semântica de Frames vem sendo aplicada ao desenvolvimento da FrameNet (FILLMORE et al., 2003), um projeto de lexicografia computacional onde itens lexicais são descritos a partir dos frames que eles evocam, com base em evidências de *corpora*. Nesse projeto, linguistas identificam e descrevem os frames, analisando os significados das palavras a partir dos sentidos que sinalizam, dados os frames de fundo que evocam, e estudando o processo de atribuição de propriedades semânticas aos seus contextos sintáticos (FILLMORE et al., 2003, p. 235). A seguir, apresenta-se a estrutura dos dados que compõem a FrameNet.

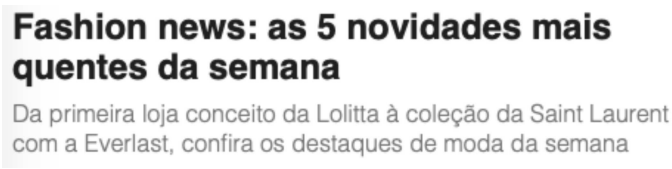
3.1.1 **Estrutura de Dados da FrameNet**

Quando o sentido de uma palavra está relacionado a um determinado frame, é possível dizer que essa palavra evoca esse frame. É por esse motivo que, ao nos deparamos com palavras como raquete ou tenista, associamos essas palavras a uma rede mais ampla de conceitos que, consequentemente, habilitam a construção de significados relacionados ao cenário de uma partida de tênis. Já a rede de conceitos e significados ativados pela palavra 'quente' – e também os frames que ela evoca – variam de acordo com o contexto em que ela é utilizada, como podemos ver nos exemplos mostrados nas Figuras 14 e 15:

Figura 14 - Diferentes significados da palavra 'quente'.

Fonte: Jornal Tribuna de Minas

Figura 15 - Diferentes significados da palavra 'quente'.



Fonte: Jornal Tribuna de Minas

Partindo desses exemplos, é possível constatar que a compreensão de ambas as sentenças contendo a palavra 'quente' requer do leitor a correta identificação dos frames evocados pelo contexto em que a palavra aparece (Figura 16). No primeiro trecho de artigo

jornalístico apresentado na Figura 14, o frame evocado é o de `Temperatura_ambiente`, que é definido como a "temperatura em um ambiente, a qual é determinada pelo Tempo e pelo Lugar[2]". No título do segundo, artigo mostrado na Figura 15, o frame evocado é o de `Popularidade`, onde "um avaliado, que pode ser uma pessoa ou um objeto, tem uma quantidade de aceitabilidade ou utilização com base na desejabilidade geral (muitas vezes não estética)[3]".

Figura 16 - Exemplo dos frames de `Temperatura_ambiente` e `Popularidade`

Fonte: Bases de dados da FrameNet Brasil

As definições dos frames (Figura 16) têm origem na avaliação das propriedades necessárias para esquematização de uma determinada cena ou situação. As palavras em destaque marcam os elementos que compõem cada frame – sejam eles personagens, objetos, circunstâncias, etc. – e que desempenham algum tipo de papel semântico nas cenas descritas, contribuindo com informações a respeito do frame evocado. Esses elementos, chamados de Elementos de Frame (EFs), podem ser de três tipos: 1. nucleares, quando representam conceitos obrigatórios para a instanciação do frame; 2. periféricos, fornecendo características adicionais para as circunstâncias em que ocorre a cena descrita pelo frame

[2] http://webtool.framenetbr.ufjf.br/index.php/webtool/report/frame/showFrame/518
[3] http://webtool.framenetbr.ufjf.br/index.php/webtool/report/frame/showFrame/1015

ou; 3. extra-temáticos, que também atuam ampliando o contexto do evento descrito na cena mas, diferentemente dos periféricos, o fazem incorporando informações fora do escopo do frame através da inclusão de atributos próprios de outros frames. Assim, se tomarmos como exemplo o frame de Viagem, – definido como um evento onde "um viajante se engaja em uma jornada, uma atividade geralmente planejada com antecedência, na qual se move de uma localização fonte para um alvo através de um caminho ou ao redor de uma área(...)[4]", – consideramos EFs nucleares (Figura 17) aqueles elementos que, de um modo geral, representam funções sintáticas mais evidentes, como o Viajante (o sujeito que faz a viagem) ou o Destino (alvo do viajante)[5]. Elementos como o Acompanhante (que viaja junto com o Viajante) ou a Bagagem (itens necessários para a viagem) são considerados elementos periféricos (Figura 17) pois, apesar de adicionarem informações à estrutura do frame, são dispensáveis para a sua constituição. Já elementos como a Finalidade (da viagem) ou a Maneira (como a viagem ocorre) operam como frames extra-temáticos na medida em que, apesar de serem parte da estrutura do frame, evocam frames próprios – como no caso do EF Maneira evocando um frame de mesmo nome que tem em sua estrutura Unidades Lexicais (ULs) que representam essa função, como o advérbio 'tranquilamente'.

A despeito da similaridade entre o conceito de EF com o conceito de casos profundos – *deep cases*, da Gramática de Casos (FILLMORE, 1977) – Fillmore explica que:

> Existem boas razões para não vincular os Elementos de Frame a nenhuma das listas familiares de funções semânticas (agente, paciente, tema, experienciador, instrumento etc.). Como é pedido aos anotadores que encontrem expressores de Elementos de Frame em sentenças reais, ter nomes de EFs mnemônicos em relação ao próprio frame torna essas identificações mais fáceis[6]. (FILLMORE, 2008, p. 51)

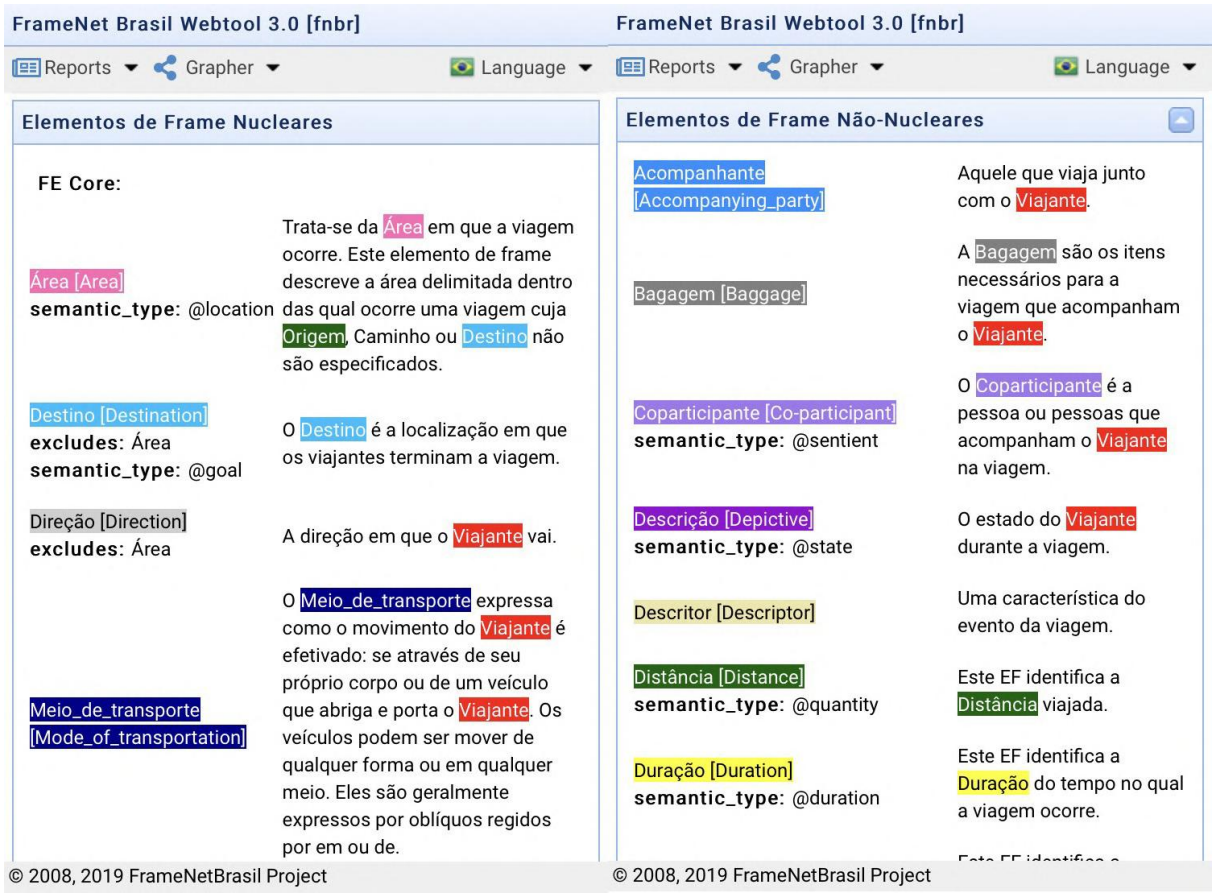
A principal das boas razões mencionadas por Fillmore na passagem acima está relacionada à noção de perspectiva, fundamental na Semântica de frames e na FrameNet. Porque os frames não só representam as cenas, mas podem adotar perspectivas distintas a elas, um mesmo EF, como Comprador, por exemplo, poderia ser mapeado tanto ao papel temático de Agente, no frame `Comércio_comprar`, quanto ao papel de Alvo, em `Comércio_vender`. Assim, qualquer mapeamento de um EF para uma função temática

[4] http://webtool.framenetbr.ufjf.br/index.php/webtool/report/frame/showFrame/315

[5] Cabe aqui destacar a existência de um tipo específico EF nuclear, chamado de EF nuclear não-expresso (em inglês *core unexpressed*), que tem esse status para garantir que a relação de Herança não seja violada, podendo ser anotado apenas no frame mãe, sem ser expresso nos frames filhos.

[6] "There are good reasons for not tying the frame elements into any of the familiar lists of semantic roles (agent, patient, theme, experiencer, instrument, etc.). Since annotators are asked to find expressors of frame elements in actual sentences, FE names that are memorable in respect to the frame itself will facilitate such identifications."

Figura 17 - EF nucleares e não-nucleares do frame `Viagem`



Fonte: Base de dados da FrameNet Brasil

mais genérica dependerá fortemente da perspectiva adotada no frame e não pode ser generalizado de modo absoluto.

Estão também descritas na base de dados da FrameNet as regras que caracterizam as relações entre EFs quanto a sua co-ocorrência, podendo ser de três tipos: *core set*, *excludes* (exclui) e *requires* (requer). Elementos de frame que fazem parte de um *core set* estabelecem entre si uma correlação que permite que a presença de apenas um deles seja suficiente para licenciar a sentença em que se evoca o frame. Se tomarmos como exemplo os EFs do frame de `Movimento`[7] (Figura 18) – que tem como *core set* os EFs Fonte, Alvo, Trajetória, Distância e Direção – e observarmos a sentença de exemplo (1).

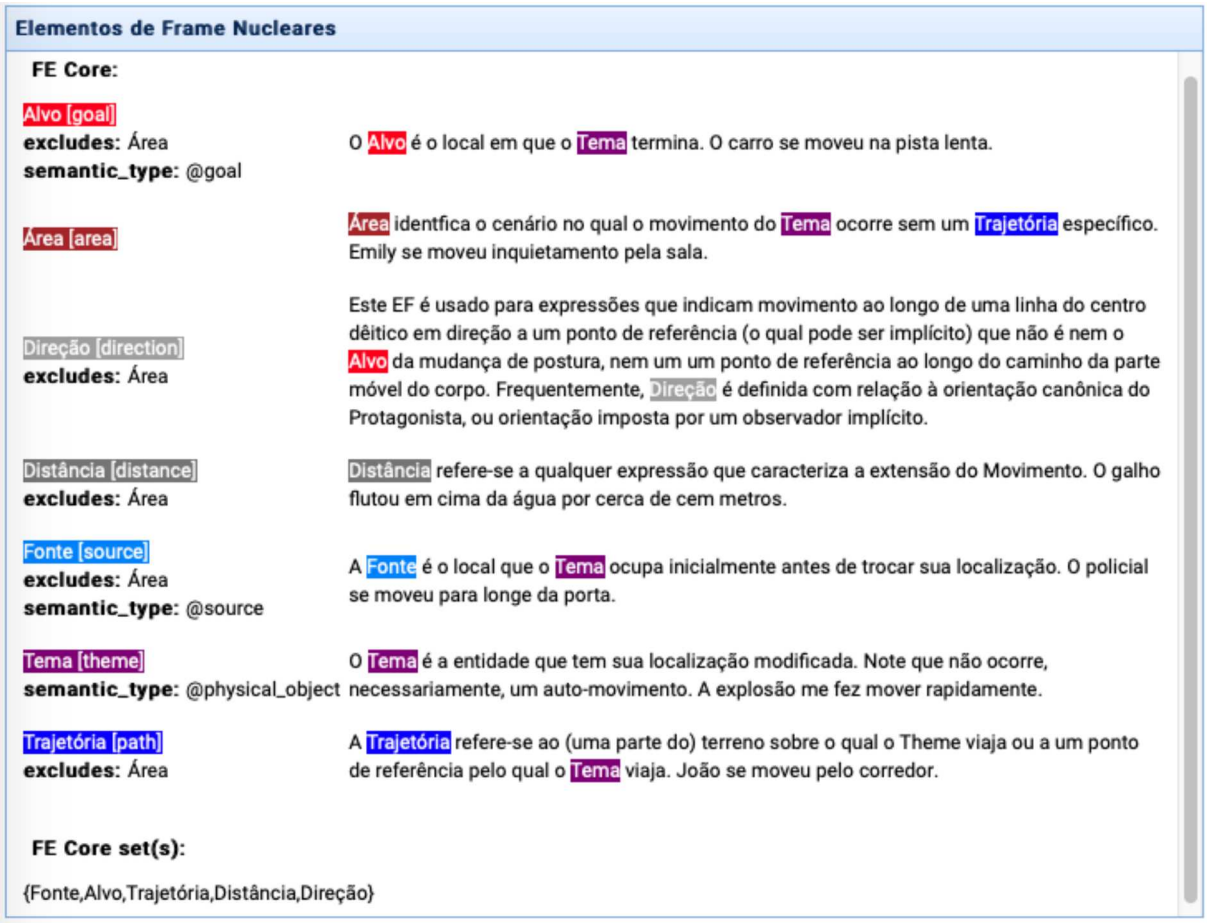
(1) a. O policial se moveu para longe da porta.

Vemos que a instanciação do EF Distância elimina a necessidade de outros EFs nucleares, como Direção e Trajetória .

Ainda tomando como referência o frame de `Movimento` e a sentença de exemplo,

[7] http://webtool.framenetbr.ufjf.br/index.php/webtool/report/frame/showFrame/3

Figura 18 - EF nucleares e *core set* do frame `Movimento`



Fonte: Base de dados da FrameNet Brasil

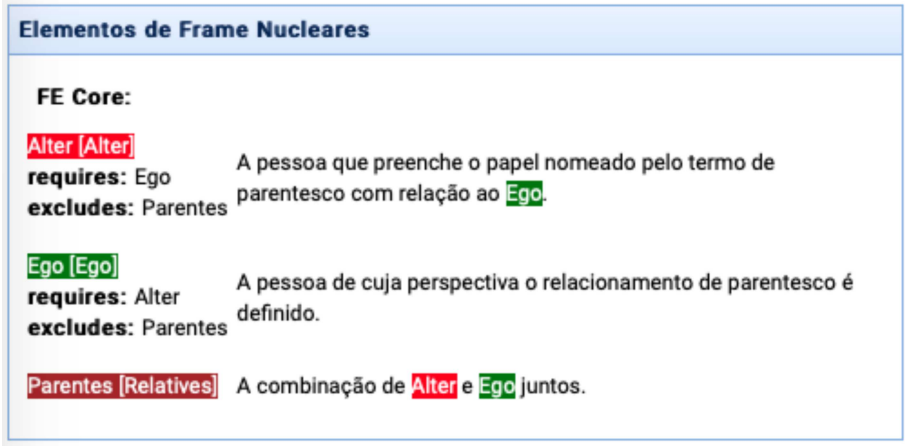
temos os casos em que a ocorrência de um determinado EF impede a ocorrência de outro. Relações de exclusão – *(excludes)* – como essa ocorrem, por exemplo, entre o EF Fonte e o EF Área, na medida em que a enunciação do local que o Tema ocupa inicialmente, antes de trocar sua localização, exclui a possibilidade de enunciação do cenário através do qual o movimento ocorre.

De maneira inversa, a relação *requires* se estabelece sempre que a enunciação de um determinado EF só puder ocorrer na presença de outro. Como exemplo, temos os EFs nucleares do frame de `Parentesco`[8] (Figura 19), que estabelecem entre si uma relação em que o EF Alter e o EF Ego requerem, igualmente, a presença um do outro para ocorrerem.

Além das relações entre EFs, a FrameNet também prevê outros tipos de relações, sobre as quais se expõe a seguir.

[8] http://webtool.framenetbr.ufjf.br/index.php/webtool/report/frame/showFrame/95

Figura 19 - EF nucleares do frame `Parentesco`



Fonte: Base de dados da FrameNet Brasil

3.1.1.1 Relações entre Frames

Na FrameNet, os frames encontram-se relacionados através de relações tipadas que conectam EFs equivalentes entre os participantes da relação. As relações entre frames podem ser de Herança, Subframe, Precedência, Perspectiva, Uso, Incoativo_de e Causativo_de. Há ainda a metarrelação Veja_também, que não representa uma conexão semântica de fato, mas serve para ajudar o consulente humano a se certificar de que está lendo o frame correto.

A relação de Herança (*Inheritance*) se estabelece quando um frame é um subtipo específico de um frame mais amplo e herda as características desse último. Nesse tipo de relação, todos os EFs, *subframes* e tipos semânticos do frame mãe serão passados para o frame filho (FILLMORE et al., 2003, p. 243). Como exemplo, temos o frame `Visitar`, evocado por ULs como *visitar.v* e *visitante.n*, que herda e especifica, a partir do frame mais genérico `Agir_intencionalmente`, os EFs AGENTE – que realiza uma ação intencional – e AÇÃO. Assim, em princípio, uma relação de herança poderia ser utilizada para a geração de paráfrases mais genéricas de sentenças que instanciam o frame filho com aquelas que instanciam o frame mãe (ELLSWORTH & JANIN, 2007). Assim, uma sentença como (2) poderia ser parafraseada como (3), fazendo uso da relação de Herança. Como a FrameNet, seguindo a tradição da Linguística Cognitiva, trabalha com heranças múltiplas, a mesma sentença poderia ser parafraseada como (4), caso a relação considerada fosse a que se estabelece entre `Visitar` e `Cenário_visita_estadia` (Figura 20).

(2) Maria visitou a Reitoria da UFJF

(3) Maria realizou uma ação.

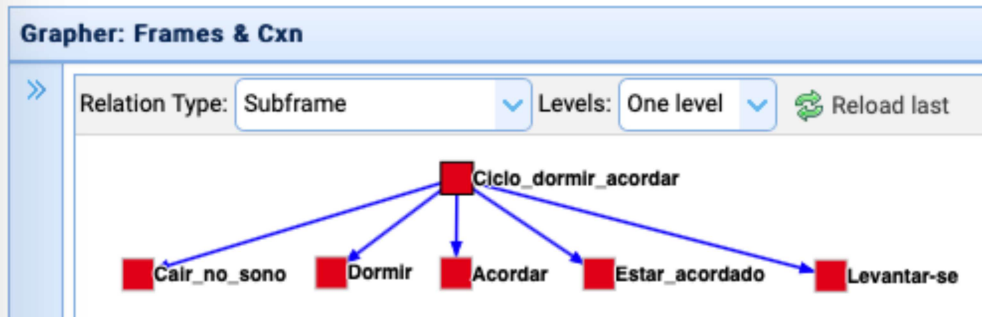
(4) Maria esteve na Reitoria da UFJF

Figura 20 - EF nucleares do frame `Parentesco`

Fonte: WebTool da FrameNet Brasil

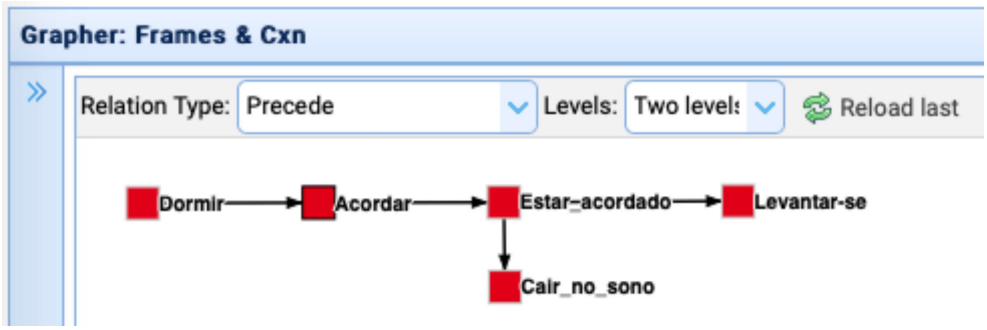
Já a relação de Sub-frame (*Subframe*) é estabelecida entre frames mais complexos e suas sub-partes, geralmente descrevendo etapas ou eventos sequenciais em que o frame mais amplo pode ser subdividido. Pode ser conjugada como uma relação de Precedência (*Precedes*), que ocorre quando a sequência de etapas ou eventos estabelece entre si uma relação temporal, seguindo uma ordem cronológica. Como exemplo, temos os frames `Cair_no_sono`, `Dormir` e `Acordar`, que estabelecem com o frame `Ciclo_dormir_acordar` – cenário em que uma entidade se encontra em um estado de consciência externa reduzida, permanece nesse estado por um certo período de tempo e normalmente retorna à consciência plena – uma relação de sub-partes (Figura 21) que, ao mesmo tempo, estabelecem entre si uma relação de precedência, na medida marcam uma ordem sequencial de eventos (Figura 22): alguém cai no sono, depois está dormindo, depois acorda e assim por diante.

Figura 21 - Relações de subframe do frame `Ciclo_dormir_acordar` no Grapher da FNBR representadas por setas azuis.



Fonte: WebTool da FrameNet Brasil

Figura 22 - Relações de precedência do frame `Acordar` no Grapher da FNBR representadas por setas pretas.



Fonte: WebTool da FrameNet Brasil

Implementação de aspecto fundamental da Semântica de frames, como já se apontou acima, a relação de Perspectiva (*Perspective_on*) conecta frames que abordam diferentes aspectos ou perspectivas sobre uma mesma cena, como no caso da relação que se estabelece entre o frame `Emoções` – em que um experienciador tem um estado emocional particular que pode ser descrito em termos de um estímulo específico que o provoca – e as diferentes perspectivas oferecidas pelo frame `Foco_no_estímulo` – que reflete a perspectiva a partir do estímulo que causa um tipo específico de emoção em um experienciador, como na sentença (5) – e o frame `Emoção_com_foco_no_experienciador` – que reflete a perspectiva do experienciador em relação ao conteúdo responsável por desencadear uma emoção específica, como em (6).

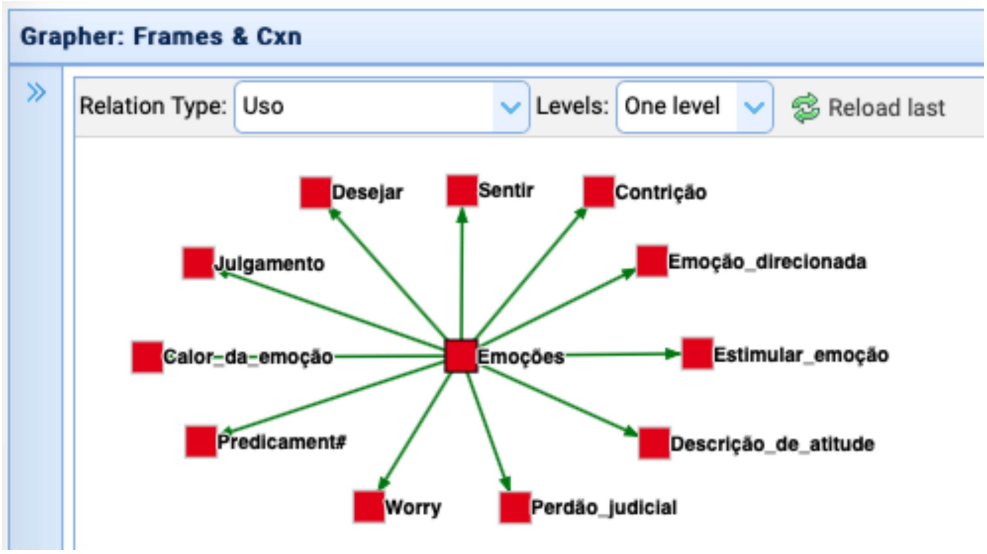
(5) A beleza das obras de arte é *surpreendente.a*

(6) Ao ver as obras de arte, ficou *impressionado.a*

A relação de Uso (*Using*), estabelecida quando um determinado frame só pode ser compreendido tendo em mente um segundo frame que atua como pano de fundo para sua estruturação e compreensão, é a mais fluida das relações, sendo na verdade uma estrutura residual de versões anteriores da FrameNet. Novamente, podemos tomar como exemplo o cenário mais abstrato estabelecido pelo frame `Emoções` (Figura 23), que é usado por frames como `Desejar` (evocado por ULs como *querer.v* e *saudade.n*) ou `Sentir` (evocado por ULs como *experieneciar.v* e *calma.n*), e serve como pano de fundo para sua estruturação na medida em que a compreensão de cenários representados por frames que tratam de sensações e desejos que afetam um experienciador – como em (7) – só é possível a partir da construção prévia de um cenário onde esse experienciador é uma entidade capaz de sentir emoções.

(7) *Sentiu.v* uma enorme *saudade.n*

Figura 23 - Relações de uso do frame `Emoções` no Grapher da FNBR representadas por setas verdes.



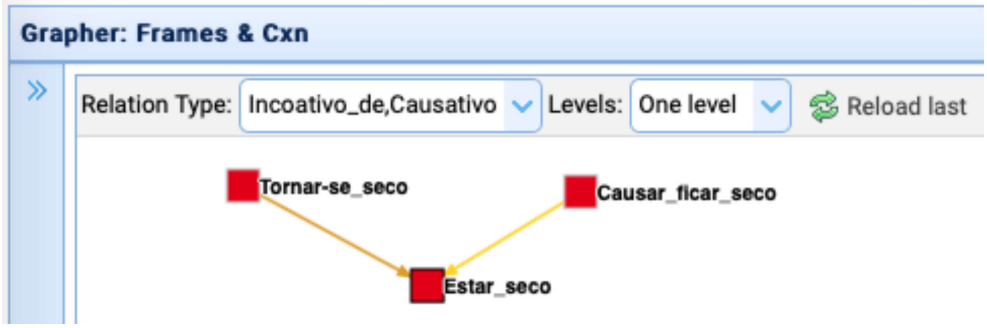
Fonte: WebTool da FrameNet Brasil

Por fim, relações de Incoativo de (*Inchoative_of*) e Causativo de (*Causative_of*) marcam, respectivamente, uma mudança de estado entre dois frames e suas relações de causalidade, como no caso do frame `Estar_seco` (Figura 24), que tem como causativo o frame `Causar_ficar_seco` – evocado por ULs como *secar.v*, presente na sentença (8) – e como incoativo o frame `Tornar-se_seco` – como em (9).

(8) O sol *secou.v* a roupa no varal.

(9) A roupa *secou.v* ao vento.

Figura 24 - Relação causativo/incoativo do frame `Estar_seco` no Grapher da FNBR representadas por setas amarelas.



Fonte: WebTool da FrameNet Brasil

3.1.1.2 Relações Qualia ternárias mediadas por frames

Para além das relações originalmente propostas pela Berkeley FrameNet, sobre as quais se tratou nas seções anteriores, outro tipo de relação foi proposto pela FrameNet Brasil, qual seja o das relações qualia ternárias mediadas por frames (TORRENT et al., 2022).

Os papéis Qualia – *Qualia roles*, (PUSTEJOVSKY, 1995) – surgem com base na ideia de que a forma como os humanos compreendem o significado das palavras pode ser descrita a partir de quatro fatores geradores, que capturam a maneira como são compreendidas as relações entre os objetos no mundo e fornecem explicações para o comportamento linguístico dos itens lexicais (PUSTEJOVSKY & JEZEK, 2016, p. 5). Essas relações se subdividem em:

- Quale Formal *(Formal_of)*, categoria básica que marca a posição de um item em um domínio mais amplo enquanto codifica informações taxonômicas sobre esse item. É essa relação que nos permite dizer que um determinado objeto é "um tipo de" *(type-of)* uma determinada categoria de objetos – por exemplo, *pizza.n* tem relação formal com *comida.n*, já que pizza é um tipo de comida.

- Quale Constitutivo *(Constitutive_of)*, que marca a relação entre um item e as partes que o constituem, permitindo dizer do que um determinado item é feito. Tomando novamente o item *pizza.n* como exemplo, podemos ter como elementos constitutivos *queijo.n*, *tomate.n* ou *farinha.n*, já que queijo e tomate são itens que podem fazer parte de uma pizza.

- Quale Agentivo *(Agentive_of)*, utilizado para marcar fatores relacionados à origem de um determinado item ou elementos que desempenham algum papel na criação do item. Usando mais uma vez o item *pizza.n*, vemos que *pizzaria.n* – local responsável pelo processo de criação da pizza – é seu agentivo.

- Quale Télico *(Telic-of)*, que formaliza as informações sobre o propósito ou a função de um determinado objeto, como no caso de *comer.v*, que representa o propósito de *pizza.n*, na medida em que pizza é um alimento.

Tais relações podem ser formalizadas como apresentado na Figura 25.

Partindo dessas relações – e considerando o papel preponderante das estruturas qualia na descrição de entidades que não apresentam padrões de valência discrepantes, tais como os nomes de entidade[9] – a base de dados da FrameNet Brasil propõe a modelagem

[9] Nomes de entidade, diferentemente de nomes de evento e verbos, não apresentam grande variação no posicionamento dos EFs do frame que evocam em sua localidade sintática. Geralmente, nesses frames, há apenas um EF, o qual é incorporado pelo radical do nome de entidade.

Figura 25 - Relações qualia ternárias para *pizza.n*.

$$\begin{bmatrix} pizza.n & \\ QUALIA & \begin{bmatrix} F = prato.n \\ T = comer.n \\ C = farinha.n, queijo.n, tomate.n \\ A = pizzaria.n \end{bmatrix} \end{bmatrix}$$

Fonte: O Autor

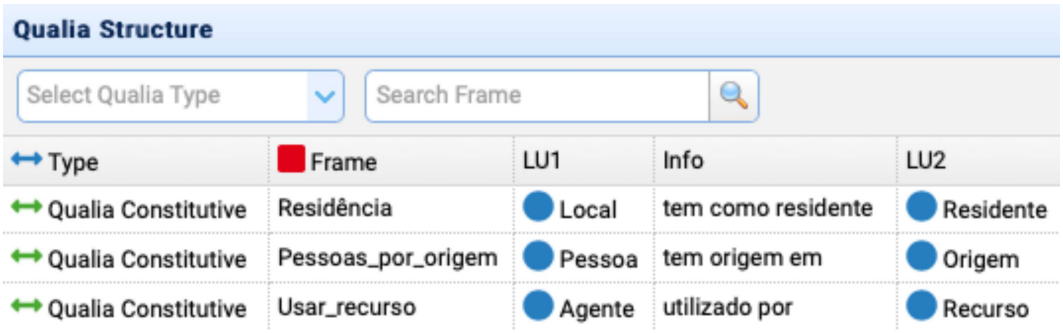
de um subtipo mais específico e granular de relações qualia[10]. Nessa modelagem, chamada de qualia ternário mediado por frames (TORRENT et al., 2022), Unidades Lexicais têm sua relação mediada pela utilização de frames específicos, que servem de pano de fundo para as relações qualia básicas, gerando informações semânticas adicionais e contribuindo para a especificação do tipo de relação entre as ULs.

Na Figura 26, temos um exemplo de representação de estrutura qualia ternária extraído da base de dados da FrameNet Brasil. Na primeira coluna, temos um dos quatro tipos básicos de qualia (formal, agentivo, constitutivo e télico); na coluna seguinte, o frame que opera como pano de fundo para a relação que será estabelecida nas colunas subsequentes; nas colunas marcadas como LU1 e LU2, temos os EFs do frame de pano de fundo listado na primeira coluna segundo os quais as ULs envolvidas na relação devem ser enquadradas; finalmente, na coluna Info, temos o tipo de relação qualia ternária estabelecida entre as ULs. Assim, observando a tabela, é possível dizer que, se tomarmos como pano de fundo o frame de `Pessoas_por_origem`, a primeira UL da relação (LU1) será uma Pessoa que tem como origem (Info) a segunda UL da relação (LU2), marcada pelo EF Origem. Tal relação pode conectar, por exemplo, na base da FrameNet Brasil, a UL *turista.n* à UL exterior.n.

Assim, frames são utilizados como mediadores de relações qualia ternárias para solucionar tanto o problema da falta de ligações diretas entre Unidades Lexicais na base de dados da FrameNet quanto problemas de baixa especificidade nas relações de qualia. Segundo BELCAVELLO et al. (2020):

[10] O que representa uma alternativa à incorporação de uma ontologia de subqualia externa, tal como a proposta na Brandeis Ontology (PUSTEJOVSKY et al., 2006)

Figura 26 - Qualia ternário mediado por frame na FrameNet Brasil.



Fonte: WebTool da FrameNet Brasil

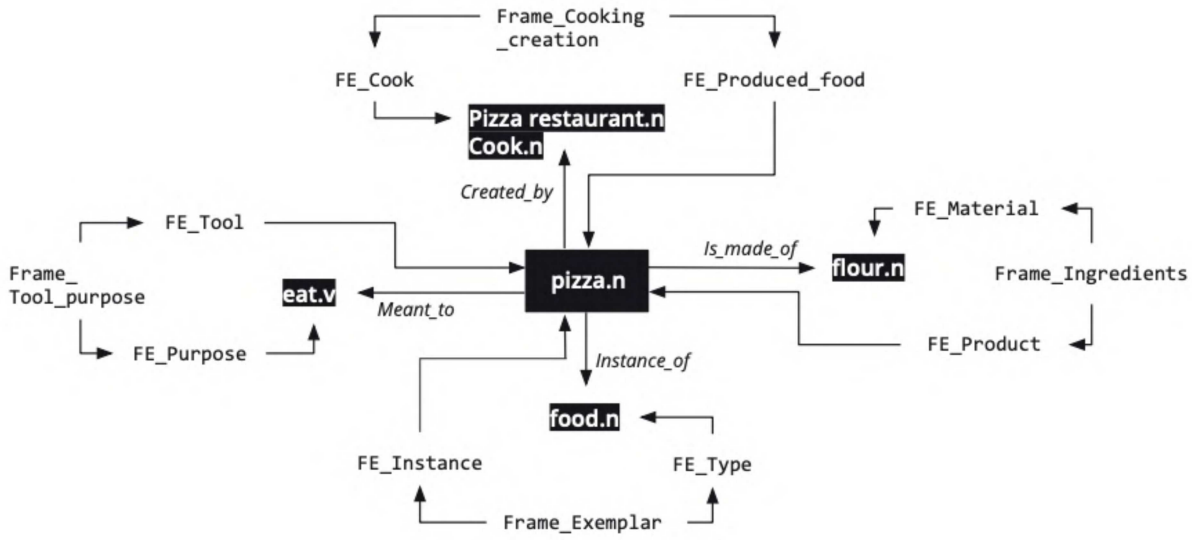
> Nesse tipo inovador de relação ternária, duas ULs, 1 e 2, estão ligadas uma a outra por meio de um quale que usa como pano de fundo a estrutura de frames como forma de adensar o papel quale em termos de informação semântica. Para cada quale, foi escolhido um conjunto de frames da base de dados FN-Br com base nos aspectos que cada quale especifica. LU1 estaria relacionada a um EF do frame que atua como pano de fundo, enquanto a LU2 estaria relacionada a outro EF do mesmo frame. O frame especificaria a semântica da relação estabelecida[11]. (BELCAVELLO et al., 2020, p. 26)

Dessa forma, modeladas as qualia ternárias, as relações exemplificadas para a Unidade Lexical *pizza.n* passam a ser representadas conforme a Figura 27.

Nesse exemplo, vemos que, no banco de dados FrameNet Brasil, a Unidade Lexical *pizza.n* estabelece relação com outras cinco Unidades Lexicais via qualia. *Pizza.n* tem uma relação agentiva (*Created_by*) com *Pizza_restaurant.n* e *Cook.n*. Essa relação é mediada pelo frame `Cooking_creation`, que relaciona *pizza.n* ao EF Produced_food e as Unidades Lexicais *Pizza_restaurant.n* e *Cook.n* ao EF Cook. *Pizza.n* também estabelece uma relação constitutiva (*is_made_of*) com a Unidade Lexical *flour.n*, que, por sua vez, é mediada pelo frame `Ingredients`, tendo *pizza.n* relacionada ao EF Produtc e *farinha.n* ao EF Material. A relação formal (*instance_of*) é estabelecida por meio do frame `Exemplar`, sendo *pizza.n* relacionada ao EF Instance e *food.n* ao EF Type. Finalmente, a relação télica (*Meant_to*) estabelece que *pizza.n* está relacionada ao EF Tool – o objeto ou processo desenvolvido especificamente para atingir um determinado propósito – no frame `Tool_purpose` e a Unidade Lexical *eat.v* ao EF Purpose, no

[11] "In this innovative type of ternary relation, two LUs, 1 and 2, are linked to each other via a given quale using the background structure of frames as a way to make the quale role denser in terms of semantic information. For each quale, a set of frames was chosen from the FN-Br database based on the aspects of such quale they specify. LU1 would be related to an FE of the background frame, whereas LU2 would be related to another FE of the same frame. The frame would specify the semantics of the relation."

Figura 27 - Exemplo de relações qualia ternárias para *pizza.n*.



Fonte: (BELCAVELLO et al., 2020)

mesmo frame.

Ainda segundo Belcavello et al. (2020), apenas Elementos de Frame nucleares e nucleares não-expressos podem ser recrutados como mediadores nas relações qualia ternárias. Isso ocorre em virtude da própria distinção entre os Elementos de Frame nucleares e os não-nucleares na metodologia FrameNet, já que apenas EFs nucleares, por serem específicos do frame, são os únicos que permitem a diferenciação entre frames (BELCAVELLO et al., 2020, p. 27).

3.1.2 **Anotação de Corpus na FrameNet**

Na FrameNet, palavras são tratadas como Unidades Lexicais, que são unidades formadas pela correlação entre um lema e seu significado em um determinado frame (RUPPENHOFER et al., 2006, p. 5). Na estrutura de dados da FrameNet, cada UL é relacionada ao frame que ela evoca e que descreve não apenas o contexto em que a UL deve ser compreendida, mas também os demais participantes e entidades envolvidas na cena evocada pela UL. Se tomarmos como exemplo o frame de `Viagem`, ele pode ser evocado pela ocorrência de ULs como *viajar.v*, *expedição.n* e *jornada.n*.

Cabe aqui destacar que um mesmo lema pode ser utilizado para a criação de mais de uma Unidade Lexical, como no caso de *pegar.v* do frame `Andar_de_veículo`, com o sentido de viajar por um meio de *transporte*; e *pegar.v* do frame `Pegar`, que significa agir de modo a adquirir algo. Além disso, ULs podem ser monolexêmicas, quando formadas por apenas uma palavra, ou polilexêmicas, como nas ULs *casa de férias.n* ou *chefe de cozinha.n*. A identificação dos padrões de valência sintática e semântica dessas Unidades

Lexicais é realizada a partir da análise das anotações de sentenças extraídas de *corpora*.

A base de dados lexicais da FrameNet é composta por dois tipos de anotações: anotações de texto corrido e anotações lexicográficas. A anotação de texto corrido identifica e descreve os frames que ocorrem em textos completos a partir da identificação das ULs que aparecem nesses textos, sem que haja uma seleção prévia, por parte do anotador, da UL que será anotada. Na anotação lexicográfica, sentenças extraídas de *corpus* são selecionadas com o intuito de mostrar a variedade de possibilidades de valências de ULs específicas. Em ambos os casos, a tarefa de anotação consiste da aplicação de etiquetas aos elementos que compõem cada uma das três camadas de anotação: 1. Elementos de Frame (EFs); 2. Função Gramatical (ou GF, sigla em inglês para *Grammatical Function*), que identifica elementos como objeto direto, objeto indireto, dependente, etc. e; 3. Tipo Sintagmático (ou PT, sigla para *Phrase Type*), que identifica o elemento como sendo um sintagma adjetival, advérbio, nome, etc.

A princípio, salvo a existência de relações de exclusão ou de *core sets*, todos os EFs nucleares precisariam ser anotados na sentença em que o frame a que pertencem é evocado. Existem, no entanto, casos em que EFs nucleares – necessários para a instanciação de um frame – encontram-se semanticamente presentes, mas sintaticamente ausentes. Casos como esse são chamados de Instanciação Nula e são tratados na FrameNet Brasil de três maneiras distintas (TORRENT et al., 2018, p. 110-111):

- Instanciação Nula Definida (ou DNI, sigla em inglês para *Definite Null Instantiation*), que ocorre quando um EF ausente na localidade sintática pode ser recuperado a partir do contexto linguístico, como nos casos de elipse e anáfora zero de argumento não-sujeito.

- Instanciação Nula Indefinida (ou INI, sigla para *Indefinite Null Instantiation*) – tratada originalmente na Berkeley FrameNet como uma propriedade de valência das ULs – ocorre, em português do Brasil, quando elementos estão ausentes no texto sem que, no entanto, seja necessária a identificação de um referente no contexto linguístico.

  Para mostrar a diferença entre os critérios de aplicação dos rótulos de INI para sentenças em inglês e português, Torrent (TORRENT et al., 2018, p. 110) destaca o comportamento dos verbos *eat.v* e *comer.v*, que licenciam o INI do EF Ingeríveis no frame `Ingestão`, e usa os seguintes exemplos comparativos:

  (10) *Would you like to try some of this delicious cake?*
  Você gostaria de provar um pouco desse delicioso bolo?

  (11) *No, thanks, I already ate.*
  Não, obrigado, eu já comi

Em inglês, a resposta para a pergunta só tem uma interpretação possível: aquela em que o ingerível é uma INI, sem um referente explícito, uma vez que, se o falante quisesse dizer que já comeu do bolo, o pronome *it* deveria aparecer após o verbo. Já em português, a sentença permite duas leituras: tanto para o caso de já ter comido o referido bolo quanto para os casos em que, previamente, tenha comido qualquer outro alimento, na medida em que os "verbos em português do Brasil - em geral - licenciam omissões de Objetos Diretos nos casos em que há tanto uma referência anafórica quanto uma existencial[12]" (TORRENT et al., 2018, p. 110).

- Instanciação Nula Construcional (CNI, *Constructional Null Instantiation*), usada na Berkeley FrameNet para os casos em que uma construção gramatical licencia a omissão do constituinte ao qual seria atribuída uma etiqueta de EF. Como o inglês exige que os sujeitos verbais sejam expressos em frases declarativas e o português do Brasil não, os CNIs do FrameNet Brasil incluem sujeitos omitidos.

Definidas as etiquetas dos EFs – derivadas dos frames da FrameNet – e definidos aqueles EFs cuja instanciação é nula, inicia-se a anotação das etiquetas referentes à camada de Função Gramatical, que pode conter uma das seguintes funções:

- Argumento externo (EXT), para marcar o termo que ocupa essa função, como na sentença (12)

  (12) $[Marcelo_{Ext}]$ $comprou^{V}$ *frutas*.

- Objeto Direto (ObjD), para os elementos que fazem parte do argumento interno mas não são regidos por preposição, como em (13).

  (13) *Marcelo* $comprou^{V}$ $[frutas_{ObjD}]$.

- Objeto Indireto (ObjInd), para o argumento interno que é regido por preposição e pode ser substituído pelo pronome *lhe*, como em (14).

  (14) *Marcelo* deu^{V} *um livro* $[paramim_{ObjInd}]$.

- Dependente (Dep), para destacar adjuntos e demais funções sintáticas não cobertas pelas outras etiquetas, como em (15).

  (15) *Marcelo* $comprou^{v}$ *frutas* $[ontem_{Dep}]$.

[12] "(...)verbs in Brazilian Portuguese – in general – license Direct Object omissions where there is either an anaphoric reference or an existential one."

- Aposto (Aposto), para marcar um elemento que exemplifica ou especifica um item de valor substantivo ou pronominal que tenha sido anotado como alvo, como em (16).

  (16) $Marcelo^{N}$, $[formadoemArtes_{Aposto}]$, *faz doutorado em linguística.*

- Determinante Possessivo (DetPoss), para marcar pronomes que acompanham item nominal indicando a relação de posse, como em (17).

  (17) *O* $[meu_{DetPoss}]$ $dinheiro^{N}$ *não é suficiente para pagar a conta.*

- Núcleo (Núcleo), que marca o núcleo nominal, como em (18).

  (18) *Marcelo comprou* $[frutas_{Nucleo}]$ $frescas^{A}$.

- Quantificador (Quant), para marcar um item que faz referência a quantidade, como em (19).

  (19) *Marcelo comprou* $[dois_{Quant}]$ $quilos^{N}$ *de maçã.*

Na camada de anotação seguinte são aplicadas as etiquetas referentes ao Tipo Sintagmático, que relaciona diferentes tipos de sintagmas – adjetival, adverbial, nominal, verbal, preposicionado, etc. – e também tipos de verbos, advérbios e numerais.

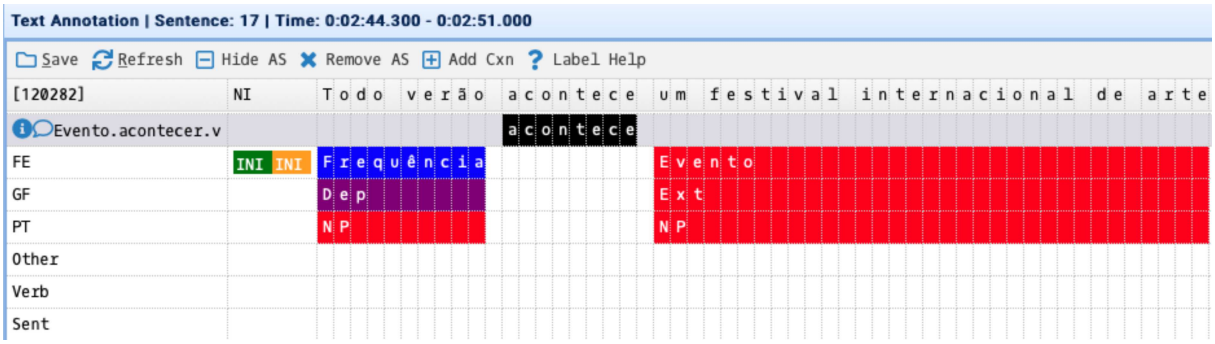
Assim, se tomarmos como exemplo a anotação da UL $acontece^{V}$ na sentença (20), extraída da base de dados da FrameNet (Figura 28), vemos que o sintagma "Todo verão" foi anotado para o EF Frequência do frame de `Evento` $acontecer^{V}$), assumindo a Função Gramatical de Dependente (Dep) e Tipo Sintagmático NP (sigla em inglês para *Noun Phrase*, ou Sintagma Nominal). A direita da UL $acontece^{V}$ está destacado o EF Evento, que tem Função Gramatical de Argumento Externo (Ext) e recebe a etiqueta PT (de *Phrase Type* ou Tipo Sintagmático) de Sintagma Nominal.

(20) Todo verão acontece um festival internacional de arte.[13]

Para alguns casos específicos, em que elementos relevantes da sentença não estão cobertos pelas três camadas de anotação iniciais, existe ainda a possibilidade de anotação de duas camada adicionais: *POS-Specific*, onde são anotados itens como verbos de suporte, partículas de aspecto e cópulas (Figura 29) e, finalmente, uma camada *Other*, onde geralmente são anotados pronomes relativos e seus antecedentes.

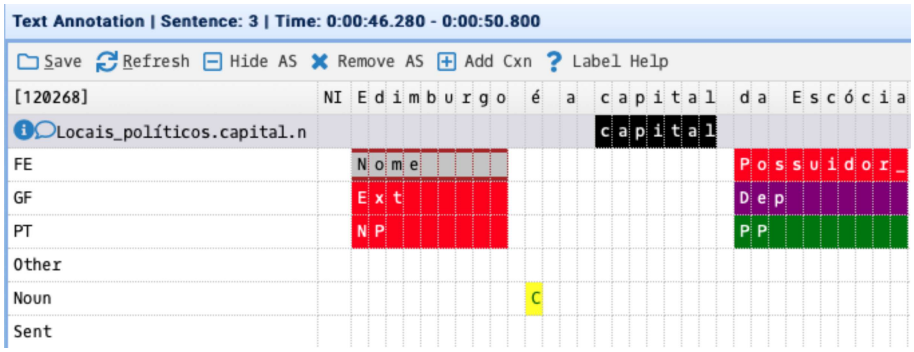
[13] Sentença número 17 do corpus multimodal da FrameNet Brasil.

Figura 28 - Exemplo de anotação de sentença na WebTool.



Fonte: WebTool da FrameNet Brasil

Figura 29 - Anotação da cópula na camada POS-Specific da sentença "Edimburgo é a capital da Escócia."



Fonte: WebTool da FrameNet Brasil

O conjunto de anotações realizadas para cada UL compõe o padrão de valência dessa UL, ou seja, suas possibilidades de comportamento sintático-semântico em termos do frame que ela evoca. O fato de tais padrões de valência serem constituídos a partir dos frames evocados pelas ULs faz com que, em alguma medida, constituam uma representação do conhecimento de mundo associado ao sentido dos itens linguísticos. A próxima seção trata desse ponto.

3.1.3 Representação de Informação Contextual na FrameNet

Como mencionado no Capítulo 3, contexto e perspectiva são aspectos centrais da teoria da Semântica de Frames – frames são, por definição, uma representação do contexto e da perspectiva a partir do qual o significado de um item lexical deve ser interpretado. Como implementação original da Semântica de Frames, a FrameNet, voltada originalmente para a análise de fenômenos linguísticos manifestos exclusivamente na modalidade textual, é capaz de fornecer representações computacionais de alguns aspectos do contexto, mas não de todos eles.

Buscando superar as limitações inerentes de uma análise semântica estruturada a partir de uma única modalidade comunicativa, a FrameNet Brasil (FNBr) vem, nos últimos anos, desenvolvendo projetos que buscam estender a estrutura original de sua rede de frames, incorporando dimensões adicionais de contexto que possibilitem uma compreensão mais detalhada e granular das informações textuais que já compõem sua base de dados. Para isso, a FNBr tem se dedicado à anotação de imagens e vídeos, explorando a ideia de que uma rede semântica enriquecida com dados multimodais e relações qualia pode aprimorar a capacidade de capturar a complexidade do significado – o que traria implicações significativas para várias aplicações em PLN, incluindo análise semântica, tradução automática e geração de língua.

Uma das inovações resultantes desse esforço em criar uma FrameNet multimodal – desenvolvida, em parte, a partir das pesquisas realizadas durante esta tese de doutorado – foi a criação de diretrizes para anotação de *datasets* multimodais que permitem que informações visuais e textuais possam ser integradas aos frames que já compõem a base de dados da FrameNet (TORRENT et al., 2022). Através dessa nova metodologia de anotação, imagens e vídeos puderam ser anotados com dados sobre frames e Elementos de Frame evocados por entidades visuais, expandindo assim as informações agregadas às Unidades Lexicais já presentes na base de dados da FrameNet Brasil. Essa abordagem multimodal permite à FNBr capturar um espectro mais amplo de significados, incorporando elementos como pistas visuais e contexto situacional que são frequentemente implícitos na comunicação.

Os resultados dessa metodologia apontam para a importância de representar computacionalmente informações contextuais de maneira estruturada, em oposição a tentativas de derivá-las apenas da manipulação da forma linguística. A ideia central é a de que imagens possam servir como simulações de contextos linguísticos (TORRENT et al., 2022) e, nesse sentido, a expansão do modelo da FrameNet para o domínio da multimodalidade é natural.

No intuito de mapear as possíveis inter-relações entre os modos comunicativos linguístico e visual, a seção seguinte se ocupa de uma revisão da literatura fundadora em gramáticas multimodais.

3.2 GRAMÁTICAS MULTIMODAIS

Um dos desafios do estudo de fenômenos linguísticos que evolvem multimodalidade diz respeito à complexidade decorrente da combinação de conteúdos visuais e textuais (COHN, 2020, p. 211). Humanos se comunicam naturalmente através da combinação de diferentes modalidades – fala, movimentos corporais, expressões faciais – e são capazes de combinar essas capacidades expressivas de maneiras criativas e altamente elaboradas. Da mesma forma, é também comum ver a comunicação escrita combinar textos com outros

recursos semióticos, como imagens estáticas ou em movimento, diagramas ou gráficos. Partindo dessa perspectiva, o termo multimodalidade é aqui adotado como "o uso de várias modalidades semióticas na criação de um produto ou evento semiótico, em conjunto com a maneira específica pela qual essas modalidades são combinadas[14]" (KRESS & VAN LEEUWEN, 2001, p. 20), considerando como textos multimodais aqueles em que:

> 1. O significado é construído a partir de diferentes recursos semióticos, cada um oferecendo potencialidades e limitações distintas; 2. A criação de significado envolve a produção de conjuntos multimodais; 3. Para estudar o significado, precisamos observar todos os recursos semióticos utilizados para formar o todo[15]. (JEWITT et al., 2016, p. 3)

Posto que a construção do significado não se dá através de uma simples soma entre as modalidades, mas, sim, pelo produto obtido quando essas são expressas em conjunto, ainda que uma das modalidades pareça ter papel acessório ou pouco relevante na construção do significado, ela pode adquirir papel de enquadre ou direcionamento de sentido. Assim, para discutirmos os processos de interpretação de modelos situacionais que envolvem conteúdo simultaneamente visual e verbal – em que leitores precisam negociar diferentes tipos de relações combinatórias entre imagens e também entre modalidades"[16]" (COHN, 2020, p. 211) – precisamos, inicialmente, estabelecer como se dão essas relações de complementaridade e criação de significado entre imagem e texto quando esses co-ocorrem em diferentes tipos de discursos multimodais para, em seguida, analisar as implicações dessas co-ocorrências em termos de relações semânticas.

Para analisar como se dão as relações entre imagem-texto, MARTINEC & SALWAY (2005) tomam como ponto de partida a taxonomia proposta por BARTHES (1977), que introduz diferentes tipos de relações de status entre as modalidades, descrevendo os três tipos fundamentais de relações de co-ocorrência imagem-texto: no primeiro, chamado de ancoramento, o texto funciona como suporte para imagem; no segundo, denominado ilustração, a imagem dá suporte ao texto, e, no terceiro, que Barthes chama de revezamento, imagem e texto têm igual importância.

Nas relações de ancoramento – quando um texto é utilizado para elucidar o sentido de uma imagem – o texto atua como guia, auxiliando o leitor na interpretação dos possíveis significados da imagem, fazendo com que ele evite alguns e receba outros e, assim, crie

[14] "(...) the use of several semiotic modes in the design of a semiotic product or event, together with the particular way in which these modes are combined."

[15] "1. Meaning is made with different semiotic resources, each offering distinct potentialities and limitations; 2. Meaning making involves the production of multimodal wholes; 3. If we want to study meaning, we need to attend to all semiotic resources being used to make a complete whole."

[16] "(...)readers must negotiate varying types of combinatorial relations between images as well as between modalities."

um processo de elucidação seletiva, "uma metalinguagem não aplicada à totalidade da mensagem icônica, mas apenas a alguns de seus sinais[17]" (BARTHES, 1977, p. 39). Assim, diante das diferentes possibilidades de interpretação de uma imagem, a fixação (ou ancoramento) de sentido gerada pelo pareamento texto-imagem atua como delimitadora do significado das informações visuais – como, por exemplo, a legenda que acompanha uma imagem e torna possível dizer "o que é essa imagem".

Nas relações de ilustração – quando a imagem está subordinada ao texto e, portanto, tem menor status – temos, por exemplo, os casos em que imagens funcionam como exemplos específicos em textos que descrevem conceitos gerais e, por isso, podem ser facilmente substituídas por uma imagem diferente sem que o pareamento imagem-texto se torne inválido – como, por exemplo, em um texto que descreve o termo "mamífero" e pode vir acompanhado da imagem de uma vaca, ou de uma baleia ou de um morcego. Nesses casos, o conteúdo visual não adiciona novas informações ao texto, mas amplia seu significado, assemelhando-se, segundo o autor, ao conceito lógico-semântico de "elaboração" *(elaboration)* descrito por MATTHIESSEN (1989), no qual um dos elementos elabora o significado de outro, especificando-o ou descrevendo-o.

Finalmente, sobre as relações de revezamento – onde imagem e texto têm igual importância como, por exemplo, no caso de histórias em quadrinhos – temos que:

> (...) o texto (na maioria das vezes um fragmento de diálogo) e a imagem mantêm uma relação de complementaridade; as palavras, da mesma maneira que as imagens, são fragmentos de um sintagma mais geral e a unidade da mensagem se dá em um nível superior, o da história. (...) Embora rara em imagens estáticas, a relação de revezamento se mostra muito importante em filmes, onde o diálogo funciona não apenas como elucidação, mas de fato avança a ação estabelecendo, na sequência de mensagens, significados que não são encontrados na própria imagem[18]. (BARTHES, 1977, p. 41)

Partindo dessa abordagem, MARTINEC & SALWAY (2005) expandem a taxonomia proposta por Barthes, destacando que uma distinção semelhante entre status e relações lógico-semânticas foi feita por MATTHIESSEN (1989) a fim de mapear as relações entre elementos. Ao investigar as maneiras como elementos em um texto se relacionam, os autores propõe um modelo para análise de uma sequência de orações interligadas – e a consequente elaboração de uma estrutura capaz de descrever os tipos de relações entre

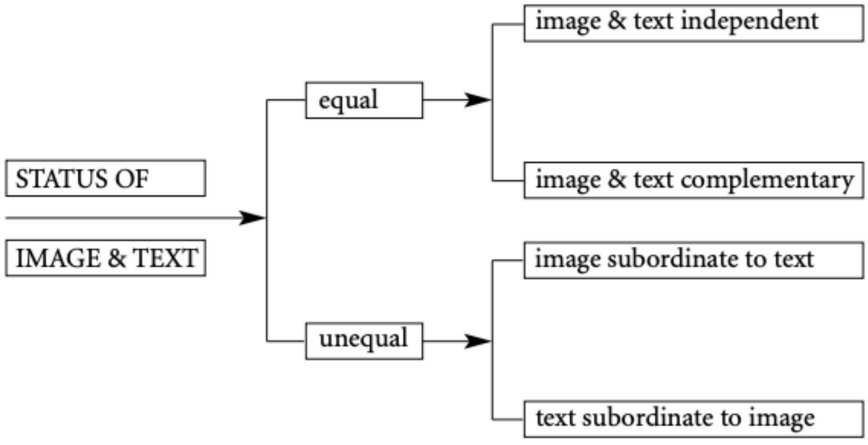
[17] "(...) a metalanguage applied not to the totality of the iconic message but only to a certain of its signs. Text is indeed the creator's right of inspection over the image"

[18] "Here text (most often a snatch of the dialogue) and image stand in complementary relationship; the words, in the same way as the images, are fragements of a more general syntagm and the unity of the message is realized at a higher level, that of the story. (...) While rare in the fixed image, this relay-text becomes very important in film, where dialogue functions not simply as elucidation but really does advance the action by setting out, in the sequence of messages, meanings that are not found in the image itself."

orações subordinadas – que mantém as dimensões de status e relação lógico-semântica claramente separadas: o status dos elementos no complexo oracional pode, portanto, ser igual ou desigual e, ao mesmo tempo, de forma independente, estar relacionado por meio de relações lógico-semânticas de expansão e projeção.

A partir desse modelo, é proposta uma análise semântica das relações imagem-texto dividindo essas relações em dois subsistemas que combinam, independentemente, status e relações lógico-semânticas, explicando como essas relações podem ser transpostas para o contexto da multimodalidade e fornecendo exemplos de possíveis combinações. As relações de status tratam da importância relativa entre texto e imagem – ou da dependência de um em relação ao outro. Assim, segundo os autores, como ocorre nas relações de subordinação entre orações, imagem e textos são considerados de status desigual quando um modifica o outro e o elemento modificador é considerado dependente ou está subordinado ao elemento modificado. Por outro lado, o status entre imagem e texto é considerado igual quando não há sinais de que imagem e texto modifiquem um ao outro e ambos podem ser entendidos individualmente – sendo considerados independentes – ou quando imagem e texto se modificam igualmente e ambos são necessários para uma comunicação bem-sucedida – sendo, assim, complementares (Figura 30) (MARTINEC & SALWAY, 2005, p. 343). As relações lógico-semânticas – que serão abordadas mais adiante – tratam de como imagens e textos podem modular ou modificar mutuamente seus significados quando combinados.

Figura 30 - Relações de status imagem-texto.

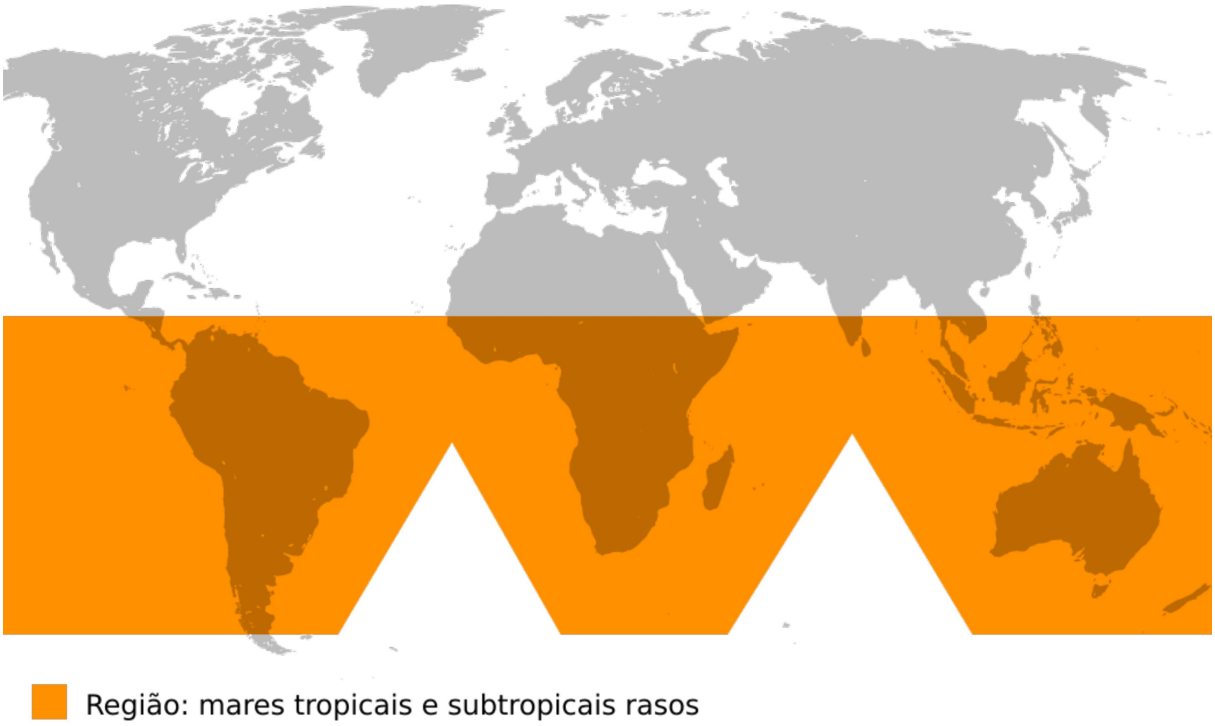


Fonte: (MARTINEC & SALWAY, 2005, p. 349)

Para exemplificar os casos em que imagem e texto são independentes – onde os conteúdos visual e textual existem em processos individuais e paralelos, mas estão relacionados por um contexto semântico – MARTINEC & SALWAY (2005) apresentam a imagem de um mapa que acompanha o verbete de uma enciclopédia (Figura 31) que trata das regiões oceânicas em que ocorrem a presença de moreias. A imagem mostra um

processo simbólico de atribuição (KRESS et al., 1996, p. 108) onde o mapa atua como portador do sentido e a faixa alaranjada como o atributo. O atributo, nesse exemplo, identifica no mapa as regiões oceânicas onde vivem as moreias. O texto, nesse caso, consiste de dois processos de identificação relacionais (MATTHIESSEN, 1989, p.122): no primeiro, identifica-se a parte do oceano onde vivem as moreias; no segundo, a região é descrita como "mares tropicais e subtropicais rasos". Como resultado, a combinação cria um novo significado ou interpretação que nenhuma das modalidades poderia ter alcançado individualmente.

Figura 31 - Relação de independência entre imagem e texto.

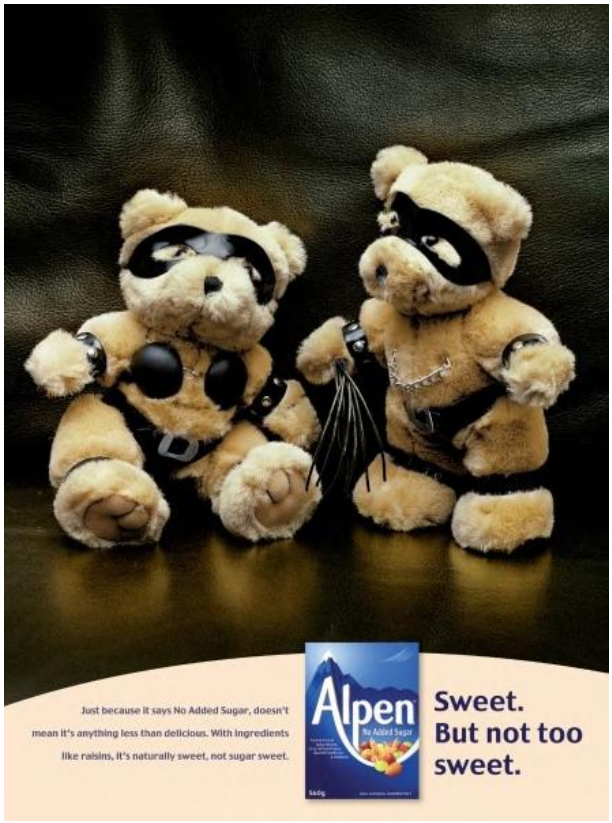


Fonte: o autor

Ao descrever a relação de complementaridade entre imagem e texto, os autores utilizam como exemplo um cartaz de propaganda de cereal matinal (Figura 32) que mostra dois ursos de pelúcia vestidos com roupas de couro, máscaras e chicote – fazendo referência à indumentária utilizada por praticantes de sado-masoquismo – operando como atributo intensificador de sentido do elemento textual "*Sweet. But not too sweet*" ("Doce. Mas não muito"). Nesse caso, onde o status da relação entre uma imagem e um texto é complementar, os dois elementos se combinam para desempenhar o papel de participantes no processo de formação de um sintagma visual-textual que expressa um significado maior que ambos isoladamente (MARTINEC & SALWAY, 2005, p. 344). Além disso, os autores também apontam para a ocorrência de uma outra instância da relação de complementaridade que se estabelece entre a imagem da caixa de cereal Alpen e o conjunto

de significado formado pelos ursos e o texto. Nessa segunda relação, o sentido de *'cute but naughty'* – algo como 'bonitinho, mas safado', expresso pelos ursos de pelúcia – é transferido para a marca de cereais, com o objetivo de tornar o produto atraente para seu público-alvo, composto por jovens.

Figura 32 - Relação de complementaridade entre imagem e texto.



Fonte: (MARTINEC & SALWAY, 2005, p. 344)

Enquanto nas relações de status independente e status complementar temos um texto inteiro sempre relacionado a uma imagem inteira, nas relações de status desigual, apenas uma parte de um texto ou de uma imagem se relaciona a outra imagem ou texto, respectivamente (MARTINEC & SALWAY, 2005, p. 346). Nesses casos, ao contrário do que ocorre nas relações de igual status, a relação de subordinação faz com que a interpretação de significado do pareamento texto-imagem ocorra de forma dependente, ou seja, o elemento subordinado não pode ser compreendido isoladamente.

Um exemplo da relação de subordinação da imagem em relação a um texto, similar ao dado pelos autores, pode ser visto nas Figura 33 e 34, que apresentam uma fotografia utilizada para ilustrar um texto jornalístico. Nesse tipo de conteúdo os componentes textuais geralmente são responsáveis por fornecer as informações relevantes sobre o evento sendo relatado. A imagem desse exemplo, isoladamente (Figura 33), não é suficiente para comunicar algo significativo e, por esse motivo, não seria, na visão dos autores, essencial

para entender a história contada pelo texto, sendo utilizada para ilustrar um aspecto do evento narrado.

Figura 33 - Imagem como ilustração para o texto.

Fonte: Business Insider Australia

Figura 34 - Relação de subordinação da imagem em relação ao texto.

Fonte: Business Insider Australia

De acordo com os autores, as características da relação imagem-texto no contexto jornalístico refletem o processo de produção do conteúdo, no qual um editor geralmente adiciona a imagem que vai ilustrar o artigo após o repórter ter concluído o texto. Nesses casos, a descrição que acompanha a imagem geralmente resume a notícia ao invés de descrever o que é retratado na imagem. Nesse exemplo, vemos que ao invés de tratar de uma pessoa específica, a imagem apresenta um indivíduo sem nome, que serve como referência visual para o grupo de pessoas a que o artigo se refere. Assim, quando vista novamente ao lado do título e seguida pelo texto jornalístico – que trata da crise mundial das companhias aéreas e seus impactos no quadro de funcionários – é que é possível

entender que o homem que arrasta uma mala enquanto caminha ao lado de um avião representa, na verdade, um membro de tripulação da referida companhia aérea afetado pelos cortes de pessoal.

Inversamente, uma indicação confiável da subordinação de um texto em relação a uma imagem é a presença de elementos textuais que só podem ser decodificados por referência a essa imagem. Dentre as possíveis maneiras de marcar essa dependência, a mais evidente ocorre nos casos de dêixis explícitas - quando o texto faz uma referência direta a imagem – comum nos textos de crítica de arte (MARTINEC & SALWAY, 2005, p. 348).

O exemplo escolhido pelos autores para demonstrar essa relação de subordinação do texto-imagem pode ser visto na Figura 35, que apresenta a pintura *A Man in a Black Cap*, de John Bettes, acompanhada de sua ficha catalográfica – contendo dados como técnicas e materiais utilizados na produção da obra, suas dimensões e ano de aquisição – e o texto da legenda que está afixada ao lado da pintura em exposição. Nesses casos, a pintura é o principal objeto de interesse e, por si só, transmite significado. O texto – adicionado posteriormente por alguém que não o artista – não é essencial e tem apenas a função de auxiliar o espectador na apreciação da obra. Entretanto, é possível perceber que trechos como "a inscrição na parte frontal indica" (*"the inscription on the front indicates"*) e "Originalmente, esse retrato era maior" (*"Originally, this portrait was larger"*) fazem pouco ou nenhum sentido quando lidos isoladamente, tornando evidente a relação de subordinação e dependência da imagem para sua compreensão.

Esses exemplos de relações de status entre textos e imagens – que formalizam os aspectos de subordinação ou dependência estabelecidos entre esses elementos quando apresentados em conjunto – são complementados por um conjunto de relações lógico-semânticas (Figura 36), que atuam como um segundo sistema de classificação, comparando o conteúdo visual apresentado pela imagem com aquilo que é referenciado no texto.

Para esse segundo sistema de classificação, MARTINEC & SALWAY (2005) utilizam os dois principais tipos de relações lógico-semânticas propostos pela gramática de MATTHIESSEN (1989): expansão e projeção. No contexto da multimodalidade, a projeção ocorre quando um conteúdo apesentado de forma textual ou visual é novamente representado na outra modalidade (MARTINEC & SALWAY, 2005, p. 354). Os casos em que isso se torna mais evidente são os diagramas que resumem os textos – onde as informações mais importantes da modalidade textual são selecionadas e modeladas de forma visual e diagramática – e histórias em quadrinhos – com balões de fala, que expressam locuções (manifestando uma projeção de texto, geralmente por processo verbal) e balões de pensamento, que expressam ideias (manifestando uma projeção de significado, geralmente por processo mental). Quanto à expansão, os três principais tipos de MATTHIESSEN (1989) – elaboração, extensão e aprimoramento – relacionam imagens e textos.

Figura 35 - Relação de subordinação do texto em relação a imagem.

John Bettes I active c.1531–1570

An Unknown Man in a Black Cap
1545
Oil on panel
470 x 410 mm
Inscribed '[...]I. 1545.' on the left; 'ÆTATIS. SV [...] on the right; on the back of the panel 'faict par Johan Bettes / Anglois' then repeated above 'faict par Johan Bettes Anglois'
('made by John Bettes, Englishman')
Purchased 1897
N01496

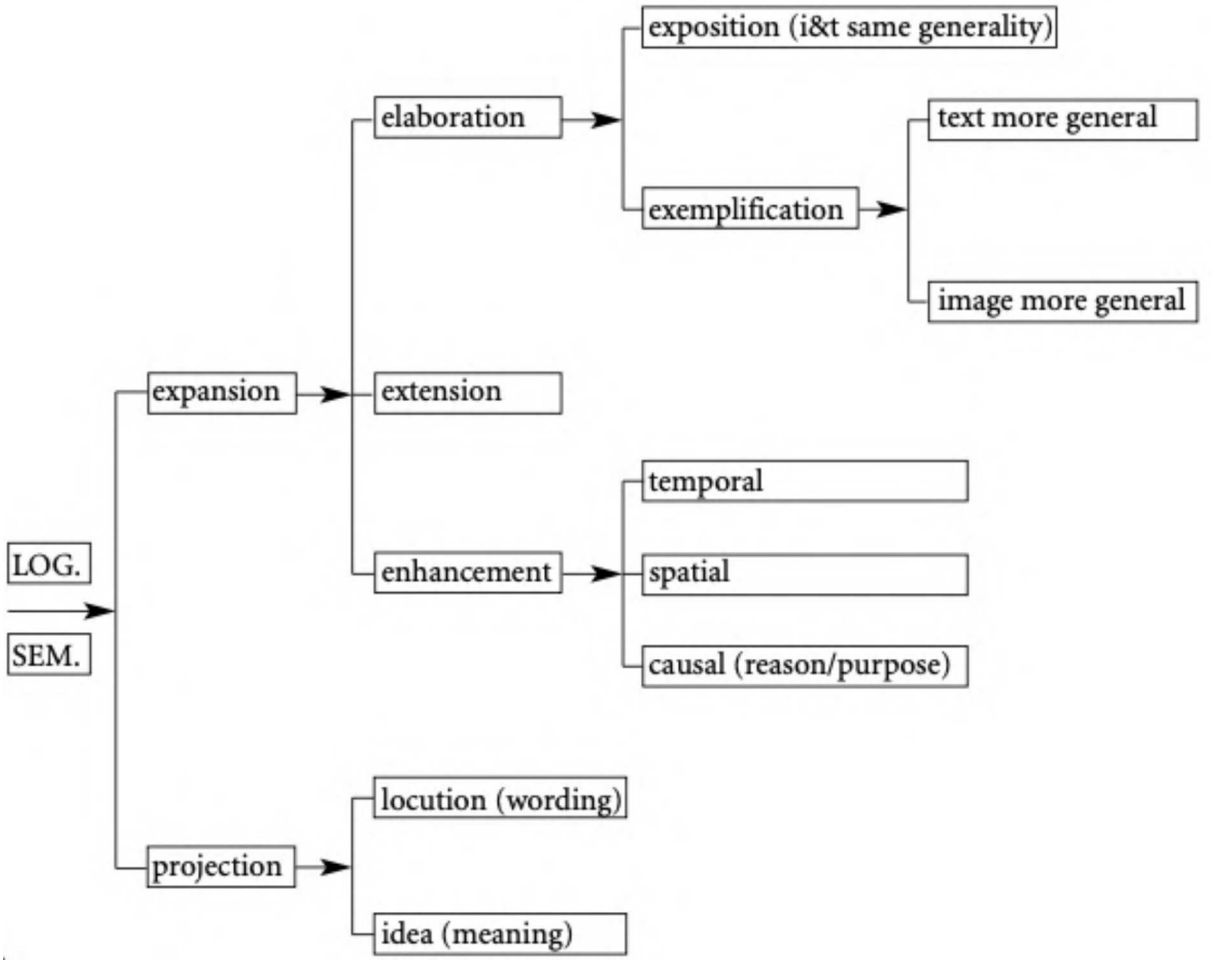
This is the earliest picture in the Tate collection. The artist's name is inscribed on the back, and the inscription on the front indicate that the work was painted 'in the year of our Lord 1545', and the sitter was aged 26. Bettes is the first recorded carrying out decorative work for Henry VIII's court in 1531-3, and he may have worked with Hans Holbein the Younger, the most famous Tudor painter. Originally, this portrait was larger, and would have had a blue background similar to the color often used by Holbein. Due to long exposure to light, the pigment (smalt) has changed to brown.

Fonte: (MARTINEC & SALWAY, 2005, p. 348)

Nesse modelo de relações, quando texto e imagem apresentam diferentes níveis de generalidade, um elabora (*elaborates*) o significado do outro através de uma relação de exemplificação (*exemplification*); quando as mesmas entidades e eventos são retratados e referidos nas duas modalidades e ambas apresentam o mesmo nível de generalidade, e uma reafirma o sentido da outra – seja apresentando-a a partir de outro ponto de vista ou apenas reforçando sua mensagem – temos uma relação de exposição (*exposition*) (MATTHIESSEN, 1989, p. 461-463). Fotografias em artigos jornalísticos, por exemplo, podem elaborar o texto – nesse caso, o título – especificando elementos textuais genéricos, como na caso da Figura 34, em que a imagem de um avião com a pintura da empresa Ryanair especifica a qual companhia aérea o título – elemento mais genérico – se refere.

Texto e imagem podem ampliar (*extends*) o significado um do outro quando uma das modalidades menciona ou descreve elementos completamente novos, adicionando novas informações à outra (MATTHIESSEN, 1989, p. 471). Um exemplo de texto que adiciona informações a uma imagem pode ser observado voltando à Figura 35. A sentença "Esse é o quadro mais antigo da Coleção Tate" (*This is the earliest picture in the Tate Collection*) fornece novas informações ao conteúdo mostrado no quadro e promove uma extensão no sentido da imagem, indo além daquilo que está representado visualmente – ou seja, os participantes envolvidos e os processos e circunstâncias de sua criação. O mesmo ocorre com a sentença "o nome do artista encontra-se inscrito no verso" (*The artist's name is inscribed on the back*), já que ele não pode ser visto na pintura quando ela está exposta.

Figura 36 - Esquema de relações lógico-semânticas.



Fonte: (MARTINEC & SALWAY, 2005, p. 354)

Um exemplo final de extensão ainda pode ser visto no trecho "Originalmente, este retrato era maior e teria um fundo azul semelhante à cor usada frequentemente por Holbein" (*Originally this portrait was larger, and would have had a blue background similar to the colour often used by Holbein*), pois acrescenta uma informação relativa a passagem do tempo, impossível de ser apreendida pela simples observação.

Finalmente, texto e imagem estabelecem entre si uma relação de aprimoramento (*enhancement*) quando modulam o significado um do outro através da qualificação de informações temporais, espaciais ou causais (MATTHIESSEN, 1989, p. 476). Observando novamente a Figura 34, vemos que há também uma relação de aprimoramento entre a imagem e o título, na medida em que é possível relacionar as demissões a que o texto faz referência como a causa do piloto deixar o avião levando sua bagagem.

No escopo dessa pesquisa, acreditamos que reconhecer como uma imagem e um texto estão relacionados é etapa fundamental para a compreensão de como um conjunto de dados multimodais codifica perspectivas e de que maneira as representações semânticas

associadas a esse conjunto de dados são influenciadas ou não por uma das modalidades. Tomando como referência os conceitos e a estrutura de relação propostas por MARTINEC & SALWAY (2005), interessam-nos, então, os casos em que a relação lógico-semântica entre imagem e texto seja a de elaboração, ou seja, quando as imagens apresentam exatamente as mesmas pessoas, objetos e eventos descritos ou referidos no texto.

Porém, interessa-nos em especial, não a definição da natureza da elaboração – se expositiva ou exemplificativa –, mas a investigação das nuances e perspectivas que uma modalidade impõe sobre a outra. Para tanto, para além de replicar a metodologia de criação do Multi30K para o português, esta tese estende o modelo de anotação da FrameNet Brasil para o domínio da multimodalidade, de modo a analisar, a partir de um modelo perspectivizado, como as modalidades verbal e visual interagem na produção de sentido no conjunto de dados. Os materiais e métodos empregados para tanto são apresentados no próximo capítulo.

4 MATERIAIS E MÉTODOS

Antes de tratarmos da metodologia usada na criação do novo conjunto de dados desenvolvido nessa pesquisa, discutiremos aqui a razão pela qual, apesar dos apontamentos realizados na seção 2.2, mantivemos o trabalho com descrições conceituais na criação no Framed Multi30k.

Dentre os diferentes métodos utilizados para descrever imagens (SHATFORD, 1986), as chamadas descrições conceituais (JAIMES & CHANG, 1999) são de maior prevalência nos datasets comumente utilizados em tarefas de PLN multimodal (UPPAL et al., 2022). Segundo (HODOSH et al., 2013):

> Descrições conceituais de imagens identificam o que é mostrado na imagem e, apesar de poderem ser abstratas (...), a compreensão da imagem está mais interessada em descrições objetivas da cena e das entidades representadas, seus atributos e relações, bem como os eventos de que participam. Por se concentrarem apenas no que está na imagem, as descrições conceituais diferem das chamadas descrições não-visuais, que fornecem informações adicionais que não podem ser obtidas apenas da imagem, por exemplo, sobre a situação, hora ou local em que a imagem foi tirada[1]. (HODOSH et al., 2013, p. 857)

Ainda sobre as descrições conceituais, os autores ressaltam que uma distinção adicional pode ser feita entre as chamadas descrições específicas – que podem identificar pessoas e locais por seus nomes – e descrições genéricas – que podem, por exemplo, descrever um indivíduo presente em uma imagem apenas pelo gênero (um homem) ou por uma ação que ele executa (um pintor); e a cena representada, como 'um parque' ou 'uma esquina'. Com exceção das entidades nomeadas que devem ser reconhecidas como tais – por exemplo, figuras públicas conhecidas (o Papa Francisco) ou locais de referência (a Torre Eiffel) –, argumentamos que o entendimento da imagem deve se concentrar nas informações capturadas por descrições genéricas, o que nos leva ao problema de como obter um conjunto de dados apropriado ao treinamento de sistemas de tradução automática multimodal, o qual contenha imagens relacionadas a descrições conceituais.

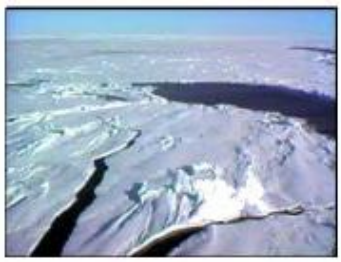
Considerados esses aspectos, vemos que, embora a internet disponibilize uma quantidade virtualmente ilimitada de imagens associadas a textos, a maior parte desses pareamentos imagem-texto não é considerada adequada para tarefas de PLN multimodal,

[1] "Conceptual image descriptions identify what is depicted in the image, and while they may be abstract (...) image understanding is mostly interested is mostly interested in concrete descriptions of the depicted scene and entities, their attributes and relations, as well as the events they participate in. Because they focus on what is actually in the image, conceptual descriptions differ from so-called non-visual descriptions, which provide additional background information that cannot be obtained from the image alone, e.g. about the situation, time or location in which the image was taken."

dada a dificuldade em controlar os aspectos do conjunto de dados que podem influenciar a performance dos sistemas resultantes. Pesquisas no campo de processamento de língua natural – como FENG & LAPATA (2010) – utilizam imagens e textos extraídos de artigos jornalísticos. Como discutimos na sessão 3.2, as imagens apresentadas nesse gênero textual geralmente são usadas ou para ilustrar as histórias – estabelecendo, assim, pouca relação conceitual com o texto a que estão associadas –, ou para sugerir inferências complexas do ponto de vista cognitivo. Assim, mesmo nos casos em que os textos utilizados para descrever as imagens fazem referência aos eventos narrados, esses normalmente tendem a se concentrar em informações que não podem ser obtidas pela leitura da imagem (Figura 37).

Figura 37 - Imagens e legendas extraídas de artigos da BBC News.

Cadbury will increase its contamination testing levels.



Satellite instruments can distinguish "old" Arctic

Fonte: (FENG & LAPATA, 2010)

Conjuntos de dados que utilizam fotografias extraídas de sites de compartilhamento de imagens com legendas criadas pelos próprios usuários também não são adequados às tarefas de PLN atualmente definidas em virtude do fato de que, quando os próprios fotógrafos acrescentam legendas às suas imagens, essas legendas tendem a fornecer informações sobre as circunstâncias em que as fotos foram tiradas ou detalhes específicos que não estão presentes na imagem como, por exemplo, os nomes das pessoas que aparecem na imagem ou o local em que a foto foi tirada (HODOSH et al., 2013, p. 858), num caso claro de ampliação da imagem pelo texto. Um exemplo desse problema pôde ser observado no desenvolvimento do *SBU Captioned Photo Dataset* (ORDONEZ et al., 2011), um conjunto de dados baseado em imagens e descrições criadas por usuários do site Flickr. Segundo o autor, dentre as imagens inicialmente coletadas, 67% foram descartadas por conterem legendas que descreviam informações que não podiam ser obtidas a partir da observação da imagem – como nomes de pessoas ou dos locais que aparecem na imagem – e 23% foram removidas por descreverem apenas um pequeno detalhe da imagem. Exemplos desse tipo de descrição – onde o conteúdo textual é vago ou insuficiente para fornecer informações sobre os aspectos visuais das entidade presentes na imagem - podem ser vistos na Figura 38.

Figura 38 - Fotos e legendas criadas pelos usuários do Flickr.

Fonte: (ORDONEZ et al., 2011)

Por outro lado, um conjunto de dados que contenha descrições semânticas precisas, porém muito longas – como o IAPR TC-12 (GRUBINGER et al., 2006) - também é pouco útil em tarefas de PLN como tradução automática e geração de respostas para perguntas com pistas visuais, na medida em que, mesmo contendo sentenças que descrevem apenas os elementos que podem ser identificados na imagem, tende a focar em descrições excessivamente detalhadas, compostas por várias sentenças, ao invés de tratar dos aspectos mais importantes de cada imagem (HODOSH et al., 2013, p. 859). A descrição da Figura 39, por exemplo, é formada por 60 palavras e traz detalhes sobre a cor do uniforme das duas comissárias de bordo – que mal podem ser vistas na imagem – e as diferentes tonalidades de marrom das montanhas ao fundo.

Figura 39 - IAPR TC-12 dataset.

a blue and white airplane is standing on a grey airport; a man and red cones are standing in front of it and two red-dressed hostesses and two passengers are directly on the stairs in front of the airplane; a brown landscape with high dark brown mountains with snow-covered summits and a light grey sky in the background;

Fonte: (GRUBINGER et al., 2006)

Assim, considerando que (i) os tipos de legendas que normalmente acompanham fotografias disponíveis na internet não descrevem de maneira satisfatória as imagens, dada a necessidade de que os sistemas de PLN encontrem correlações entre as representações

semânticas geradas para as modalidades visual e verbal, mas também que, por outro lado, (ii) as descrições conceituais não são, de fato, isentas de perspectiva e inferência, optamos por produzir uma nova extensão de uma família padrão de datasets – Flickr30K – porém com informações semânticas perspectivizadas agregadas aos metadados. Esse esforço envolveu tanto a criação de descrições das imagens através de duas tarefas de *crowdsourcing* – uma para criação de legendas e outra para tradução inglês-português de descrições já existentes – elaboradas com base nas diretrizes adotadas em tarefas de anotação semelhantes (RASHTCHIAN, 2010; HODOSH et al., 2013; ELLIOTT et al., 2016), e de duas tarefas de anotação de imagens e descrições para frames.

4.1 DATASETS-BASE

Nesta seção, apresentamos os *datasets* multimodais-multilíngues que usamos como base para construção do nosso novo dataset.

Os *datasets* Multi30K (ELLIOTT et al., 2016) e Flickr30K Entities (PLUMMER et al., 2015) – ambos extensões do Flickr30K (YOUNG et al., 2014) – serviram de base para a criação de fontes para o Framed Muklti30K (FM30k). O Flickr30K contém 31.783 fotos de atividades e eventos cotidianos, cada uma emparelhada com cinco diferentes legendas em inglês que descrevem entidades e eventos em cada imagem.

O dataset Multi30K é a expansão multilíngue do Flickr30K para múltiplos idiomas. Para o FM30K, repetimos a metodologia utilizada nas tarefas de expansão para o alemão. Essa expansão é composta por 31.104 traduções em alemão das descrições originais em inglês – uma por imagem, produzidas por tradutores profissionais – e 155.070 descrições originais em alemão, criadas por falantes nativos, sem relação com as descrições originais.

O dataset Flickr30K Entities, por sua vez, estende o Flickr30K ao adicionar correlações imagem-texto, com *bounding boxes* anotadas manualmente, criando pareamentos entre entidades – elementos visuais presentes nas imagens, como pessoas e objetos – e os sintagmas nominais que as descrevem em cada uma das sentenças. O *dataset* apresenta, também, 244.035 cadeias de co-referência que ligam menções de uma mesma entidade nas imagens aos sintagmas nominais correspondentes nas cinco descrições associadas àquela imagem.

Combinados, Multi30K e Flickr30K Entities fornecem um conjunto de dados como o exemplificado na Figura 40 .

Na sequência, descrevemos como foram realizadas as tarefas envolvidas na criação do Framed Multi30K, a partir dos *datasets* base descritos acima.

Figura 40 - Imagem 213216174.jpg do *dataset* Flickr30K, com uma das descrições originais em inglês e sua tradução para alemão. As correlações entre as regiões da imagem e seus descritores na sentença em inglês são indicadas por cores.

Fonte: Multi30K (ELLIOTT et al., 2016) e Flickr30K Entities (PLUMMER et al., 2015)

4.2 TAREFAS PARA CRIAÇÃO DO FRAMED MULTI30K

A primeira etapa na criação do FM30K foi, então, a coleta e organização dos dois *datasets* que expandimos. Todo o conteúdo original do Flickr30K foi armazenado em um novo banco de dados criado especificamente para este projeto, vinculando as 158.915 descrições em inglês às 31.783 imagens que elas descrevem, e atribuindo os nomes das imagens como IDs das sentenças. A etapa seguinte foi incluir as 31.014 traduções para o alemão do Multi30K e vinculá-las às suas contrapartes em inglês. Para isso, usamos as divisões de dados do Multi30K para obter triplas contendo as sentenças em cada idioma e a imagem que descrevem. Em seguida, consultamos o Multi30K e vinculamos a sentença em alemão ao ID da sentença em inglês. Com esta parte do banco de dados, sinalizamos as sentenças em inglês que tinham uma tradução para o alemão como referências para nossa tarefa de tradução.

Para as tarefas de anotação de frames e FEs, os dados do Flickr30K Entities foram importados para a mesma estrutura. Cada *bounding box* foi vinculada a uma única imagem e a até cinco sentenças, dependendo de estar ou não fundamentada em um sintagma nominal contido na sentença. Essa relação entre um sintagma e uma *bounding box* sempre especifica onde na sentença a entidade da imagem está sendo referida. Essa estrutura

permitiu-nos ter FEs anotadas sobre uma tupla (*bounding box*, trecho da sentença) ou apenas sobre caixas delimitadoras.

4.2.1 Traduções para o Português Brasileiro

As traduções do inglês para o português – *Portuguese Translations*, que abreviamos para PTT – foram feitas por 28 estudantes universitários, alunos do Bacharelado em Letras - Tradução inglês-português da UFJF, com proficiência avançada em inglês, divididos em dois grupos. Um grupo permanente de anotação, composto por 12 anotadores contratados para esta tarefa, foi responsável por criar 23.074 das 31.014 descrições PTT , ou 74,3% do total. O restante das descrições – 25,7% das descrições PTT – foi criado por um grupo de 16 estudantes que participaram de oficinas práticas de anotação linguística em troca de horas de crédito acadêmico. Ambos os grupos passaram por um treinamento na tarefa de anotação durante uma oficina de 15 horas, durante as quais foram instruídos sobre o uso da ferramenta de anotação e puderam participar de sessões práticas com subconjuntos de teste do corpus original. Além disso, os anotadores também participaram de reuniões semanais de alinhamento, durante as quais era possível fazer perguntas sobre a tarefa e buscar esclarecimentos sobre quaisquer questões surgidas durante o processo de anotação. A qualidade das anotações foi avaliada tanto manualmente – por meio de verificações periódicas de subconjuntos das anotações – quanto por métodos automatizados – para questões como erros de digitação, gramática e ortografia. Aos anotadores de melhor desempenho, foram oferecidas posições permanentes na equipe de anotação, garantindo que anotações de alta qualidade compusessem a maior parte do *dataset*. Anotadores de baixo desempenho receberam *feedback* adicional durante as reuniões semanais e, nos poucos casos de anotações que não atenderam às diretrizes, essas anotações eram descartadas e os anotadores eram removidos da tarefa.

Para garantir o alinhamento entre as traduções do Multi30K em alemão e as traduções para o português brasileiro, selecionamos o mesmo subconjunto de 31.014 descrições originais em inglês usado pelo Multi30K para a tarefa de tradução para o alemão. Os tradutores de português brasileiro também seguiram a mesma metodologia usada na tarefa de tradução do conjunto de dados Multi30K – ao ver a imagem e a descrição original em inglês, os anotadores foram solicitados a produzir uma tradução correta e fluente da descrição da imagem para o português brasileiro.

4.2.2 Descrições Originais em Português Brasileiro

As descrições originais em português – que chamamos de *Portuguese Originals*, abreviadas como PTO – foram criadas por 148 estudantes universitários da Licenciatura em Letras e do mestrado e doutorado em Linguística da UFJF, falantes nativos do português do Brasil, também divididos em dois grupos. A equipe de anotação permanente contava com

22 estudantes e produziu 81.834 (51,5%) das 158.915 descrições PTO – cinco por imagem do conjunto de dados Flickr30K. O restante das legendas foi criado por 126 estudantes matriculados em oficinas de anotação prática que concediam créditos acadêmicos, com uma média de 612 descrições PTO criadas por aluno. Mais uma vez, ambos os grupos receberam instruções na tarefa por 15 horas antes de se envolverem com o trabalho.

Seguindo novamente a mesma metodologia utilizada para a criação de descrições em alemão, os estudantes foram apresentados a uma versão traduzida da interface de coleta de dados desenvolvida originalmente por HODOSH et al. (2013), com instruções traduzidas do inglês para o português. Para evitar fadiga dos anotadores e garantir a qualidade das descrições, cada estudante recebeu uma cota semanal de aproximadamente cinquenta sentenças por hora de trabalho.

Para garantir a qualidade das novas descrições originais, foram utilizados métodos de inspeção manuais e automatizados. Primeiramente, procuramos por quaisquer descrições duplicadas da mesma imagem e as substituímos por uma nova sentença, criada por outro anotador. Este foi o caso de 76 descrições – menos de 0,1% to total. Foram também substituídas sentenças compostas por menos de quatro palavras – pouco mais de 1% do total. Por fim, outras 168 sentenças foram editadas ou refeitas por apresentarem problemas como caracteres especiais, erros de digitação, ou adjuntos entre parênteses e barras indicando conjunções. Problemas como a falta de pontos finais, espaçamento e capitalização incorretos foram corrigidos de forma automática.

Por estarmos interessados em descrições conceituais, foi solicitado aos anotadores que, sem ter conhecimento adicional sobre quaisquer informações relacionadas ao contexto em que as fotografias foram produzidas, descrevessem as pessoas, animais, objetos e quaisquer atividades que fossem efetivamente mostradas nas imagens. Para garantir que houvesse um número mínimo de variação na forma como cada uma das fotos foi descrita, várias legendas foram coletadas para cada imagem. Como consequência, é pouco provável que as descrições de uma mesma imagem sejam paráfrases diretas uma da outra, ou seja, a mesma entidade, evento ou situação foi descrita de várias maneiras – idoso/adulto/homem/pessoa; criança/menino/garoto; praia/areia/mar –, e mesmo que todos os anotadores mencionem o homem, nem todos mencionam o chapéu ou a areia (Figura 41).

Antes de iniciar a tarefa, cada anotador encarregado de criar legendas originais em português do Brasil recebeu um documento contendo uma lista de instruções, uma imagem de referência acompanhada dos critérios mínimos exigidos para cada descrição (Figura 42) e duas imagens de exemplo acompanhadas de cinco descrições em português do Brasil semelhantes às que existem no Flickr30K em outras línguas. As instruções, organizadas na forma de tópicos, orientavam os anotadores a:

Figura 41 - Imagem do Flickr30K com legendas em PT-BR.

Fonte: Flickr 30K

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;

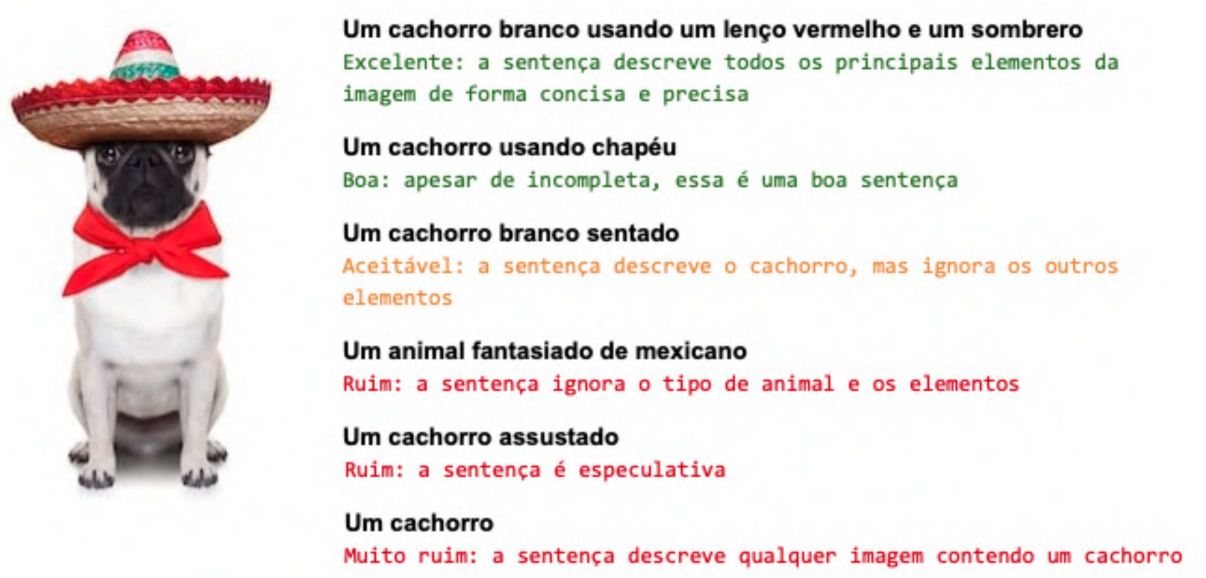
2. Fornecer uma descrição precisa das atividades sendo executadas na imagem, das pessoas ou animais que executam a atividade e de quaisquer objetos envolvidos nela;

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

Após as instruções, os anotadores foram orientados a acessar uma interface de anotação simples, desenvolvida especialmente para essa tarefa, composta pela imagem a ser descrita, uma caixa de texto onde a descrição deve ser feita e um botão 'enviar' que, quando clicado, salva a descrição criada e apresenta ao anotador uma nova imagem extraída do subconjunto de 1500 designadas para cada anotador – escolhidas aleatoriamente entre as 31.014 que compõem o *corpus* Flickr30K.

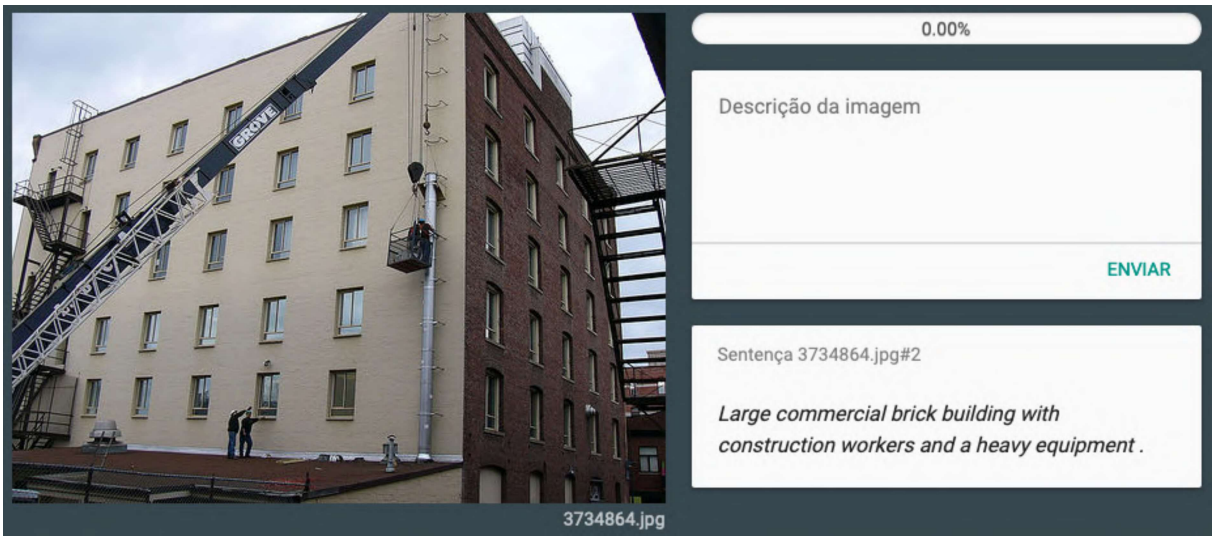
Para os anotadores encarregados da tarefa de tradução, descrita na subseção anterior, as instruções pediam que se traduzisse para o português do Brasil a legenda originalmente escrita em inglês para a imagem apresentada. Na interface dessa tarefa, uma caixa de texto adicional – que é removida na versão destinada a criação de descrições – apresenta ao anotador uma das cinco descrições em inglês que compõem o Flcikr30K (Figura 43).

Figura 42 - Imagem do Flickr30K com legendas em PT-BR.



Fonte: O autor

Figura 43 - Interface de criação de descrições e tradução.



Fonte: O autor

4.2.3 Rotulagem Automática de Papéis Semânticos de Frames em Descrições de Imagens

Para enriquecer as descrições das imagens no FM30K com informações semânticas extraídas da FrameNet, optamos por usar um instância pré-treinada do LOME (XIA et al., 2021). O LOME – sigla em inglês para *Large ontology multilingual extraction* – é um sistema desenvolvido para extrair informações multilíngues que utiliza um *parser* treinado em dados semânticos para identificar entidades e eventos em modalidades textuais. Além

disso, o LOME também realiza a resolução de co-referência – ou seja, identifica em uma sentença itens lexicais que fazem referência a uma mesma entidade, – tipagem de entidades – que é a atribuição de um rótulo para cada entidade, como, por exemplo, o rótulo 'pessoa' – e previsão de relações temporais de eventos – em uma sentença que descreve múltiplos eventos, em que ordem eles ocorrem. Nesta implementação, foram incluídas também informações sobre frames e Elementos de Frame extraídas da base de dados da FrameNet Brasil.

Uma das principais vantagens do LOME é o fato de que ele localiza e rotula os itens lexicais que evocam frames ou que são Elementos de Frame, ao passo que sistemas similares dependem de que informações de localização dos itens lexicais sejam fornecidas como parte do *input* para rotulação semântica dos dados. O *parser*, treinado sobre vetores do modelo de língua XLM-R (CONNEAU et al., 2020), também permite que o LOME aprenda representações em uma língua e as extrapole para outras. Para este trabalho, usamos uma versão do LOME que foi treinada não apenas em anotações *fulltext* da Berkeley FrameNet 1.7, mas também em 8.558 anotações *fulltext* da FrameNet Brasil, o que nos permitiu incluir informações sobre frame e Elementos de Frame provenientes tanto da Berkeley FrameNet 1.7 quanto da FrameNet Brasil.

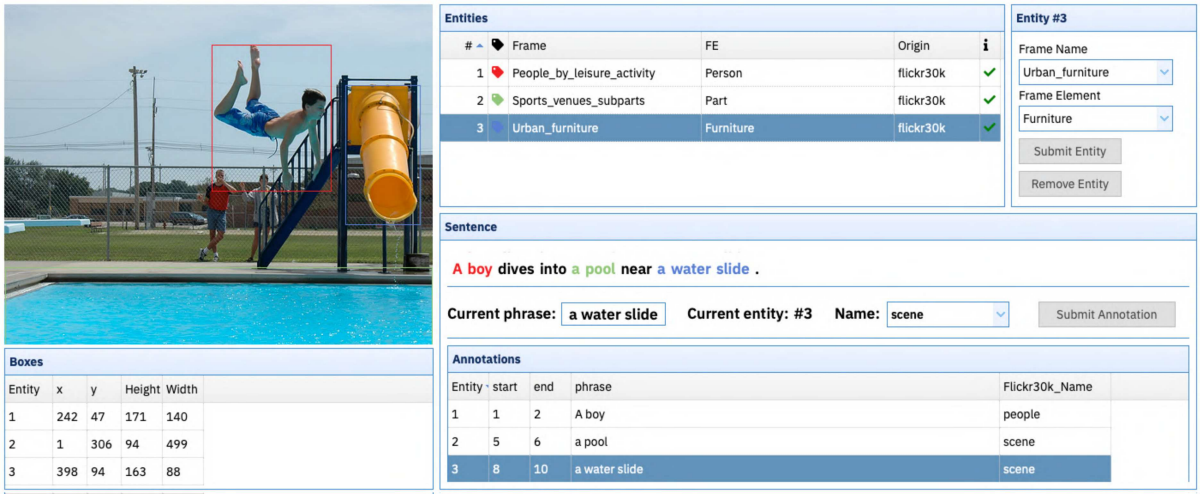
O *parser* para frames semânticos da FrameNet no LOME possui dois componentes principais: um que identifica sequências de interesse e outro que classifica essas sequências como evocando um frame ou instanciando um elemento de frame. O modelo primeiro transforma as sentenças de entrada em uma lista de vetores, onde cada posição armazena o vetor de um *token* da sentença. Essa lista de vetores é usada como *input* a um marcador BIO – sigla em inglês para *Beggining-Inside-Outside* – que identifica sequências de interesse. Essas sequências são rotuladas por um módulo de classificação. Esse processo é executado duas vezes para cada sentença: na primeira iteração, o marcador identifica sequências evocadoras e rotula seus frames; a segunda iteração faz o mesmo, porém condicionada aos rótulos de frame já previstos para identificar e rotular Elementos de Frame. Como os Elementos de Frame pertencem a um único frame, durante o treinamento, a perda é considerada máxima quando o modelo relaciona um elemento ao frame incorreto. Dessa forma, o LOME também é capaz de aprender quais elementos pertencem a cada frame.

4.2.4 **Anotação Humana de Entidades para Frames e Elementos de Frame**

O *dataset* Flickr30K Entities foi manualmente anotado para associar *bounding boxes* referentes às entidades presentes nas 31.783 imagens no Flickr30K com seus respectivos descritores – sintagmas nominais – nas sentenças ENO. Para a tarefa de atribuir frames e EFs às correlações imagens-sintagmas nominais, selecionamos 29.920 conjuntos de *bounding boxes* presentes em 31.104 descrições ENO, tendo o cuidado de selecionar dentre as cinco descrições disponíveis aquela que foi utilizada na criação das traduções.

A tarefa de anotação de imagens foi realizada utilizando a Charon (BELCAVELLO et al., 2022), uma ferramenta criada na FrameNet Brasil para anotação de *corpora* multimodais para dados provenientes da FrameNet. A Figura 44 mostra a interface para o modo de anotação de imagens estáticas, usada para anotar pares imagem-texto.

Figura 44 - Interface da ferramenta de anotação usada para atribuir frames e Elementos de Frame ao Flickr30K Entities.



Fonte: O autor

O canto superior esquerdo da interface de anotação exibe a imagem de referência do Flickr30K. O painel 'Boxes' mostra as coordenadas das caixas delimitadoras anotadas manualmente – obtidas do conjunto de dados Flickr30K Entities – e o número associado a cada entidade sendo anotada para aquela imagem. O painel 'Entities' exibe as correlações entre cada entidade na imagem, sua frase nominal correspondente – codificada por cores com a descrição vista no painel 'Sentence' – e o frame e elemento de frame atribuídos àquele par imagem-texto pelo anotador, usando o painel 'Entity' no canto superior direito da interface.

A Figura 44 também mostra a anotação resultante para a sentença *"A boy dives into a pool near a water slide"*. Nessa sentença, os sintagmas nominais *"A boy"*, *"a pool"* e *"a water slide"* são correlacionadas com três entidades distintas na imagem. Ao sintagma nominal *"A boy"*, correspondente à Entidade 1, o anotador atribuiu o frame `People_by_leisure_activity` e o EF PERSON. Para o sintagma *"a pool"*, correspondente à Entidade 2, o anotador selecionou o frame `Sports_venues_subparts` e o EF PART. Finalmente, para o sintagma *"a water slide"*, correspondente à Entidade 3, o anotador escolheu o frame `Urban_furniture` e o EF FURNITURE.

Uma versão ligeiramente modificada da interface de anotação na Figura 44 foi usada para a tarefa de atribuição de frames e EFs para imagens sem a presença de descrições.

Nessa versão, o painel ‘Sentence’ não era exibido para os anotadores. As duas versões da interface de anotação permitiram a produção de anotações de frame e EF para *bounding boxes* nas imagens em duas condições: em presença da descrição associada à imagem e em ausência de descrição. A motivação por trás desse design de anotação foi analisar a influência da descrição na representação semântica das imagens resultantes da anotação. Resultados experimentais preliminares relatados por VIRIDIANO et al. (2022) indicam que tal influência é estatisticamente relevante e serão descritos no capítulo 5.

Para ambas as condições de anotação, os anotadores foram instruídos a associar um frame e um EF a cada uma das *bounding boxes* listadas no painel ‘Entities’, contanto que tais entidades fossem visíveis na imagem. Cada entidade é associada a um conjunto de coordenadas – mostradas no painel ‘Boxes’ – correspondentes a cada uma das *bounding boxes*.

Os anotadores foram explicitamente instruídos a não anotar entidades que não estivessem visíveis na imagem. Por exemplo, no caso da Figura 45, que tem como descrição a sentença *“A man standing on a stage playing a guitar and harmonica waving to the crowd.”*, o conjunto de dados Flickr30K Entities atribui uma correlação imagem-texto entre *“the crowd”* e uma entidade que não é visível na imagem – a audiência do show. Em casos como este, os anotadores foram explicitamente instruídos a não atribuir frames e FEs a essas correlações. Para aquelas correlações onde a entidade é mostrada na imagem – *“A man”*, *“a stage”*, *“a guitar”* e *“harmonica”* – frames e elementos de frame foram atribuídos.

Figura 45 - Imagem 4827958485.jpg pareada com a descrição “Um homem em pé no palco tocando guitarra e gaita acenando para a multidão.” no *dataset* Flickr30K Entities, onde o sintagma “a multidão” está correlacionada a uma entidade que não é mostrada na imagem.

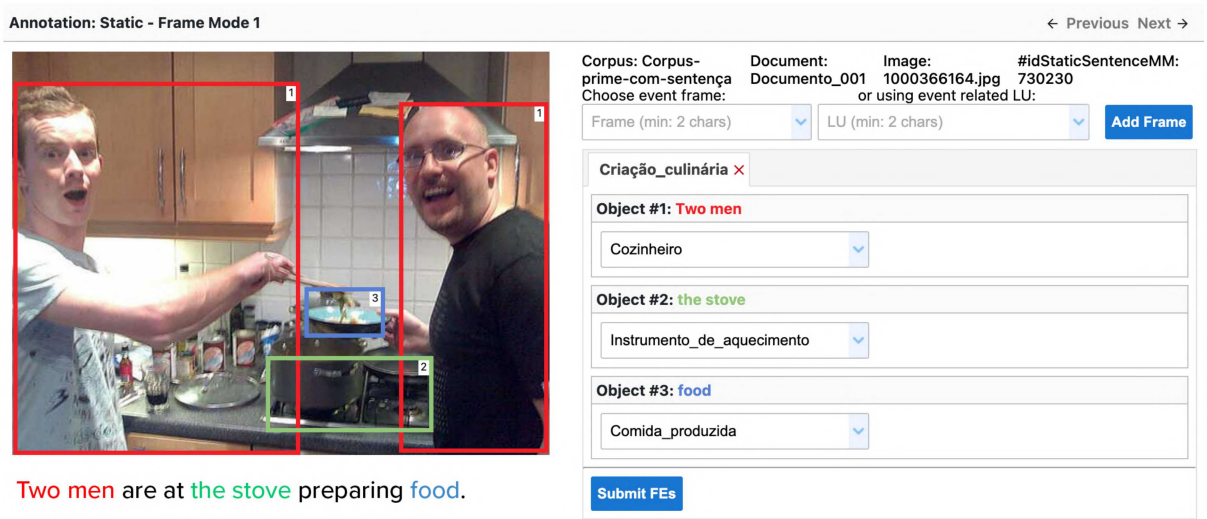
Fonte: Flickr30K Entites

Ambas as condições da tarefa de anotação de imagem – Imagem com Descrição (IcD) e Imagem sem Descrição (IsD) – foram realizadas pela mesma configuração das equipes de anotadores descritas nas seções 4.2.4, de modo que um mesmo anotador não anotou uma mesma imagem em duas condições distintas. A equipe permanente para esta tarefa contou com 16 estudantes remunerados da maneira já descrita. O grupo de estudantes inscritos nos *workshops* de anotação era composto por 32 estudantes.

4.2.5 **Anotação Humana de Eventos para Frames e Elementos de Frame**

Além das tarefas da atribuição de frames e Elementos de Frame para as entidades provenientes do *dataset* FLickr30K Entites, uma terceira tarefa de anotação, focada nos eventos presentes nos pareamentos imagem-texto, também foi realizada. Para essa tarefa de atribuição de frames de evento para imagens na presença das descrições (IcD-Ev) foram realizadas pela mesma equipe permanente de anotadores, composta por 16 estudantes, remunerados da maneira já descrita. A Figura 46 mostra a interface da ferramenta de anotação de eventos para imagens em presença da descrição.

Figura 46 - Interface da ferramenta de anotação eventos usada para atribuir frames e Elementos de Frame ao Flickr30K Entities.



Fonte: o autor

No lado esquerdo da interface, temos a imagem a ser anotada acompanhada da sentença que a descreve, com *bounding boxes* numeradas e nas mesmas cores dos sintagmas nominais a elas relacionados na descrição. A parte superior do lado direito da interface traz as informações sobre o *corpus*, documento, imagem e sentença que está sendo anotada. Logo abaixo, duas caixas de busca – uma para busca textual pelo nome dos frames e outra para a busca por Unidades Lexicais – que permitem ao anotador acessar a base de dados da FrameNet e buscar pelos frames de evento que descrevem o conjunto texto-imagem alvo

da anotação. Após a busca e escolha do frame, o anotador passa a ter acesso, no canto inferior direito da interface, à lista de sintagmas nominais presentes na descrição, e aos objetos – entidades – a que eles se relacionam na imagem. Com isso, é possível selecionar, com base no frame escolhido e utilizando a caixa de busca relacionada a cada objeto, o Elemento de Frame correspondente a cada uma dessas entidades.

Nessa figura, é possível ver a anotação resultante para a sentença *"Two men are at the stove preparing food"*. Para essa descrição, os sintagmas nominais *"Two men"*, *"the stove"* e *"food"* são correlacionadas, através de suas cores, com as três entidades correspondentes na imagem. Para o frame de evento selecionado pelo anotador para essa imagem – o frame `Criação_culinária`, evocado pelo verbo *"preparing"* na descrição – o sintagma nominal *"Two men"*, corresponde ao Elemento de frame Cozinheiro. Já o objeto descrito pelo sintagma *"the stove"* corresponde ao EF Instrumento_de_aquecimento. Finalmente, para o sintagma *"food"*, correspondente à entidade 3, o anotador selecionou o EF Comida_produzida.

Nos casos como o apresentado na Figura 47, em que uma descrição contém dois ou mais verbos evocadores de eventos – respectivamente, ‘*'dress.v"*, *"stand.v"* e ‘*'study.v"* – os anotadores foram orientados a anotar os frames e Elementos de Frame evocados por cada um dos eventos. Para as entidades que representam Elementos de Frame em apenas um dos frames evocados, mas que não estão necessariamente relacionadas aos outros frames evocados, foi atribuído o rótulo *NULL*. Estes são os casos, por exemplo, da entidade *"a blue coat"*, que representa o EF Vestuário no frame de evento `Trajar`, evocado por *"dress.v"*, mas que não tem nenhuma função no frame `Postura`; ou da entidade *"a busy sidewalk"*, que representa o EF Localização no frame `Postura`, evocado por *"stand.v"*, que não corresponde a nenhum EF no frame `Escrutínio`.

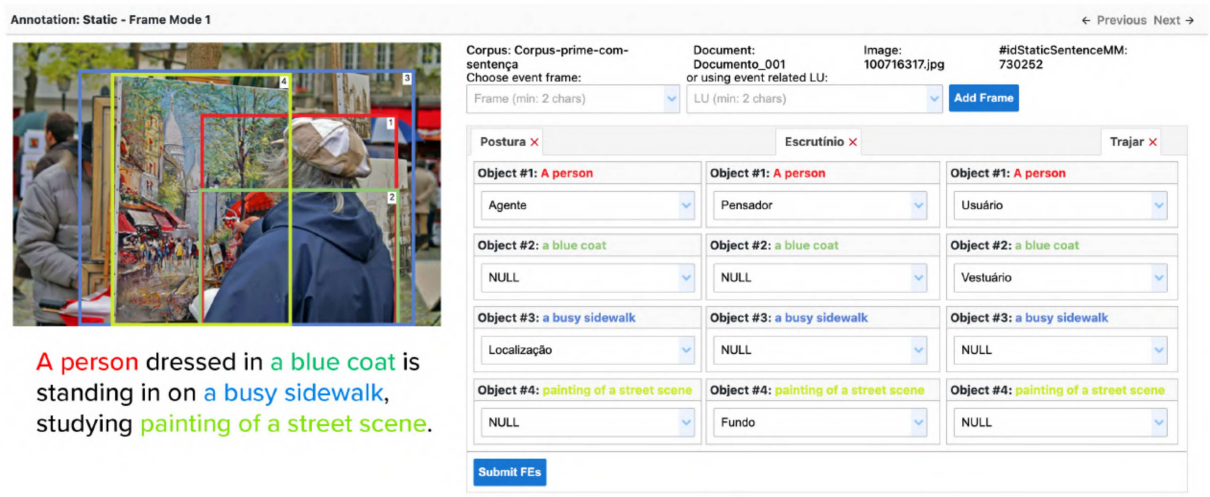
4.3 MÉTRICAS UTILIZADAS NA ANÁLISE DESCRITIVA DO DATASET

O FM30K foi analisado para dois conjuntos de métricas. De um lado, aplicaram-se à descritiva do novo *dataset* as mesmas métricas usadas por ELLIOTT et al. (2016) para a expansão do Multi30k em alemão. De outro, com vias a comparar as representações semânticas geradas via anotação para as descrições e imagens, foi utilizada a similaridade de cosseno. Ambas são apresentadas a seguir.

4.3.1 **Métricas propostas em ELLIOTT et al. (2016)**

ELLIOTT et al. (2016), ao apresentarem o Multi30k, fornecem estatísticas do então novo *dataset*. Tais estatísticas têm por objetivo permitir uma comparação entre os *dataset* base – o Flickr30k – e sua expansão para o alemão. Foram medidos, além do número de caracteres e extensão média de cada descrição, o número de *tokens* – ocorrências individuais de palavras no *corpus*, – o numero de *types* – ou seja, as formas únicas de

Figura 47 - Anotação de eventos para descrições que evocam dois ou mais eventos.



Fonte: o autor

palavras, – e o número de *singletons* – *types* que aparecem exatamente uma vez em um *corpus*.

4.3.2 Similaridade de Cosseno

Para comparar representações semânticas em diferentes idiomas e condições de anotação, seguimos a mesma metodologia de VIRIDIANO et al. (2022), que consiste em calcular similaridades de cosseno entre pares de dados usando vetores derivados dos frames anotados para os textos ou imagens. Esses vetores não são *embeddings* como os obtidos de modelos de língua ou imagem, mas sim computados usando um algoritmo de Ativação Propagada – *Spread Activation* – sobre a rede de frames da FrameNet.

Primeiro, o grafo direcionado da FrameNet é construído usando os frames e suas relações. Então, para cada sentença ou imagem, os frames anotados servem como nós de ativação no grafo. Esses nós iniciais recebem um valor máximo de energia que atua como um peso para aquele frame. O algoritmo opera através de iterações, ativando nós relacionados à medida que propaga a energia pelos nós da rede. A propagação e consequente perda de energia a cada iteração fazem com que o algoritmo – quando não há mais energia – encerre o processo, atribuindo a cada frame que foi ativado um determinado peso. São esses os pesos usados para construir o vetor de frames que será utilizado para calcular similaridades cosseno.

Ao comparar sentenças em diferentes línguas, uma etapa importante é a seleção de pares de comparação. Por serem traduções, as PTT foram comparadas com seus originais em inglês, resultando em 31.014 pares e respectivos escores de similaridade de cosseno. Para as PTO, os conjuntos de cinco descrições em cada língua foram alinhados

considerando as combinações que minimizariam a diferença média no número de palavras em cada par. Isso foi feito para evitar a comparação entre descrições curtas e descrições mais longas – que, em virtude do maior número de palavras, tenderiam a evocar mais frames. A diferença média no número de palavras para os pares alinhados usando este método foi de 2,3, com 47% dos pares tendo o mesmo comprimento ou apenas uma palavra a mais. Com essas configurações no algoritmo de Ativação Propagada para cada par, foram obtidos 158.915 escores de similaridade cosseno.

Agora passamos à discussão do FM30K em relação tanto aos conjuntos de dados originais que ele expande quanto às representações semânticas construídas a partir das atividades desenvolvidas nessa tese.

5 ANÁLISE DESCRITIVA DO DATASET RESULTANTE

Nesse capítulo, destacaremos como o acréscimo de descrições originais e traduções para o português, a adição de informações semânticas mais granulares aos pareamentos imagem-texto e as diferentes desenhos experimentais para a anotação de frames e EFs para imagens impactaram as informações existentes no *dataset* produzido neste trabalho.

5.1 ANÁLISE QUANTITATIVA

Nesta seção, apresentamos as análises quantitativas realizadas para o FM30k, divididas em dois conjuntos. Primeiramente, apresentamos as estatísticas da expansão do Multi30k para o português e da anotação automática das descrições ENO, PTO e PTT pelo Lome. Na sequência, apresentamos a análise contrastiva, via similaridade de cosseno, das representações semânticas geradas para línguas e semioses distintas.

5.1.1 **Estatísticas das Descrições Originais e Traduzidas**

A Tabela 1 apresenta contagens de *tokens*, *types*, número de caracteres, comprimento médio das sentenças e *singletons*, para as sentenças PTO e PTT do FM30K, em comparação com as de outros idiomas incluídos no *dataset* Multi30k (ELLIOTT et al., 2016).

As sentenças traduzidas podem ser organizadas em dois grupos diferentes: (i) Português Brasileiro, Inglês e Francês, e (ii) Alemão e Tcheco, ambos compartilhando entre si contagens semelhantes de *tokens, types* e *singletons*. O primeiro grupo contém *corpora* com mais *tokens*, sentenças mais longas, mas *types* menos diversificados. O segundo grupo tem os atributos opostos. As sentenças PTT do FM30K são, em média, mais longas do que todas as outras traduções em termos de palavras, com exceção do Francês. As traduções em Português Brasileiro também têm 23% mais *singletons* do que o Inglês, mas menos da metade do Tcheco. Em comparação com o Inglês e o Francês, as sentenças PTT têm uma boa variedade de *tokens* e comprimentos compatíveis.

Em relação às descrições originais, as em Português são mais longas do que as em Inglês e Alemão, tanto em termos de caracteres quanto de palavras. Em média, há um aumento de 9% no comprimento das sentenças em número de palavras, e de 21,6% em termos de caracteres, quando comparado aos dados ENO. Esses números são de 39,5% e 27,7% em comparação ao Alemão. Ao observar as contagens de *types* e *singletons*, PTO e ENO são muito próximos um do outro, enquanto o Alemão tem mais que o dobro do número de *types* e mais de três vezes o número de *singletons.*

Esses números não apenas mostram diferenças linguísticas, mas também como descrições originais podem ser bastante diferentes das traduções. Por exemplo, a diferença no comprimento médio das sentenças entre o par Português e Alemão é de mais de uma palavra.

	Tokens	Types	Caracteres	Comp. Médio	Singletons
Traduções (31.014)					
Português	374.097	12.212	1.749.365	12,0	5.426
Inglês	369.848	10.500	1.523.855	11,9	4.406
Alemão	345.326	19.363	1.834.937	11,1	11.226
Francês	387.536	12.129	1.796.038	12,4	5.232
Techco	281.138	23.166	1.346.020	9,0	12.369
Descrições (158.915)					
Português	2.127.452	19.135	9.798.114	13,4	7.660
Inglês	1.950.410	20.278	8.057.457	12,3	7.788
Alemão*	1.482.389	44.033	7.669.557	9,6	25.229

Tabela 1 – Estatísticas do corpus para Traduções e Descrições em Português comparadas a outros idiomas no Multi30k. * O corpus de descrições originais em Alemão possui 155.070 sentenças.

Em linhas gerais, a expansão do Multi30k para o português, seguindo a metodologia proposta por ELLIOTT et al. (2016), resultou na adição de uma nova língua para o *dataset*. Tal adição não destoa das demais, na medida em que tais variações eram esperadas e não comprometem a inserção do Português do Brasil à família Multi30K.

Para além da expansão para o português, como pontuado na seção 3.2.3, o FM30K inclui ainda análises de frames feitas com o LOME para todas as 158.915 descrições originais em inglês – ENO – no Flickr30K, assim como para as 31.104 descrições PTT e as 158.915 descrições PTO. O número de rótulos de frame e Elementos de Frame associados a cada grupo de descrições é mostrado na Tabela 2.

	frames/FEs	**média p/ sent**
ENO	2.073.114	13,0
PTO	2.131.036	13,4
PTT	372.972	12,0
total	4.577.122	-

Tabela 2 – Número de frames e EFs incluídos para cada idioma do conjunto de dados FM30K e as médias por sentença.

Na sequência, debruçamo-nos sobre as análises contrastivas realizadas para as representações semânticas geradas para as descrições a partir dos frames a elas associados.

5.1.2 Similaridades de Cosseno entre Descrições

A Tabela 3 mostra as similaridades de cosseno médias normalizadas para PTT e PTO quando comparadas a ENO. Como esperado, sentenças traduzidas estão considera-

velmente mais próximas de suas sentenças fontes em termos das representações semânticas a elas associadas, as quais foram geradas a partir das análises automáticas do LOME. Esse efeito também pode ser observado nas distribuições na Figura 48, onde a distribuição PTT × ENO é deslocada, quando comparada ao gráfico mais centralizado ENO × PTO.

Para avaliar a significância das diferenças entre as distribuições o Teste-t de Student foi utilizado (STUDENT, 1908). O teste é utilizado para aferir se a diferença entre dois conjuntos de observações é estatisticamente significativa ou não. A hipótese nula é a de que a média dos dois conjuntos de observações é a mesma. Quando $p < 0.01$, a hipótese nula deve ser rejeitada, isto é, os conjuntos não possuem a mesma média. O valor p é encontrado utilizando-se a tabela de valores selecionados, a estatística de teste t e o número de graus de liberdade. Para cada teste realizado, a forma $t(número de graus de liberdade) = valor$, $p < 0.01$ ou $p > 0.01$ será utilizada. Nos casos em que os conjuntos de observação possuem um número de amostras diferentes, foi utilizada a versão independente do teste. Em todas instâncias, assumiu-se que a variância dos conjuntos era a mesma.

A primeira aplicação dos teste foi feita para os conjuntos de similaridades de cosseno entre traduções em português (PTT) e ENO e descrições originais criadas apenas com a referência da imagem (PTO) e ENO. As sentenças PTT têm similaridades significativamente mais altas (M = 0,77, DP = 0,14) do que as descrições PTO (M = 0,53, DP = 0,16), com estatística de teste t(189929) = -216,41, $p <$ 0,001.

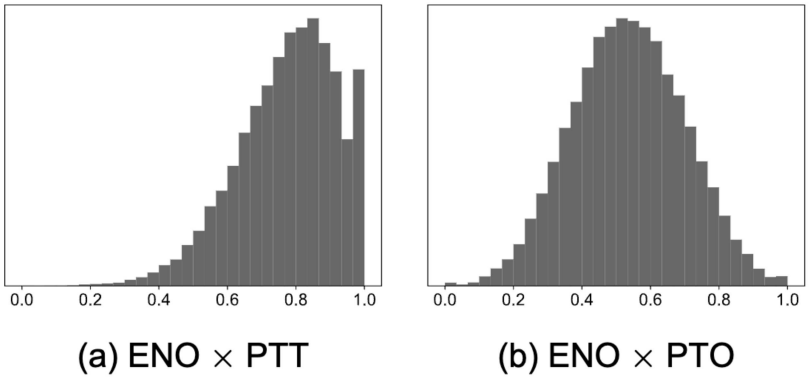
	ENO	
	Sim. Média	DP
PTT	0.77	0.14
PTO	0.53	0.16

Tabela 3 – Similaridade para configurações de anotação de frames de imagem com e sem a presença de legendas.

Combinadas, essas informações mostram que representações semânticas de descrições de imagens podem mudar consideravelmente em função do fato de a descrição ser ou não uma tradução. Isso destaca a importância de o *dataset* conter ambos os tipos de sentenças em cenários multilíngues: as traduções são mais facilmente relacionadas às suas fontes, mas as sentenças originais aumentam a variedade de maneiras pelas quais a mesma cena pode ser representada, incorporando novas perspectivas sobre a imagem, para além daquela inicialmente dada pelo anotador que criou a descrição ENO que foi traduzida.

Considerando-se o fato de que o FM30K tem o potencial de ser utilizado para tarefas diversas de PLN, as descrições PTO podem ser empregadas em diversas tarefas para além da tradução automática, tais como *Visual Q&A*, por exemplo.

Figura 48 - Distribuição dos valores de similaridade entre ENO e PTT e PTO.



Fonte: O autor

5.1.3 Similaridades de Cosseno entre Modos Semióticos

Outra análise importante para o *dataset* FM30K é a comparação entre as representações semânticas geradas para os diferentes modalidades – imagem e língua escrita – a partir de diferentes condições de anotação de imagens. Conforme discutido na seção 4.2.4, frames e EFs foram atribuídos a entidades sob duas condições de anotação distintas: uma anotação de imagens na presença da descrição (IcD) e outra sem a presença da descrição (IsD). Ademais, uma terceira rodada de anotações foi realizada, na qual as imagens foram anotadas para frames de evento, na presença das descrições ENO (IcD-Ev).

O número total de anotações geradas a partir de cada condição é dado nas Tabelas 4, 5 e 6.

	frames/EFs	**média p/ anotador**
Permanentes	49.475	2.248,9
Workshop	38.063	302,0
Total	87.538	-

Tabela 4 – Contagem de frames e EFs anotados no *dataset* FM30K por cada equipe de anotação, e médias por anotador sob a condição 'Imagem na presença de Descrição' (IcD).

	frames/EFs	**média p/ anotador**
Permanentes	65.538	2.979,0
Workshop	16.484	130,8
Total	82.022	-

Tabela 5 – Contagem de frames e EFs anotados no *dataset* FM30K por cada equipe de anotação, e médias por anotador sob a condição 'Imagem sem Descrição' (IsD).

A pequena variação – inferior a 6,3% – no número total de EFs atribuídos em cada condição de anotação se deve ao fato de que, na condição IsD, o anotador poderia não

	frames/EFs	**média p/ anotador**
Permanentes	83.077	6.923,08

Tabela 6 – Contagem de frames e EFs anotados no conjunto de dados FM30K e médias por anotador para frames de eventos evocados na condição 'Imagens acompanhadas de descrição' (IcD-Ev).

anotar alguma *bounding box* por não conseguir atribuir a ela um frame de entidade, o que poderia ter sido guiado pela descrição. Já para a condição IcD-Ev, a não anotação de alguma *bounding box* pode ter se devido ao fato de um determinado elemento mostrado na imagem não poder ser conectado a um EF do frame de evento escolhido para a anotação.

Para cada um dos cenários de anotação, os vetores semânticos de cada conjunto de entidades ou eventos de uma imagem foram calculados usando o mesmo método de *Spread Activation* descrito na 4.3.2. Como os conjuntos de entidades ou eventos vinculados a uma imagem também estão ligados a uma sentença ENO, essa sentença foi usada para calcular 4 conjuntos de 29.831 escores de similaridade. Suas médias são mostradas na Tabela 7 e as distribuições na Figura 49.

	ENO	
	Sim. Média	DP
IcD	0,49	0,17
IsD	0,42	0,18
IcD-Ev	0,36	0,18
IcD+IcD-Ev	0,55	0,17

Tabela 7 – Similaridade para configurações de anotação de frames em imagem com e sem a presença de descrições.

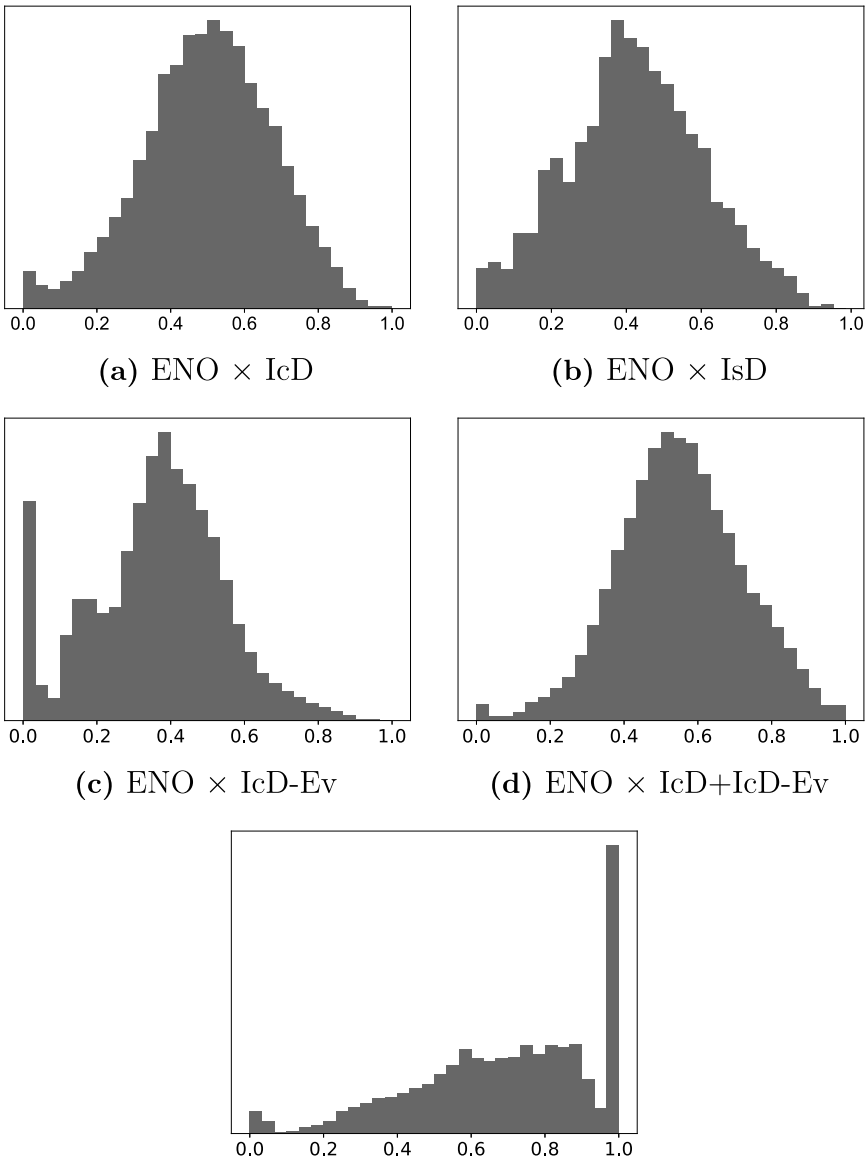
Adicionalmente, calculamos a similaridade de cossenos entre as representações semânticas geradas para as imagens nas condições de anotação IcD e IsD, de modo a prover a análise com evidências adicionais sobre os impactos das distintas condições de anotação sobre as representações semânticas produzidas para uma mesma imagem. A média normalizada é mostrada na Tabela 8 e a distribuição na Figura 49e.

	IcD	
	Sim. Média	DP
IsD	0,69	0,24

Tabela 8 – Similaridade entre configurações de anotação de frames de entidade em imagem com e sem a presença de descrições.

A maior similaridade de cossenos se verifica entre as representações semânticas

Figura 49 - Distribuição dos valores de similaridade entre ENO e IcD, IsD, IcD-Ev e IcD+IcD-Ev.



(a) ENO × IcD

(b) ENO × IsD

(c) ENO × IcD-Ev

(d) ENO × IcD+IcD-Ev

(e) IsD × IcD

Fonte: O autor

geradas para as sentenças ENO e o conjunto de anotações de imagem para frames de entidade e evento produzidas em presença da descrição ENO com a qual são comparadas. Tal resultado era esperado, uma vez que nossa hipótese era a de que a presença da modalidade verbal de alguma forma influenciaria a anotação de imagens, de modo a aproximar as representações semânticas. Como é esperado que cada descrição conceitual, conforme a metodologia proposta por HODOSH et al. (2013), inclua os participantes mostrados na foto e atividade em que estão engajados, é natural que as anotações combinadas de entidade e evento para a imagem tenham maior chance de se sobrepor àquela realizada pelo LOME para a descrição ENO.

As similaridades de cosseno entre as representações semânticas geradas a partir da combinação das condições de anotação IcD e IcD-Ev (M = 0,55, DP = 0,17) e ENO são, em média, bastante próximas àquelas entre ENO e PTO (M = 0,53, DP = 0,16), porém menores do que aquelas entre ENO e PTT (M = 0,77, DP = 0,14). Apesar de próximas, ainda há uma diferença estatística significativa entre as comparações de ENO com IcD+IcD-Ev e PTO, com estatística de Teste-t de Student $t(187972) = 14{,}95$ e $p < 0{,}001$. Os *scores* indicam que, para o FM30k, a manutenção de uma mesma modalidade comunicativa – língua verbal – tem um efeito positivo na similaridade das representações semânticas, desde que a sentença original ENO seja apresentada como referência, seja para a criação da tradução da sentença – PTT –, seja para a criação das representações semânticas das imagens que descreve – IcD e IcD-Ev.

Na sequência, as maiores similaridades semânticas mostradas na Tabela 7 são aquelas encontradas entre ENO e IcD (M = 0,49, DP = 0,17) e IsD (M = 0,42, DP = 0,18), respectivamente. Quando comparadas com PTT e PTO, tais similaridades em relação a ENO são consideravelmente menores e estatisticamente relevantes, com estatística de Teste-t de Student $t(59660) = 25{,}24$ e $p < 0{,}001$. Esses resultados também são aparentes na Figura 49, onde as distribuições são mais semelhantes (em contraste com a Figura 48).

Também se observa na Figura 49 que as anotações IcD+IcD-Ev (M = 0,55, DP = 0,17) possuem similaridades semânticas maiores que as de IcD-Ev (M = 0,36, DP = 0,18) em relação às sentenças ENO. Essa diferença é significante estatisticamente, com com estatística de teste $t(57253) = 127.86$, $p < 0{,}001$ para o Teste-t de Student. No caso da distribuição de ENO × IcD-Ev, a média é afetada pelo número alto de comparações com escore de similaridade igual a 0. Tal resultado indica que em um número relativamente maior de instâncias, as anotações IcD-Ev evocam um conjunto de frames completamente distinto pelo evocado pela sentença em inglês.

A diferença entre as similaridades de IcD e IsD mostra que há mudanças na representação semântica quando os anotadores anotam imagens na presença de sentenças. Similarmente ao PTT, a inclusão das duas modalidades restringe o número de possíveis interpretações e aproxima as representações semânticas às das descrições originais. Tais dados corroboram a hipótese de que os distintos desenhos experimentais geram representações semânticas distintas para as imagens, o que reforça a importância da abordagem perspectivista na constituição do FM30k.

Ainda que as anotações na condição IcD-Ev também tenham sido produzidas na presença das descrições ENO, sua similaridade de cosseno é consideravelmente menor em relação àquela de IcD. Isso indica que as escolhas do anotador quanto aos frames de entidade presentes nas imagens eram menos variáveis em relação aos frames presentes em ENO do que aquelas relativas aos frames de evento. Tal distinção também era esperada,

dada a natureza do modelo da FrameNet, que é muito granular na descrição de eventos. Mais uma vez, isso demonstra a importância de saber como uma anotação foi criada e de usar dados produzidos em diferentes contextos para acomodar múltiplas perspectivas.

5.2 ANÁLISE QUALITATIVA

Passamos agora às análises qualitativas aplicadas ao Multi30k. Em específico, debruçamo-nos sobre os possíveis enviesamentos negativos de etnia e de gênero do *dataset* resultante, correlacionando-os às condições de anotação.

5.2.1 **Perfil dos Anotadores**

Sobre a variante linguística utilizada na criação das descrições originais e legendas, todas foram produzidas por falantes nativos do português brasileiro, inseridos no ambiente acadêmico do Ensino Superior, fazendo com que a variedade de português brasileiro prevalente no *dataset* seja aquela equivalente a de falantes urbanos altamente escolarizados em um ambiente comunicativo monitorado.

Em relação aos aspectos sociodemográficos dos anotadores recrutados, o *dataset* contém dados produzidos por trinta e nove estudantes de graduação do curso de Licenciatura em Letras e Linguística (54,9%), dezesseis estudantes com Licenciatura em Letras e Linguística concluída (22,5%), nove estudantes cursando Mestrado em Linguística (12,7%) e sete estudantes cursando Doutorado em Linguística (9,9%). Destes, quatro estavam na faixa etária de 18 a 24 anos (5,6%), quarenta e um entre 20 e 29 anos (57,8%), quinze na faixa dos 30 e 39 anos (21,2%), seis entre 40 e 49 anos (8,4%), e cinco acima dos 50 anos de idade (7%).

Em termos de gênero, o grupo foi constituído por quarenta e cinco anotadores que se identificaram como mulheres (63,4%), vinte e quatro anotadores identificados como homens (33,8%) e dois anotadores identificados como não-binários (2,8%). Quanto à etnia, quarenta e sete anotadores se autodeclararam brancos ou caucasianos (66,2%), vinte se identificaram como negros ou pardos (28,2%) e apenas um como indígenas (1,4%). Três anotadores, representando aproximadamente 4,2% do grupo, optaram por não divulgar sua origem racial ou étnica.

Em relação à proficiência em inglês, trinta e sete anotadores se autodeclararam proficientes ou com nível avançado de proficiência (52,1%), enquanto dezoito declararam ter nível intermediário (25,4%) e onze declararam ter nível básico ou iniciante de proficiência em inglês (15,5%). Apenas cinco anotadores (7%) relataram não ter proficiência na língua. Estes percentuais podem ser vistos em forma de gráfico na Figura 50.

Como se pode notar, o perfil de anotadores que produziu o FM30k é diverso no que tange as variáveis demográficas de identidade de gênero, etnia e idade, o que pode acabar

Figura 50 - Gráficos dos dados sócio-demográficos dos anotadores que trabalharam na criação do FM30K.

Fonte: Dashboard do projeto Reinventa

por impactar os possíveis enviesamentos étnico-racias e de gênero contidos nas anotações de imagem.

5.2.2 Enviesamentos Étnico-raciais e de Gênero

Comparações entre o número de vezes em que determinados frames foram evocados por uma mesma imagem nas duas condições de anotação de entidades realizadas durante essa tese – anotação para frames em presença da descrição (IcD) e na ausência da descrição (IsD) – mostram como estereótipos e preconceitos presentes nos dados de origem podem vir a ser perpetuados durante a expansão de conjuntos de dados, corroborando os resultados que discutimos na Seção 2.2 acerca da perpetuação de enviesamentos danosos em *datasets* amplamente utilizados em tarefas computacionais, ainda que eles sejam sujeitos à curadoria humana.

Como exemplo de enviesamento étnico-racial motivado pela descrição, temos a Figura 51, que tem como uma de suas descrições a sentença "**Uma moça asiática**

vestindo um avental verde serve bebidas em uma bandeja." – no original, *"**An Asian girl** in a green hat and apron is serving drinks on a tray."*. Ao compararmos os frames atribuídos para a mesma entidade em destaque na imagem, vimos que, no primeiro caso, quando o frame foi atribuído na presença da descrição – que associa a entidade ao sintagma "Asian girl", – o anotador optou pelo frame `Pessoa_por_etnia` (Figura 52). No entanto, a anotação feita na ausência da descrição, a mesma entidade foi relacionada ao frame `Pessoa_por_vocação` (Figura 53), claramente mais relevante para o contexto apresentado na cena retratada pela imagem.

Figura 51 - Exemplo de enviesamento étnico-racial na anotação de frames na presença da descrição.

Fonte: Imagem 4756096571.jpg do *dataset* Flickr30K

Dentre as ocorrências do frame `Pessoa_por_etnia` nas anotações manuais feitas para entidades, 303 ocorrências se deram nas anotações feitas na presença da descrição (IcD), ao passo que apenas 36 ocorreram nas anotações sem a presença de descrição (IsD). No caso das atribuições feitas para as anotações IsD, cabe destacar que as pistas contextuais que motivam a atribuição do frame `Pessoa_por_etnia` são, frequentemente, itens de vestuário ou outros objetos presentes na imagem – como no caso da Figura 54 – ao invés dos traços fisionômicos das pessoas retratadas.

O mesmo se verifica para outros grupos étnicos, tais como pessoas indígenas e negras, conforme pode ser visto nas Figuras 55, em que o frame de `Pessoa_por_etnia` foi utilizado na anotação com presença de descrição, enquanto os frames de `Pessoa` e `Pessoa_por_idade` foram usados na condição sem presença de descrição, respectivamente.

Por outro lado, na condição de anotação de imagem sem a presença da descrição, a presença de indivíduos desses mesmos grupos étnicos gerou a anotação para o frame de `Pessoa_por_etnia` em contextos bastante diversos, como mostrados na Figura 56.

Por fim, cabe ressaltar o fato de que, dentre as 303 ocorrências do frame de

Figura 52 - Descrição do frame `Pessoa_por_etnia` na base de dados da FrameNet Brasil.

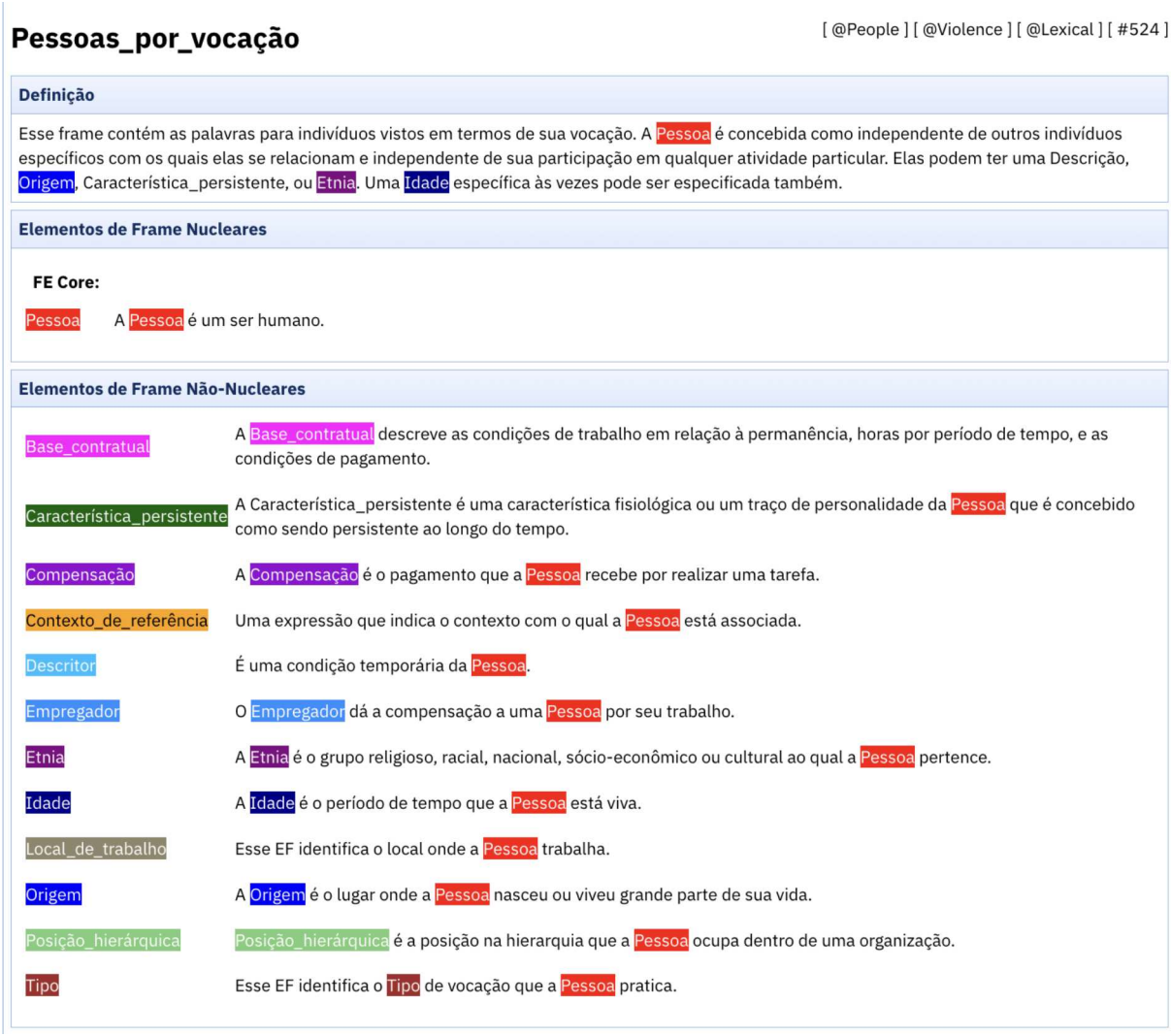
Fonte: FrameNet Webtool

`Pessoa_por_etnia` na anotação em presença de descrição, apenas 20 – ou 6,6% – se referiram a uma pessoa branca. Um exemplo desse tipo de imagem pode ser visto na Figura 57. Esse achado confirma estudo amostral de VAN MILTENBURG (2016), o qual demonstrou que as descrições constantes do Flickr30k tendem a mencionar etnias muito mais frequentemente para pessoas não brancas.

Outro exemplo do enviesamento na atribuição de frames a entidades quando estas são apresentadas na presença da descrição pode ser visto na Figura 58. Quando acompanhada da sentença *"**A female blacksmith** is shoeing a horse."*, a entidade destacada na imagem – uma "Uma mulher ferreira" – foi associada pelo anotador ao frame `Pessoa_por_gênero`. Entretanto, na anotação feita na ausência da descrição (IsD), o frame atribuído foi o de `Pessoa_por_vocação`, contextualizando a pessoa retratada em relação a atividade que ela desempenha e aos demais elementos presentes na cena.

Casos como esse chamam nossa atenção em virtude da diferença no número de ocorrências do frame `Pessoa_por_gênero` para cada uma das condições de anotação. Para as anotações de imagens na presença da descrição, foram registradas 4.738 de `Pessoa_por_gênero` (15,3%). Em contraste, para a condição de anotação de entidades sem a presença da descrição, apenas 2.577 casos foram registrados (8,3%). Acreditamos que essa diferença não seja apenas um reflexo das crenças e expectativas do criador das descrições – no exemplo da Figura 58, a crença de que ferreiros geralmente são homens -

Figura 53 - Descrição do frame `Pessoa_por_vocação` na base de dados da FrameNet Brasil.



Fonte: FrameNet Webtool

mas sim a indicação de um viés de gênero que, como vimos, não se reflete nas anotações feitas por brasileiros sem o enviesamento prévio imposto à imagem pela descrição.

As sensíveis diferenças no emprego dos frames de `Pessoa_por_etnia` e `Pessoa_por_gênero` nas duas condições de anotação – em presença ou em ausência da descrição – apontam para o fato de que, para os anotadores do FM30k, a relação entre descrição e imagem se deu, conforme a classificação proposta por MARTINEC & SALWAY (2005), de maneira desigual, em que a imgaem estava subordinada ao texto. Tal fato levanta um sinal de alerta para esforços futuros de anotação semântica de *datasets* multimodais: não fosse o fato de esta tese ter contrastado as duas condições de anotação, as representações semânticas geradas para as imagens seriam, majoritariamente, enquadradas a partir dos elementos evocadores de frames constantes das descrições mostradas juntamente às imagens.

Figura 54 - Uma mulher asiática, em traje tradicional, evocando o frame `Pessoa_por_etnia`.

Fonte: Imagem 4918525947.jpg do *dataset* Flickr30K

Figura 55 - Exemplo de imagem de indígena e pessoa negra anotada para o frame `Pessoa_por_etnia` em presença da descrição.

Fonte: Imagens 3215589470.jpg e 3131220160.jpg do *dataset* Flickr30K

Figura 56 - Exemplo de imagem de indígena e pessoa negra anotada para o frame `Pessoa_por_etnia` na ausência da descrição.

Fonte: Imagens 7355163918.jpg e 104824673.jpg do *dataset* Flickr30K

Figura 57 - Exemplo de imagem onde o frame `Pessoa_por_etnia` foi atribuído a uma pessoa de pele branca.

Fonte: Imagem 7397183064.jpg do *dataset* Flickr30K

Figura 58 - Exemplo de enviesamento de gênero na anotação de frames na presença da descrição.

Fonte: Imagem 11382381.jpg do *dataset* Flickr30K

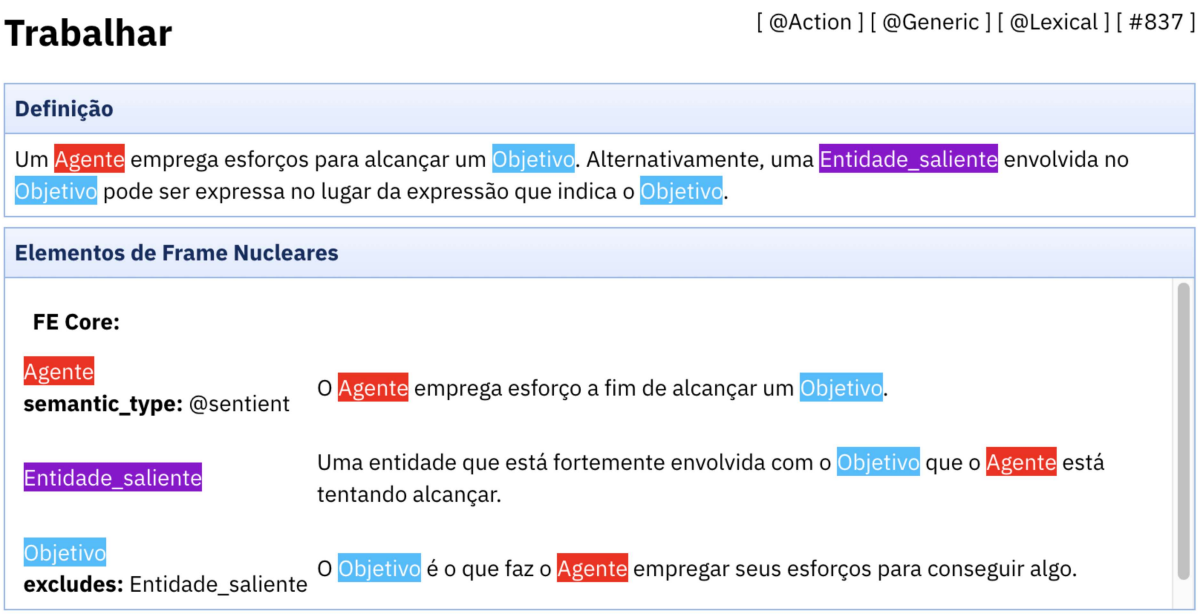
Tal fato representaria um falseamento das reais condições em que os modos comunicativos visual e verbal podem se relacionar em usos linguísticos reais.

Ademais, as análises aqui desenvolvidas – tanto a quantitativa quanto a qualitativa – reforçam a importância de se desenharem tarefas de criação de *datasets* perspectivizados, ou seja, que contenham distintos olhares possíveis sobre os dados a serem tratados (BASILE et al., 2021). As duas condições de anotação, somadas à diversidade demográfica dos anotadores e à natureza inerentemente granular e multiperspectivizada da Semântica de Frames permitiram a geração de diferentes representações semânticas para as imagens. A partir de categorias de uma FrameNet é possível representar um indivíduo semanticamente a partir de sua etina, ou de seu gênero, ou de sua profissão. Cada uma dessas representações se ancora em um frame herdeiro do frame de `Pessoa`. O mesmo não seria possível se as categorias de anotação fossem aquelas usadas pelo Flickr30k Entities, em que só existe a categoria genérica "pessoa", ou mesmo as de outras famílias de *datasets* multimodais. O MS-COCO (LIN et al., 2014) também só conta com uma categoria de anotação para pessoas – *'person'* - enquanto o Open Images (KUZNETSOVA et al., 2020) tem a mesma categoria – *'person'* - subdividida em *'man, woman, boy, girl'*. Em outras palavras, um *tagset* oriundo de um modelo semântico que não abraçasse a diversidade de perspectivas por *default* não permitira a geração de representações semânticas tão diversas para um mesmo conjunto de dados.

Por fim, a anotação das mesmas imagens para frames de evento traz ainda outra perspectiva para o FM30k, uma vez que permite correlacionar os distintos enquadramentos dados às entidades, tomadas como evocadoras de frames autônomos, com os papéis que elas assumem nos eventos. Voltando à imagem reproduzida na Figura 58, enquanto

as condições de anotação de imagem para frames de entidade na presença ou não da descrição enquadraram a pessoa representada na foto pelo seu gênero ou por sua profissão, respectivamente, a anotação para frames de evento permite a representação de que essa pessoa é o AGENTE do frame de `Trabalhar`, reproduzido na Figura 59.

Figura 59 - Descrição do frame `Trabalhar` na base de dados da FrameNet Brasil.



Fonte: FrameNet Webtool

As análises aqui apresentadas corroboram a hipótese enunciada na introdução desta tese, qual seja a de que diferentes condições de anotação no que concerne à interação entre modalidades – com ou sem presença das descrições – levam a distintas representações semânticas para as imagens. Acreditamos, portanto, haver cumprido o proposto para esta tese.

6 CONCLUSÕES

Dentre as contribuições dessa tese de doutorado, destacam-se a expansão do *dataset* Multi30K para o português brasileiro, o que não apenas introduz uma das dez línguas mais faladas do mundo à família Flickr30K, mas também contribui para reduzir a sub-representação da língua portuguesa no campo do Processamento de Língua Natural (PLN). Além disso, essa expansão estimula a pesquisa multilíngue e multimodal, abrindo novas possibilidades para tarefas compartilhadas em Tradução Automática Multimodal e promovendo um cenário de pesquisa mais diverso em PLN.

Destacamos também como contribuição dessa tese a introdução na FrameNet da anotação de frames para imagens estáticas – uma modalidade visual ainda não explorada, que marca uma ruptura com trinta anos de anotações baseadas exclusivamente em texto – e a criação de uma nova ferramenta de anotação multimodal, disponível online e de forma gratuita para pesquisadores de outras framenets ao redor do mundo.

Para além desta tese o trabalho de pesquisa desenvolvido ao longo do doutorado gerou cinco trabalhos completos publicados nas edições de 2020 (BELCAVELLO et al., 2020), 2022 (BELCAVELLO et al., 2022; TORRENT et al., 2022b; VIRIDIANO et al., 2022) e 2024 (VIRIDIANO et al., 2024) da Language Resources and Evaluation Conference (LREC), uma das mais relevantes conferências da área de Linguística Computacional, cuja mediana h5 é de 126. Houve ainda uma publicação de artigo no volume temático *Representation of Context* do periódico Frontiers in Psychology (TORRENT et al., 2022). Somam-se às publicações, diversas apresentações de trabalho.

Em um compromisso com a ciência aberta, o FM30K já se encontra disponível nos repositórios da FrameNet Brasil no GitHub[1] e no HuggingFace[2].

Finalmente, ao consideramos o atual cenário de pesquisas em Linguística Computacional e Visão Computacional, vemos que muitos estudos que utilizam *datasets* da família Flickr30K ainda produzem modelos e métricas que, acreditamos, carecem de análises detalhadas das semânticas resultantes da integração das modalidades visual e textual. Nesse sentido, a adição de dados semânticos da FrameNet a esses *datasets* tem o potencial de aumentar significativamente sua granularidade e informatividade.

Como trabalhos futuros, planejamos incluir em nosso *dataset* as anotações das imagens para frames de evento sem a presença de descrição. Com isso, esperamos garantir mais uma perspectiva sobre cada imagem, aprimorando as descrições semânticas das situações apresentadas e enriquecendo ainda mais o conjunto de dados.

Planejamos, também, dar continuidade ao desenvolvimento de uma nova tarefa de anotação voltada para a expansão das correlações existentes entre entidades e sintagmas

[1] $github.com/FrameNetBrasil/framed-multi30k$
[2] $huggingface.co/datasets/FrameNetBrasil/Framed_Multi30k$

nominais, fornecidas pelo Flickr30K Entities, para o português. Nessa nova tarefa – já testada com estudantes da Universidade de Leipzig durante o período de Doutorado Sanduíche – os anotadores serão convidados a alinhar os sintagmas nominais das legendas originais em inglês com suas respectivas traduções para o português, o que nos permitirá usar as *bounding boxes* anotadas manualmente para o Flickr30K Entites para também atribuir frames e EFs aos pares de imagens e traduções em português.

Por fim, a partir da experiência de anotação de um vasto número de imagens para frames e Elementos de Frame, o que, não raro, levou à criação de novos frames, uma pergunta de pesquisa digna de uma nova tese se coloca: Em que medida a expansão da cobertura da FrameNet Brasil para o domínio visual impactará na própria natureza dos frames criados. Como recurso lexicográfico em sua origem, a FrameNet Brasil carrega, na definição de seus frames, o enviesamento imposto pelo léxico da língua portuguesa. A nova pergunta diz respeito, portanto, aos impactos desta nova perspectiva sobre a própria base de dados trazida pela anotação de imagens.

REFERÊNCIAS

ANTHIS, J. R., Lum, K., EKSTRAND, M., FELLER, A., D'AMOUR, A., & TAN, C. (2024). The Impossibility of Fair LLMs. **arXiv e-prints, arXiv-2406.**

AROYO, L., & WELTY, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. **AI Magazine,** 36(1), 15-24.

BANERJEE, S., & LAVIE, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: **Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization** (pp. 65-72).

BARRAULT, Loïc; BOUGARES, Fethi; SPECIA, Lucia; LALA, Chiraag; ELLIOTT, Desmond, et al.. Findings of the Third Shared Task on Multimodal Machine Translation. **THIRD CONFERENCE ON MACHINE TRANSLATION (WMT18)**, Oct 2018, Brussels, Belgium. pp.308-327

BARTHES, R. Rhetoric of the Image, in **R. Barthes (ed.) Image–Music–Text**, London. 1977.

BELCAVELLO, F.; VIRIDIANO, M.; DINIZ DA COSTA, A.; MATOS, E. E.; TORRENT, T. T. (2020). Frame-Based Annotation of Multimodal Corpora: Tracking (A)Synchronies in Meaning Construction. In: **Proceedings of the LREC International FrameNet Workshop 2020**. Marseille, France: ELRA, p. 23-30.

BELCAVELLO, F.; VIRIDIANO, M.; MATOS, E.; TORRENT, T. T. (2022). Charon: A FrameNet Annotation Tool for Multimodal Corpora. In: **Proceedings of The 16th Lingusitic Annotation Workshop (LAW-XVI) within LREC2022.** Marseille, France: ELRA, p. 91-96.

BENDER, E. M., GEBRU, T., MCMILLAN-MAJOR, A., & SHMITCHELL, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. **In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency** (pp. 610-623).

BEUKEBOOM, C. J., FORGAS, J., Vincze, O., & LASZLO, J. (2014). Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. **Social cognition and communication**, 31, 313-330.

BIRHANE, A., PRABHU, V. U., & KAHMBWE, E. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. **arXiv preprint arXiv:2110.01963.**

BURNS, K., HENDRICKS, L. A., SAENKO, K., DARRELL, T., & ROHRBACH, A. (2018). Women also snowboard: Overcoming bias in captioning models. **arXiv preprint arXiv:1803.09797.**

BASILE, V., CABITZA, F., CAMPAGNER, A., & FELL, M. (2021). Toward a perspectivist turn in ground truthing for predictive . Toward a perspectivist turn in ground truthing for predictive computing. In: **Proceedings of the AAAI Conference on Artificial Intelligence.**

CHO, K., VAN MERRIËBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWECK, H., & BENGIO, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. **arXiv preprint arXiv:1406.1078**, 2014.

COHN, Neil; MAGLIANO, Joseph P. Editors' Introduction and Review: Visual Narrative Research: An Emerging Field in Cognitive Science. **Topics in Cognitive Science**, v. 12, n. 1, p. 197-223, 2020.

CONNEAU, A., KHANDELWAL, K., GOYAL, N., CHAUDHARY, V., WENZEK, G, GUZMÁN, F., GRAVE, E., OTT, M., ZETTLEMOYER, L., and STOYANOV, V. "Unsupervised cross-lingual representation learning at scale."**arXiv preprint arXiv:1911.02116** (2019).

DÁNNELS, D., TORRENT, T. T., SIGILIANO, N.S., and DOBNIK, S. (2022) Beyond strings of characters: Resources meet NLP–Again. In: VOLODINA, E., DANNELLS, D., BERDICEVSKIS, A., FORSBERG, M. and VIRK, S. (Orgs.). **Live and Learn** (pp.29-36). Göterborgs: Instituitionen för Svenska, Flerspråkighet och Språkteknologi.

ELLIOTT, D., & KELLER, F. (2014, June). Comparing automatic evaluation measures for image description. In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics** (Volume 2: Short Papers) (pp. 452-457).

ELLIOTT, Desmond et al. Multi30k: Multilingual english-german image descriptions. **arXiv preprint arXiv:1605.00459**, 2016.

ELLIOTT, Desmond et al. Findings of the second shared task on multimodal machine translation and multilingual image description. **arXiv preprint arXiv:1710.07177**, 2017.

ELLSWORTH, Michael; JANIN, Adam. Mutaphrase: Paraphrasing with framenet. In: **Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing**. 2007. p. 143-150.

FENG, Yansong; LAPATA, Mirella. How many words is a picture worth? automatic caption generation for news images. In: **Proceedings of the 48th annual meeting of the Association for Computational Linguistics**. 2010. p. 1239-1249.

FILLMORE, Charles J. The case for case reopened. **Syntax and semantics**, v. 8, n. 1977:59-82, 1977.

FILLMORE, C. J. Frame semantics. **In: Linguistics in the Morning Calm.** Seoul, South Korea: Hanshin Publishing Co. 1982.

FILLMORE, Charles J.; JOHNSON, Christopher R.; PETRUCK, Miriam RL. Background to framenet. **International journal of lexicography**, v. 16, n. 3:235-250, 2003.

FILLMORE, Charles J. Border conflicts: FrameNet meets construction grammar. In: **Proceedings of the XIII EURALEX international congress**. 2008.

GARG, M., WAZARKAR, S., SINGH, M., & BOJAR, O. (2022, June). Multimodality for NLP-Centered Applications: Resources, Advances and Frontiers. In: **Proceedings of the Thirteenth Language Resources and Evaluation Conference** (pp. 6837-6847).

GRUBINGER, Michael et al. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In: **International workshop ontoImage**. 2006.

HODOSH, Micah; YOUNG, Peter; HOCKENMAIER, Julia. Framing image description as a ranking task: Data, models and evaluation metrics. **Journal of Artificial Intelligence Research**, v. 47, p. 853-899, 2013.

HOFFMANN, J., BORGEAUD, S., MENSCH, A., BUCHATSKAYA, E., CAI, T., RUTHERFORD, E., ...& Sifre, L. (2022). Training compute-optimal large language models. **arXiv preprint arXiv:2203.15556.**

JAIMES, Alejandro; CHANG, Shih-Fu. Conceptual framework for indexing visual information at multiple levels. In: **Internet Imaging. International Society for Optics and Photonics**, 1999. p. 2-15.

JEWITT, Carey; BEZEMER, Jeff; O'HALLORAN, Kay. **Introducing multimodality**. Routledge, 2016.

MCKEVITT, Paul. MultiModal semantic representation. In: **First Working Meeting of the SIGSEM Working Group on the Representation of MultiModal Semantic Information**. 2003. p. 1-16.

KILICKAYA, M., AKKUS, B. K., CAKICI, R., ERDEM, A., ERDEM, E., & IKIZLER-CINBIS, N. (2017). Data-driven image captioning via salient region discovery. **IET Computer Vision,** 11(6), 398-406.

KRESS, Gunther R. et al. **Reading images: The grammar of visual design**. Psychology Press, 1996.

KRESS, Gunther; VAN LEEUWEN, Theo. Multimodal discourse. **The Modes and Media of Contemporary Communication.(Cappelen, London 2001)**, 2001.

KUZNETSOVA, A., ROM, H., ALLDRIN, N., UIJLINGS, J., KRASIN, I., PONT-TUSET, J., ... & FERRARI, V. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. **International journal of computer vision**, 128(7), 1956-1981.

LALA, Chiraag; SPECIA, Lucia. Multimodal lexical translation. In: **proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. 2018.

LAKOFF, G. (1987). Women, fire, and dangerous things (Vol. 10). Chicago: University of Chicago press.

LIN, T. Y., MARIE, M., BELONGIE, S., HAYS, J., PERONA, P., RAMANAN, D., ... & ZITNICK, C. L. (2014). Microsoft coco: Common objects in context. In **Computer Vision–ECCV 2014: 13th European Conference**, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740-755). Springer International Publishing.

Li, X., Lan, W., Dong, J., & Liu, H. (2016, June). Adding chinese captions to images. In **Proceedings of the 2016 ACM on international conference on multimedia retrieval** (pp. 271-275).

MARTINEC, Radan; SALWAY, Andrew. A system for image–text relations in new (and old) media. **Visual communication**, v. 4, n. 3:337-371, 2005.

MATTHIESSEN, C. **Introduction to functional grammar**. 1989.

MØLLER, A. G., PERA, A., DALSGAARD, J., & AIELLO, L. (2024, March). The parrot dilemma: Human-labeled vs. llm-augmented data in classification tasks. In **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics** (Volume 2: Short Papers) (pp. 179-192).

NAVEED, H., KHAN, A. U., QIU, S., SAQIB, M., ANWAR, S., USMAN, M., ... & MIAN, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.

ORDONEZ, Vicente; KULKARNI, Girish; BERG, Tamara L. Im2text: Describing images using 1 million captioned photographs. In: **Advances in neural information processing systems**. 2011. p. 1143-1151.

PAPINENI, Kishore et al. BLEU: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th annual meeting on association for computational linguistics**. Association for Computational Linguistics, 2002. p. 311-318.

PLUMMER, B. A., WANG L., CERVANTES, C. M., CAICEDIO, J. C., HOCKENMAIER, J., and LAZEBNIK, S. (2015). Flickr30k Entities: **Collecting region-to-phrase correspondences for richer image-to-sentence models.**

PUSTEJOVSKY, J. **The Generative Lexicon**. Cambridge, USA: MIT Press. 1995.

PUSTEJOVSKY, James et al. Towards a Generative Lexical Resource: The Brandeis Semantic Ontology. In: **LREC 2006.** p. 1702-1705.

PUSTEJOVSKY, J.; JEZEK, E. Qualia Structure. In: PUSTEJOVSKY, J.; JEZEK, E. (Orgs.). **Integrating Generative Lexicon and Lexical Semantic Resources**. **LREC 2016-Slovenia.** 2016. 139 p.

RADFORD, A., KIM, J.W., HALLACY, C., RAMESH, A., GOH, G., AGARWAL, S., SASTRY, G., ASKELL, A., MISHKIN, P., CLARK, J. and KRUEGER, G. (2021, July). Learning transferable visual models from natural language supervision. In **International conference on machine learning** (pp. 8748-8763). PMLR.

RASHTCHIAN, Cyrus et al. Collecting image annotations using Amazon's Mechanical Turk. In: **Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk**. Association for Computational Linguistics, 2010. p. 139-147.

REITER, E. (2018). A structured review of the validity of BLEU. **Computational Linguistics**, 44(3), 393-401.

ROGERS, A. (2021). Changing the world by changing the data. **arXiv preprint arXiv:2105.13947.**

RUPPENHOFER, Josef et al. **FrameNet II: Extended theory and practice**. 2006.

SANABRIA, Ramon et al. How2: a large-scale dataset for multimodal language understanding. **arXiv preprint arXiv:1811.00347**, 2018.

SHATFORD, Sara. Analyzing the subject of a picture: a theoretical approach. **Cataloging & classification quarterly**, v. 6, n. 3, p. 39-62, 1986.

SCHUHMANN, C., VENCU, R., BEAUMONT, R., KACZMARCZYK, R., MULLIS, C., KATTA, A., COOMBES, T., JITSEV, J., and KOMATSUZAKI, A. "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs."**arXiv preprint arXiv:2111.02114** (2021).

STUDENT. The probable error of a mean. **Biometrika**, p. 1-25, 1908.

TORRENT, T. T.; MATOS, E.; LAGE, L.; LAVIOLA, A.; TAVARES, T.; ALMEIDA, V. G.; SIGILIANO, N. (2018). Towards continuity between the lexicon and the constructicon in FrameNet Brasil In: LYNGFELT, B.; BORIN, L.; OHARA, K. H.; TORRENT, T. T. (Orgs.). **Constructional Approaches to Language**. Amsterdam: John Benjamins Publishing Company.

TORRENT, T. T.; MATOS, E. E.; SIGILIANO, N. S. (2020). Construction grammar across borders. **Constructions and Frames,** v. 12, p. 1-7.

TORRENT, T. T.; MATOS, E. E. S.; BELCAVELLO, F.; VIRIDIANO, M.; GAMONAL, M. A.; COSTA, A. D.; MARIM, M. C. (2022). Representing Context in FrameNet: A Multi-Dimensional, Multimodal Approach. **Frontiers in Psychology,** v. 13, article 838441.

TORRENT, T. T.; LORENZI, A.; MATOS, E. E.; BELCAVELLO, F.; VIRIDIANO, M.; GAMONAL, M. A. (2022). Lutma: A Frame-Making Tool for Collaborative FrameNet Development. In: **Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022.** Marseille, France: ELRA, p. 100-107.

TROTT, S; TORRENT, T. T.; CHANG, N; SCHNEIDER, N. (2020). (Re)construing Meaning in NLP. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: ACL, p.5170-5184.

UPPAL, S., BHAGAT, S., HAZARIKA, D., MAJUMDER, N., PORIA, S., ZIMMERMANN, R., & ZADEH, A. (2022). Multimodal research in vision and language: A review of current and emerging trends. **Information Fusion,** 77, 149-171.

VAN MILTENBURG, E. (2016). Stereotyping and bias in the flickr30k dataset. **arXiv preprint arXiv:1605.06083.**

VAN MILTENBURG, E., ELLIOTT, D., and VOSSEN, P. 2017. Cross-linguistic differences and similarities in image descriptions. In **Proceedings of the 10th International Conference on Natural Language Generation.**

VAN MILTENBURG, C. W. J. (2019). **Pragmatic factors in (automatic) image description.**

VEDANTAM, R., LAWRENCE ZITNICK, C., & PARIKH, D. (2015). Cider: Consensus-based image description evaluation. In **Proceedings of the IEEE conference on computer vision and pattern recognition** (pp. 4566-4575).

VIRIDIANO, M.; TORRENT, T. T.; CZULO, O.; LORENZI, A.; MATOS, E.; BELCAVELLO, F. (2022). . In: **Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022.** Marseille, France: ELRA, p. 108-116.

VIRIDIANO, M.; LORENZI, A.; TORRENT, T. T.; MATOS, E. E.; PAGANO, A.; SIGILIANO, N. S.; GAMONAL, M. A.; ABREU, H. A.; DUTRA, L. V.; SAMAGAIO, M.; CARVALHO, M. ET AL. (2024). Framed Multi30K: A Frame-Based Multimodal-Multilingual Dataset. In: **Proceedings the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**: Torino: ELRA/ICCL, p.7438–7449.

XIA, P., QIN, G., VASHISHTA, S., CHEN, Y., CHEN, T., MAY, C., HARMAN. C., RAWLINS, K., WHITE, A. S., and VAN DURME, B. "LOME: Large ontology multilingual extraction."arXiv preprint **arXiv:2101.12175 (2021)**

YOUNG, Peter et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. **Transactions of the Association for Computational Linguistics**, v. 2, p. 67-78, 2014.

ZHAO, J., WANG, T., YATSKAR, M., ORDONEZ, V., & CHANG, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.** Association for Computational Linguistics, Copenhagen, Denmark, pages 2979–2989.