v. 30, n. 1, p. 74–93, 2025
ISSN: 2238-3824 · DOI: 10.35699/2238-3824.2025.57569
Submetido em: 14/02/2024 · Aprovado em: 31/03/2025

caligrama
revista de estudos românicos

Representações multimodais de conteúdos do gênero jornalístico: ganhos e desafios da expansão dos datasets da ReINVenTA

Multimodal representations of content in the journalistic genre: gains and challenges of expanding ReINVenTA datasets

Frederico Belcavello
Universidade Federal de Juiz de Fora (UFJF) | Juiz de Fora | MG | BR
College of Arts and Science, Case Western Reserve University (CWRU) | Cleveland | OH | EUA
CNPq
CAPES
fred.belcavello@ufjf.br
https://orcid.org/0000-0001-5808-5201

Marcelo Viridiano
Universidade Federal de Juiz de Fora (UFJF) | Juiz de Fora | MG | BR
College of Arts and Science, Case Western Reserve University (CWRU) | Cleveland | OH | EUA
CNPq
CAPES
marcelo.viridiano@case.edu
https://orcid.org/0000-0002-9706-8663

Resumo: Este artigo discute os ganhos e desafios da expansão do *dataset* da ReINVenTA para a inclusão gêneros multimodais jornalísticos, explorando as especificidades e relações entre elementos visuais e textuais neste novo gênero, e buscando aprimorar a semântica das representações multimodais da atual base de dados. Dois novos *corpora* são propostos: um de imagens e textos jornalísticos, e outro de telejornais, com foco nas matérias televisivas. A metodologia envolve a extração e rotulação automática de dados visuais e textuais, com validação humana para garantir a precisão e mitigar vieses, e anotação integrada de áudio falado e imagens de conteúdos audiovisuais jornalísticos conforme as características peculiares do gênero.

Palavras-chave: Semântica de Frames; multimodalidade; jornalismo; FrameNet; dataset multimodal.

Abstract: This article examines the benefits and challenges of expanding the ReINVenTA dataset to include multimodal journalistic genres, exploring the specificities and relationships between visual and textual elements in this new genre while aiming to enhance the semantic representation of multimodal data in the existing database. Two new corpora are proposed: one corpus consisting of journalistic images and texts and another corpus focused on television news broadcasts, particularly news reports. The methodology involves the automatic extraction and labeling of visual and

textual data, complemented by human validation to ensure accuracy and mitigate biases, as well as the integrated annotation of spoken audio and images from audiovisual journalistic content, considering the peculiar characteristics of the genre.

Keywords: Frame Semantics; multimodality; journalism; FrameNet; multimodal dataset.

1 Introdução

Um dos desafios do estudo de fenômenos linguísticos que envolvem multimodalidade diz respeito à complexidade decorrente da combinação de conteúdos visuais e textuais (Cohn e Magliano, 2020, p. 211). Humanos se comunicam naturalmente por meio da integração de múltiplas modalidades, combinando elementos verbais e não verbais de forma criativa e sofisticada. Para que modelos computacionais sejam capazes de capturar essa complexidade, é essencial que consigam processar dados multimodais de maneira eficiente. Nesse contexto, *datasets* multimodais – especialmente aqueles curados por humanos – desempenham um papel fundamental, pois permitem que modelos de Processamento de Língua Natural (PLN) e Visão Computacional analisem informações combinadas de texto, imagem, áudio e vídeo de forma integrada, possibilitando avanços em tarefas como descrição automática de imagens, legendagem e tradução automática multimodal (Rogers, 2021; Møller *et al*., 2024).

Os conteúdos noticiosos e informativos em circulação no domínio do jornalismo podem ser vistos hoje em dia como exemplares da multimodalidade em muitos suportes midiáticos. Se durante séculos a prática jornalística esteve intimamente relacionada à expressão textual, nas últimas décadas tem-se observado uma transformação profunda, caracterizada por uma integração cada vez maior entre textos e recursos visuais. Durante o século XX, os avanços tecnológicos permitiram a incorporação gradual de elementos como fotografias, ilustrações e infográficos, que passaram a enriquecer as narrativas informativas publicadas em jornais e revistas. Essa evolução culminou em uma revolução digital no século XXI, na qual o uso de imagens se intensificou e diversificou de maneira significativa, povoando não apenas o papel, mas também, e principalmente, as telas.

Paralelamente, o surgimento e a consolidação de plataformas digitais ampliaram o alcance e a influência dos vídeos jornalísticos, promovendo uma mudança paradigmática na forma de disseminar informações. Em contraste com a predominância de conteúdos textuais que caracterizava os primórdios da internet, na década de 1990, e os estágios iniciais das redes sociais, nos anos 2000, o ambiente digital contemporâneo privilegia uma comunicação mais dinâmica e visualmente atraente. Nesse sentido, imagens e vídeos não são meros complementos estéticos, mas constituem instrumentos estratégicos essenciais para a captação da atenção do público e para a efetiva transmissão de informações em meio a um fluxo de dados cada vez mais intenso e diversificado.

Partimos da premissa de que o desenvolvimento de tarefas computacionais no campo do processamento multimodal pode se beneficiar de uma compreensão mais aprofundada

dos significados que emergem da interação entre informações textuais e visuais no conteúdo jornalístico. Assim, esta expansão do dataset da ReINVenTA busca contribuir para a criação de modelos mais eficazes, através da construção de um novo conjunto de dados, com curadoria humana, no qual relações semânticas estruturadas a partir da base de dados da FrameNet possam ser atribuídas a pareamentos de imagem e texto extraídos de conteúdos jornalísticos. Isso possibilita um mapeamento mais granular das relações multimodais desse gênero e amplia as aplicações da FrameNet Brasil nesse domínio.

Uma FrameNet pode ser definida como um modelo computacional da cognição linguística, implementado na forma de um banco de dados relacional (Fillmore *et al.*, 2003). Nele, itens lexicais, bem como outras estruturas linguísticas, são modelados em termos de *frames* – ou sistemas de conceitos – que evocam. Assim, palavras como *informação.n* e noticiar.v evocam o frame de Informação, incluído na base de dados da FrameNet Brasil conforme a Figura 1. Note-se que há três elementos nucleares para este frame – Informação, Pensador e Tópico – e dois elementos periféricos – Fonte e Meio_de_coleta. Enquanto os três primeiros são necessários à instanciação do frame, os dois últimos são opcionais, como demonstra o exemplo (1), em que os elementos linguísticos na sentença são marcados por cores conforme as apresentadas na Figura 1, sendo os optativos indicados entre parênteses.

(1) Eu recebi as informações atualizadas sobre o projeto (do meu orientador) (em resposta ao e-mail que enviei ontem).

Em uma abordagem multimodal, norteia os trabalhos desenvolvidos no âmbito da ReINVenTA o pressuposto de que, assim como os itens linguísticos podem evocar *frames*, elementos visuais presentes em imagens podem também fazê-lo ou atuar de maneira complementar aos *frames* evocados pela língua. Dessa forma, iniciou-se a criação de *corpora* e de *datasets* multimodais anotados e com padrão ouro (*gold standard datasets*) que possam ser posteriormente empregados para tarefas de aprendizagem de máquina dedicadas à rotulação semântica automática de objetos multimodais, tanto na semiose linguística quanto em outras. Detalhamos esses *datasets* na próxima seção.

Figura 1: O frame de Informação na base de dados da FrameNet Brasil

Fonte: FrameNet Brasil Webtool

2 Os *datasets* da ReINVenTA

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 Processamento de Língua Natural e Visão Computacional como *Visual Question Answering*, *Visual Commonsense Reasoning*, *Image and Video Captioning*, *e Multimodal Machine Translation*, dentre outras (Garg *et al*., 2022). Os termos 'multimodal' e 'multimodalidade', apesar de ensejarem debates teóricos sobre as nuances de sua abrangência e precisão de sua aplicação, referem-se sempre ao processo de composição de mensagens em que dois ou mais modos comunicativos ou modos semióticos são combinados para a produção de sentido. Apesar de frequentemente associarmos modos comunicativos ou semióticos com os canais ou suportes midiáticos – tais como texto, imagens, ou sons – percebemos modo ou modalidade de forma mais ampla, como um recurso (ou conjunto de recursos) reconhecido experiencialmente para a criação de significado. Isso quer dizer que, às vezes, podemos ter multimodalidade com diferentes modos visuais, ou diferentes modos sonoros, por exemplo.

Assim, no escopo deste artigo, tomando a cognição humana como fundamento, usamos o termo multimodalidade para nos referirmos à capacidade de um sistema ou modelo de processar dados obtidos simultaneamente a partir de diferentes modalidades comunicativas, ou seja, para nos referirmos à integração dos múltiplos modos de comunicação ou informação utilizados por esses sistemas e modelos para interpretar e analisar dados. Nesse sentido, chamamos de *datasets* multimodais os conjuntos de dados que combinam duas ou mais dessas modalidades comunicativas.

2.1 O conjunto de imagens e descrições

Para a composição inicial do conjunto de imagens estáticas que integram parte do *dataset* da ReINVenTA, foram utilizados três *datasets* multimodais como referência. O primeiro deles, o Flickr 30k (Young *et al*., 2014), contém 31.014 imagens – fotografias de atividades, eventos e cenas cotidianas extraídas do Flickr – cada uma acompanhada por cinco descrições em inglês, totalizando 158.915 descrições. Essas descrições foram criadas através de uma tarefa na qual os participantes, sem acesso a informações contextuais adicionais sobre as imagens, foram orientados a descrever as entidades e eventos apresentados em cada imagem – pessoas, objetos, ambientes e atividades sendo desenvolvidas – buscando produzir um tipo específico de descrições, chamadas de "descrições conceituais" (Hodosh *et al*., 2013, p. 857), ou seja, descrições que, embora possam conter inferências sobre o contexto da cena retratada, concentram-se apenas nas informações que podem ser obtidas a partir da imagem.[1] A esse conjunto

[1] Descrições conceituais podem ser melhor compreendidas em oposição ao que Hodosh e outros (2013, p. 857) definem como "descrições não-visuais", ou seja, 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 na medida em que referência a elementos visuais que não podem ser identificados na imagem.

inicial de dados foram adicionadas duas extensões. A primeira, Flickr 30K Entities (Plummer *et al*., 2015), introduz *bounding boxes*[2] para estabelecer não apenas a correspondência entre as entidades presentes nas imagens e os itens lexicais com os quais se relacionam nas descrições – processo denominado *region-to-phrase correspondences*, ou correspondências entre região e sintagma – mas também a correlação entre os diferentes sintagmas que, ao longo das cinco descrições, referem-se a uma mesma entidade ou conjunto de entidades na imagem. Esse último aspecto, denominado cadeias de correferência – em inglês, *coreference chains*, – permite mapear menções a um mesmo elemento visual em diferentes descrições, aprimorando a relação entre texto e imagem no *dataset*. A segunda, uma expansão multilíngue chamada Multi30K (Elliott *et al*., 2016), incorpora ao dataset 155.070 novas descrições originais em alemão – criadas seguindo a metodologia utilizada na criação do Flickr 30K – e 31.014 traduções para o alemão – produzidas por tradutores profissionais falantes nativos da língua alemã a partir de uma das cinco descrições originais em inglês para cada imagem. A partir desses conjuntos de dados desenvolve-se, no âmbito da ReINVenTA, o Framed Multi30K (Viridiano *et al*., 2024). Trata-se de um conjunto de dados que estende esses *datasets* de referência com o acréscimo de cinco novas descrições originais em português para cada imagem e com a tradução para o português das descrições em inglês utilizadas pelo Multi30K. Além disso, aprimora sua granularidade semântica ao atribuir, através de processos de rotulação automática e anotações manuais, relações entre as entidades representadas em cada imagem e os frames e elementos de *frame* existentes na base de dados da FrameNet Brasil.

Apesar de seu *status* de *corpora benchmark* para tarefas que envolvem processamento simultâneo de conteúdos visuais e linguísticos (Uppal *et al*., 2022), cabe aqui destacar as críticas de autores como Van Miltenburg (2016) à premissa de neutralidade das descrições conceituais, ou seja, à ideia de que é possível criar descrições objetivas baseadas exclusivamente nos elementos visuais de uma imagem, desconsiderando os processos de interpretação e recontextualização inerentes à cognição humana. Segundo Van Miltenburg (2016, p. 1), tais descrições representam apenas uma simplificação conveniente para a criação de *datasets* voltados ao desenvolvimento de modelos que dependem de um mapeamento direto entre os elementos visuais e suas descrições. O autor argumenta que, mesmo quando instruídos explicitamente a fornecer descrições simples, completas e objetivas das entidades proeminentes em uma imagem – evitando inferências sobre o que está acontecendo na cena – os anotadores frequentemente introduzem enviesamentos linguísticos e inferências infundadas em suas descrições.

Críticas dessa natureza, sobre a validade e a aplicabilidade dos pareamentos entre imagem e texto do Flickr30K e suas extensões – construídos artificialmente para atender a demandas de tarefas computacionais específicas – promovem questionamentos sobre sua representatividade e adequação para aplicações que vão além do domínio técnico para o qual foram projetados. Essas questões nos motivaram a buscar uma nova expansão desse *corpus*, incorporando dados que contenham descrições em contextos mais amplos, que emergem de situações reais de uso da linguagem e desempenham funções discursivas

[2] *Bounding boxes* são retângulos delimitadores utilizados na anotação de imagens para marcar a localização de objetos ou regiões de interesse dentro de uma cena. Essas caixas são amplamente empregadas em visão computacional para tarefas como detecção de objetos, rastreamento e reconhecimento de padrões.

específicas, como no caso do pareamento entre imagem e texto no contexto jornalístico, abordagem que exploraremos na seção 3.

2.2 O conjunto de vídeo: imagens e áudio falado

Entender e modelar as relações estabelecidas entre as imagens e o áudio falado em vídeos foi a motivação que desencadeou uma série de projetos desenvolvidos pela equipe da FrameNet Brasil no âmbito da ReINVenTA. Começando com Belcavello *et al*. (2020), o princípio balizador foi o de que há uma combinação entre elementos presentes na imagem e elementos presentes no áudio falado que concorrem para a construção de sentido e o estabelecimento de significado em termos multimodais, ou seja, na totalidade da percepção de um vídeo e não em cada modalidade separadamente.

Em termos da Semântica de *Frames* (Fillmore, 1982), a hipótese confirmada em Belcavello (2023) foi a de que de maneira equivalente ao material linguístico, as imagens – e os elementos nelas contidos – também podem: i. evocar *frames*; ii instanciar elementos de *frames*; ou iii. trabalhar em conjunto com o material linguístico presente no áudio falado no processo de preenchimento de elementos de frame, estabelecendo relações frame a frame ou relações qualia entre itens lexicais. Tal conclusão foi materializada por meio da constituição do *dataset* Frame² – *Frame Squared* – (Belcavello *et al*., 2024) construído sobre os 230 minutos de vídeo, correspondentes aos 10 episódios da primeira temporada do programa de viagens Pedro Pelo Mundo.3 Ao material audiovisual se sobrepõem as anotações para as categorias da FrameNet tanto do texto quanto da imagem conduzidas manualmente pela equipe de anotadores da FrameNet Brasil. Dessa forma, Frame² é um *dataset* composto por objetos multimodais: uma combinação imagem e texto estabelecida em um determinado espaço de tempo que carrega consigo as anotações e as relações entre os dados anotados, conforme mediado pela estrutura semântica modelada no banco de dados da FrameNet Brasil (Torrent *et al*., 2022). Isso significa que as informações sobre os frames, seus elementos de *frame*, suas relações com outros frames e as relações entre unidades lexicais estão incluídas no dataset Frame².

A metodologia adotada para a tarefa de anotação que construiu o *dataset* se dividiu em duas subtarefas: anotação de texto (gerado a partir da transcrição do áudio falado no *corpus*) e anotação de imagem. A partir dos experimentos reportados em Belcavello (2023), concluiu-se que a anotação de texto deveria ser feita primeiro, uma vez que neste gênero multimodal a organização inicial do sentido se dá prioritariamente a partir do texto, deixando a imagem, frequentemente, no papel de ilustração, ancorada ao texto. Assim, anotadores poderiam

[3] O programa estreou em 2016 no canal a cabo GNT, do grupo Globo, que é dedicado a produções audiovisuais sobre entretenimento e estilo de vida. Quatro temporadas de "Pedro pelo Mundo" foram ao ar até 2019. A primeira temporada tem 10 episódios de 23 minutos cada. A segunda, a terceira e a quarta também são compostas por 10 episódios cada, mas com 48 minutos de duração. Para fins desse *dataset*, o *corpus* foi limitado aos 10 episódios da primeira temporada. O enredo de cada episódio se resume a entrar em contato e explorar aspectos sociais, econômicos e culturais de um local que passou por algum tipo de transformação recente. Assim, o que o espectador vê é Pedro Andrade, o apresentador, tentando se conectar com os moradores locais, em vez de simplesmente propor uma visão turística de locais de interesse. O formato do programa combina passagens, sequências com locução em *off*, entrevistas curtas e sequências de videoclipes. Assim, oferece material rico como exemplo de composição audiovisual complexa para a criação de significado.

proceder de duas maneiras: (i) anotar primeiro todas as sentenças de um episódio e depois começar a anotar as imagens; ou (ii) completar a anotação das sentenças que correspondem a uma sequência,[4] anotar em seguida as imagens presentes na respectiva sequência e, depois, passar para as sentenças da sequência seguinte. Na prática, o que se verificou foi que os anotadores, sem exceção, optaram pela primeira maneira, anotando todas as sentenças do episódio para, apenas depois, passarem a anotar as imagens. Considerando que são 2.195 sentenças para um total de 10 episódios, cada anotador, portanto, dedicou-se a um lote de aproximadamente 200 sentenças para depois iniciar a anotação de objetos visuais.

Deve-se salientar, no entanto, que as diretrizes fundamentais para fazer dessa tarefa um processo de anotação multimodal foram: i) ao anotar texto para *corpora* audiovisuais, anotadores devem sempre assistir ao vídeo e ver as frases em seu contexto multimodal; e ii) da mesma forma, ao anotar imagens, anotadores devem sempre ouvir o áudio falado e, também, ler as sentenças transcritas disponibilizadas no espaço de trabalho de anotação de vídeo. Assim, considerando sempre tanto as imagens quanto o áudio falado conjuntamente no processo de anotação, garante-se o aspecto multimodal da anotação e, por consequência, do *dataset*.

As 2.195 sentenças do *corpus* geraram 11.796 *annotation sets* (AS) de texto, enquanto as imagens foram anotadas para 6.841 objetos visuais (VOs). Até onde sabemos, esse é o primeiro conjunto de dados que combina uma abordagem multimodal e semântica de *frames* para anotação de vídeo de objetos visuais. O *dataset* Frame² é uma expansão da FrameNet para o domínio multimodal a partir de um corpus de vídeo. O objetivo era oferecer um novo recurso padrão ouro, de granulação refinada, enriquecido semanticamente, para tarefas de PNL multimodal.

A abordagem multimodal do conjunto de dados mantém a ancoragem linguística na maneira como os elementos nele contidos podem ser analisados, explorados e usados. No entanto, a pesquisa realizada para culminar nesse conjunto de dados mostra que o caminho para abordar a imagem nos processos de criação de significado é amplo e oferece outras possibilidades que valem a pena ser exploradas. Algumas dessas possibilidades já estão sendo exploradas e são reportadas neste volume de *Caligrama*.

Outro ponto amadurecido a partir da experiência com a construção do Frame² é o de que é necessário levar em conta o gênero audiovisual em questão para cada *corpus*, cada tarefa de anotação e, assim, cada *dataset*. Dessa forma, a constituição de um novo dataset a partir de conteúdos do domínio jornalístico traz novas perspectivas, as quais debatemos na seção seguinte.

3 A Multimodalidade no Contexto Jornalístico

É de fácil percepção empírica o crescimento significativo do uso de imagens no jornalismo nas últimas décadas, refletindo uma abordagem mais integrada e visual nas publicações informativas contemporâneas. O desenvolvimento tecnológico vem possibilitando o cresci-

[4] Para os fins da tarefa, no âmbito do *corpus* Pedro Pelo Mundo, uma sequência foi definida como um conjunto de cenas que apresenta uma unidade distinta em termos do tópico apresentado como um subtópico do tema do episódio. Exemplo: a sequência do kilt no episódio de Edimburgo; a sequência da carne de tubarão no episódio de Reiquiavique.

mento do uso de fotografias, ilustrações, infográficos e quaisquer aparatos visuais ao longo de todo o século XX, com uma explosão digital no século XXI. Da mesma forma, a presença de vídeos de conteúdo jornalístico em websites e redes sociais alcançou patamares bastante superiores à realidade notadamente textual que se conformava tanto no início da internet, na década de 1990, quanto no início das redes sociais nos anos 2000. Em geral, em todo o lócus digital, imagens tornaram-se fundamentais para captar a atenção de usuários, em meio a um fluxo significativamente denso de informações no ambiente da web.

Dado que a construção de significado não resulta de uma simples soma entre as modalidades, mas emerge da interação entre elas, mesmo quando uma modalidade parece desempenhar um papel secundário ou marginal, esta pode exercer uma função essencial de enquadramento (*framing*) ou direcionamento de sentido dentro de uma narrativa discursiva mais ampla, explorando processos de elaboração, extensão ou aprimoramento do significado (Matthiessen, 1989). A relação entre imagem e texto nesses contextos demanda que leitores ou telespectadores negociem diferentes tipos de relações combinatórias, mobilizando estratégias interpretativas mais complexas do que aquelas previstas para as descrições conceituais tradicionais.

3.1 A relação entre imagem e texto no jornalismo impresso e digital

No contexto jornalístico, a construção das relações entre imagem e texto, moldada por propósitos comunicativos e editoriais específicos, diferencia-se significativamente da abordagem que adotamos na construção do Framed Multi30K. Para analisar como se estabelecem estas relações, autores como Martinec e Salway (2005) tomam como ponto de partida a taxonomia proposta por Barthes (1977). Essa taxonomia introduz diferentes tipos de relações de *status* entre as modalidades, descrevendo certos tipos fundamentais de relações de coocorrência imagem-texto. Nos interessam dois tipos de relações: o ancoramento, no qual o texto fornece contexto e interpretação para a imagem; e a ilustração, que ocorre quando a imagem apoia e amplia o significado do texto.

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. 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, em uma descrição ou legenda que acompanha uma imagem e torna possível dizer "o que é aquela imagem". Na Figura 2, o título ancora o significado da imagem ao direcionar a interpretação do leitor. Sem a descrição, a fotografia poderia ser interpretada de diferentes maneiras: o homem poderia ser um palestrante, um ativista ou mesmo um artista se apresentando. O texto, ao especificar que ele é um candidato e que está discursando sobre economia em um comício, delimita o significado da imagem e reduz a ambiguidade interpretativa, reforçando um enquadramento específico para o evento representado.

Figura 2: Exemplo de ancoramento do sentido da imagem pelo texto.

Em Contagem, Lula discursa sobre questões econômicas e condições financeiras dos brasileiros.

Candidato à presidência cumprirá agenda nesta quarta-feira em Juiz de Fora. Lula é o primeiro pré-candidato ao Planalto a passar pela cidade em 2022.

Fonte: Tribuna de Minas (2022a).

Nas relações de ilustração – quando a imagem está subordinada ao texto – 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. Na Figura 3, vemos o exemplo de três imagens distintas que poderiam ser utilizadas para ilustrar uma sentença que faça referência a "materiais escolares", sem que a imagem interfira na interpretação do texto. Nesses casos, o conteúdo visual não adiciona novas informações ao texto, mas apenas ilustra seu significado.

Figura 3: Exemplos de imagens que servem como ilustração para uma mesma sentença sobre materiais escolares.

Fonte: Leonard (2021); Fernandez (2021); Ricciardi (2021).

Além das relações de ancoramento e ilustração, a complementaridade entre imagem e texto desempenha papel central na construção do significado em discursos multimodais (Martinec; Salway, 2005, p. 342). Segundo os autores, a complementaridade ocorre quando ambas as modalidades contribuem para a mensagem de maneira interdependente, fornecendo informações distintas que se entrelaçam para formar um significado coeso. Diferentemente da ilustração, em que a imagem pode ser substituída sem comprometer o conteúdo textual, e do ancoramento, no qual o texto fixa a interpretação da imagem, a complementaridade pressupõe que nenhum dos elementos – imagem ou texto – é plenamente compreensível sem o outro. No jornalismo, essa relação é frequentemente observada em

manchetes que introduzem um enunciado cujo sentido completo só se realiza em conjunto com a imagem associada.

Essa interação entre texto e imagem no jornalismo revela um nível de complexidade que ultrapassa a mera relação de apoio entre as modalidades. As escolhas e decisões de pareamento, no contexto jornalístico, operam sob uma lógica comunicativa caracterizada por relações semânticas mais ricas e complexas. Segundo Otto *et al*. (2019), estas relações podem ser analisadas a partir de três dimensões: informação mútua intermodal, correlação semântica e *status* hierárquico. A informação mútua intermodal (*Cross-modal mutual information*) mede o grau de sobreposição entre os conceitos representados na imagem e no texto: quanto maior a informação mútua intermodal, mais diretamente o texto descreve o conteúdo visual. A correlação semântica (*Semantic correlation*) avalia a coerência entre as informações das duas modalidades, podendo variar entre relações altamente coesas e casos de contradição. Já o *status* indica a relação de hierarquia e dependência entre imagem e texto, diferenciando, por exemplo, casos em que o texto ancora o significado da imagem, em que a imagem ilustra o texto ou em que ambos contribuem de maneira equivalente para a construção do significado.

No caso das descrições conceituais há uma alta sobreposição entre as informações visuais e textuais, ou seja, uma elevada informação mútua intermodal, com um foco na correspondência direta entre entidades e eventos representados. Em contraste, no jornalismo, os títulos frequentemente apresentam uma baixa informação mútua intermodal, pois não necessariamente descrevem o conteúdo da imagem de maneira exata, mas uma alta correlação semântica, uma vez que ambos os modos comunicativos colaboram para construir uma narrativa coesa. Além disso, enquanto as descrições conceituais tendem a ser subordinadas à imagem – como no ancoramento –, os títulos jornalísticos podem assumir diferentes *status*: podem ancorar a interpretação da imagem, funcionar como uma ilustração textual do evento representado ou até mesmo estabelecer relações interdependentes, nas quais o significado só é plenamente compreendido na articulação entre as duas modalidades. Essas diferenças evidenciam que, no jornalismo, as relações texto-imagem não são meramente descritivas, mas orientadas por intencionalidades discursivas, enquadramentos editoriais e dinâmicas interpretativas que vão além da mera correspondência referencial entre texto e imagem.

Para explorar essas relações, optamos por constituir um novo *corpus* composto por pareamentos de imagem e texto extraídos do portal *online* do jornal Tribuna de Minas.[5] Cada entrada no corpus contém uma fotografia, um marcador de editoria, que situa a matéria dentro de uma seção específica do jornal – no exemplo da Figura 4, a editoria "Boa Viagem", destinada a conteúdos relacionados a turismo –, um título, que sintetiza a ideia principal da reportagem, um bigode,[6] que complementa e expande a informação fornecida pelo título, e a data de publicação. O pareamento entre esses elementos neste novo corpus reflete um vínculo discursivo direcionado ao leitor, estruturado para orientar a interpretação da imagem

[5] Disponível em: https://tribunademinas.com.br/. Acesso em: 24 abr. 2025.

[6] No contexto jornalístico, o 'bigode' é um elemento textual complementar ao título da matéria, geralmente posicionado logo abaixo dele. Sua função é fornecer uma breve contextualização ou um resumo da reportagem, antecipando ao leitor o tema central do conteúdo sem necessariamente repetir o que está no título. O bigode pode acrescentar informações essenciais, destacar um aspecto relevante da notícia ou reforçar o apelo da matéria para engajamento do público.

dentro de um contexto noticioso e editorial, muitas vezes mobilizando estratégias de ancoramento e complementaridade semântica.

Figura 4: Exemplo de pareamento imagem-texto no jornalismo digital, extraído da Tribuna de Minas.

Fonte: Tribuna de Minas (2025).

A extração dos dados deste novo *corpus* será realizada utilizando modelos baseados em inteligência artificial, combinando algoritmos de visão computacional para análise das imagens e métodos de processamento de linguagem natural para a estruturação dos textos. Para o processamento visual das imagens serão utilizados modelos de visão computacional como, por exemplo, o modelo de visão computacional da OpenAI, GPT-4o Vision (OpenAI, 2024), capazes de extrair das imagens as entidades e as relações que estas estabelecem entre si, categorizando elementos como pessoas, objetos, lugares e eventos, e estabelecendo conexões entre eles. Essas descrições detalhadas de cada cena incluem a disposição espacial dos elementos e seu contexto geral, além de um resumo do evento representado na imagem, destacando ações e interações relevantes. As descrições nos possibilitam aplicar técnicas de detecção de objetos a partir de um conjunto aberto de classes – como o GroundingDINO (Liu *et al*., 2024) – para geração automática de *bounding boxes* que permitem o mapeamento preciso das entidades visuais. Posteriormente, essas entidades poderão ser associadas a frames e elementos de *frame* na base de dados da FrameNet Brasil, replicando a metodologia utilizada no Framed Multi30K e estabelecendo correspondências entre as estruturas semânticas da FrameNet e as representações visuais do conteúdo jornalístico. Para atribuição automática de rótulos semânticos baseados em semântica de frames aos títulos e aos bigodes utilizaremos uma versão do LOME – *Large Ontology Multilingual Extraction* (Xia *et al*., 2021) – treinada em dados da FrameNet, permitindo mapear as entidades mencionadas no texto aos *frames* semânticos que estas evocam.

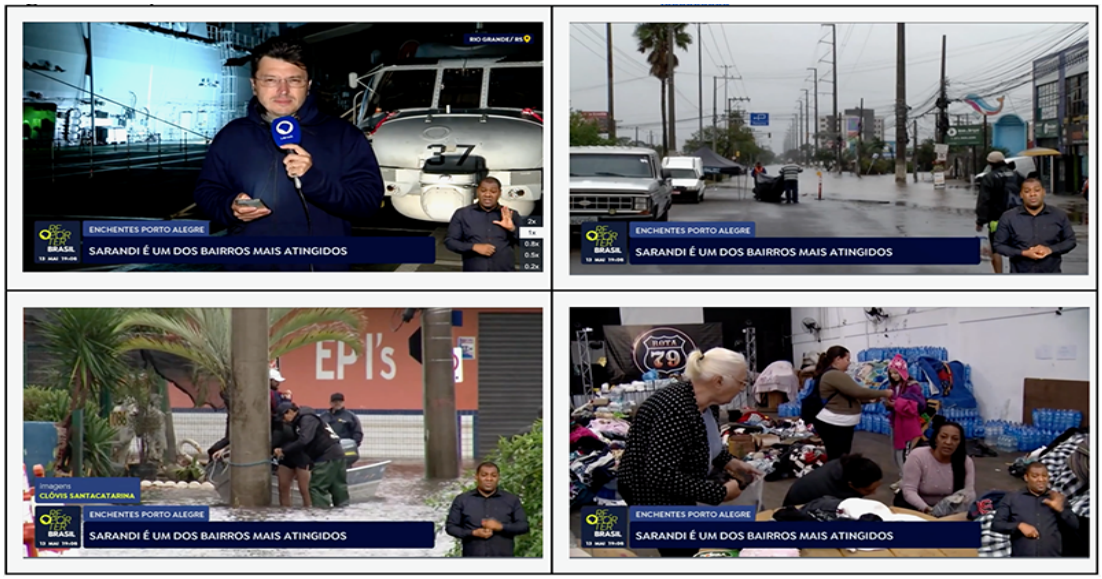
Todo o processo de extração automática de dados visuais e rotulação para *frames* semânticos será, posteriormente, validado por especialistas humanos. Essa validação é fundamental não apenas para garantir a precisão e coerência das relações extraídas, mas também para mitigar os riscos inerentes à curadoria automatizada de dados que, sem supervisão crítica, pode perpetuar vieses, reforçar desigualdades estruturais e distorcer representações socioculturais (Prabhu & Birhane, 2021).

3.2 Imagem e texto em telejornalismo

Para explorar as possibilidades de combinação entre imagem e som em produtos audiovisuais jornalísticos, optamos por construir um novo corpus para compor um novo *dataset* a partir de um telejornal. Dos formatos noticiosos que compõem um telejornal, nos interessa particularmente aquele conhecido como matéria telejornalística, reportagem ou VT.[7] Angelo (2014) apresenta um detalhado escrutínio sobre o que caracteriza uma matéria televisiva como um gênero textual telejornalístico. Em linhas gerais, é possível defini-la como qualquer segmento de um telejornal que apresenta uma notícia ou desenvolve um tema por meio de material pré-gravado, caracterizado na maioria das vezes por um texto narrado em *off*, sincronizado com imagens ilustrativas do que é falado, frequentemente intercalado com pequenos trechos de entrevistas – chamadas de sonoras, especialmente quando só se vê trecho de resposta do entrevistado – e com eventuais aparições de um repórter dirigindo-se à audiência, olhando para a câmera, enquanto fala parte da informação que compõe o todo da matéria.

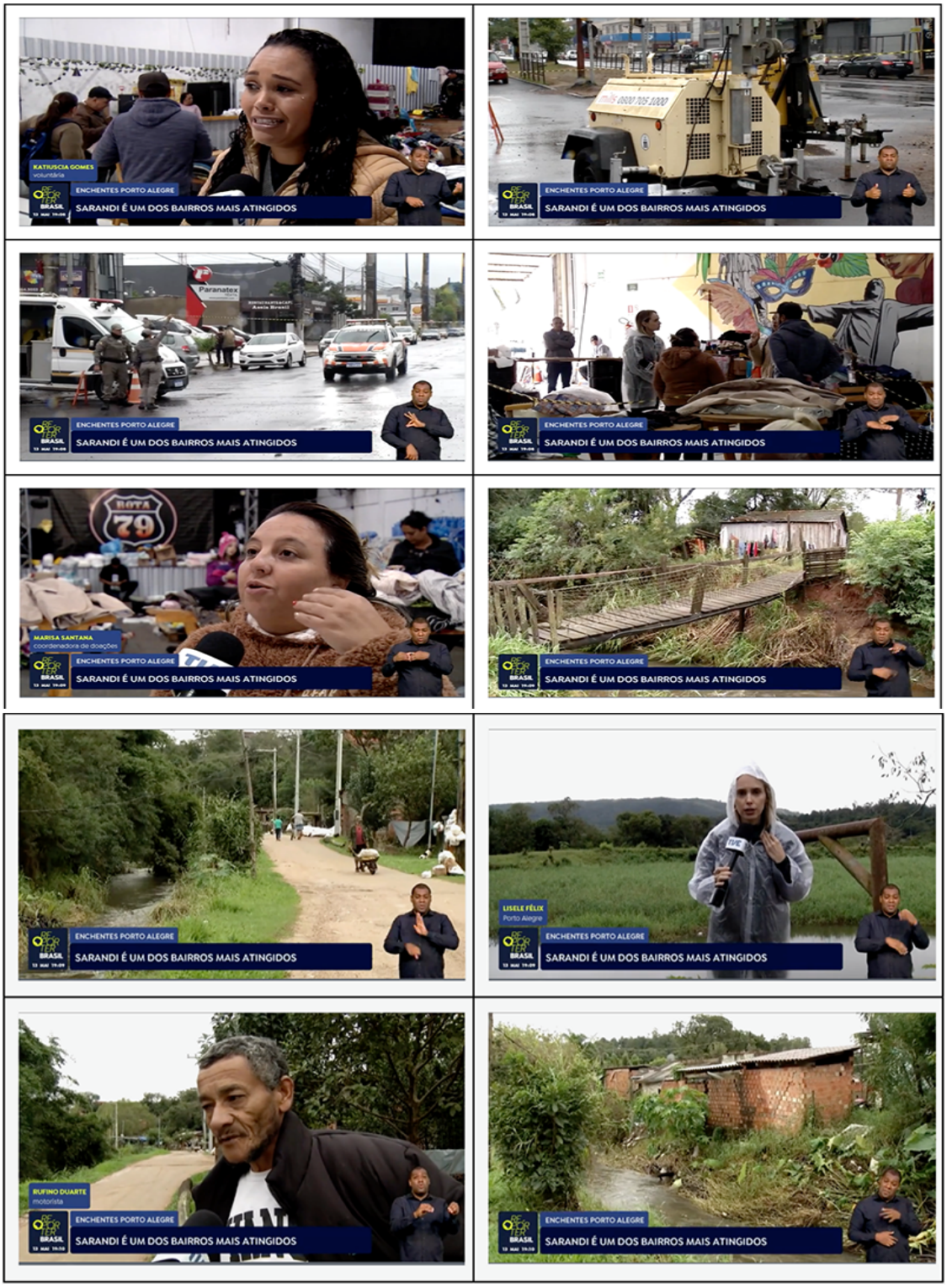
A Figura 5 apresenta uma sequência visual que representa uma matéria televisiva[8] prototípica. A única exceção à prototipia é o fato de o apresentador do telejornal não estar em estúdio, mas em externa – primeiro quadro. A matéria em questão trata da enchente que atingiu o bairro Sarandi, em Porto Alegre, no Rio Grande do Sul. Os quadros dois, três e quatro; seis, sete e oito; dez, onze e catorze (da esquerda para a direita, de cima para baixo) mostram imagens exibidas enquanto há narração em *off* da repórter. Os quadros cinco, nove e treze apresentam falas de entrevistados – sonoras. O quadro doze apresenta a passagem da repórter.

Figura 5: Exemplo de matéria televisiva – Chuvas no RS: subida do Guaíba deixa bairros de POA em alerta



[7] O termo "VT" vem de "vídeo-tape", que era usado no jornalismo quando os segmentos de notícias eram gravados em fitas magnéticas de vídeo. Originalmente, esses segmentos eram pré-gravados para serem transmitidos posteriormente, por isso ficaram conhecidos como VTs. Embora as fitas de vídeo não estejam mais em uso, o termo se manteve, especialmente no jornalismo brasileiro.

[8] A matéria está disponível em: https://tvbrasil.ebc.com.br/reporter-brasil/2024/05/chuvas-no-rs-subida-do-guaiba-deixa-bairros-de-poa-em-alerta. Acesso em: 24 abr. 2025.



Fonte: TV Brasil (2024).

Essa descrição poderia representar boa parte dos segmentos presentes em *Pedro pelo mundo*, no *dataset* Frame[2]. Porém, as matérias telejornalísticas, em geral, têm duração mais curta, apresentam mais narração em *off*, sonoras curtas e poucas imagens desacompanhadas de narração. As matérias telejornalísticas representam um gênero de altíssima circula-

ção, dado que fazem parte do cotidiano televisivo diariamente, em vários horários diferentes. Além disso povoam, atualmente, o ambiente *web*, não apenas como parte integrante de telejornais, mas também como vídeos avulsos publicados em *websites* e plataformas de informação ou em redes sociais. Dessa forma, a combinação entre imagem e áudio falado presente nas matérias emerge como um espécime multimodal altamente representativo dos processos de comunicação humana contemporâneos e, como tal, rico material para investigação e composição de um dataset para uso em um modelo computacional.

A Figura 6 mostra uma sequência de imagens acompanhada do texto narrado em off que compõe trecho da matéria já citada:

Figura 6: Exemplo de pareamento imagem texto em matéria televisiva

		
Na avenida Assis Brasil, na altura do número sete mil, a água voltou a subir.	Em uma semana, foram muitos resgates nessa área.	Agora, o voluntariado está organizando doações em um espaço cedido por uma casa de shows.

Fonte: TV Brasil (2024).

Considerando o modo produtivo do telejornalismo, segundo o qual para se construir uma matéria grava-se material bruto, depois um repórter escreve um texto, grava o texto lido e entrega para um editor de vídeo montar a matéria, entendemos que os exemplos retratados na Figura 5 são casos em que a imagem ilustra o texto. Dessa forma, a metodologia usada para a anotação do *corpus* Pedro Pelo Mundo e do *dataset* Frame[2] – segundo a qual a anotação começa pelo texto – é perfeitamente adequada para a nova tarefa.

Para compor um corpus telejornalístico buscamos um telejornal de abrangência nacional que disponibilizasse seu material veiculado de forma simples para que fosse capturado. Encontramos isso no telejornal Repórter Brasil, veiculado pela emissora pública nacional de televisão TV Brasil, mantida pela EBC, Empresa Brasileira de Comunicação. Por orientação da Central de Pesquisas da EBC fizemos o download de dez edições do telejornal diretamente de seu website. Trata-se de um total de 178 vídeos, veiculados em 10 dias diferentes no período entre 13 de maio de 2024 e 16 de julho de 2024. Há duas edições veiculadas em segundas-feiras, duas em terças, duas em quartas, duas em quintas e duas em sextas. Os dias não são subsequentes, configurando, assim, uma semana construída – constructed week samples (Riffe, Aust e Lacy, 1993). O tempo total dos vídeos é de 394 minutos e 50 segundos. A duração média dos vídeos é de 2 minutos e 13 segundos. O mais longo dura 6 minutos e 11 segundos, e o mais curto dura 31 segundos.

Todo esse conteúdo, naturalmente, não se refere a matérias prototípicas. Ainda não dimensionamos qual porcentagem ou qual a minutagem total representa matérias televisivas. Mas sabemos que há muito tempo dedicado a cabeças – a abertura feita pelos apresen-

tadores do telejornal antes da exibição de uma matéria; muito tempo dedicado a notas secas – notícias enunciadas pelos apresentadores em estúdio sem recorrer a imagens ilustrativas, ou seja, tendo os apresentadores como cabeças falantes;[9] há um pouco de entrevistas ao vivo, ou seja, não editadas, em que o que se vê são pessoas conversando num formato que preserva perguntas e respostas na íntegra e o que se vê são apenas as pessoas falando e ouvindo, sem ilustrações externas à conversa; e há ainda os vídeos de previsão do tempo, que possuem uma dinâmica peculiar de interação do que se fala com mapas e esquemas visuais ou infográficos.

Essas variações suscitam a possibilidade de se aproveitar o material para outras anotações, como a anotação dos gestos dos apresentadores, por exemplo, ou mesmo o desenvolvimento de uma nova metodologia para anotar previsão do tempo em telejornal.

Na seção seguinte discutimos desafios, possibilidades e ganhos com a construção de novos *datasets e a expansão do modelo.*

4 Ganhos e Desafios da Expansão do Modelo

Em relação ao novo *corpus* de imagens e textos extraídos de artigos jornalísticos, a incorporação de conteúdos jornalísticos ao *corpus* de imagens estáticas da ReINVenTA representa um avanço significativo para a anotação baseada na Semântica de *Frames*, especialmente no que tange à relação entre elementos visuais e estruturas semânticas da FrameNet. Diferentemente dos *datasets* tradicionalmente utilizados em tarefas de Processamento de Língua Natural, construídos com metodologias que privilegiam a neutralidade e objetividade e frequentemente apresentam enunciados descontextualizados e sem intencionalidade discursiva clara, os textos jornalísticos estabelecem vínculos semânticos mais ricos e diversos entre imagem e linguagem, possibilitando não só o enriquecimento da rede de relações já presentes nos dados da ReINVenTA, mas também a exploração das especificidades de um novo campo discursivo, permitindo a criação de novas relações na base de dados e aprimorando a granularidade da FrameNet através da incorporação de elementos de um domínio ainda pouco coberto pela rede.

No que tange ao *corpus* e ao *dataset* de vídeo cabe destacar que a metodologia proposta em Belcavello (2023) e adotada até então, apesar de criteriosa e bem fundamentada, é bastante custosa. O tempo dedicado à anotação manual tanto de texto quanto de imagens é extenso, e pode acarretar cansaço dos anotadores e, por consequência, eventual queda da qualidade das anotações. A estruturação do *corpus* Repórter Brasil, no entanto, traz uma vantagem para evitar essas questões. Como cada uma das dez edições do jornal está subdividida em pelo menos 13 vídeos – chegando até 23 – a metodologia de se anotar vídeo a vídeo pode ser aplicada – todo o texto do vídeo primeiro e, em seguida, as imagens – ao invés de edição a edição. Isso pode otimizar a anotação passo a passo, diminuir o risco de exaustão, acelerar o processo e, até mesmo, assegurar ainda mais a qualidade da anotação.

Esses novos *datasets* também oferecem ganhos em termos da solidez do gênero. Matérias jornalísticas representam um pareamento de forma e sentido razoavelmente está-

[9] O termo *talking head* (cabeça falante) é geralmente usado para descrever personalidades da TV – especialmente âncoras de notícias ou especialistas – cuja imagem na tela se limita a um enquadramento fechado, exibindo apenas a cabeça e os ombros do falante.

vel e muito profuso na sociedade contemporânea. Ainda que tenham ocorrido evoluções e atualizações ao longo do tempo, as formas utilizadas pelo jornalismo para contar notícias, debater temas, ou veicular informações se mantêm razoavelmente sedimentadas. No caso das matérias telejornalísticas, mantém-se a estrutura narrativa de um repórter que narra um texto em *off*, entrecortado por sonoras e passagem, há cerca de cinquenta anos. Expandir o modelo a partir desse formato pavimenta o caminho para outras expansões futuras que podem se valer de *corpora* semelhantes.

Dentre os principais desafios para a composição do *dataset* de imagens estáticas destaca-se a recorrência de textos contendo imagens que cumprem apenas o papel de ilustração editorial, sem desempenhar papel significativo na construção do significado, e servindo apenas para preencher o espaço visual e atrair a atenção do leitor. Um exemplo típico desse fenômeno ocorre em textos que fazem referência a uma personalidade conhecida, cujo nome aparece como principal destaque no título que a acompanha (Figura 7). Em casos como esse, a imagem não adiciona novas camadas interpretativas ao texto, nem estabelece relações semânticas para além da identificação da entidade que foi nomeada no título. Pareamentos desse tipo oferecem poucas possibilidades de anotação na base de dados da FrameNet, limitando a utilidade dessas instâncias dentro do *corpus* – o que reforça a necessidade de curadoria dos dados para selecionar apenas pares imagem e texto que efetivamente contribuam para a modelagem de relações multimodais relevantes.

Figura 7: Exemplo em relação multimodal pouco relevante.

Nara Vidal lança novo romance, 'Eva'

Obra mergulha no universo dos abusos em relacionamentos com mulheres e discute sentimento de posse no amor.

Fonte: Tribuna de Minas (2022b).

Cabe também ressaltar os desafios relativos às restrições impostas pela Lei do Direito Autoral (Lei nº 9.610/1998), que regula o uso de imagens e estabelece os direitos dos fotógrafos sobre suas obras. Segundo a legislação, a reprodução e redistribuição de fotografias exige autorização expressa do autor, salvo em casos específicos previstos em lei, como o uso jornalístico informativo sem fins comerciais. No contexto da construção do *corpus*, isso impõe limitações na coleta e disponibilização de imagens, especialmente quando não há licenciamento explícito ou quando a fotografia foi produzida por terceiros. Essas restrições exigem estratégias alternativas, como a obtenção de licenças adequadas e o desenvolvimento de metodologias de anonimização que permitam a anotação sem comprometer os direitos de imagens das pessoas retratadas.

Apesar da natureza dos direitos ser diferente, a liberação de conteúdo audiovisual também encontra desafios. A escolha por um telejornal veiculado por uma TV pública federal

para servir de material de pesquisa e nutrir um projeto de desenvolvimento sediado em uma universidade federal não foi casual. Esse foi um caminho trilhado com o intuito de facilitar o acesso e a liberação do material. Ainda assim, pudemos perceber que o sistema de gerenciamento do acervo adotado pela EBC não estava preparado para fornecer o volume de material que solicitamos. Dessa forma, fomos orientados a nos engajarmos nós mesmos na extração do material publicado no *website* da TV Brasil, diferente do que aconteceu no processo do Frame[2], quando o canal GNT forneceu os vídeos após solicitação.

5 Conclusão

Os projetos relatados neste artigo referem-se à ampliação da composição do *gold standard dataset* da ReINVenTA para incluir gêneros multimodais que circulam na esfera jornalística, nomeadamente composições de foto, legenda e manchete ou chamada de capa, bem como matérias telejornalísticas. Para além da inclusão de um outro domínio da atividade humana no *dataset*, esta proposta representa um considerável avanço na representação computacional da semântica multimodal. O desdobramento do projeto original da ReINVenTA aqui relatado não é mera extensão do *corpus* que compõe o *gold standard dataset*, mas constitui-se no descortinamento de novos níveis de análise para as combinações semânticas multimodais, as quais se relevam mais complexas nos gêneros eleitos para a análise. Assim, do ponto de vista teórico, os novos *datasets* avançam no sentido de levar as análises multimodais desenvolvidas com base na Semântica de *Frames* pela ReINVenTA para a esfera jornalística, contribuindo para a compreensão, dentro de um modelo semântico refinado e estruturado computacionalmente, de como o aparato multimodal dos gêneros dessa esfera é mobilizado para a criação de efeitos de sentido. Do ponto de vista tecnológico e de inovação, reforçam um aparato de *datasets* ainda mais robusto, na medida em que incorporam gêneros da esfera jornalística que aumentam a complexidade das relações semânticas entre os modos comunicativos analisados. Como consequência, têm o potencial de melhorar o desempenho do algoritmo de rotulação semântica multimodal desenvolvido até o momento, contribuindo para acelerar e melhorar outras tarefas vindouras.

Agradecimentos

A pesquisa apresentada neste artigo foi desenvolvida pela ReINVenTA – Rede de Pesquisa e Inovação em Visão e Análise de Texto de Objetos Multimodais. A ReINVenTA é fomentada pela FAPEMIG, por meio do financiamento RED 00106/21, e pelo CNPq, por meio dos financiamentos 408269/2021-9 e 420945/2022-9. A pesquisa de Belcavello foi financiada pela programa de doutorado sanduíche no exterior (PDSE) CAPES – processo 88881.362052/2019-01 e pela bolsa de pós-doutorado no exterior CNPq PDE Chamada 26/2021 – processo 200270/2023-0. A pesquisa de Marcelo Viridiano foi financiada pelo programa de doutorado CAPES PROEX – processo 88887.816219/2023-00, pela bolsa de doutorado sanduíche no exterior CAPES PROBRAL – processo 88887.628830/2021-00, e pela bolsa de pós-doutorado no exterior CNPq PDE Chamada 26/2021 – processo 200270/2023-0.

Referências

ANGELO, M. H. *Gêneros textuais e telejornalismo*: caminhos da produção escrita de matérias televisivas. 2014. 286 p. Tese (Doutorado em Linguística) – Faculdade de Letras, Universidade Federal de Juiz de Fora, 2014.

BARTHES, Roland. Rhetoric of the Image. *In*: BARTHES, Roland (ed.) *Image-Music-text*. London: Fontana, 1977[1964]. p. 33-51.

BELCAVELLO, Frederico; VIRIDIANO, Marcelo; COSTA, Alexandre Diniz da; MATOS, Ely E. S.; TORRENT, Tiago T. Frame-Based Annotation of Multimodal Corpora: Tracking (A) Synchronies in Meaning Construction. *In*: LANGUAGE RESOURCES AND EVALUATION CONFERENCE (LREC 2020). *Proceedings of the International FrameNet Workshop 2020*: Towards a Global, Multilingual FrameNet. Marseille: ELRA, 2020. p. 23-30**.**

BELCAVELLO, Frederico. *FrameNet Annotation for Multimodal Corpora:* Devising a Methodology for the Semantic Representation of Text-image Interactions in Audiovisual Productions. 2023. 134 p. Tese (Doutorado em Linguística) – Faculdade de Letras, Programa de Pós-graduação em Linguística, Universidade Federal de Juiz de Fora, Juiz de Fora, 2023.

BELCAVELLO, Frederico *et al*. Frame^2: A FrameNet-based Multimodal Dataset for Tackling Text-image Interactions in Video. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC). *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024). Torino: ELRA and ICCL, 2024. p. 7429-7437. Disponível em: https://aclanthology.org/2024.lrec-main.655/. Acesso em: 22 abr. 2025.

COHN, N., MAGLIANO, J. 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.

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

FERNANDEZ, Leohoho. *Green and yellow scissors on white graphing paper*. 2021. Fotografia. Disponível em: https://unsplash.com/photos/green-and-yellow-scissors-on-white-graphing-paper-J_galDuu4kc. Acesso em: 22 abr. 2025.

FILLMORE, C. J. Frame semantics. *In*: THE LINGUISTIC SOCIETY OF KOREA. *Linguistics in the Morning Calm*. Seoul: Hanshin, 1982. p. 111-137.

FILLMORE C. J., PETRUCK, M. R., RUPPENHOFER, J., & WRIGHT, A. FrameNet in Action: The case of attaching. *International journal of lexicography*, 16 (3), 297-332. 2003.

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

HODOSH, M., YOUNG, P., & HOCKENMAIER, J. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47, 2013, p. 853-899.

LEONARD, Cathryn. *Person holding pencil writing on notebook*. 2021. Fotografia. Disponível em: https://unsplash.com/photos/person-holding-pencil-writing-on-notebook-RdmLSJR-tq8. Acesso em: 22 abr. 2025.

LIU, S., ZENG, Z., REN, T., LI, F., ZHANG, H., YANG, J., ZHANG, L. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *In*: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024. p. 38-55.

MARTINEC, R.; SALWAY, A. A System for Image–text Relations in New (and old) Media. *Visual communication*, v. 4, n. 3, p. 337-371, 2005. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/1470357205055928. Acesso em: 22 abr. 2025.

MATTHIESSEN, C. *Introduction to functional grammar.* London: Hodder Arnold, 1989.

MØLLER, A. G., PERA, A., DALSGAARD, J., & AIELLO, L. The Parrot Dilemma: Human-labeled vs. llm-augmented data in Classification Tasks. *In*: Graham, Y.; Purver, P. (ed.) *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* (Volume 2: Short Papers), 2024. p. 179-192. Disponível em: https://aclanthology.org/volumes/2024.eacl-long/. Acesso em: 22 abr. 2025.

OPENAI. ChatGPT-4o: Multimodal AI Model, 2024. [Online]. Disponível em: https://openai.com. Acesso em: 22 abr. 2025.

OTTO, Christian; SPRINGSTEIN, Matthias; ANAND, Avishek; EWERTH, Ralph. Understanding, categorizing and predicting semantic image-text relations. *In*: INTERNATIONAL CONFERENCE ON MULTIMEDIA RETRIEVAL, 2019, New York. *Proceedings* [...]. New York: Association for Computing Machinery, 2019. p. 168-176. Disponível em: https://doi.org/10.1145/3323873.3325049. Acesso em: 23 abr. 2025.

PRABHU, V. U., & BIRHANE, A. Large datasets: A pyrrhic win for computer vision. *In*: *Institute of Electrical and Electronics Engineers/Computer Vision Foundation Conference on Applications of Computer Vision*. 2021.

PLUMMER B. A., WANG, L., CERVANTES, C. M., CAICEDO, J. C., HOCKENMAIER, J., & LAZEBNIK, S. Flickr30k entities: Collecting Region-to-phrase Correspondences for Richer Image-to-sentence models. *In*: 2015 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION. *Proceedings of the IEEE international conference on computer vision*, Chile, 2015. p. 2641-2649.

RICCIARDI, Dean. *Pink blue and green pens*. 2021. Fotografia. Disponível em: https://unsplash.com/photos/pink-blue-and-green-pens-uWh-hYisqAw. Acesso em: 22 abr. 2025

RIFFE, D.; AUST, C. F.; LACY, S. R. The Effectiveness of Random, Consecutive Day and Constructed Week Sampling in Newspaper Content Analyses. *Journalism Quarterly*, v. 70, n. 1, p. 133-139, spring, 1993.

ROGERS, A. Changing the World by Changing the Data. *arXiv preprint arXiv:2105.13947*. 2021. DOI: https://doi.org/10.48550/arXiv.2105.13947.

SANABRIA, Ramon *et al*. How2: a large-scale dataset for multimodal language understanding. Cornell University, 2018.doi: https://doi.org/10.48550/arXiv.1811.00347.

TORRENT, T.; MATOS, E. E. da S.; BELCAVELLO, F.; VIRIDIANO, M.; GAMONAL, M. A.; COSTA, A. D. da; MARIM, M. C. Representing Context in FrameNet: A Multidimensional, Multimodal Approach. *Frontiers in Psychology*, v. 13, *2022.* Disponível em: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.838441. Acesso em: 22 abr. 2025. DOI: 10.3389/fpsyg.2022.838441. ISSN 1664-1078.

SALLES, Renato. Em Contagem, Lula discursa sobre questões econômicas e condições financeiras dos brasileiros. *Tribuna De Minas,* Juiz de Fora, 10 maio 2022a. Disponível em: https://tribunademinas.com.br/noticias/politica/eleicoes-2022/10-05-2022/em-contagem-lula-discursa-sobre-questoes-economicas-e-condicoes-financeiras-dos-brasileiros.html. Acesso em: 23 abr. 2025.

BOA Viagem. *Tribuna De Minas,* Juiz de Fora, 11 fevereiro 2025. Disponível em: https://tribunademinas.com.br/especiais/boa-viagem. Acesso em: 23 abr. 2025.

MAZOCOLI, Elisabetta. Nara Vidal lança novo romance "Eva". *Tribuna De Minas,* Juiz de Fora, 7 abr. 2022b. Disponível em: https://tribunademinas.com.br/noticias/cultura/07-04-2022/nara-vidal-lanca-novo-romance-eva.html. Acesso em: 23 abr. 2025.

CHUVAS no RS: subida do Guaíba deixa bairros de POA em alerta. *Tv Brasil*. Brasília, 6 maio 2024. Disponível em: https://tvbrasil.ebc.com.br/reporter-brasil/2024/05/chuvas-no-rs-subida-do-guaiba-deixa-bairros-de-poa-em-alerta. Acesso em: 23 abr. 2025.

UPPAL, S., BHAGAT, S., HAZARIKA, D., MAJUMDER, N., PORIA, S., ZIMMERMANN, R., & ZADEH, A. Multimodal research in vision and language: A Review of Current and Emerging Trends. *Information Fusion*, v. 77, p. 149-171, 2022.

VAN MILTENBURG, E. Stereotyping and bias in the flickr30k dataset. *arXiv preprint arXiv:1605.06083*, 2016. DOI: https://doi.org/10.48550/arXiv.1605.06083.

VIRIDIANO, M., LORENZI, A., TORRENT, T. T., MATOS, E. E., PAGANO, A. S., SIGILIANO, N. S., de FREITAS, M. H. P. Framed Multi30K: A Frame-Based Multimodal-Multilingual Dataset. *In:*THE 2024 JOINT INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, LANGUAGE RESOURCES AND EVALUATION (LREC-COLING 2024). *Proceedings* [...]. [*S. l.*] 2024. p. 7438-7449.

XIA, P., QIN, G., VASHISHTA, S., CHEN, Y., CHEN, T., MAY, C., HARMAN. C., RAWLINS, K., WHITE, A. S., VAN DURME, B. *LOME*: Large Ontology Multilingual Extraction. arXiv preprint arXiv:2101.12175. 2021. DOI: https://doi.org/10.18653/v1/2021.eacl-demos.19.

YOUNG, P., LAI, A.; HODOSH, M.; HOCKENMAIER, J. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference Over Event Descriptions. *Transactions of the Association for Computational Linguistics*, v. 2, 67-78, 2014.