

e-ISSN 2238-3824

**Faculdade de Letras**  
**Universidade Federal de Minas Gerais**

# calígrama

**revista de estudos românicos**

**v. 30, n. 1**  
**janeiro – abril 2025**

# **caligrama**

**revista de estudos românicos**

**Universidade Federal de Minas Gerais**

Reitora: Sandra Regina Goulart Almeida; Vice-Reitor: Alessandro Fernandes Moreira

**Faculdade de Letras**

Diretora: Sueli Maria Coelho; Vice-Diretor: Georg Otte

**Editora-chefe**

Aléxia Teles Guimarães, Universidade Federal de Minas Gerais, Brasil

**Comissão editorial**

André Vinícius Lopes Coneglian, Universidade Federal de Minas Gerais, Brasil

Evandro Landulfo Teixeira Paradelo Cunha, Universidade Federal de Minas Gerais, Brasil

Giulia Bossaglia, Universidade Federal de Minas Gerais, Brasil

Laureny Aparecida Lourenço da Silva, Universidade Federal de Minas Gerais, Brasil

Lia Araujo Miranda de Lima, Universidade Federal de Minas Gerais, Brasil

**Conselho editorial interno**

Ana Maria Chiarini, Universidade Federal de Minas Gerais, Brasil

César Nardelli Cambraia, Universidade Federal de Minas Gerais, Brasil

Haydée Ribeiro Coelho, Universidade Federal de Minas Gerais, Brasil

Lúcia Castello Branco, Universidade Federal de Minas Gerais, Brasil

Lúcia Monteiro de Barros Fulgêncio, Universidade Federal de Minas Gerais, Brasil

Márcia Cristina de Brito Rumeu, Universidade Federal de Minas Gerais, Brasil

Marcos Antônio Alexandre, Universidade Federal de Minas Gerais, Brasil

Maria Antonieta A. de Mendonça Cohen, Universidade Federal de Minas Gerais, Brasil

Maria Juliana Gambogi Teixeira, Universidade Federal de Minas Gerais, Brasil

Sara Rojo, Universidade Federal de Minas Gerais, Brasil

**Conselho editorial externo**

Aldina Quintana, The Hebrew University of Jerusalem, Israel

Célia Marques Telles, Universidade Federal da Bahia, Brasil

David Bunis, The Hebrew University of Jerusalem, Israel

Geraldine Rogers, Universidad Nacional de La Plata, Argentina

João Bosco Cabral dos Santos, Universidade Federal de Uberlândia, Brasil

Leila de Aguiar Costa, Universidade Federal de São Paulo, Brasil

Leonardo Francisco Soares, Universidade Federal de Uberlândia, Brasil

Lilián Guerrero, Universidad Nacional Autónoma de México, México

Luis Urbano Afonso, Universidade de Lisboa, Portugal

Manoel Mourivaldo Santiago-Almeida, Universidade de São Paulo, Brasil

Maria Célia Pereira Lima-Hernandes, Universidade de São Paulo, Brasil

Maria Eugênia Olímpio de Oliveira, Universidade Federal da Bahia, Brasil

Maria Maura Cezario, Universidade Federal do Rio de Janeiro, Brasil

Mariangela Rios de Oliveira, Universidade Federal Fluminense, Brasil

Martine Kunz, Universidade Federal do Ceará, Brasil

Mauricio Sartori Resende, Universidade Federal de Minas Gerais, Brasil

Mirta Groppi, Universidade de São Paulo, Brasil

Márcia Paraquett, Universidade Federal da Bahia, Brasil

Pedro Dolabela Chagas, Universidade Federal do Paraná, Brasil

Raquel Meister Ko. Freitag, Universidade Federal de Sergipe, Brasil

Rita de Cássia Ribeiro de Queiroz, Universidade Estadual de Feira de Santana, Brasil

Roberto Mulinacci, Università degli Studi di Bologna, Itália

Roberto Vecchi, Università degli Studi di Bologna, Itália

Saulo Neiva, Université Blaise Pascal (Clermont II), França

Sebastião Carlos Leite Gonçalves, Universidade do Estado de São Paulo, Brasil

Sergio Romanelli, Universidade Federal de Santa Catarina, Brasil

Silvia Inês Cárcamo, Universidade Federal do Rio de Janeiro, Brasil

Thomas Hoelbeek, Vrije Universiteit Brussel, Bélgica

Walter Carlos Costa, Universidade Federal de Santa Catarina, Universidade Federal do Ceará, Brasil

**Secretaria:** Lilian Souza dos Anjos, Julia Netto

**Editor de arte:** Emerson Eller

**Projeto gráfico:** Stéphanie Paes

**Revisão:** Floriane Abreu, Izabelly Silva Duque, Kathleen Oliveira

**Diagramação:** Floriane Abreu, Kathleen Oliveira

# caligrama

revista de estudos românicos

Editores convidados

André V. Lopes Coneglian (UFMG)

Tiago Torrent (UFJF)

FALE  
FACULDADE  
DE LETRAS

UF *m* G

Caligrama | Belo Horizonte | v. 30 | n. 1 | jan.–abr. 2025 | 154 p. | e-ISSN 2238-3824





This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Os conceitos emitidos em artigos assinados são de responsabilidade exclusiva de seus autores.

Ficha catalográfica elaborada pela Biblioteca da Faculdade de Letras da UFMG

Caligrama: Revista de Estudos Românicos, v. 1, dez. 1988 – . Belo Horizonte, Faculdade de Letras da UFMG. il. ; color.; online.

Histórico

Título anterior: Estudos Românicos, 1981–1985 (n. 1–3).

Periodicidade semestral, a partir do v. 15, n. 1, jan./jun. 2010

Passou a ser online a partir de 2015.

Periodicidade quadrimestral, a partir do v. 23, n. 1, jan./abr. 2018.

ISSN: 0103-2178

e-ISSN: 2238-3824

1. Línguas românicas – Estudo e ensino – Periódicos. 2. Literatura românica – História e crítica – Periódicos. I. Universidade Federal de Minas Gerais. Faculdade de Letras.

CDD: 440.05

Faculdade de Letras da UFMG  
Seção de Periódicos, sala 2017  
Av. Pres. Antônio Carlos, 6627 – Pampulha  
31270-901 – Belo Horizonte, MG – Brasil  
Tel.: (31) 3409-6009  
[www.lettras.ufmg.br/periodicos](http://www.lettras.ufmg.br/periodicos)  
[periodicosfaleufmg@gmail.com](mailto:periodicosfaleufmg@gmail.com)

# Sumário

## Apresentação

- 8 Multimodalidade: abordagens cognitivas e representações computacionais

*Multimodality: Cognitive Approaches and Computational Representation*

Tiago Timponi Torrent; André V. Lopes Coneglian

## Dossiê: Multimodalidade: abordagens cognitivas e representações computacionais

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

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

André V. Lopes Coneglian; Adriana Pagano

- 51 Permanência semântica entre áudio original e legenda: um estudo sobre anotação semântica multimodal em obra audiovisual

*Semantic Permanence between Original Audio and Subtitles: A Study on Multimodal Semantic Annotation in Audiovisual Work*

Mariana Mota Silva Souza; Maucha Andrade Gamonal; Adriana Silvina Pagano

- 74 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; Marcelo Viridiano
- 94 A FrameNet Brasil Approach to Annotation of Pragmatic Frames Evoked by Turn Organization Gestures  
*Uma abordagem da FrameNet Brasil para anotação de frames pragmáticos evocados por gestos de organização de turno*  
Helen de Andrade Abreu; Ely Edison da Silva Matos
- 110 Multimodal Frame Semantics: Expanding the Analytical Categories of FrameNet Brasil Multimodal Datasets  
*Semântica de Frames Multimodal: expandindo as categorias de análise dos datasets multimodais da FrameNet Brasil*  
Natália Sathler Sigiliano

## Varia

- 140 Uma casa não é um lar: o afeto e o espaço físico em Madame Bovary (1856)  
*A House is Not a Home: The Affection and the Physical Space in Madame Bovary (1856)*  
Laís Marx Umpierre; Luísa Freire

# Apresentação

# Multimodalidade: abordagens cognitivas e representações computacionais

## *Multimodality: Cognitive Approaches and Computational Representation*

**Tiago Timponi Torrent**

Universidade Federal de Juiz de Fora  
(UFJF) | Juiz de Fora | MG | BR  
Conselho Nacional de Desenvolvimento  
Científico e Tecnológico (CNPq)  
tiago.torrent@ufjf.br  
<https://orcid.org/0000-0001-5373-2297>

**André V. Lopes Coneglian**

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

**Resumo:** Este artigo apresenta uma introdução ao conceito de multimodalidade, discutido sob duas perspectivas principais: a metateórica, que compreende a multimodalidade como um campo de investigação sobre a produção de significado por meio de múltiplas formas semióticas; e a fenomenológica, que a entende como a integração de diferentes modalidades expressivas (fala, gesto, imagem, entre outras) em práticas comunicativas. A partir dessa base conceitual, o texto destaca a ausência histórica de atenção à multimodalidade nos campos da Linguística e da Ciência da Computação, refletida em modelos teóricos e computacionais que privilegiam formas linguísticas isoladas e convencionalizadas. Frente a esses desafios, apresentam-se os projetos no desenvolvidos no âmbito da ReINVenTA, uma rede de pesquisa dedicada à construção e anotação de datasets multimodais com base na Semântica de Frames, visando integrar linguística cognitiva e modelos computacionais. A conclusão aponta para a necessidade de abordagens interdisciplinares que reconheçam a linguagem como um fenômeno social, interacional e intrinsecamente multimodal.

**Palavras-chave:** Multimodalidade; visão computacional; datasets anotados; semântica de frames.

**Abstract:** This article introduces the concept of multimodality, discussed from two main perspectives: the metatheoretical, which understands multimodality as a field of inquiry into the production of meaning through multiple semiotic forms; and the phenomenological, which views it as the integration of different expressive modalities (speech, gesture, image, among others) in



communicative practices. Based on this conceptual foundation, the text highlights the historical lack of attention to multimodality in the fields of Linguistics and Computer Science, reflected in theoretical and computational models that favor isolated and conventionalized linguistic forms. In response to these challenges, the article presents projects developed within the scope of ReINVenTA, a research network dedicated to the construction and annotation of multimodal datasets based on Frame Semantics, aiming to integrate cognitive linguistics and computational models. The conclusion emphasizes the need for interdisciplinary approaches that recognize language as a social, interactional, and inherently multimodal phenomenon.

**Keywords:** Multimodality; computer vision; annotated datasets; frame semantics.

## 1 Introduzindo a noção de “multimodalidade”

Quem quer que se proponha a falar de “multimodalidade” precisa especificar de que exatamente é que se fala. Assim como inúmeros termos linguísticos, como “gramática”, “sentença”, “semântica” e “pragmática”, o termo “multimodalidade” pode ter diferentes acepções.

Em um sentido metateórico, “multimodalidade” pode fazer referência a uma orientação teórica, a uma metodologia de investigação, ou, ainda, a todo um campo de investigação (cf. Jewitt, 2014). Independentemente do recorte que se faça pelo uso do termo, é indiscutível o fato de que, qualquer pessoa que se interessa por multimodalidade está interessada nos meios (multi)semióticos pelos quais indivíduos constroem significado na vivência da linguagem. Assim, podem-se invocar Bateman, Wildfeuer e Hiippala (2017, p. 8-9, tradução nossa) que caracterizam multimodalidade como uma “orientação de pesquisa [...] que busca dar conta do que acontece quando diversas formas de comunicação se combinam com a finalidade de ‘produzir significados’ – independentemente de como ou onde isso seja feito”.<sup>1</sup>

Por outro lado, em um sentido descritivo fenomenológico, o termo pode abarcar a combinação de diferentes sistemas expressivos, como fala, gesto e imagem, entre outros, em um todo comunicativo. Nesse caso, convergem diferentes recursos como sons, gestos, olhar, representações ortográficas e imagéticas. Como delimitar, então, o fenômeno **multimodalidade**?

Considerando um conjunto de práticas comunicativas e acionais que se caracterizam como multimodais, tais como um programa jornalístico televisivo em que se misturam textos escritos, falados e imagens, conversa face a face acompanhada de gestos, olhares, expressões

---

<sup>1</sup> Texto original: “... a research orientation [...] that seeks to address what happens when diverse communicative forms combine in the service of ‘making meanings’ – however and wherever, this is done”.

faciais. Bateman, Wildfeuer e Hiippala (2017) apresentam uma caracterização de multimodalidade de acordo com sua natureza fenomenológica.

A multimodalidade é uma forma de caracterizar situações comunicativas (entendidas de maneira ampla) que dependem da combinação de diferentes “formas” de comunicação para serem eficazes – o programa de televisão utiliza linguagem falada, imagens e textos; o livro recorre à linguagem escrita, imagens, diagramas, composição da página, entre outros elementos; a conversa na cafeteria integra a linguagem oral com uma variedade de posturas e capacidades corporais; e o jogo eletrônico pode apresentar representações de qualquer uma dessas modalidades, além de incluir movimentos e ações (Bateman, Wildfeuer e Hiippala, 2017, p. 8, tradução nossa).<sup>2</sup>

O que os autores chamam de “formas de comunicação” pode ser entendido como “modalidade”, ou seja, como um sistema expressivo (cf. Cohn e Schilperoord, 2024). Assim, se uma prática comunicativa ou uma ação se configura pela mobilização de diferentes “formas”, então essa é uma prática multimodal.

Construindo uma interface entre semiótica (neo-peirciana) e análise da conversa, Engle (1998) defende que “sinais multimodais” (*multimodal signals*) são produzidos e compreendidos como sendo unidades integradas de comunicação, o que ela chama de “signos compostos” (*composite signals*). A consequência de se tratar artefatos multimodais como signos compostos é o fato de que as modalidades (ou os “canais”<sup>3</sup>, termo que usa a autora<sup>4</sup>) não podem ser interpretadas isoladamente, mas, sim, como um todo integrado que produz significado.

Nessa mesma linha, com a atenção voltada especificamente para a interação face a face e a articulação entre fala e gesto, diz Enfield (2009) que os movimentos que interactantes fazem no ato de comunicação não são semioticamente simples, mas são compostos (*composite*). Sem usar o termo “multimodalidade”, o autor apresenta o mesmo tipo de fenômeno que Bateman, Wildfeuer e Hiippala (2017) apresentam.

A natureza composta desses elementos é amplamente variada em sua forma: uma palavra combinada com outras palavras, uma sequência de palavras associada a um contorno entoacional, um diagrama acompanhado de uma legenda, um ícone articulado com outro ícone, uma enunciação oral acompanhada de um gesto manual (Enfield, 2009, p. 1).<sup>5</sup>

<sup>2</sup> Texto original: “Multimodality is a way of characterising communicative situations (considered very broadly) which rely upon combinations of different ‘forms’ of communication to be effective – the TV programme uses spoken language, pictures and texts; the book uses written language, pictures, diagrams, page composition and so on; talking in the cafeteria brings together spoken language with a host of bodily capabilities and postures; and the computer game might show representations of any of these things and include movement and actions as well”.

<sup>3</sup> Em inglês, *channels*.

<sup>4</sup> Devido a limitações de espaço e de escopo, não se pode explicitar, neste artigo, uma possível inter-relação entre a noção semiótica de multimodalidade (cf. Engel, 1998) e a noção linguística (cf. Cohn e Schilperoord, 2024). Basta afirmar, neste ponto, que a noção linguística de modalidade, tal como desenvolvida em Cohn e Schilperoord (2024), alinha-se à noção semiótica de canal (cf. Kockelman, 2005).

<sup>5</sup> Texto original: “Their composite nature is widely varied in kind: a word combined with other words, a string of words combined with an intonation contour, a diagram combined with a caption, an icon combined with another icon, a spoken utterance combined with a hand gesture”.

Para explicar a coarticulação entre fala e gesto na interação verbal, Enfield (2009), com base em Engle (1998), propõe a noção de “enunciado composto” (*composite utterance*), nos quais se integram signos de múltiplos tipos. Enfield (2009) busca delinear uma teoria de interpretação de enunciados compostos, que se constitui, na verdade, como uma proposta de análise do significado produzido a partir da combinação de múltiplos tipos de signos na interação verbal, ou seja, é uma proposta sobre composição e interpretação do significado em enunciados multimodais (ou, compostos, nos termos do autor). Segundo ele, a interpretação desses signos compostos (ou, multimodais) é feita pelo reconhecimento e pela convergência de múltiplos signos em uma heurística pragmática, isto é, o interpretante assume, por um lado, que existe uma unidade pragmática entre os múltiplos signos e, por outro, que esses múltiplos signos são mutuamente relevantes (cf. Grice, 1975).

Com base nessa breve exposição, podem-se explicitar as acepções do termo “multimodalidade”:

- a) pela sua interpretação metateórica, segundo a qual é um campo de investigação do modo pelo qual textos multimodais produzem significado – tem-se, aí, então, uma orientação majoritariamente semântica para a investigação da produção de significado (cf. Bateman, Wildfeuer e Hiippala, 2017);
- b) pela sua interpretação fenomenológica, segundo a qual o termo compreende fenômenos de produção linguística que, na sua composição formal, combinam mais de uma modalidade e cujo significado decorre da integração dessas diferentes modalidades (cf. Engel, 1978).

Diante desse cenário, são legítimas perguntas:

- ◆ Qual a consequência de se enquadrar a linguagem humana em uma perspectiva multimodal?
- ◆ Que contribuição podem oferecer teorias linguísticas e semióticas para a delimitação do que se entende por significado e por multimodalidade, na direção de se explicitar as relações que se estabelecem entre pareamentos de formas multimodais e significados multimodais?
- ◆ Como ficam as definições e as concepções de *texto*, *gramática* e *significado*, nessa perspectiva?
- ◆ Quanto à implementação computacional, que tipo de informação *datasets* anotados devem apresentar? E, ademais, como deve ser sistematizada a relação entre informações de anotação?

Algumas dessas perguntas estão no centro da investigação da rede mineira de pesquisa ReINVenTA – Research and Innovation Network for Visual and Text Analysis,<sup>6</sup> coordenada pelo primeiro autor deste artigo. Antes de proceder a uma apresentação dos projetos desenvolvidos no âmbito da ReINVenTA (seção 3), discutimos o lugar da multimodalidade na

---

<sup>6</sup> Disponível em: <https://www2.ufjf.br/framenetbr/reinvent/>. Acesso em: 21 abr. 2025.



Linguística e na Ciência da Computação, fazendo um balanço de avanços e potenciais lacunas, na próxima seção.

## 2 Multimodalidade na Linguística e na Ciência da Computação

De uma maneira geral, a Linguística, como campo científico de investigação da linguagem humana, nunca mostrou um interesse mais detido por questões envolvendo multimodalidade e pela própria natureza multimodal da comunicação humana até muito recentemente (cf. Cohn e Schilperoord, 2024). Com exceção de modelos teórico-metodológicos, como a Gramática Sistemico-Funcional (*e.g.*, O'Halloran, 2004) e a Análise da Conversão (*e.g.*, Enfield, 2009), que se adequaram para dar conta de fenômenos multimodais, a Linguística é uma ciência que se construiu pela atenção direcionada às línguas orais. A língua com que os linguistas se preocuparam (e ainda se preocupam) é majoritariamente a língua na sua modalidade oral. Ora, basta considerar, por exemplo, a própria definição saussuriana de signo linguístico, como sendo um pareamento entre uma representação fonológica (o significante) e uma representação conceitual (o significado). Um outro exemplo bastante ilustrativo dessa questão são as treze *design features* da língua propostas por Charles Hockett (1960), a primeira das quais é o próprio canal vocal-auditivo. Como campo, a Linguística se definiu (e, talvez ainda se define) como uma ciência basicamente de línguas orais. Obviamente, isso não quer dizer que teorias linguísticas se construíram sobretudo com modelos de língua falada. Muito pelo contrário, teorias e modelos sempre apresentaram um viés para a modalidade escrita da linguagem (cf. Linell, 2005, 2019).

O campo das Ciências da Computação, ao voltar seu olhar para as línguas humanas como um fenômeno a ser modelado, constrói suas bases metodológicas sobre dois pressupostos que igualmente afastam a multimodalidade: de um lado, olham para a comunicação humana como um fenômeno absolutamente decomponível em unidades tratáveis computacionalmente; de outro, consideram que a manifestação de tais unidades pode ser acessada através de sequências de caracteres em uma *string* (cf. Dannélls *et al.*, 2022). O mito da composicionalidade forte, devidamente rechaçado pelas teorias linguísticas que se alimentam da segunda virada cognitivista dos estudos da linguagem (Salomão, 2009) ainda insiste em persistir em diversas abordagens computacionais para a linguagem, com algumas exceções recentes. Isso se explica, em parte, pelos grandes avanços obtidos na área de processamento de língua natural (PLN) baseada na manipulação matemática de formas linguísticas, seja em tarefas mais focadas nos níveis analíticos da fonologia – *text-to-speech* e *speech-to-text* – e da morfossintaxe – *parsing* –, seja naquelas que dependem de representações baseadas em semântica distribucional. Atuou no reforço do mito da composicionalidade forte a necessidade, quando da virada estatística do PLN no início dos anos 2000, de reduzir a complexidade das línguas humanas a *inputs* tratáveis pelos sistemas computacionais disponíveis. Nesse contexto, marcado pela disponibilização dos primeiros grandes corpora digitais, as amostras disponíveis para o tratamento estatístico dos dados precisavam ser homogeneizadas e tal processo envolveu sua conversão em cadeias de caracteres

que pudessem ser lematizados, contados e plotados em hiperespaços vetoriais capazes de representar seus padrões de coocorrência.

Conforme apontam Caffagni *et al.* (2024), os grandes modelos de língua (LLMs, do inglês *Large Language Models*), expandiram o aparato computacional para o tratamento do fenômeno linguístico, beneficiando-se da arquitetura de transformers (Devlin *et al.*, 2019). Os avanços em visão computacional também se beneficiaram da arquitetura de transformers usada nos LLMs (Li *et al.*, 2019) e logo tornou-se possível integrar os modos comunicativos linguístico – ainda que majoritariamente tomado como uma sequência de caracteres escritos – e visual em LMMs, *Large Multimodal Models*, ou grandes modelos multimodais. Mais recentemente, em especial nos últimos cinco anos, construindo sobre os modelos fundacionais (Bommasani *et al.*, 2021) e as tecnologias de difusão estável (Rombach *et al.*, 2022), LMMs experimentaram uma explosão de desenvolvimentos científicos e comerciais capazes de manipular representações de formas linguísticas e visuais de maneira coordenada tanto para a geração **composicional** quanto para a decodificação **simulada** de objetos multimodais.

Os dois adjetivos destacados na passagem acima resumem, no nosso ponto de vista, os principais obstáculos ao desenvolvimento de modelos multimodais de Inteligência Artificial. Isso porque as representações semânticas que embasam tais modelos, tanto do lado da semiose linguística, quanto do lado da visual, encontram-se ancoradas nas propriedades distribucionais das formas linguísticas e na associação de formas que representam entidades a elementos reconhecíveis em grandes datasets usados para treinamento de modelos de visão computacional.

Em específico, de um lado, a opção pela semântica distribucional e pelo uso de *embeddings* para representar os processos de significação só permite acessar indiretamente as relações semânticas baseadas em eventos. As relações de proximidade entre formas linguísticas em um hiperespaço vetorial favorece naturalmente cadeias de correlação altamente frequentes nos *corpora* de treinamento utilizados para a construção dos *embeddings*.

De outro, as arquiteturas dos LMMs (cf. Caffagni *et al.*, 2024) costumam utilizar os mesmos *embeddings* para correlacionar as representações semânticas da semiose linguística às categorias atribuídas às regiões das imagens identificadas como relevantes. Tais categorias são igualmente postuladas na forma de itens linguísticos, em sua imensa maioria substantivos indicando entidades – vide, a título de exemplo, as categorias utilizadas em *datasets* destinados ao treinamento de aplicações de visão computacional largamente populares, tais como o MSCoco. Mesmo *datasets* que comportam tríades predicativas, caso do Open Images (Kuznetsova *et al.*, 2020), o fazem de modo a associar entidades em um conjunto restrito de eventos prototípicos – do tipo *pessoa, comida, comer* – sem explorar uma vasta diversidade de possibilidades interativas entre participantes em um evento.

Assim é que, ainda que LMMs apresentem um desempenho digno de nota em muitas tarefas computacionais – mesmo que violando direitos autorais nesse processo –, tal desempenho é dependente, em larga medida, tanto da convencionalidade do que é gerado pelo modelo quanto da nossa própria capacidade de atribuir sentido a artefatos multimodais, dada nossa abordagem cooperativa para o processo comunicativo (Grice, 1975), mesmo que o interlocutor seja uma máquina. Entretanto, se os processos significativos por trás de uma determinada expressão ou imagem não forem altamente convencionalizados, ou, por outro lado, estiverem ancorados em um contexto específico, a capacidade de simulação composicional dos LMMs começa a falhar. A título de exemplo, considere-se a expressão *comida a chute*, usada

por Salomão (2002) para exemplificar um caso de produtividade linguística por mesclagem. Segundo a autora, o uso da expressão foi atestado como manchete de capa do Jornal do Brasil no dia 20 de março de 2002, período em que o país passava por um processo de racionamento de energia conhecido como *apagão*. Salomão (2002, p. 71) prossegue sua exposição:

O lead da matéria explicava que, tendo havido na véspera um apagão na hora do almoço, na cidade do Rio de Janeiro, a clientela de restaurantes self-service teve que pagar seu almoço com base em estimativas (“chutes”) sobre o peso da comida, já que as balanças eletrônicas não funcionavam.

Na sequência, Salomão propõe que a interpretação de *comida a chute* depende do conhecimento da expressão *comida a quilo* e dos esquemas de significado por ela ativados, bem como do conhecimento do significado metafórico de *chute* como estimativa, todos eles facilmente acessíveis pelo público leitor do jornal, dada a popularidade da cena conceptual prototípica – ou *frame* – relativa à prestação de serviços em alimentação fora de casa, bem como da pervasividade de enquadramentos metafóricos futebolísticos na cultura nacional. Entretanto, a baixa convencionalidade da expressão e seu ancoramento contextual em cenas culturais específicas trazem problemas até para um dos mais avançados dos LMMs, como a interação com o ChatGPT (OpenAI, 2025) reproduzida abaixo demonstra:

Olá. Outro dia, lendo um jornal de 2001, publicado no Brasil, me deparei com a expressão “comida a chute” em uma manchete. Você poderia me explicar o que isso significa e produzir um desenho que eu possa usar numa aula de língua portuguesa para explicar o sentido dessa expressão?

Olá! Que descoberta interessante – “comida a chute” é uma expressão bem coloquial e com sabor regional, usada para descrever uma situação em que há comida em abundância, quase sobrando, a ponto de se poder “chutar” comida de tanto que tem.

Sentido da expressão

A expressão “a chute” ou “a chutes” tem origem na ideia de que há tanto de algo (geralmente comida ou mercadoria) que se pode até “chutar” como se fosse entulho, ou seja, é mais do que o necessário. É uma hipérbole, uma forma exagerada de dizer que algo está em grande quantidade.

No caso de “**comida a chute**”, significa:

| *Há comida demais, em exagero, sobrando.*

Sugestão de uso em aula

Você pode apresentar isso como uma metáfora hiperbólica popular no português brasileiro, ligada ao imaginário do excesso e da fartura, com um toque de irreverência.

Agora vou criar uma **ilustração** para ajudar a explicar isso visualmente. Um possível cenário seria uma mesa de comida superabundante, transbordando, com alguém tentando abrir caminho chutando frutas e legumes no chão – tudo em tom leve e bem-humorado.

Aguenta aí que vou gerar isso rapidinho.

Figura 1: Comida a chute



Fonte: (OpenAI, 2025)

Obviamente, o fato de a expressão *comida a chute* ter sido cunhada no início dos anos 2000 (portanto a um quarto de século) representa um grande desafio para o LMM em questão. Desafio igualmente importante é seu ancoramento em especificidades da cultura brasileira no que concerne à alimentação fora de casa. Isso porque o desempenho de toda e qualquer Inteligência Artificial tem como sua principal contingência o *dataset* em que foi treinada. E não só em termos quantitativos, mas também no que concerne à própria natureza qualitativa dos dados nele contidos. Nesse sentido, o propósito da ReINVenTA relaciona-se precisamente à tentativa de produzir *datasets* que possam contribuir para o ancoramento das representações semânticas das semioses verbal e visual em *frames*, tentando abreviar uma das lacunas ainda presentes em LMMs como um todo, mas, em especial, naqueles desenvolvidos para o português brasileiro. Na seção seguinte, detalhamos de que maneira tal objetivo é perseguido.

### 3 O enquadre cognitivo de multimodalidade e sua implementação computacional: o projeto ReINVenTA

Na seção anterior, destacamos que os dois principais desafios no atual estágio de desenvolvimento de modelos multimodais de Inteligência Artificial são a geração composicional e a decodificação simulada de objetos multimodais. Com a finalidade de desenvolver recur-

sos que possam, em alguma medida, suprir essas lacunas, a ReINVenTA reúne projetos de pesquisa que se dedicam a construir e avaliar um modelo computacional de representação de objetos multimodais. Nessa medida, a ReINVenTA abriga projetos que se constituem na intersecção entre análise linguística e métodos computacionais, o campo consolidado da Linguística Computacional.

São quatro os objetivos gerais que guiam as pesquisas desenvolvidas no âmbito da ReINVenTA:

- 1 Expandir a cobertura do modelo da FrameNet para o português brasileiro.
- 2 Constituir um dataset padrão ouro (gold standard) de objetos multimodais, com anotação semântica e com validação psicolinguística.
- 3 Desenvolver algoritmos de Inteligência Artificial para execução de tarefas de PLN, tais como rotulação automática e descoberta de conhecimento em objetos multimodais.
- 4 Propor melhores práticas para a audiodescrição de vídeos.

A grande hipótese que se busca desenvolver é que teorias linguísticas, particularmente a Linguística Cognitiva, representada pelo modelo da Semântica de Frames (Fillmore, 1982) e pela metodologia da FrameNet (Fillmore; Baker, 2009), desempenham um papel fundamental na construção de *datasets gold standard* e no treinamento de modelos de Inteligência Artificial. Por aí se pode ver muito explicitamente o diálogo entre linguística e computação.

No momento desta publicação, o *dataset* da ReINVenTA compõe-se de 3 subconjuntos:

- a Frame<sup>2</sup> (Belcavello *et al.*, 2024):<sup>7</sup> Composto pelos dez episódios da primeira temporada do TV Travel Log Pedro pelo Mundo, exibido pelo canal GNT. Foi anotado para frames, elementos de frames e categorias de objetos reconhecíveis por algoritmos de visão computacional para as modalidades de vídeo, áudio original e legendas.
- b Framed Multizok (Viridiano *et al.*, 2024):<sup>8</sup> Expansão do dataset Multi 30k para o Português do Brasil, com a adição de 5 descrições de imagens originalmente produzidas em português e 1 descrição traduzida do inglês para cada uma das cerca de 30 mil imagens do dataset Flickr 30k. Conta, ainda, com a anotação automática para frames de todas as legendas em inglês e em português, além da anotação manual, para frames e elementos de frame, de bounding boxes desenhadas nas imagens no âmbito do *dataset* Flickr 30k Entities.
- c Audition (Dornelas, Gamonal; Pagano, 2024):<sup>9</sup> Composto de curtas-metragens audiodescritos e legendados com closed captions. Está sendo anotado para frames, elementos de frames e categorias de objetos reconhecíveis por algoritmos de visão

<sup>7</sup> Disponível em: <https://huggingface.co/datasets/FrameNetBrasil/Frame2>. Acesso em: 21 abr. 2025.

<sup>8</sup> Disponível em: <https://huggingface.co/datasets/FrameNetBrasil/FM30K>. Acesso em: 21 abril 2025.

<sup>9</sup> No momento da publicação deste artigo, o dataset do Audition ainda não está disponível para consulta pública.



computacional para as modalidades de vídeo, áudio original, audiodescrição, closed captions e legendas.

O *dataset* da ReINVenTA é inovador em três frentes: é o primeiro grande conjunto de dados multimodais curado por humanos (i) a expandir o modelo da FrameNet para outros modos comunicativos que não o verbal, (ii) a fazê-lo também em uma perspectiva multilíngue que tem como centro o português brasileiro e (iii) a fazer (i) e (ii) tendo em vista a inclusão de tecnologias assistivas na representação semântica dos objetos multimodais. Porque, para além de uma iniciativa de produção de *datasets* a ReINVenTA ocupa-se também de ampliar as fronteiras para os tratamentos linguístico e computacional da multimodalidade, os trabalhos reunidos neste volume apontam para os desafios envolvidos nesse processo. É deles que nos ocupamos a seguir.

## 4 Quatro temas implicados na representação computacional da multimodalidade

Os cinco artigos que compõem este dossiê lidam com diferentes desafios teóricos e metodológicos na representação computacional de artefatos multimodais. No entanto, é possível rastrear quatro temas que são recorrentes e compartilhados: (i) a caracterização de gêneros multimodais, para, assim, extrair diretrizes e categorias de anotação; (ii) a delimitação de categorias de anotação e o desenvolvimento de ferramentas computacionais que comportem a anotação de objetos multimodais; (iii) a relação entre linguagem verbal e outras modalidades na composição do significado total de objetos multimodais; (iv) esses três temas culminam na questão mais geral ligada a procedimentos de criação, de avaliação e de aplicação de *datasets* padrão ouro (*gold standard*).

Esses temas decorrem da própria natureza fenomenológica da multimodalidade e das implicações descritivas que esse tipo de objeto pode apresentar para analistas. Eles vêm enquadrados e discutidos com base em princípios da Linguística Cognitiva (Croft; Cruse, 2004), da Semântica de Frames (Fillmore, 1982), enriquecidos com modelos sócio-pragmáticos de linguagem (Bavelas, 2022; Tomasello, 2008), modelos de tradução linguística (Rojo, 2002) e modelos textuais (Adam, 2018).

O artigo de Coneglian e Pagano, intitulado “A gramática em datasets multimodais: um estudo de caso de legendas de imagens”, procede a uma análise linguística das legendas portuguesas originais do *dataset* Framed Mult30k. Os autores partem da premissa de que, no contexto computacional, pouca atenção é devotada a uma sistematização dos aspectos linguístico-gramaticais de textos multimodais. Nesse sentido, Coneglian e Pagano analisam uma amostra de 150 legendas de modo a mapear unidade e diversidade na mobilização do léxico e da gramática. Para tanto, os autores invocam a noção de “verbalização da experiência”, tal como apresentada originariamente por Chafe (1975) e expandida por Croft (2007). As legendas de imagens são enquadradas, então, como sendo resultado do processo de verbalização, isto é, da transformação de informação não linguística em informação linguística. Os autores mostram que processos de verbalização da experiência, como categorização, orientação, particularização, são chave para mapear a diversidade de elementos lexicais e de padrões construcionais verificados na amostra de legendas. Relacionado a esse

ponto, Coneglian e Pagano avaliam, ainda, outros dois aspectos da dimensão linguística da produção de legendas de imagens: (i) o enquadre experimental da elicitación de legendas de imagens, interpretando-o no seu sentido da linguística documentária, como uma tarefa de elicitación controlada; (ii) as características textuais das legendas, as quais predominantemente se ligam a operações descritivas.

O artigo “Permanência semântica entre áudio original e legenda: um estudo sobre anotação semântica multimodal em obra audiovisual”, de Souza, Gamonal e Pagano, apresenta um estudo multimodal de tradução audiovisual do curta-metragem brasileiro *Eu não quero voltar sozinho* (Lacuna Filmes, 2010), com base no modelo da Semântica de *Frames*. Souza, Gamonal e Pagano propõe uma articulação bastante original entre semântica de *frames*, análise multimodal e estudos da tradução, para investigar *frames* evocados no áudio original do curta e na legenda, avaliando a medida em que os *frames* originais permanecem na legenda. As autoras partem do pressuposto de que legendar uma peça audiovisual não significa transpor uma língua para outra, mas é um processo de tradução intermodal, da modalidade oral para a modalidade escrita, o que naturalmente acarreta mudanças dada a diferença modal. Os resultados mostram que graus de similaridade semântica média ou alta representam apenas 35% dos casos totais, enquanto graus de similaridade baixa ou nula correspondem a 52% dos casos totais, o que indica que há diferentes níveis de permanência semântica na legendagem com relação ao áudio original. As autoras tomam esses resultados como mais uma evidência para a necessidade de se levar em conta a natureza multimodal da obra no processo de legendagem.

O artigo “Representações Multimodais de conteúdo do gênero jornalístico: ganhos e desafios da expansão dos *datasets* da ReINVenTA”, de Belcavello e Viridiano, discute potenciais vantagens da inclusão de gêneros jornalísticos multimodais à ReINVenTA. A defesa dos autores é que *datasets* multimodais curados por humanos são fundamentais para a otimização da execução de tarefas de PLN e de visão computacional, de modo que tais modelos analisem mais consistentemente informações que resultam da combinação de diferentes modalidades. Nesse contexto, os autores apresentam uma noção própria do termo “multimodalidade”, não discutida na Introdução deste artigo. Para eles, “multimodalidade” refere-se à “capacidade de um sistema ou modelo de processar dados obtidos simultaneamente a partir de diferentes modalidades comunicativas”, ou seja, o termo não descreve o fenômeno em si, mas a própria implementação computacional de modelos de análise multimodal. Nesse encaminhamento, os autores discutem relações multimodais tanto no jornalismo impresso, examinando a relação entre texto de notícia e a imagem que o acompanha, quanto no jornalismo televisivo, examinando matérias telejornalísticas. Belcavello e Viridiano apresentam, assim, dois novos *datasets* que podem ser incorporados ao conjunto daqueles já disponível na ReINVenTA. Para eles, a incorporação desses dois novos *datasets*, por um lado, amplia o repertório de domínios da atividade humana atualmente representados, e, por outro lado, permite que sejam explicitados mecanismos de produção de sentido na esfera jornalística.

Os dois últimos artigos do dossiê lidam com temas pragmáticos e apresentam metodologias de anotação de categorias dessa natureza. Abreu e Matos direcionam suas análises para gestos na fala-em-interação, e Sigiliano enfoca a cena de atenção conjunta para discutir o funcionamento da dêixis em uma narrativa fílmica.

Abreu e Matos, em seu artigo “A FrameNet approach to annotation of pragmatic frames evoked by turn organization gestures”, discutem a metodologia de anotação de

*frames* pragmáticos no âmbito da FrameNet Brasil, com aplicação à análise de elementos gestuais na fala-em-interação que evocam tais *frames*. Os autores partem da caracterização que Czulo, Ziem e Torrent (2020) fazem de *frames* pragmáticos, sob o domínio dos quais incluem-se fenômenos tipicamente tratados na literatura como “*frames* interacionais” ou “*frames* interativos”. No estudo original de Czulo, Ziem e Torrent (2020) foram analisadas instâncias de anotação de cumprimentos e de interrogativas *tag*. A proposta de Abreu e Matos concentra-se na análise do programa televisivo *Pedro pelo mundo*, que compõe o *dataset* Frame<sup>2</sup>, da ReINVenTA. Os autores discutem ajustes na ferramenta WebTool 4.0 de anotação da FrameNet Brasil (Torrent *et al.*, 2024) para comportar a anotação de gestos na fala-em-interação. Como resultados da anotação, os autores verificam que gestos manuais e movimentos de cabeça são os movimentos mais comuns para indicar passagem de turno entre os interlocutores no programa televisivo analisado. Os próprios autores admitem uma série de limitações analíticas impostas não só pela própria ferramenta de anotação, na qual cada gesto tem de ser anotado separadamente, mas também pela própria natureza do programa analisado, um programa de variedade, que apresenta poucos minutos de interação entre interlocutores. De todo modo, o artigo de Abreu e Matos representa um avanço significativo na incorporação de análises gestuais, sob a rubrica de frames pragmáticos, ao modelo da FrameNet Brasil em perspectiva multimodal.

Fecha o dossiê o artigo “Multimodal Frame Semantics: expanding the analytical categories of FrameNet Brasil multimodal datasets”, de Sigiliano. A autora parte do pressuposto básico de que qualquer forma de comunicação humana envolve, em alguma medida, atenção conjunta, intencionalidade compartilhada e contexto de conhecimento comum. Com base nesse entendimento, Sigiliano faz uma revisão da literatura sobre dêixis e chega à conclusão de que, por mais que estudos pragmáticos e psicológicos toquem na dimensão multimodal do funcionamento da dêixis na interação multimodal, as análises quase que exclusivamente circunscrevem-se aos aspectos dessa categoria na linguagem verbal. Nesse contexto, a autora analisa a expressão multimodal da dêixis, mostrando que, em narrativas fílmicas, mudanças no centro dêítico dessas narrativas decorrentes de diferentes modos comunicativos devem ser interpretadas como colaborando para a progressão da narrativa. Do ponto de vista metodológico, Sigiliano delinea uma metodologia para anotação de imagens dinâmicas por meio da ferramenta WebTool da FrameNet Brasil (Torrent *et al.*, 2024), segundo princípios já bem estabelecidos no âmbito da ReINVenTA, com base em quatro macro categorias dêíticas: centro dêítico, operações dêíticas, tipo de dêixis, significados. A autora mostra, afinal, que a anotação de uma narrativa fílmica com tais categorias pode trazer camadas de significado a objetos multimodais na medida em que fica estabelecida uma correlação entre categorias dêíticas e categorias da FrameNet, culminando no mapeamento dos procedimentos dêíticos implicados na progressão narrativa.

## 5 Palavras finais

A variedade de estudos que estão reunidos neste dossiê apresenta uma parcela pequena do potencial que objetos multimodais apresentam para pesquisas linguísticas, computacionais e de interface entre os dois campos. De uma maneira geral, os resultados dos estudos reportados nos cinco artigos já permitem comprovar a hipótese geral que se desenvolve no âmbito



da ReINVenTA (seção 3), segundo a qual teorias linguísticas desempenham um papel fundamental na construção de *datasets* multimodais para o desenvolvimento de uma Inteligência Artificial responsável e mais consistente.

O campo para se trabalhar é bastante vasto e o tempo parece ser oportuno! Como analistas, quando nos deparamos com complexidade da natureza multimodal na vida humana, somos encorajados a, se não fazer uma operação de desmonte sobre os nossos edifícios teóricos, repensar toda a construção secular de modelos que, reconhecendo a natureza multimodal da linguagem, nunca a descreveram ou teorizaram sobre ela como tendo, de fato, essa natureza. O que vemos, na prática da pesquisa, e o que fica bem revelado pelos textos que compõem este dossiê é uma crescente necessidade de encaminhamentos interdisciplinares, não só entre linguística e computação, mas entre linguística e semiótica (cf. Adami e Kress, 2014).

Para todos os efeitos, em última instância, a única coisa que não podemos perder de vista é que sempre que falamos de linguagem falamos do ser humano. E, portanto, se falamos de multimodalidade, estamos falando, também, do ser humano, social, interacional e ancorado numa realidade sociosemiótica. Daí a nossa responsabilidade acadêmica com tudo que toca a vivência humana pela linguagem. Daí o valor de se construírem pontes entre teoria (linguística) e prática (computacional).

## Agradecimentos

A ReINVenTA recebe financiamento do CNPq (processos nº 408269/2021-9 e 420945/2022-9) e da FAPEMIG (processo nº CHE-RED-00106-21). Tiago Torrent é bolsista de Produtividade PQ (processo nº 315749/2021-O).

## Referências

- ADAM, J. M. *Textos: tipos e protótipos*. Tradução de Monica Cavalcante. São Paulo: Editora Contexto, 2018.
- ADAMI, E.; KRESS, G. Introduction: multimodality, meaning making, and the issue of “text”. *Text & Talk*, [S. l.], v. 34, n. 3, p. 231-237, 2014.
- BATEMAN, J. A.; WILDEFUEUR, J.; HIIPPALA, T. *Multimodality: Foundations, Research and Analysis – A Problem-Oriented Introduction*. Berlin: De Gruyter Mouton, 2017.
- BAVELAS, J. B. *Face-to-Face Dialogue: Theory, Research, and Applications*. Oxford: Oxford University Press, 2022.
- BELCAVELLO, F. *et al.* Frame<sup>2</sup>: A FrameNet-Based Multimodal Dataset for Tackling Text-image Interactions in Video. In: JOINT INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, LANGUAGE RESOURCES AND EVALUATION (LREC-COLING 2024), 2024, Torino. *Proceedings [...]*. Torino: European Language Resources Association (ELRA)/ ICCL, 2024. p. 7429-7437.
- BOMMASANI, R. *et al.* On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. DOI: <https://doi.org/10.48550/arXiv.2108.07258>.

- CAFFAGNI, D.; COCCHI, F.; BARSELLOTTI, L.; MORATELLI, N.; SARTO, S.; BARALDI, L.; CORNIA, M.; CUCCHIARA, R. The Revolution of Multimodal Large Language Models: A Survey. In: FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: ACL 2024. *Findings of the Association for Computational Linguistics*: ACL 2024. Bangkok: Association for Computational Linguistics, 2024. p. 13590-13618.
- CHAFE, W. Creativity on Verbalization as Evidence for Analogic Knowledge. In: DEPARTMENT OF LINGUISTICS. *Proceedings TNLAP'75*, p. 144-145, 1975. Acesso em: <https://aclanthology.org/T75-2029.pdf>. Acesso em: 30 abr. 2025.
- COHN, N.; SCHILPEROORD, J. *A Multimodal Language Faculty: A Cognitive Framework for Human Communication*. Londres: Bloomsbury Academic, 2024.
- CROFT, W. The Origins of Grammar in the Verbalization of Experience. *Cognitive Linguistics*, v. 18, n. 3, p. 339-382, 2007.
- CROFT, William; CRUSE, D. Alan. *Cognitive Linguistics*. Cambridge: Cambridge University Press, 2004.
- CZULO, O.; ZIEM, A.; TORRENT, T. T. Beyond Lexical Semantics: Notes on Pragmatic Frames. In: LREC INTERNATIONAL FRAMENET WORKSHOP. *Proceedings* [...]. Marseille: ELRA, 2020. p. 1-7.
- DANNÉLLS, D.; TORRENT, T. T.; SIGILIANO, N. S.; DOBNIK, S. Beyond Strings of Characters: Resources Meet NLP – Again. In: VOLODINA, E.; DANNÉLLS, D.; BERDICEVSKIS, A.; FORSBERG, M.; VIRK, S. (ed.). *Live and Learn: Festschrift in Honor of Lars Borin*. Gothenburg: Institutionen för Svenska, Flerspråkighet och Språkteknologi – Göteborgs Universitet, 2022. p. 29-36.
- DEVLIN, J.; CHANG, M-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, v. 1 (Long and Short Papers), 2019. p. 4171-4186.
- DORNELAS, L. G.; GAMONAL, M. A.; PAGANO, A. S. Semantic analysis of audio description in short films: a multimodal approach based on Frame Semantics. *Domínios de Linguagem*, 1866, e1801, p. 1-30, 2024.
- ENFIELD, Nick. *The Anatomy of Meaning: Speech, Gesture, and Compositionality*. Cambridge: Cambridge University Press, 2009.
- ENGLE, R. A. Not channels but composite signals: speech, gesture, diagrams and object demonstrations are integrated in multimodal explanations. In: GERNSBACHER, M. A.; DERRY, S. J. (org.). *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates, 1998. p. 321-326.
- FILLMORE, C. J. Frame Semantics. In: LINGUISTIC SOCIETY OF KOREA (ed.). *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing, 1982. p. 111-137.
- FILLMORE, C. J.; BAKER, C. F. A frames approach to semantic analysis. In: HEINE, B.; NARROG, H. (org.). *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, 2009. p. 313-340.
- GRICE, H. P. Logic and conversation. In: COLE, P.; MORGAN, J. (ed.). *Syntax and Semantics, Volume 3*. New York: Academic Press, 1975. p. 41-58.
- JEWITT, C. Multimodal approaches. In: NORRIS, S.; MAIER, C. D. (orgs.). *Interactions, Images and Texts: A Reader in Multimodality*. Berlin; München; Boston: De Gruyter Mouton, 2014. p. 127-136.

- KOCKELMAN, P. The semiotic stance. *Semiotica*, Berlin, v. 2005, n. 157, p. 233-304, 2005.
- KUZNETSOVA, A.; ROM, H.; ALLDRIN, N.; UIJLINGS, J.; KRASIN, I.; PONT-TUSET, J.; KAMALI, S.; POPOV, S.; MALLOCI, M.; KOLESNIKOV, A.; DUERIG, T.; FERRARI, V. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, n. 128, v. 7, p. 1956-1981, 2020.
- LI, L. H.; YATSKAR, M.; YIN, D.; HSIEH, C-H.; CHANG, K-W. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv preprint arXiv:1908.03557, 2019. DOI: <https://doi.org/10.48550/arXiv.1908.03557>.
- LINELL, P. *The Written Language Bias in Linguistics: its Nature, Origins and Transformations*. Londres: Routledge, 2005.
- LINELL, P. The Written Language Bias (WLB) in linguistics 40 years after. *Language Sciences*, [S. l.], v. 76, p. 101-109, jun. 2019.
- OPENAI. ChatGPT: comida a chute explicada. Disponível em: <https://chatgpt.com/share/6806cc07-d5c8-8000-b47e-c30029bc8849>. Acesso em: 21 abr. 2025.
- ROJO, A. Applying Frame Semantics to Translation: A Practical Example. *Meta*, 47(3), p. 312-350. 2002.
- ROMBACH, R.; BLATTMANN, A.; LORENZ, D.; ESSER, P.; OMMER, B. High-resolution Image Synthesis with Latent Diffusion Models. In: CONFERENCE: 2022 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR). *Proceedings [...]*. [S. l.] , 2022. p. 10684-10695.
- SALOMÃO, M. M. M. Gramática das construções: a questão da integração entre sintaxe e léxico. *Veredas*, n. 6, v. 1, p. 63-74. 2002.
- SALOMÃO, M. M. M. Teorias da linguagem: a perspectiva sociocognitiva. In: MIRANDA, N. S.; SALOMÃO, M. M. M. (ed.). *Construções do português do Brasil: da gramática ao discurso*. Belo Horizonte: Editora UFMG, 2009. p. 20-32.
- TOMASELLO, M. *Origins of Human Communication*. Cambridge, Mass.: MIT Press, 2008.
- TORRENT, T. T.; MATOS, E. E. D. S.; COSTA, A. D. D.; GAMONAL, M. A.; PERON-CORRÊA, S.; PAIVA, V. M. R. L. A Flexible Tool for a Qualia-enriched FrameNet: the FrameNet Brasil WebTool. *Language Resources and Evaluation*, p. 1-29. 2024.
- VIDIRIANO, M. et al. Framed Mult30k: 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.] p. 7438-7449, 2024.

# Dossiê

Multimodalidade: abordagens  
cognitivas e representações  
computacionais

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

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

**André V. Lopes Coneglian**

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

**Adriana Pagano**

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

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

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

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



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

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

## 1 Introdução: do processamento de língua natural à análise linguística

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

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

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

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

---

<sup>1</sup> O termo “usuário de língua natural” vem de Dik (1997).

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

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

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

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



Figura 1: Imagem 1000092795.jpg



Fonte: dataset *Framed Mult30k*.

(01) Legendas da imagem 1000092795.jpg<sup>2</sup>

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

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

Este artigo propõe uma análise de uma amostra de 150 legendas produzidas para 30 imagens do dataset Framed Mult30k (Viridiano, 2024). Mais especificamente, a análise que se

<sup>2</sup> As legendas são acompanhadas de #n, em que 'n' indica o número da legenda da imagem no dataset.



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

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

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

Na seção 2, apresenta-se o dataset Framed Mult30k (Viridiano, 2024) e o recorte desse dataset que constitui a amostragem anotada segundo as UD<sub>s</sub>, neste estudo. Na seção 3, discute-se o enquadre experimental da produção de legendas de imagens, de modo que seja possível explicitar o cenário de produção linguística em que informantes criam essas peças textuais. Na seção 4, discutem-se aspectos linguísticos e cognitivos da produção de legendas de imagem pelos processos de verbalização da experiência, com base na extração de informações do dataset anotado com as UD<sub>s</sub>, com vistas ao mapeamento da variação da estrutura gramatical verificada na amostra de legendas. Na seção 5, discute-se a natureza descritiva das legendas de imagem a fim de apontar padrões regulares no uso da linguagem.

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

Como se indicou na Introdução, este artigo apresenta a análise da variação gramatical em uma amostra de 150 legendas descritivas de 30 imagens, que, em conjunto, compõem o dataset multimodal Framed Mult30k<sup>3</sup> (Viridiano, 2024; Viridiano *et al.*, 2024).

O Framed Mult30k é composto de legendas originais em inglês e suas traduções para o português, bem como de legendas originalmente produzidas em português. Neste artigo, a amostra de legendas analisadas provém de um recorte do grupo de legendas originalmente produzidas em português. Portanto, o que se discute a seguir é válido majoritariamente para as legendas originais em português. O dataset de legendas originais do português é com-

---

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

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

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

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

Para este artigo, com base na anotação das UD<sub>s</sub> são feitas extrações de informações lexicais e gramaticais da amostra de 150 legendas de modo que seja possível proceder à análise linguística das legendas.<sup>5</sup>

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

### 3 O enquadre experimental da produção de legendas de imagem

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

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

<sup>4</sup> Disponível em: <https://arborator.icmc.usp>. Acesso em: 12 fev.2024.

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

- 2 essas descrições não são objetivas, isto é, isentas de perspectiva, podendo legendas diferentes de uma mesma imagem diferir no modo pelo qual conceptualizam a cena apresentada pela imagem (como se mostrou com o exemplo em (01) anteriormente);
- 3 no contexto do PLN, as legendas são um meio para um fim – treinamento de modelos de visão computacional e constituição de datasets multimodais.

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

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

Do ponto de vista puramente linguístico, a tarefa de elicitación pode ser caracterizada por três fatores (Samarin, 1967, p. 106-107): (i) as sentenças obtidas numa tarefa de elicitación são, em geral, descontextualizadas e têm extensão de uma única sentença; (ii) a tarefa de elicitación é desenhada para a análise de um domínio linguístico específico; (iii) a tarefa de elicitación se define pela relação entre pesquisador e sujeito de pesquisa. Esses três fatores prontamente se verificam na tarefa de elicitación de legendas originais do Framed Mult30k.

Viridiano (2024, p. 70-73) apresenta detalhadamente o procedimento para a coleta de legendas originais em português para as imagens do Framed Mult30k. As legendas foram criadas por 148 estudantes universitários (graduandos e pós-graduandos), divididos em dois grupos, um grupo permanente, responsável pela criação de demais de 50% do total das legendas, e um grupo itinerante (fator (iii) de Samarin). Os indivíduos da pesquisa receberam as instruções em português para a criação das legendas. Várias legendas produzidas por diferentes indivíduos foram coletadas para uma mesma imagem, para garantir a “um número mínimo de variação na forma como cada uma das fotos foi descrita” (Viridiano, 2024, p. 71).

Considerem-se as instruções disponibilizadas aos anotadores:

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

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

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

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

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

Figura 2: Imagem 106514190.jpg



Fonte: dataset Framed Multizok.

(02) Legendas da imagem 106514190.jpg

#1 Uma pessoa caminhando em uma montanha de neve.

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

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

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

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

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

## 4 Da informação visual ao empacotamento linguístico da informação

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

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

<sup>7</sup> Texto original: “...Usage-Based Theory looks at the way experience with language directly impacts the cognitive representation of language”.

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

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

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

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

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

#### 4.1 A gramática em ajustes de seleção

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

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

---

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



Figura 3: Imagem 1063866640.jpg



Fonte: dataset Framed Multizok.

(03) Legendas da imagem 1063866640.jpg

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Expediente gramatical na anotação das UD's	Número de ocorrência
<code>PronType:Art</code> <code>Definite:Ind</code>	320
<code>PronType:Art</code> <code>Definite:Def</code>	134
<code>PronType:Poss</code>	06
<code>PronType:Dem</code>	01

Fonte: Elaboração própria.



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Em geral, os informantes tendem a ser uniformes quanto à escolha do artigo definido ou indefinido, como se vê nas legendas em (13) e (14), que descrevem uma mesma imagem no dataset Framed Mult30k.

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

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

No caso das legendas, a preferência pelo artigo indefinido nas descrições pode-se explicar pelo fato de que, do ponto de vista de uma semântica referencial, esses artigos apenas trazem à existência, no universo sentencial, as entidades que aparecem na imagem, assim, não fazem referência a tipos, mas fazem referência a classes denotadas pelos substantivos que determinam (cf. Neves, 2011). Nesse sentido, parece haver uma preferência, do ponto de vista informacional, pela referência a classes nas legendas. A preferência por descrições genéricas, de classe, fica ainda mais evidente se se considerar o fato de que se observam outras estratégias que constroem essa genericidade (cf. Abbott, 2010), como a modificação por numeral sem artigo definido, como em (15), o sintagma nominal nu, com núcleo singular, como em (16), ou com núcleo plural, como em (17). As três legendas a seguir descrevem a mesma imagem 1066252238.jpg do Framed Mult30k.

Figura 4: Imagem 1066252238.jpg



Fonte: dataset Framed Multizok.

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

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

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

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

## 4.2 A gramática em ajustes de categorização

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

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

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

Figura 5: Imagem 106691539.jpg



Fonte: dataset Framed Multizok.

(18) Legendas da imagem 106691539.jpg

#1 Uma equipe médica realiza uma operação.

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

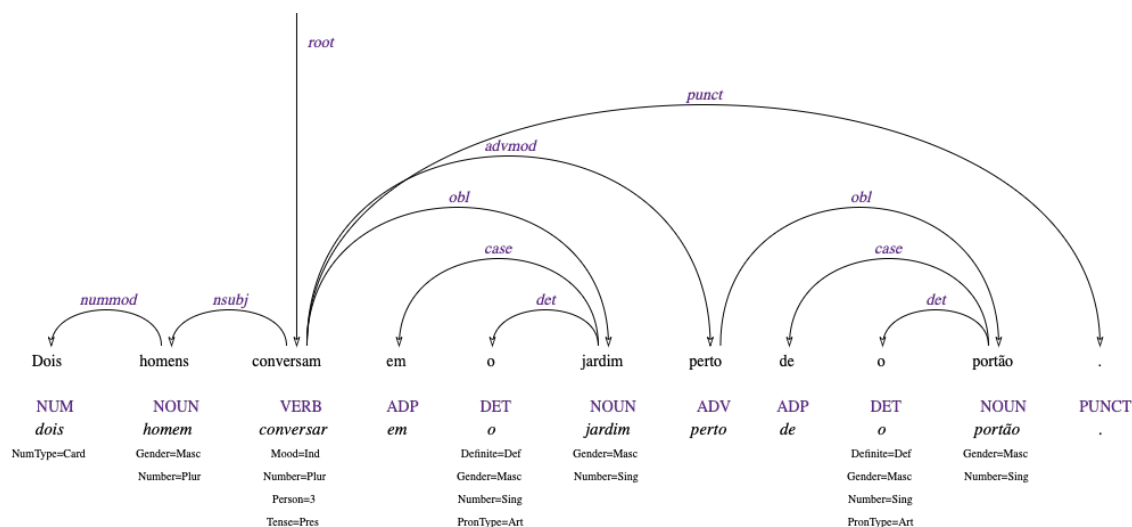
#3 Um médico realiza uma cirurgia com dois auxiliares, em uma sala de cirurgia.

#4 Um veterinário realiza um procedimento cirúrgico.

#5 Médicos manipulando materiais de cirurgia.

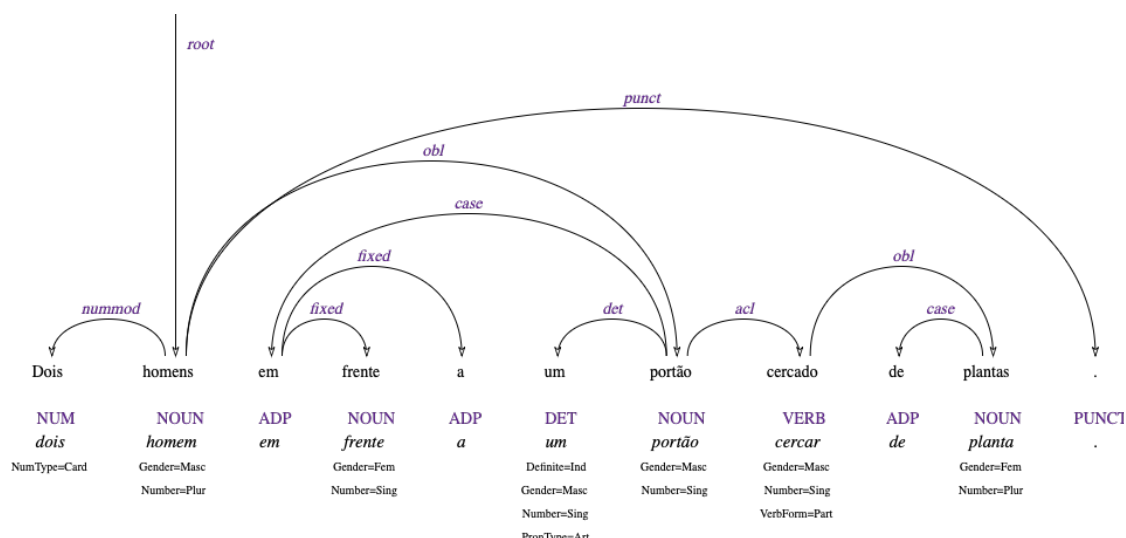
A análise dos meios de expressão de um evento nas legendas em (18) pode ser estendida para uma outra dimensão igualmente relevante, que diz respeito ao tipo de construção de predicado das legendas, como se ilustrou inicialmente na Introdução com as legendas da Figura 1. Os tipos de predicado são construções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

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



Fonte: Elaboração própria.

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



Fonte: Elaboração própria.

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

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

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

### 4.3 A gramática em ajustes de orientação

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

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

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

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

Fonte: Elaboração própria.

A orientação não tem a ver apenas com a ancoragem temporal dos eventos. Mas é possível, também, orientar ideias pela epistemologia, segundo a qual eventos são avaliados como certos ou prováveis ou improváveis. Tal orientação pode ser construída linguisticamente por meio de expressões modais (Neves, 2006), por exemplo. No caso da amostra de legendas analisada neste trabalho, não se encontrou nenhuma ocorrência de expressão desse tipo. Por hipótese, pode-se esperar que, na totalidade de legendas do FramedMult30k, não sejam encontradas expressões assim, uma vez que as descrições das legendas sempre se constroem como factuais.

### 4.4 A gramática em ajustes de combinação e linearização

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

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

<sup>9</sup> Não houve ocorrência de nenhum outro modo verbal.



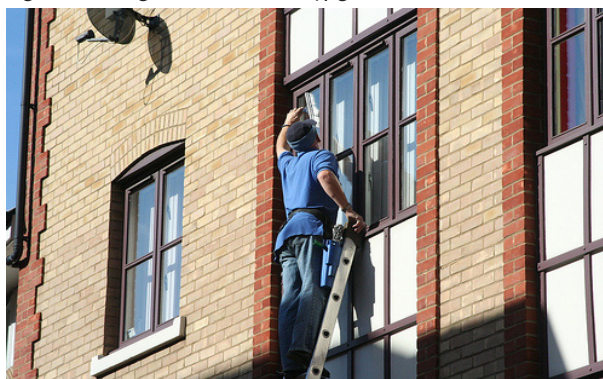
é recuperada para a construção de um discurso no cenário interacional. Nesse ponto, a produção das legendas é diferente, como se discute na seção 3. As legendas são produzidas em contextos controlados de elicitación, fora de um contexto interacional. E elas não resultam de experiência armazenada, mas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

## 5 A natureza descritiva das legendas de imagem

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

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

Figura 8: Imagem 1000344755.jpg



Fonte: dataset Framed Multizok.



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

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

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

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

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

Figura 9: Imagem 100207720.jpg



Fonte: dataset Framed Multizok.

- (23) #1 Mulher de óculos e blusa preta na cozinha polvilhando açúcar em um bolo de chocolate.
- (24) #3 Uma mulher de cabelos castanhos decora um bolo de chocolate polvilhando açúcar de confeitiro sobre ele.

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

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

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

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

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

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

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

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

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

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

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

Figura 10: Imagem 1001465944.jpg



Fonte: dataset Framed Multizok.

Figura 11: Imagem 1065323785.jpg do dataset



Fonte: dataset Framed Multizok.

(28) Quatro legendas da imagem 1001465944.jpg

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

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

Figura 7: Imagem 100197432.jpg



Fonte: dataset Framed Multizok.

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

(30) Quatro legendas da imagem 100197432.jpg

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

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

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

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

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

## Considerações finais

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

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

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

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

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

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

## Referências

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- CROFT, W. The origins of grammaticalization in the verbalization of experience. *Linguistics*, v. 48, n. 1, p. 1-48, 2010.
- CROFT, W. *Ten lectures on construction grammar and typology*. New York: Brill, 2020.
- DE MARNEFFE, Marie-Catherine *et al.* Universal dependencies. *Computational Linguistics*, v. 47, n. 2, p. 255-308, 2021.
- DIK, S. *The theory of functional grammar*. 2. ed. Berlin: Mouton de Gruyter, 1997.
- FRAMED MULTI30K. Imagem 1000092795.jpg. In: FRAMED MULTI30K. Banco de dados. [S. l.]: [S. n.]. Disponível em: Acesso em: <https://github.com/FrameNetBrasil/framed-multi30k>. Acesso em: 25 abr. 2025.
- GIVÓN, T. *Functionalism and grammar*. Amsterdam: John Benjamins, 1995.
- HALLIDAY, M.; HASAN, R. *Cohesion in English*. London: Longman, 1976.
- KEMMER, S.; BARLOW, M. Introduction: a usage-based conception of language. In: BARLOW, M.; KEMMER, S. (orgs.) *Usage-based models of language*. Sanford: CSLI Publications, 2000. p. vii-xxviii.
- JAIMES, A.; CHANG, S.-F. A Conceptual Framework for Indexing Visual Information at Multiple Levels. *IS&T/SPIE Internet Imaging*, v. 3964, s.p., 2000.
- MAURI, C. Ad hoc categorization in linguistic interaction. In: MAURI, C. *et al.* (org.) *Building categories in interaction: linguistic resources at work*. Amsterdam: John Benjamins, 2021. p. 9-34.
- MILTENVERG, E. Stereotyping and bias in the Flickr30K dataset. In: MULTIMODAL CORPORA: COMPUTER VISION AND LANGUAGE PROCESSING (MMC 2016). *Proceedings [...]*. [S. l.] p. 1-4, 2016. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-MCC-2016-proceedings.pdf>, Acesso em: 14 fev. 2024.
- NEVES, M. H. M. *Gramática de usos do português*. 2. ed. São Paulo: Editora Unesp, 2011.
- PARDO, T *et al.* Porttinari: a Large Multi-genre Treebank for Brazilian Portuguese. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DE INFORMAÇÃO E DA LINGUAGEM HUMANA. *Anais...* Porto Alegre: SBC, 2021. p. 1-10.
- PEZZATI, E. *A ordem das palavras no português*. São Paulo: Parábola, 2014.
- RINKE, E. A combinação de artigo definido e pronome possessivo na história do português. *Estudos de Linguística Galega*, v. 2, p. 121-139, 2010.
- SAMARIN, W. J. *Field Linguistics*. New York: Holt, 1967.
- SLOBIN, D. Language and Thought Online: Cognitive Consequences of Linguistic Relativity. In: GENTNER, D.; GOLDIN-MEADOW, S. (org.) *Language in Mind: Advances in the Study of Language and Thought*. Cambridge, MA: The MIT Press, 2003. p. 157-192.
- STRAKA, M. *et al.* UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In: CALZOLARI, Nicoletta *et al.* (org.) *c (LREC'16)*. European Language Resources Association, 2016. p. 4290-4297. Disponível em: <https://aclanthology.org/L16-1680.pdf>. Acesso em: 08 jan. 2025.
- TALMY, L. *Toward cognitive semantics*. Cambridge, MA: The MIT Press, 2000. 2 v.



VIRIDIANO, M. *Framed Mult30k*: um dataset multimodal-multilíngue baseado em semântica de frames. 2024. 107f. Tese (Doutorado em Estudos Linguísticos) – Universidade Federal de Juiz de Fora, 2024. Disponível em: <https://repositorio.ufjf.br/jspui/handle/ufjf/16854>. Acesso em: 14 abr. 2025.

VIRIDIANO, M. *et al.* Framed Mult30k: a frame-based multimodal multilingual dataset. *In*: THE 2024 JOINT INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, LANGUAGE RESOURCES AND EVALUATION (LREC-COLING 2024). *Proceedings* [...]. Torino, Italia, 2024. p.7438-7449. Disponível em: <https://aclanthology.org/2024.lrec-main.656.pdf>. Acesso em: 14 abr. 2025.

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

# Permanência semântica entre áudio original e legenda: um estudo sobre anotação semântica multimodal em obra audiovisual

## *Semantic Permanence between Original Audio and Subtitles: A Study on Multimodal Semantic Annotation in Audiovisual Work*

**Mariana Mota Silva Souza**

Universidade Federal de Minas Gerais  
(UFMG) | Belo Horizonte | MG | BR  
marianassouza.mota@gmail.com  
<https://orcid.org/0009-0005-0171-6312>

**Maucha Andrade Gamonal**

Universidade Federal de Juiz de Fora  
(UFJF) | Juiz de Fora | MG | BR  
CAPES  
mauchaandrade@gmail.com  
<https://orcid.org/0000-0002-5810-3913>

**Adriana Silvina Pagano**

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

**Resumo:** Com base na Semântica de *Frames* (Fillmore, 1982) e no Modelo da Primazia do Frame (Czulo, 2013), este estudo realiza uma análise multimodal das sentenças transcritas do áudio original em português, das legendas em inglês e das imagens correspondentes aos seis primeiros minutos do curta-metragem brasileiro *Eu não quero voltar sozinho*, de Daniel Ribeiro. A pesquisa objetiva investigar a similaridade semântica entre os *frames* evocados nas sentenças equivalentes, avaliar as escolhas tradutórias e analisar as relações semânticas estabelecidas entre texto e imagem. Adota a metodologia da FrameNet Brasil para anotações semânticas multimodais na plataforma Webtool, e a similaridade semântica entre *frames* é calculada por meio do algoritmo de *spread activation* e da métrica de similaridade de cossenos soft. As medidas organizam as sentenças em similaridade nula, parcial ou total, conforme a proposta de Samagaio (2023). As estratégias tradutórias de legendagem são analisadas conforme Díaz-Cintas e Remael (2021) e avaliadas quanto às relações semânticas entre texto e imagem. Os resultados evidenciam que as estratégias de tradução empregadas alteram a evocação de *frames* nas línguas analisadas, impactando a similaridade semântica entre sentenças. Este estudo contribui para o campo da tradução audiovisual e da



Semântica de *Frames*, ao explorar a construção de sentido em objetos multimodais e demonstrar como estratégias tradutórias influenciam na relação entre texto, imagem e *frames* evocados.

**Palavras-chave:** legendagem; anotação semântica multimodal; Semântica de *Frames*.

**Abstract:** Based on Frame Semantics (Fillmore, 1982) and the Frame Primacy Model (Czulo, 2013), this study performs a multimodal analysis of the sentences transcribed from the Portuguese original audio, the English subtitles and the images corresponding to the first six minutes of the Brazilian short film *Eu não quero voltar sozinho* (“I Don’t Want to Go Home by Myself”), by Daniel Ribeiro. This research aims to investigate the semantic similarity between the frames evoked in their equivalent sentences, evaluate the translation choices and analyze the semantic relations established between text and image. It adopts the FrameNet Brasil methodology for multimodal semantic annotations on the Webtool platform, and the semantic similarity between frames is calculated using the spread activation algorithm and the soft cosine similarity metric. The measures organize the sentences in null, partial or total similarity, as proposed by Samagaio (2023). The subtitling translation strategies are analyzed according to Díaz-Cintas and Remael (2021) and evaluated in terms of the semantic relations between text and image. Our results show that the translation strategies employed alter the evocation of frames in the languages analyzed, impacting the semantic similarity between sentences. This study contributes to the field of audiovisual translation and to Frame Semantics, by exploring the construction of meaning in multimodal objects and demonstrating how translation strategies influence the relationship between text, image and evoked frames.

**Keywords:** Subtitle; multimodal semantic annotation; Frame Semantics.

# 1 Introdução

A Semântica de *Frames*, desenvolvida por Fillmore (1982), é uma teoria linguística inserida nos estudos cognitivos da linguagem que postula que o significado das palavras é estruturado por meio de cenas, denominadas frames, evidenciando as relações entre experiência e linguagem. Essa abordagem tem sido utilizada nos Estudos da Tradução, em que Rojo e Ibarretxe-Antuñano (2013) afirmam que, assim como a linguagem, a tradução possui uma natureza cognitiva. Nesse contexto, Czulo (2017) propôs o Modelo da Primazia do Frame, segundo o qual o tradutor busca maximizar a similaridade semântica entre o texto fonte e o texto alvo.

Este artigo apresenta um estudo multimodal da tradução audiovisual, fundamentado na Semântica de Frames e no Modelo da Primazia do Frame, a partir da análise de um trecho do curta-metragem brasileiro *Eu não quero voltar sozinho*, produzido pela Lacuna Filmes em 2010 e dirigido por Daniel Ribeiro. A obra acompanha Leonardo, um adolescente com deficiência visual, em suas descobertas sobre sexualidade e amor ao conhecer Gabriel, um novo aluno na escola.

A pesquisa justifica-se pela escassez de estudos em português brasileiro que relacionem legendagem à Semântica de Frames. Investigações anteriores concentram-se na tradução de elementos culturais (Rojo, 2002; Bertoldi, 2016), audiodescrição fílmica (Souza *et al.*, 2022; Dornelas *et al.*, 2024) e criação de recursos tradutórios (Boas, 2013). Em relação à legendagem, destacam-se os trabalhos de Freitas (2022), que anotou o áudio original e a legenda de um corpus sem incluir uma análise multimodal, e de Samagaio (2023), que analisou a similaridade semântica em textos transcritos de áudio e legenda de um episódio da série *Pedro pelo mundo*, mas também sem incorporar o aspecto multimodal.

A presente pesquisa expande esses estudos ao incluir a análise multimodal, comparando frames evocados no texto transcrito do áudio original, na legenda em inglês e no vídeo. Os objetivos gerais do trabalho incluem contribuir para os Estudos da Tradução, para as pesquisas em Semântica de Frames e para a análise semântica de objetos multimodais, especialmente no campo da tradução audiovisual.

Os objetivos específicos da pesquisa são: (i) verificar os frames evocados na anotação de texto corrido do áudio original transcrito e da legenda, bem como na anotação de imagens dinâmicas do excerto do curta-metragem; (ii) avaliar a permanência semântica na legenda em inglês ao comparar os frames do áudio original com os da legenda; (iii) determinar como estratégias de tradução de legendagem influenciam a métrica de similaridade semântica entre o texto original e a legenda; e (iv) analisar as relações semânticas entre texto e imagem.

A metodologia consiste na anotação semântica da FrameNet Brasil voltada à multimodalidade. Para isso, foi utilizado o software Webtool. Além disso, aplicaram-se o algoritmo de spread activation e o cálculo de similaridade de cossenos soft para aferir a similaridade semântica entre os frames evocados nas sentenças equivalentes do áudio original e da legenda.

Este trabalho contribui para a ampliação do diálogo entre Semântica de Frames e Estudos da Tradução, investigando a interação entre texto e imagem em legendagem com base em uma abordagem multimodal. Ao considerar a interseção entre essas áreas, espera-se fornecer subsídios teóricos e metodológicos para futuras pesquisas que explorem a tradução audiovisual sob a perspectiva da linguística cognitiva.

## 2 Abordagem multimodal nos Estudos da Tradução e na Semântica de *Frames* e FrameNet Brasil

### 2.1 Estudos da Tradução e Legendagem

Os Estudos da Tradução, conforme Holmes (1972), podem ser categorizados em estudos puros e aplicados. No âmbito dos estudos puros, destacam-se os teóricos e os descritivos. Os estudos descritivos podem ser orientados ao produto, ao processo ou à função. O presente trabalho se insere no campo dos Estudos da Tradução como um estudo descritivo orientado ao produto, pois o objetivo é analisar o produto final de uma tradução audiovisual: a legenda em inglês do curta-metragem. Além disso, o estudo contribui para o campo dos estudos aplicados ao auxiliar na construção de um *corpus* anotado que pode ser empregado em tarefas de processamento de linguagem natural, como a tradução automática.

No que diz respeito aos tipos de tradução, Pagano *et al.* (2020) apontam que eles podem ser divididos em tradução no sentido restrito e interpretação no sentido amplo. A tradução no sentido restrito abrange tanto a tradução escrita quanto a tradução audiovisual. Esta última subdivide-se em linguagem visual-gestual e linguagem vocal-auditiva. A linguagem vocal-auditiva, por sua vez, inclui modalidades como a revocalização e a legendagem, sendo a última o foco de estudo desta pesquisa.

A legendagem envolve diversos aspectos que influenciam o processo e o produto da tradução, como o tempo de exibição do texto na tela, a sincronia com a fala e as restrições de caracteres, além das adaptações culturais necessárias. Díaz-Cintas e Remael (2007) definem a legendagem como:

uma prática tradutória que consiste na apresentação de um texto escrito, geralmente na parte inferior da tela, que busca narrar o diálogo original dos locutores, bem como os elementos discursivos que aparecem na imagem (cartas, encartes, grafite, inscrições, cartazes e similares) e as informações contidas na trilha sonora (músicas, *voices off*) (Díaz-Cintas; Remael, 2007, p. 8).

A legendagem, portanto, não se limita a uma simples transposição de uma língua para outra. Como destacam Díaz-Cintas e Remael (2021), trata-se de um processo de tradução da linguagem oral para a linguagem escrita. Isso demanda reformulações e interpretações do discurso para adaptá-lo ao meio escrito, o que frequentemente resulta na perda de certas características da oralidade, por exemplo. Os limites temporais e espaciais impostos pelas legendas originam uma linguagem marcada pela simplificação. Díaz-Cintas e Remael (2021) observam que:

A gramática, a sintaxe e os itens lexicais tendem a ser simplificados e apagados, enquanto os aspectos interacionais e a entonação são mantidas apenas até certo ponto (por exemplo, através da ordem das palavras, perguntas retóricas, interjeições ocasionais e frases incompletas) (Díaz-Cintas; Remael, 2021, p. 89).

Essas restrições estruturais refletem-se nas convenções específicas seguidas pelos tradutores, e o resultado tem impacto na experiência do espectador. De acordo com as conven-

ções, para evitar que as legendas desviem o foco da narrativa audiovisual, elas devem ocupar até duas linhas de texto e ocupar no máximo um sexto da tela. Além disso, as legendas são localizadas no centro da tela, na parte inferior, podendo ser deslocadas para não se sobreponham a outros textos na tela. Devem ser sincronizadas com o áudio e ter duração de cinco a seis segundos. Para facilitar a leitura, as fontes não devem possuir serifas, como Arial ou Helvetica, e devem estar em cores como branco ou amarelo.

As reduções mencionadas inserem a necessidade de estratégias de redução textual. Segundo Díaz-Cintas e Remael (2021), a reprodução fiel das falas na legenda não é necessária, pois ela é interpretada em conjunto com as imagens e os sons do produto audiovisual. Além disso, como eles destacam, as legendas devem respeitar o tempo que o espectador leva para absorver o texto e permitir que experiencie também as imagens e a trilha sonora.

Para os autores, a redução pode ser parcial, condensando o texto, ou total, eliminando elementos textuais. No nível da palavra, Díaz-Cintas e Remael (2021) destacam a simplificação de perífrases verbais, generalização de enumerações, uso de sinônimos ou expressões equivalentes. Já no nível da sentença, observam-se várias mudanças, como no tipo de oração (se negativa ou afirmativa, interrogativa ou imperativa etc.), alteração de voz ativa para passiva, fusão de orações, uso de dêiticos no lugar de grupos nominais e junção de duas ou mais sentenças.

Omissões também desempenham um papel relevante na legendagem. No nível da palavra, itens como modificadores (adjetivos e advérbios), interjeições e vocativos podem ser omitidos sem comprometer a mensagem principal. No nível da sentença, exclusão de orações inteiras podem ocorrer em situações como sobreposição de diálogos ou quando as falas apresentam baixa contribuição para a narrativa. Segundo Díaz-Cintas e Remael (2021), o tradutor deve priorizar a integridade da mensagem e a experiência do espectador ao optar por tais omissões.

Com base nas estratégias listadas, este estudo reforça a importância de compreender a legendagem como um processo criativo e ao mesmo tempo técnico, que demanda decisões criteriosas capazes de conciliar a fidelidade ao conteúdo a partir das limitações do formato.

## **2.2 Semântica de *Frames***

A Semântica de *Frames*, teoria linguística proposta por Charles Fillmore em 1982 é descrita pelo próprio autor como “um programa de pesquisa sobre semântica empírica e um modelo descritivo para apresentar os resultados de tal pesquisa” (Fillmore, 2009, p. 25). Em termos práticos, a Semântica de *Frames* constitui tanto uma base estruturada para investigação sobre significação linguística, baseada em corpora, quanto um modelo para a apresentação e descrição sistemática dos resultados obtidos.

O objetivo central de Fillmore com a Semântica de *Frames* era compreender as relações entre a linguagem e a experiência, investigando como as sociedades categorizam o mundo e como essas categorias são codificadas linguisticamente. Para isso, Fillmore (2009) introduziu o conceito de *frame*, definido como:

qualquer sistema de conceitos relacionados de tal modo que, para entender qualquer um deles, é preciso entender toda a estrutura na qual se enquadram;

quando um dos elementos dessa estrutura é introduzido em um texto, ou em uma conversa, todos os outros elementos serão disponibilizados automaticamente (Fillmore, 2009, p. 25).

Os *frames*, portanto, são representações esquemáticas de situações que englobam elementos interconectados. Sua compreensão está intrinsecamente ligada ao entendimento de instituições sociais, experiências humanas e aspectos culturais. Assim, segundo a teoria, os significados linguísticos são inseparáveis de seu contexto, representado por *frames* semânticos.

Esses *frames* podem ser evocados pelo texto por meio de unidades lexicais (LU, do inglês *Lexical Unit*), que combinam uma forma linguística (lema) e um significado, ou podem ser invocados, por inferência, pelo intérprete a partir de pistas contextuais, mesmo na ausência de material linguístico.

A teoria de Fillmore está inserida no campo de estudos da linguística cognitiva, uma abordagem que integra a língua às demais faculdades cognitivas humanas, como a memória e o raciocínio (Rojo; Ibarretxe-Antuñano, 2013). Essa teoria refuta o gerativismo chomskyano, predominante na década de 1970, que se concentrava nos padrões linguísticos como internos à estrutura da língua e ignorava o papel da semântica e do contexto na construção de significados.

A Semântica de *Frames* apresenta uma aplicação particularmente relevante no campo dos Estudos da Tradução. Como explicam Rojo e Ibarretxe-Antuñano (2013):

A Linguística Cognitiva defende a natureza cognitiva da tradução como um processo de mediação entre dois mundos conceituais diferentes. Além disso, a sua visão integrada da língua e da cognição, juntamente com o papel crucial da cultura, ajuda a reforçar a ligação entre o comportamento do tradutor e as estratégias cognitivas que conduzem a tal comportamento, fortalecendo assim a ligação entre o produto e o processo de tradução (Rojo; Ibarretxe-Antuñano, 2013, p. 19).

De acordo com as autoras, o tradutor utiliza não apenas seu conhecimento linguístico das línguas de origem e de chegada, mas também suas experiências, seu conhecimento de mundo e sua compreensão das culturas que permeiam ambas as línguas. Esse processo envolve adaptar o texto de forma que ele seja acessível e compreensível à audiência da língua-alvo. Isso ressalta a importância de teorias como a Semântica de *Frames* para compreender a tradução como um processo cognitivo e culturalmente situado, indo além da transposição lexical.

Um modelo particularmente relevante que articula a teoria com os estudos da tradução é o Modelo de Primazia do *Frame* desenvolvido por Czulo (2017). Após extensa revisão da literatura, Czulo avalia que, no processo tradutório, o tradutor busca preservar o conteúdo semântico da língua de origem na língua de chegada. Sob essa ótica, o modelo analisa os processos cognitivos envolvidos na tradução e se propõe a mapear os *frames* evocados em cada idioma, visando a uma comparabilidade máxima.

O modelo sugerido pelo autor reforça o potencial da Semântica de *Frames* e da sua contraparte prática a FrameNet para contribuir tanto para a descrição dos processos tradutórios como também para o entendimento do processo de tradução como atividade cognitiva situada culturalmente.



## 2.3 FrameNet Brasil numa abordagem multimodal


A FrameNet é um projeto de Lexicografia Computacional iniciado por Fillmore em 1997 no *International Computer Science Institute*, em Berkeley, com o objetivo de descrever a língua inglesa tanto semântica quanto sintaticamente. Essa descrição é baseada na análise linguística de sentenças extraídas de corpora, orientada pelos princípios da Semântica de *Frames*. O projeto tem ampla aplicabilidade em áreas como tradução automática e Processamento de Língua Natural (PLN) (Torrent *et al.*, 2024).

Com os anos, o recurso foi adaptado para diversas línguas, incluindo o português brasileiro. A versão brasileira, a FrameNet Brasil (FN.Br), é desenvolvida no laboratório de Linguística Computacional da Universidade Federal de Juiz de Fora. A FN.Br se destaca pelas iniciativas multimodais e pela colaboração com instituições parceiras no desenvolvimento de aplicações baseadas na teoria.

O projeto ReINventa (*Research and Innovation Network for Visual and Textual Analysis of Multimodal Objects*) se destaca nesse sentido. Conduzido em colaboração com universidades parceiras, os objetivos incluem a entrega de *dataset* de objetos multimodais anotados semanticamente e a criação de algoritmos de inteligência artificial para rotulação automática e descoberta de conhecimento em objetos multimodais (Torrent *et al.*, 2022).

No glossário de uma *framenet*, destacam-se alguns conceitos, como *frames*, elementos de *frame*, relações entre *frames* e unidades lexicais. Os *frames* são representações linguístico-computacionais do conhecimento humano motivados lexicograficamente. Para que sejam modelados, são definidos por meio de seus elementos, denominados Elementos de *Frame* (FE, do inglês *frame element*). Estes podem ser nucleares e não nucleares. Os FE nucleares são essenciais para a compreensão do *frame*, diferenciando-o de outros, enquanto os FE não nucleares acrescentam informações complementares. A Figura 1 ilustra o *frame* Chegar<sup>1</sup> modelado na FN.Br.

Figura 1: Definição e FE nucleares do *frame* Chegar

 @Generic #48 Arriving [en] PDF	
<b>Definition</b>	
Um <b>Tema</b> se move na direção de um <b>Alvo</b> . O <b>Alvo</b> pode ser expresso ou pode ser entendido a partir do contexto, mas é sempre implícito no próprio verbo.	
<b>Frame Elements</b>	
<b>Core</b>	
<b>Alvo</b>	O <b>Alvo</b> é qualquer expressão que diz onde o tema acaba, ou iria acabar, como resultado do movimento. @location
<b>Tema</b>	O <b>Tema</b> é o objeto que se move. Pode ser uma entidade que se move sob seu próprio poder, mas não precisa ser. @physical_object

Fonte: Captura de tela da webtool 4.02

<sup>1</sup> Por convenção, o nome do *frame* é registrado em fonte Courier New.

O evento de chegar é definido por um Tema que se move em direção de um Alvo. Alvo e Tema são FE nucleares, como ilustrados na sentença (1) em que a LU *chegar.v* em destaque é antecedida pelo FE Tema “eu”. Depictivo e Frequência são FE não nucleares instanciados nesta sentença.

- (1) Às vezes, raro, esse ano eu acho que [eu<sub>TEMA</sub>] **CHEGUEI**<sup>CHEGAR</sup> [atrasada<sub>DEPICTIVO</sub>] [duas vezes<sub>FREQUÊNCIA</sub>], porque com essas obras do Metrô, com congestionaman ... congestionamento [#7873]

Quando um FE nuclear não está expresso em uma sentença, ele pode ser classificado em tipos de instanciação nula, que é uma maneira de registrar a motivação de sua ausência. Na sentença (1), o FE Alvo não é instanciado linguisticamente e não é recuperável na sentença. Desse modo, prevê-se essa anotação semântica como uma Instanciação Nula Indefinida (INI, do inglês *Indefinite Null Instantiation*). Os outros dois casos de instanciação nula previstos numa FrameNet é a Nula Definida (DNI) e a Construcional (CNI). A DNI ocorre quando o elemento de *frame* pode ser recuperado pelo contexto linguístico, como é o caso de anáforas e elipses. E a CNI ocorre em casos em que a construção linguística licencia a ausência do FE, como no modo imperativo.

Um aspecto fundamental da FrameNet é a capacidade de estabelecer relações semânticas entre si, o que define o projeto em uma rede semântica. As relações tradicionais previstas entre os *frames* são: Herança, Uso, Subframe, Perspectiva, Precedência e Causativo\_de / Incoativo\_de.

Na relação Herança, um *frame* gera outro mais específico. *Percepção\_ativa*, por exemplo, é herdeiro dos *frames* *Agir\_intencionalmente* e *Percepção* e é herdado por *Fazer\_turismo*. Uma consequência da relação de herança é que os *frames* mãe e filhos possuem FE correspondentes iguais ou semelhantes entre si. *Percepção\_ativa* tem por FE nucleares *Fenômeno* e *Perceptor*, enquanto os FE nucleares de *Percepção* são *Fenômeno* e *Perceptor*.

Na relação Uso, um *frame* faz referência a outro, mais abstrato. *Percepção\_ativa* estabelece relação de uso com *Atenção*, pois considera-se que não seja possível compreender o primeiro sem o conhecimento sobre o segundo. Nesse tipo de relação, não há, necessariamente, correspondência entre os FE dos dois *frames*.

Já na relação de *Subframe*, um *frame* é subevento de um que representa um evento mais complexo. Frequentemente, há uma relação de sequência temporal ou estados de coisas entre mãe e filho. A relação de Precedência é um tipo de relação *Subframe* que ocorre quando os subeventos são ordenados temporalmente. *Atividade\_em\_andamento*, por exemplo, é precedido por *Atividade\_iniciar* e precede *Atividade\_terminar*. Todos os três são *Subframes* de *Atividade*.

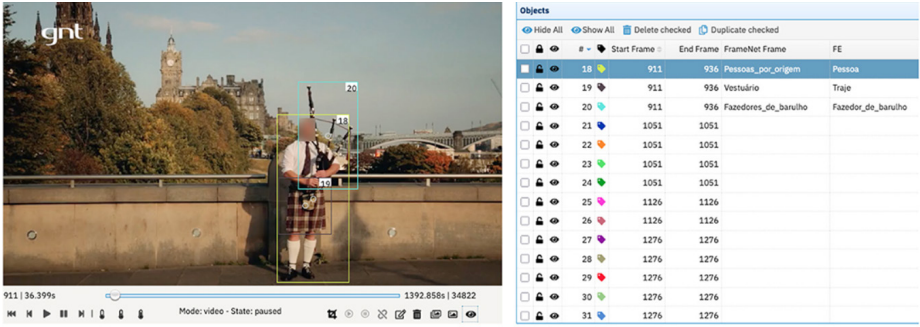
Na relação de perspectiva, os *frames* compartilham um mesmo cenário, mas focalizam aspectos diferentes. Por exemplo, *Comércio\_transferência\_de\_mercadorias* pode ser visto da perspectiva do comprador em *Comércio\_comprar* ou do vendedor em *Comércio\_vender*.

As relações *Causativo\_de* / *Incoativo\_de* ocorrem entre os *frames* chamados de causativos, incoativos e estativos. Nos *frames* causativos, o sujeito é o agente causador da ação; nos incoativos, o sujeito não é o causador da ação, e os estativos representam os casos em que o

sujeito é o experienciador de um estado. A relação Causativo\_de ocorre entre *frames* que causam outros *frames*, como é o caso de Causar\_emoção, um *frame* causativo, que é causativo de Emoção\_com\_foco\_no\_experienciador, um *frame* incoativo. Já a relação Incoativo\_de ocorre em Experienciar\_ferimento\_corporal, que é incoativo do estativo Estar\_ferido.

Ao adotar uma abordagem multimodal, a FB.Br assume que a linguagem verbal assim como imagens, gestos e outros modos semióticos atuam de forma interdependente na construção de sentido. Bateman *et al.* (2017) propõem uma taxonomia de classificação dos gêneros multimodais e discutem a centralidade de ferramentas computacionais que auxiliam na análise de conjuntos de dados multimodais. Nesse sentido, Torrent *et al.* (2022) apresentam um panorama do aperfeiçoamento da FN.Br de modo a incluir uma abordagem multimodal e multidimensional para representar sentido e garantir inferências contextuais na integração de objetos modais diversos.

Figura 2: Anotação de vídeo na FN.Br



Fonte: Torrent *et al.* (2022, p. 12)

A Figura 2 é um exemplo em que os autores mostram a anotação semântica da imagem do vídeo de um programa televisivo brasileiro, associada à sentença (2) proferida pelo apresentador do programa.

(2) Quando a gente pensa na Escócia, a primeira coisa que vem à mente é homem de saia, uísque escocês e gaita de fole.

Os autores discutem que a sentença é proferida alguns segundos antes de a imagem ser exibida na tela, o que torna necessário um ajuste de *frames*, homem [de saia], que, primeiramente, é anotado no *frame* Pessoa, a partir da exibição da imagem, sinaliza a necessidade de reajuste de anotação para outro mais específico, o Pessoa\_por\_etnia. Esse exemplo reforça a relevância de representar informações contextuais capturadas por informações não advindas de textos verbais, como é o caso do enquadramento situacional proporcionado pela imagem mostrada.

### 3 Material de análise e anotação semântica multimodal

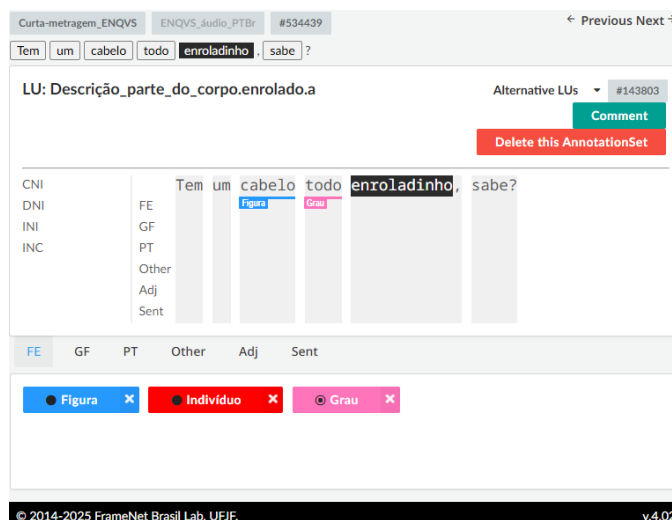
O material de análise utilizado neste estudo consiste no texto transcrito do áudio original em português, nas legendas em inglês e nas imagens dinâmicas do curta-metragem brasileiro *Eu não quero voltar sozinho*, de 2010, dirigido e roteirizado por Daniel Ribeiro e produzido pela Lacuna Filmes. A trama acompanha Leonardo, um adolescente com deficiência visual que, ao longo da narrativa, explora questões relacionadas ao amor, à amizade e à descoberta de sua sexualidade. A duração total do curta é 17 minutos, e o excerto analisado neste trabalho se concentra nos seis minutos e quatro segundos iniciais.

A coleta e o processamento do material incluem a transcrição do áudio original e a organização das legendas em inglês em um editor de texto, segmentando-as em sentenças alinhadas. Posteriormente, as sentenças foram inseridas na ferramenta para anotação multimodal da FN.Br, atualmente integrada na Webtool 4.02. Por meio dela, foram realizadas a anotação estrutural das sentenças do áudio em português. Nesta etapa, cada sentença recebe uma etiqueta especificando a origem do texto (no caso, áudio original e legenda) e faz-se a delimitação do tempo inicial e final.

Feita a anotação estrutural, o próximo passo é a anotação semântica de texto. Para essa etapa, foi feita a anotação de texto corrido (no inglês, *full text annotation*), que consiste em anotar todas as LU de cada sentença. Uma vez concluída a anotação de texto verbal, inicia-se a anotação de imagens dinâmicas. Nesta etapa, selecionam-se objetos visuais nas imagens do vídeo com base na anotação de texto corrido. Para cada objeto visual, identificam-se *frame*, FE e LU, esta última associada a um modelo de visão computacional utilizado para tarefas de reconhecimento automático de objeto visual nas imagens que vem sendo refinado com informações dos *frames* semânticos.

A Figura 3 exibe a anotação de texto da LU *enroladinho.a* na sentença “*Tem o cabelo todo enroladinho*”. Para essa LU, o *frame* *Descrição\_parte\_do\_corpo.enrolado.a* foi evocado.

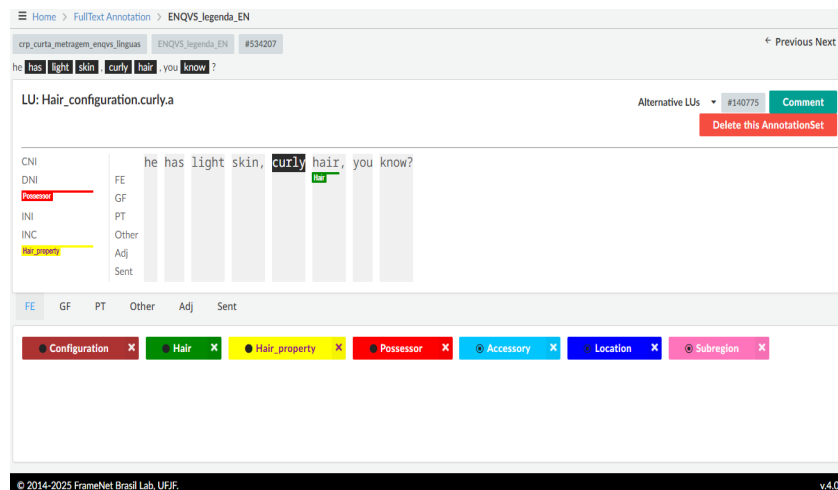
Figura 3: Anotação semântica da LU *enroladinho.a*



Fonte: Captura de tela da FN.Br 4.02

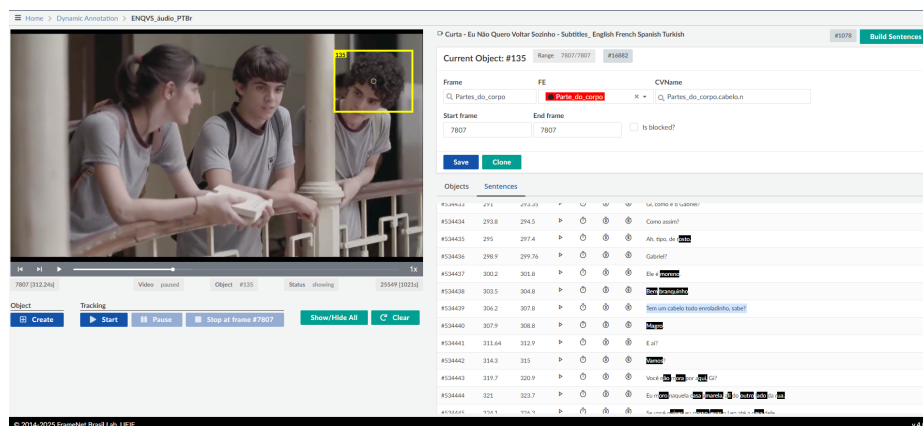
A legenda dessa mesma cena é *He has light skin, curly hair, you know?* A Fig. 4 mostra a LU *curly*.a anotada em *frame* mais específico que o escolhido para a sentença em pt.br, o *Hair\_configuration*.

Figura 4: Anotação semântica da LU *curly*.a



Fonte: Captura de tela da FN.Br 4.02

Figura 5: Anotação de imagem dinâmica do objeto visual *cabelo.n*



Fonte: Captura de tela da FN.Br 4.02

A Figura 5 apresenta uma tela do *software* de anotação, em que o cabelo do personagem Gabriel foi marcado como objeto visual, tendo como referência a sentença da Fig. 3. Para isso, foi selecionado o *frame* *Partes\_do\_corpo* e FE *Parte\_do\_corpo*. A LU é *cabelo.n* e está vinculada ao *frame* *Partes\_do\_corpo*.

## 4 A medida de similaridade semântica entre áudio original e legenda

Após a anotação do material linguístico, a medida de similaridade semântica foi realizada utilizando o algoritmo de *spread activation* em conjunto com a similaridade de cossenos. O algoritmo calculou o decaimento de energia entre nós em um gráfico semântico com valores no intervalo de  $[0,1]$ . A metodologia seguiu dois estágios principais: o pré-ajuste, no qual foi calculado o valor de energia inicial para cada nó, e a propagação de energia, em que o decaimento foi determinado ao longo da rede semântica. O estágio de pós-ajuste não foi realizado, pois o gráfico da FrameNet é acíclico, como explicado por Viridiano *et al.* (2022). Nesse modelo, cada nó representa um *frame* evocado diretamente por uma sentença ou indiretamente por meio de relações entre *frames*, e os valores atribuídos refletem a proximidade entre eles na rede semântica.

A similaridade de cossenos foi então utilizada para medir o grau de similaridade semântica entre sentenças equivalentes no áudio original em português e nas legendas em inglês. Os resultados foram categorizados em três níveis. A similaridade total de valor 1 ocorreu quando todos os *frames* evocados pelas duas sentenças foram equivalentes. A similaridade parcial foi registrada quando alguns *frames* eram equivalentes e outros não, com valores entre 0 e 1. Por fim, a similaridade nula foi subdividida em dois tipos: o valor 0 foi atribuído a sentenças sem qualquer coincidência ou comparabilidade entre os *frames* evocados, enquanto o valor -1 foi aplicado a casos específicos, como sentenças sem anotação de *frames* ou que não foram traduzidas para o inglês. Essa distinção refina a categorização proposta por Samagaio (2023), que havia atribuído valor 0 a todos esses casos.

A análise dos resultados levou em conta estratégias de tradução descritas por Díaz-Cintas e Remael (2021), garantindo um olhar detalhado sobre as relações semânticas nas legendas. A classificação geral como similaridade semântica nula foi mantida para as sentenças sem tradução ou sem ativação de *frames*, seguindo Samagaio (2023). No entanto, ao adotar valores distintos de 0 e -1, a pesquisa oferece maior precisão ao modelo proposto, sem comprometer a comparabilidade com estudos anteriores.

## 4 Resultados e discussão

### 4.1 Anotação semântica multimodal em números: áudio original, legenda e vídeo

O total de *frames* anotados para o texto da legenda em inglês foi inferior ao número registrado na anotação do áudio original. Esse resultado era esperado considerando que o número de sentenças na legenda do curta é menor que o total de falas transcritas do áudio original, como mostra a Tabela 1.

Tabela 1: Levantamento das anotações de texto corrido

Anotação do áudio original		Anotação da legenda	
Frames distintos	Frames totais	Frames distintos	Frames totais
105	235	104	203
45%	100	51%	100%

Fonte: Elaborada pelos autores

As tabelas 2 e 3 mostram os *frames* mais frequentes nas anotações de texto. Foram considerados mais frequentes os *frames* com 5 ocorrências ou mais.

Tabela 2: *Frames* mais frequentes no áudio original

Frame	Ocorrências	Porcentagem
Negação	16	7%
Desejar	9	4%
Frequência	8	3%
Evento_desejável	5	2%
Dizer	5	2%
Conhecimento	5	2%
Outros	188	80%
Total	235	100%

Fonte: Elaborada pelos autores

Tabela 3: *Frames* mais frequentes na legenda

Frame	Ocorrências	Porcentagem
Motion	8	4%
Direction	6	3%
Frequency	6	3%
Body_parts	6	3%
Experiencer_ focused_emotion	5	2%
Certainty	5	2%
Animals	5	2%
Outros	162	81%
Total	203	100%

Fonte: Elaborada pelos autores



A partir das tabelas 2 e 3, é possível observar que estratégias de tradução audiovisual como a omissão de itens lexicais e o uso de contrações interferiram na permanência semântica entre áudio original e legenda do curta. O *frame* *Negação*, por exemplo, foi o mais frequente na anotação do português, mas sua contraparte para o inglês no *frame* *Negation* não está na lista dos mais frequentes.

Tabela 4: Levantamento da anotação de vídeo

Frames	Ocorrências	Porcentagem
Partes_do_corpo	13	8%
Pessoas	8	5%
Conhecimento	7	5%
Conhecer_pessoas	7	5%
Familiaridade	6	4%
Atividade	6	4%
Atividade_pausar	5	3%
Desejar	5	3%
Transportar	5	3%
Outros	93	60%
Total	155	100%

Fonte: Elaborada pelos autores

Quanto à anotação semântica de imagens do vídeo, foram registrados 155 frames semânticos no recorte escolhido. A Tabela 4 especifica os *frames* mais frequentes atribuídos aos objetos visuais. A respeito dos *frames* registrados a partir das LUs identificadas nas imagens, o *frame* *Pessoas* assume destaque com 120 ocorrências em comparação com 8 ocorrências na anotação de objetos visuais. Isso ocorreu, pois a maioria dos objetos visuais anotados foram os personagens do curta, descritos como pessoas pela LU.

## 4.2 Similaridade semântica em números: áudio original e legenda

Segundo classificação de Samagaio (2023), similaridades com valores entre 0,01 e 0,40 foram consideradas baixas; similaridades entre 0,41 e 0,69 foram consideradas médias, e similaridades entre 0,70 e 0,99 foram consideradas altas. Samagaio (2023) constatou que a similaridade nula ocorreu em três casos: quando não houve comparabilidade ou equivalência entre os *frames* evocados nas sentenças, quando uma das sentenças não evocou *frame*, e quando a sentença não foi traduzida na legenda. No trabalho de Samagaio (2023), todos esses casos foram representados pelo valor 0, no entanto, nesta pesquisa, os dois últimos casos receberam o valor -1. A similaridade total foi representada pelo valor 1.

A Tabela 5 mostra a frequência dos níveis de similaridade semântica obtidos no cálculo da similaridade de cossenos. A similaridade parcial representou 54% dos casos. A similaridade nula foi a segunda mais frequente, com 33% do total, somando as similaridades de

valor 0 e -1. No entanto, considerando apenas a similaridade nula de valor 0, em que ambas as sentenças evocaram *frames*, esta passa a ser a menos frequente, com 4% dos casos.

Tabela 5: Níveis de similaridade semântica

Similaridade	Ocorrências	Porcentagem
1	13	13%
de 0,01 a 0,40	20	19%
de 0,41 a 0,69	22	21%
de 0,70 a 0,99	15	14%
0	4	4%
-1	30	29%
Total	104	100%

Fonte: Elaborada pelos autores

Nota-se que, embora considerado por Czulo (2017) no modelo da Primazia do *Frame* que o tradutor busque a permanência semântica máxima entre texto original e texto traduzido, o nível de similaridade menos frequente foi o de similaridade total de valor 1, com 13 ocorrências, equivalentes a 13% do total. Somada à porcentagem das sentenças com similaridade alta, 27% do total obteve similaridade igual ou maior que média, indicando comparabilidade parcial entre os *frames* das sentenças.

#### 4.2.1 Similaridade semântica total

A similaridade semântica total foi indicada com o valor 1. Nesses casos, os *frames* evocados foram idênticos nas duas línguas, como se pode observar no exemplo (2):

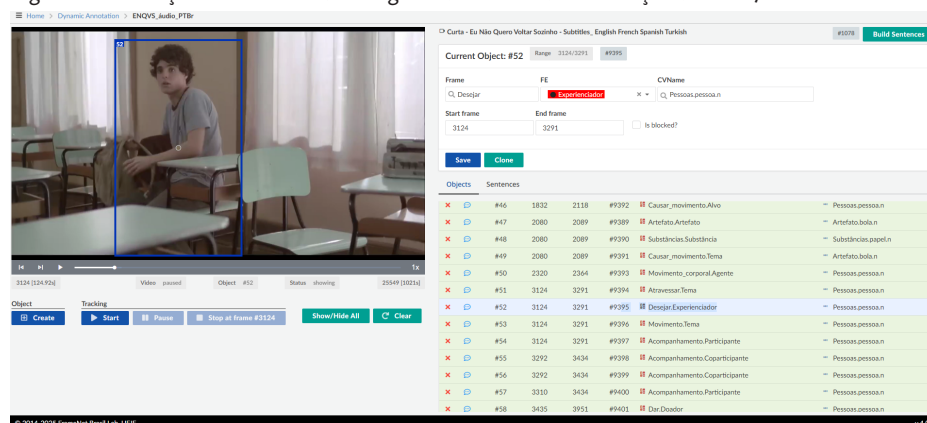
(2) **QUER**<sup>DESEJAR</sup> **VIR**<sup>MOVIMENTO</sup> com a gente? (628267)<sup>2</sup>

*Do you* **WANT**<sup>DESIRING</sup> *to* **COME**<sup>MOTION</sup> *with us?* (628060)

Nesse caso, a interação entre a sentença e a imagem foi a mesma para áudio original e legenda, em termos de *frames* evocados. A Figura 6 mostra a anotação para a sentença 628267.

<sup>2</sup> As sentenças nos exemplos contêm o número identificador no *corpus* da FN.Br.

Figura 6: Anotação semântica de imagem dinâmica da sentença id 628267



Fonte: Captura de tela de anotação semântica multimodal na FN.Br 4.02

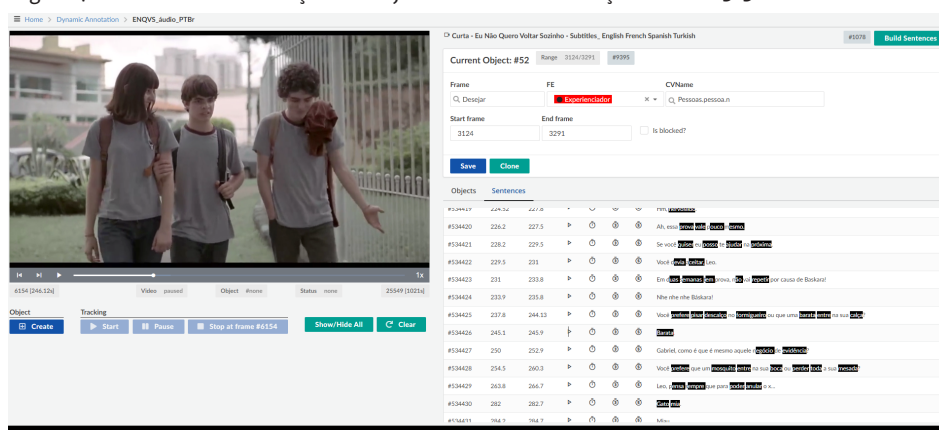
Na Figura 6, a personagem Gabriel foi delimitada como um objeto no *frame* *Desejar*, instanciando o elemento de *frame* *Experienciador*, pois, na sentença que embasa essa anotação, esse FE é instanciado no sujeito “[você] quer vir com a gente?”, que se refere à personagem. Como todos os *frames* estão presentes nas duas línguas, as duas sentenças interagem com o vídeo da mesma maneira, do ponto de vista dos *frames* evocados.

(3) **BARATA**<sup>ANIMAIS</sup>. (628313)

*The* **COCKROACH**<sup>ANIMALS</sup>! (628106)

Da mesma forma, houve sentença com similaridade total que não gerou anotação no vídeo, pois as entidades instanciadas pelos FE não possuíam objetos visuais que as representassem na tela, como é o caso no exemplo (3).

Figura 7: Cena sem demarcação de objeto visual na sentença de id 628313



Fonte: Captura de tela de anotação semântica multimodal na FN.Br 4.02

Nessa cena, mostrada na Figura 7, as personagens discutem se preferem pisar descalços no formigueiro ou que uma barata entre em suas calças. Como é uma discussão de uma situação hipotética, a barata não aparece na tela.

### 4.2.2 Similaridade semântica nula

A similaridade semântica nula ocorre em três casos: (i) não existe convergência de *frames* entre as sentenças; (ii) uma das sentenças não evoca *frame*, e (iii) não existe uma sentença equivalente em uma das línguas.

O primeiro caso resulta em similaridade semântica com valor 0 e pode ser observada no exemplo (4):

(4) **SUBIR**<sup>ATRAVESSAR</sup>. (628266)

**DOWN**<sup>DIRECTION</sup>. (628059)

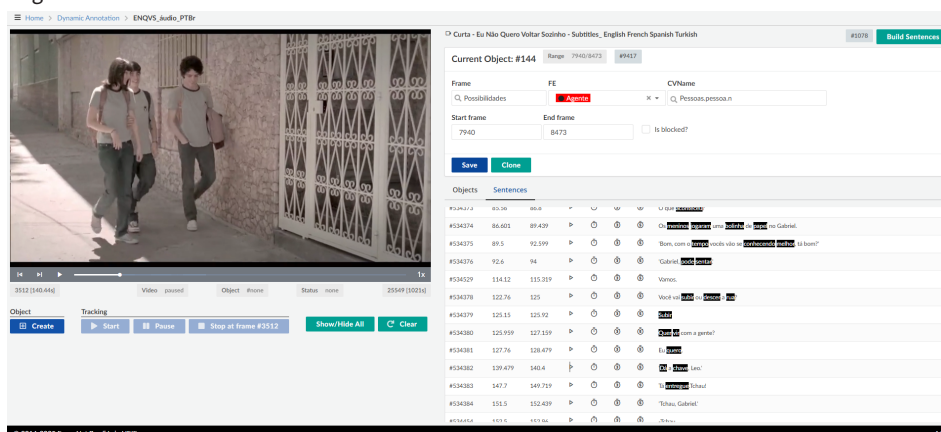
O exemplo (4) ilustra a natureza multimodal da tradução audiovisual, que deve levar em consideração as imagens na tela em busca de sentidos na língua alvo. No áudio original, com intenção de convidar o estudante novato a se juntar aos dois, a personagem Giovana faz uma pergunta a ele, exibida no exemplo 5 com a respectiva legenda produzida para o inglês:

(5) Você vai **SUBIR**<sup>ATRAVESSAR</sup> ou **DESCER**<sup>ATRAVESSAR</sup> a **RUA**<sup>VIAS</sup>?

Is your **HOUSE**<sup>BUILDINGS</sup> **UP**<sup>DIRECTION</sup> or **DONW**<sup>DIRECTION</sup> the **STREET**<sup>ROADWAYS</sup>?

As respostas em português e em inglês, mostradas no exemplo (5), parecem contraditórias se se leva em conta a cena seguinte, que mostra os personagens descendo a rua. Em português, o áudio, certamente, justifica-se por eles terem realizado um caminho mais longo, inicialmente subindo a rua, mas, em inglês, a ênfase, certamente, foi dada à cena que aparece em seguida, como mostra a Figura 8.

Figura 8: Cena da descida até a casa de Leonardo

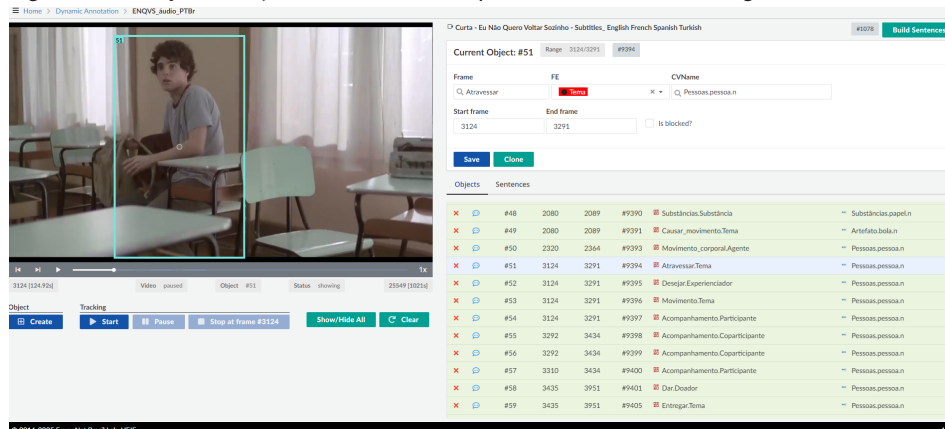


Fonte: Captura de tela de anotação semântica multimodal na FN.Br 4.02

O exemplo (4) apresentado anteriormente ilustra caso em que a anotação dos *frames* das imagens do curta não se alinha com aqueles evocados na legenda, como mostra a Figura 9. Nesse caso, a personagem Gabriel instancia o FE Tema do *frame* Atravessar evocado pela LU *subir.v*. De verbo no áudio, a opção tradutória na legenda foi um advérbio, que evocou

Direction, por esse motivo, Gabriel não é mais marcado como entidade que instancia um FE, pois o *frame* em questão apresenta como FE nucleares Ponto\_de\_referência e Trajetória.

Figura 9: Marcação de objeto visual não correspondente entre áudio e legenda



Fonte: Captura de tela de anotação semântica multimodal na FN.Br 4.02

O segundo caso de similaridade nula possui valor -1 e ocorre quando não há equivalente de uma sentença na outra língua, como nos exemplos (6) e (7):

(6) **QUARTO** NÚMEROS\_ORDINAIS **ANDAR** SUBPARTES\_DE\_PRÉDIOS! (628246)

(7) Eu **VOU** MOVIMENTO no **DÉCIMO** NÚMEROS\_ORDINAIS! (628247)

Essas sentenças não foram traduzidas para o inglês, configurando uma estratégia de tradução de legendagem prevista por Díaz-Cintas e Remael (2021) chamada omissão. As falas acontecem em uma cena na qual os alunos brincam sobre o barulho da máquina de escrever de Leonardo parecer o som de um elevador. Além de a sentença ser proferida rapidamente, no meio de várias outras falas, e não ter destaque na composição da cena, ela é dita por personagens figurantes que não aparecem na tela. Por esse motivo, as sentenças não estão atreladas a objetos visuais evocadores de *frames*, portanto a ausência na legenda se justifica por fatores inerentes à multimodalidade da obra audiovisual.

(8) Oi Gabriel! (628258)

Hi Gabriel! (628052)

(9) **DESENCANA** PREOCUPAR\_SE. (628290)

Nevermind. (628082)

(10) Como assim? (628321)

O terceiro caso também é indicado com o valor -1 e ocorre quando pelo menos uma das sentenças não evoca *frame*. De acordo com a metodologia de anotação da FN.Br, expres-

sões com enfoque pragmático<sup>3</sup> e nomes próprios são tipos de casos não anotados. As sentenças em (8), (9) e (10) mostram alguns exemplos.

### 4.2.3 Similaridade semântica parcial

Similaridades de cosseno entre 0 e 1 indicam uma convergência parcial dos *frames*. Vários casos de similaridade parcial ocorreram por consequência das estratégias de tradução da legendagem. O exemplo (11) mostra um caso de simplificação de perífrase verbal:

(11) E aí, **JÁ**<sup>VETOR\_TEMPO</sup> tá **CONSEGUINDO**<sup>CAPACIDADE\_DE\_AÇÃO</sup> se **ACHAR**<sup>LOCALIZAR</sup> na **ESCOLA**<sup>LOCAIS\_POR\_USO?</sup> (628293)

So, are you **FINDING**<sup>LOCATING</sup> your **WAY**<sup>SELF\_MOTION</sup> around **SCHOOL**<sup>LOCALE\_BY\_USE?</sup> (628086)

Consequentemente, o frame *Capacidade\_ação* não é evocado pela legenda, devida à omissão da LU *conseguir.v*. Da mesma forma, a LU *já.adv* é omitida na legenda, eliminando, assim, *Vetor\_tempo*. Sua ausência não altera a relação estabelecida entre a sentença e as imagens, pois não houve FE correspondente a um objeto visual nesse caso. A similaridade entre as sentenças foi média, equivalente a 0,68.

(12) **SE**<sup>OCORRÊNCIA\_CONDICIONAL</sup> você **QUISER**<sup>DESEJAR</sup> eu **POSSO**<sup>POSSIBILIDADES</sup> te **AJUDAR**<sup>ASSISTÊNCIA</sup> na **PRÓXIMA**<sup>SEQUÊNCIA</sup>. (628308)

I **CAN**<sup>POSSIBILITY</sup> **HELP**<sup>ASSISTANCE</sup> you for the **NEXT**<sup>RELATIVE\_TIME</sup> **TEST**<sup>EXAMINATION</sup>. (628100)

O exemplo (12) ilustra a estratégia de simplificação de oração composta e obteve similaridade baixa, equivalente a 0,31. Essa estratégia eliminou a marcação de Leonardo como objeto visual instanciador do FE Experienciador de *Desejar*. Outro exemplo de similaridade baixa (0,32) exhibe a junção de sentenças, o que aproveita o limite de caracteres por segundo em uma sentença no lugar de duas. Além disso, as LU *próximo.a* e *next.a*, apesar de estabelecerem relação de equivalência nas sentenças do áudio original e da legenda, não evocaram os mesmos *frames*. A primeira evocou *Sequência*, e a última evocou *Relative\_time*. No entanto, os dois são comparáveis pela rede semântica: tanto *Sequence*, que é o equivalente de *Sequência* no inglês, quanto *Relative\_Time* herdam de *Relation*. Dessa forma, as duas sentenças evocam o *frame Relation* (ou *Relação*) indiretamente.

## 4.3 Discussão

A análise dos resultados revelou que há diferentes níveis de permanência semântica na legendagem e diferentes maneiras em que as sentenças podem interagir com as imagens na tela. Em casos de permanência semântica máxima, a interação das sentenças com os objetos

<sup>3</sup> Para informações sobre modelagem de *frames* ditos pragmáticos, ver o trabalho de Andrade e Matos (2025) publicado neste dossiê.



na tela é a mesma, em termos de *frames* evocados. A similaridade nula em caso de omissão de sentença faz com que não haja *frames* evocados na imagem correspondentes na legenda.

Nos casos de total divergência dos *frames*, ou parcial como na similaridade média, os *frames* na imagem pela legenda serão diferentes dos *frames* evocados pelas falas do áudio original. Assim como constatado por Samagaio (2023) na análise do corpus Pedro Pelo Mundo, as estratégias de tradução audiovisual descritas por Díaz-Cintas e Remael (2021) influenciaram a permanência semântica entre texto original e traduzido.

Os pares de sentenças com similaridade semântica diferente de total mostraram que estratégias como a omissão de conteúdo menos importante, a simplificação de sentenças compostas, a junção de sentenças, a simplificação de perífrase verbal e a mudança do tipo de oração provocaram divergências na evocação de *frames*, resultando na similaridade parcial ou nula.

Além disso, a análise evidenciou que o processo de tradução audiovisual envolve a análise do conteúdo não verbal em conjunto com o linguístico. Em alguns casos, as imagens justificam a mudança do significado de uma sentença e a omissão de sentenças, como em instâncias em que não há objetos visuais correspondentes aos *frames* evocados nas falas.

Com esses resultados, avalia-se que as estratégias para a tradução de uma obra audiovisual influenciam a permanência semântica máxima entre texto original e traduzido almejada pelo tradutor, segundo o modelo da Primazia do Frame (Czulo, 2017).

Nota-se que as relações estabelecidas entre os *frames* de sentenças equivalentes através da rede semântica da FrameNet possibilitaram uma similaridade média ou alta em 35% dos casos, enquanto 52% do total, mais da metade, obtiveram similaridade baixa ou nula, sendo significativamente afetados pelas estratégias utilizadas.

A constatação dos casos em que houve omissão total ou parcial da sentença foi compatível com os resultados de Freitas (2022), que observou uma diminuição do número de *frames* do texto da legenda em comparação com o texto em português referente ao áudio original. Em adição a essa discussão, constatou-se através dos exemplos apresentados que a relação estabelecida entre o texto da legenda em inglês e as imagens do vídeo foi diferente da estabelecida entre as imagens e o texto original em português, em decorrência dessas omissões, bem como da evocação de *frames* diferentes em ambos os textos.

## 5 Conclusão

Em concordância com os achados de Samagaio (2023), conclui-se por meio da discussão dos resultados que as estratégias de tradução audiovisual utilizadas no processo de legendagem do curta-metragem, como listadas por Díaz-Cintas e Remael (2021), interferem diretamente no nível de permanência semântica da legenda em relação ao áudio original.

Tendo em vista o modelo da Primazia do *Frame* proposto por Czulo (2017), notou-se que, apesar de a similaridade total ter sido a menos frequente, com 4% dos casos, houve comparabilidade parcial em 54% dos casos, aproximando semanticamente as sentenças do áudio original e da legenda mesmo quando não houve permanência semântica máxima. Além disso, o uso do *frame* como uma representação cognitiva no processo tradutório evidencia a natureza cognitiva da tradução, como afirma Rojo e Ibarretxe-Antuñano (2013).

Ademais, as relações semânticas estabelecidas entre texto e imagem foram diferentes para o áudio original e para a legenda. No caso analisado de similaridade total, os *frames* evocados no vídeo foram os mesmos evocados no áudio original e na legenda. Porém, nos casos de similaridade parcial, os objetos visuais anotados no vídeo seriam diferentes caso a

anotação tivesse sido feita com base na legenda, em vez do áudio original. Em alguns casos, as mudanças feitas no texto na legenda, como a omissão de falas, foram justificadas pela análise da interação imagem-texto entre o áudio original e o vídeo.

Por esse motivo, concluiu-se também que o processo da legendagem deve levar em consideração a multimodalidade da obra. Além de estar limitada por fatores espaciais e temporais, como o espaço máximo que a legenda deve ocupar na tela e o limite de tempo de cada legenda para que o espectador possa acompanhar a obra cinematográfica, a legenda deve estabelecer uma relação de coerência com as imagens. Além disso, visto que a redução do texto original é inevitável, as relações semânticas entre a imagem e o texto podem servir para guiar o tradutor no momento de decidir o que é relevante ou não para ser traduzido na legenda.

Nesse sentido, esta pesquisa contribui para o melhor entendimento da legendagem e do processo de tradução audiovisual, por meio da Semântica de *Frames*, e este trabalho criou um corpus paralelo multimodal anotado semanticamente, que pode ser utilizado em tarefas de processamento de língua natural e em tarefas multilíngues.

Ao desenvolver um *corpus* paralelo multimodal de legendagem anotado semanticamente com potencial para uso em tarefas de processamento de língua natural, a pesquisa amplia a rede semântica da FrameNet Brasil e aumenta o seu conjunto de *corpora* para tarefas multilíngues, além de promover melhor entendimento sobre a legendagem e sobre o processo de tradução audiovisual. Dessa forma, o estudo contribui para os Estudos da Tradução no âmbito da tradução audiovisual e para o campo da Semântica de *Frames* e convida novas investigações que aprofundem as evocações semânticas de vídeo partindo da anotação de texto da legenda em vez do áudio original.

## Referências

BATEMAN, J.; WILDFEUER, J.; HIIPPALA, T. *Multimodality: Foundations, Research and Analysis - A Problem-oriented Introduction*. Walter de Gruyter GmbH & Co KG. 2017. DOI <https://doi.org/10.1515/9783110479898>. Disponível em: <https://www.degruyterbrill.com/document/doi/10.1515/9783110479898/html>. Acesso em: 10 jan. 2025.

BERTOLDI, A. Semântica de *Frames* e tradução: um estudo da equivalência de tradução de termos culturalmente marcados. *Letras & Letras*. 2016, v.32. 149. 10.14393/LL63-v32n1a2016-8. Disponível em: <https://seer.ufu.br/index.php/letraseletras/article/view/33045>. Acesso em: 10 jan. 2025.

BOAS, H. C. Frame semantics and translation. In: ROJO, A; IBARRETXEANTUNANO, I. *Cognitive Linguistics and Translation: Advances in Some Theoretical Models and Applications*. Berlin/Boston: De Gruyter Mouton, 2013. p. 125-158. Disponível em: <https://www.degruyterbrill.com/document/doi/10.1515/9783110302943/html>. Acesso em: 10 jan. 2025.

CINTAS, J. D.; REMAEL, A. *Translation Practices Explained*. Oxfordshire: Routledge, 2007.

CINTAS, J. D.; REMAEL, A. *Subtitling: Concepts and Practices*. Oxfordshire: Routledge, 2021. (Translation Practices Explained).

CZULO, O. Aspects of a Primacy of Frame Model of Translation. In: HANSEN-SCHIRRA, S.; CZULO, O.; HOFMANN, S. (ed.). *Empirical Modelling of Translation and Interpreting*. Berlin, Germany: Language Science Press, 2017. p. 465.

DORNELAS, L. D., GAMONAL, M. A., PAGANO, A. S. Análise semântica de audiodescrição em curta metragem: uma abordagem multimodal a partir da Semântica de Frames. *Domínios de Linguagem*, Uberlândia, v. 1866, p. 2-30, 2024. DOI: <https://doi.org/10.34019/1808-9461.2022.v23.38564>. Disponível em: <https://seer.ufu.br/index.php/dominiosdelinguagem/article/view/75272>. Acesso em: 10 jan. 2025.

FILLMORE, C. Frame Semantics. In: THE LINGUISTIC SOCIETY OF KOREA (ed.). *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Co, 1982. p. 111-137.

FILLMORE, C. J. Semântica de Frames. *Cadernos de Tradução*. Porto Alegre, n. 25, jul-dez, 2009.

FREITAS, K. C. G. A tradução audiovisual sob a perspectiva da semântica de frames: Uma análise dos significados evocados pelo texto do áudio original e sua legenda em L2. 2022. Trabalho de Conclusão de Curso. Faculdade de Letras, Universidade Federal de Minas Gerais (UFMG). Belo Horizonte, 2022.

HOLMES, J. S. The Name and Nature of Translation Studies. *Translation Section of the Third International Congress of Applied Linguistics*, Copenhagen, p. 66-80, ago. 1972. Disponível em: <https://archive.org/details/Holmes1972TheNameAndNatureOfTranslationStudies/mode/2up>. Acesso em: 10 jan. 2025.

PAGANO, A. S.; TEIXEIRA, A. L. R.; MAYER, F. A. Accessible Audiovisual Translation. In: JI, M.; LAVIOSA, S. (ed.). *The Oxford Handbook of Translation and Social Practices*. Oxford: Oxford University Press, 202. p. 66-82. DOI: <https://doi.org/10.1093/oxfordhb/9780190067205.013.4>.

ROJO, A. Applying Frame Semantics to Translation: a practical example. *Meta: Translators' Journal*, 47(3), p. 312-350. 2002. DOI: <https://doi.org/10.7202/008018ar>.

ROJO, A.; IBARRETXE-ANTUÑANO, I. Cognitive linguistics and translation studies: Past, present and future. *Cognitive Linguistics and Translation*, Berlin, p. 3-30, 2013. Disponível em: <https://www.degruyterbrill.com/document/doi/10.1515/9783110302943/html>. Acesso em: 10 jan. 2025.

SAMAGAIIO, M. M. A permanência semântica na tradução audiovisual: uma análise da legendagem pela ótica da Semântica de Frames. 2023. 184p. Dissertação (Mestrado em Linguística) – Universidade Federal de Juiz de Fora, Juiz de Fora, Minas Gerais.

SOUZA, D., GAMONAL, M. A., PAGANO, A. S. A audiodescrição sob a perspectiva da Semântica de Frames: um estudo exploratório. *Revista Gatilho*, Juiz de Fora, v. 23, p. 101-125, 2022. DOI: <https://doi.org/10.34019/1808-9461.2022.v23.38564>. Disponível em: <https://periodicos.ufjf.br/index.php/gatilho/article/view/38564>. Acesso em: 10 jan. 2025.

TORRENT, T. T., MATOS, E. BELCAVELLO, F., VIRIDIANO, M., GAMONAL, M. A., COSTA, A.D. & MARIM, M. C. Representing Context in FrameNet: A Multidimensional, Multimodal Approach. *Frontiers in Psychology*, v. 13, 1-20, 2022. DOI: <https://doi.org/10.3389/fpsyg.2022.838441>. Disponível em: <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.838441/full>. Acesso em: 10 jan 2025.

TORRENT, T. T.; HOFFMANN, T.; ALMEIDA, A. L.; TURNER, M. *Copilots for Linguists: AI, Constructions, and Frames*. Cambridge, Reino Unido: Cambridge University Press, 2024.

WEBTOOL 4.02. Definição e FE nucleares do *frame* Chegar. In: WEBTOOL 4.02. *Banco de dados*. [S. l.]: [s. n.]. Disponível em: <https://webtool.frame.net.br/report/frame/1455>. Acesso em: 10 jan. 2025.

WEBTOOL 4.02. Anotação Semântica da LU enroladinho.a. *In: WEBTOOL 4.02. Banco de dados.* [S. l.]: [s. n.]. Disponível em: <https://webtool.frame.net.br/annotation/fe/sentence/534439>. Acesso em: 10 jan. 2025.

WEBTOOL 4.02. Anotação semântica da LU curly.a. *In: WEBTOOL 4.02. Banco de dados.* [S. l.]: [s. n.]. Disponível em <https://webtool.frame.net.br/annotation/dynamicMode/1078>. Acesso em: 10 jan. 2025.

WEBTOOL 4.02. Anotação de imagem dinâmica do objeto visual cabelo.n. *In: WEBTOOL 4.02. Banco de dados.* [S. l.]: [s. n.]. Disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/10785>. Acesso em: 10 jan. 2025.

WEBTOOL 4.02. Anotação semântica de imagem dinâmica da sentença id 628267. *In: WEBTOOL 4.02. Banco de dados.* [S. l.]: [s. n.]. Disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/573>. Acesso em: 10 jan. 2025.

WEBTOOL 4.02. Cena sem demarcação de objeto visual na sentença de id 628313. *In: WEBTOOL 4.02. Banco de dados.* [S. l.]: [s. n.]. Disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/573>. Acesso em: 10 jan. 2025.

WEBTOOL 4.02. Cena da descida até a casa de Leonardo. *In: WEBTOOL 4.02. Banco de dados.* [S. l.]: [s. n.]. Disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/573>. Acesso em: 10 jan. 2025.

WEBTOOL 4.02. Marcação de objeto visual não correspondente entre áudio e legenda. *In: WEBTOOL 4.02. Banco de dados.* [S. l.]: [s. n.]. Disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/573>. Acesso em: 10 jan. 2025.

# 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

<b>Informação</b>		@Generic	#712	Information [en]	PDF
<b>Definition</b>					
Um <b>Pensador</b> sabe ou saberá sobre informações a respeito de um <b>Tópico</b> . Neste frame, muitas ULs codificam um específico <b>Meio de coleta</b> e/ou uma <b>Fonte</b> , mas elas também podem ser expressadas separadamente.					
<b>Frame Elements</b>					
<b>Core</b>					
<b>Informação</b>	A <b>Informação</b> que o <b>Pensador</b> possui ou está prestes a possuir.				
<b>Pensador</b>	O <b>Pensador</b> é conhecedor da <b>Informação</b> .				@human
<b>Tópico</b>	O <b>Tópico</b> é sobre o que a informação se trata em geral.				
<b>Peripheral</b>					
<b>Fonte</b>	O <b>Pensador</b> obtém a informação de uma <b>Fonte</b> .				
<b>Meio de coleta</b>	Uma ação ou método usado pelo <b>Pensador</b> que resulta na aquisição de <b>Informação</b> .				

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.<sup>1</sup> A esse conjunto

---

<sup>1</sup> 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*<sup>2</sup> 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 Multiz0K (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 Multiz0K (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 Multiz0K. 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

---

<sup>2</sup> *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* Frame<sup>2</sup> – *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.<sup>3</sup> 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, Frame<sup>2</sup> 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 Frame<sup>2</sup>.

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

---

<sup>3</sup> 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,<sup>4</sup> 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* Frame<sup>2</sup> 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 Frame<sup>2</sup> 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-

---

<sup>4</sup> 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.<sup>5</sup> 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,<sup>6</sup> 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

<sup>5</sup> Disponível em: <https://tribunademinas.com.br/>. Acesso em: 24 abr. 2025.

<sup>6</sup> 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.



**BOA VIAGEM**

### **Confira um roteiro de passeios gastronômicos fora do comum em São Paulo**

Experiências acontecem dentro de importantes espaços turísticos da capital

11 DE FEVEREIRO DE 2025

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 MultizOK 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.<sup>7</sup> 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.

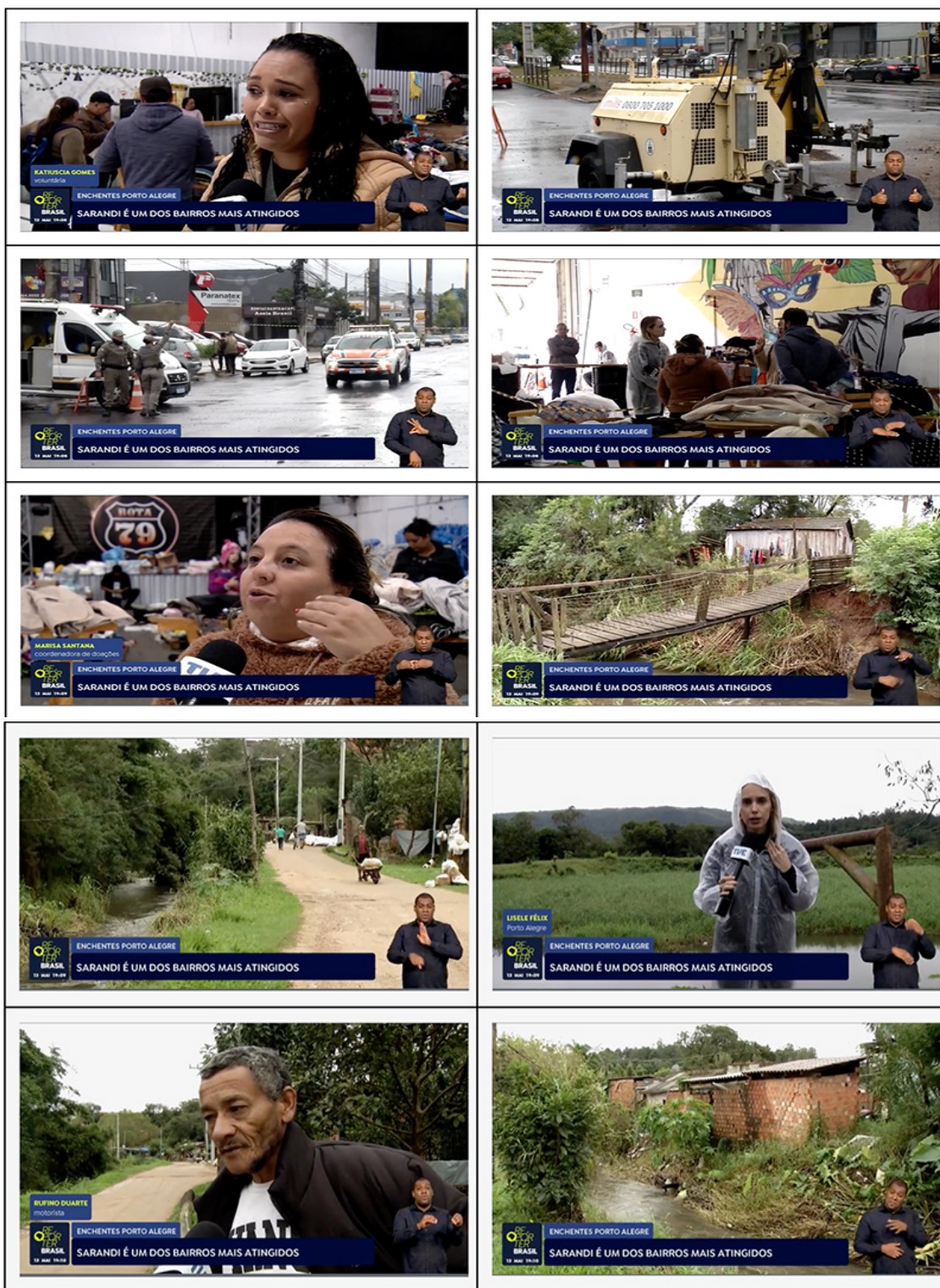
A Figura 5 apresenta uma sequência visual que representa uma matéria televisiva<sup>8</sup> prototípica. A única exceção à prototípia é o fato de o apresentador do telejornal não estar em estúdio, mas em externa – primeiro quadro. A matéria em questão trata da enchente que atingiu o bairro Sarandi, em 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



<sup>7</sup> 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.

<sup>8</sup> 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<sup>2</sup>. 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<sup>2</sup> – 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;<sup>9</sup> 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-

---

<sup>9</sup> 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<sup>2</sup>, 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<sup>2</sup>: 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. Multizok: 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](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 Mult30K: 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.



# A FrameNet Brasil Approach to Annotation of Pragmatic Frames Evoked by Turn Organization Gestures

## *Uma abordagem da FrameNet Brasil para anotação de frames pragmáticos evocados por gestos de organização de turno*

**Helen de Andrade Abreu**

Federal University of Juiz de Fora (UFJF)

Juiz de Fora | MG | BR

FAPEMIG

CNPq

[helen.abreu@estudante.ufjf.br](mailto:helen.abreu@estudante.ufjf.br)

<https://orcid.org/0000-0003-1228-7144>

**Ely Edison da Silva Matos**

Federal University of Juiz de Fora (UFJF)

Juiz de Fora | MG | BR

[ely.matos@ufjf.br](mailto:ely.matos@ufjf.br)

<https://orcid.org/0000-0002-9464-9315>

**Abstract:** This paper presents the manner in which FrameNet Brasil has developed annotation of pragmatic frames, with a focus on those evoked by interactive gestures used in turn organization during face-to-face conversation. As corpus, we chose the interviews shown in the episodes of the first season of the Brazilian television series *Pedro Pelo Mundo*. We used Webtool 4.0, the newest version of this tool, for annotation of a total of 47 interactive gestures. As a result, we were able to demonstrate that these interactive gestures evoke specific pragmatic frames – turn passing, turn confirmation and turn taking – and are tools which make conversation a multimodal means of communication. Our research shows that the annotation of pragmatic frames is not only possible, but also useful for pragmatic investigations and for expanding FrameNet multimodal analysis.

**Keywords:** pragmatic frames; conversational turn organization; multimodality.

**Resumo:** Este artigo apresenta como a FrameNet Brasil desenvolveu a anotação de frames pragmáticos, com foco nos que são evocados por gestos interativos usados na organização de turnos durante conversas face a face. Como *corpus*, escolhemos as entrevistas mostradas nos episódios da primeira temporada da série de televisão brasileira *Pedro Pelo Mundo*. Utilizamos a Webtool 4.0, a versão mais nova dessa ferramenta, para anotação de um total de 47 gestos interativos. Como resultado, fomos capazes de demonstrar que esses gestos interati-



vos evocam frames pragmáticos específicos – passagem de turno, confirmação de turno e tomada de turno – e são ferramentas que tornam a conversa um meio multimodal de comunicação. Nossa pesquisa mostra que a anotação de frames pragmáticos não é apenas possível, mas também útil para investigações pragmáticas e para expandir a análise multimodal da FrameNet.

**Palavras-chave:** frames pragmáticos; organização de turno conversacional; multimodalidade.

## 1 Introduction

Charles Fillmore first developed the concept of frames in a series of articles (Fillmore, 1975, 1982, 1985, 2008; Fillmore *et al.*, 2003a, 2003b, 2009) encompassing three decades. He defined a frame as a “system of concepts” (Fillmore, 1982, p. 111) which works in a way that, for a person to understand one of these concepts, it is necessary to understand the whole system or structure in which it is interconnected. In this way, when one of these concepts is evoked, all the other concepts in the system are automatically made available to our conceptualization. When a person mentions the fact that they have purchased something, for example, the whole structure (the frame) in which there is a buyer, goods, a seller, an amount of money used in the purchase, etc. (the frame elements) is made available to our conceptualization, allowing us to understand the topic of the conversation and, furthermore, to elaborate upon that.

Using his frame semantics theory as a basis, Fillmore (Fillmore *et al.*, 2003a, 2003b) decided to develop a lexicographic database which would permit annotation of frames that would in turn be used for both linguistic research and machine learning development. For this purpose, he created the FrameNet initiative in Berkeley, California, in 1997 (Fillmore *et al.*, 2003a, 2003b). Within the initiative, data were to be collected from corpora composed exclusively of real use of English. Since then, other FrameNets have been developed for studies in other languages (Subirats-Rüggeberg; Petruck, 2003; You; Liu, 2005; Boas; Ziem, 2018; Gruzitis *et al.*, 2018; Ohara *et al.*, 2018; Hahm *et al.*, 2020). To work with Brazilian Portuguese, FrameNet Brasil was developed at the Federal University of Juiz de Fora in 2009.

FrameNet Brasil developed its own annotation tool, Webtool (Torrent; Ellsworth, 2013; Torrent *et al.*, 2024), which has recently been updated to the version Webtool 4.0. Among the other FrameNets, FrameNet Brasil stands out, because it has developed into a multimodal database (Torrent; Ellsworth, 2013; Belcavello *et al.*, 2022; Viridiano *et al.*, 2022; Luz *et al.*, 2023). Webtool 4.0 can work with frames within texts as well as those evoked by visual images – both static and dynamic – thus making it possible for the database to work with textual frames and visual frames separately or together. This has greatly enriched the realm of research possibilities within FrameNet Brasil.

As part of FrameNet Brasil, ReINVenTA (Research and Innovation Network for Vision and Text Analysis) is a network comprised of research groups and laboratories from these universities: UFJF, UFMG, UFU and PUC-MG. This network was created for the development

of a computational model for annotation of multimodal objects. At present, ReINVenTA's dataset is composed of three corpora – Audition (Dornelas; Gamonal; Pagano, 2024); Framed Multi 30K (Viridiano *et al.*, 2022; Viridiano *et al.*, 2024); and Frame2 (Belcavello *et al.*, 2024), on which our project is based.

## 1.1 Frame<sup>2</sup>

Frame<sup>2</sup> is a multimodal dataset (Belcavello *et al.*, 2024). For the construction of this dataset, a Brazilian television series, *Pedro Pelo Mundo* (*Pedro Around the World*), was selected. FrameNet was granted written permission by the copyright owners of the series, so that the first season of the series could be used for research purposes.

The first season of *Pedro Pelo Mundo* was aired in Brazil in 2016 on the cable television channel GNT. It was composed of 10 episodes with duration of 23 minutes each, which resulted in 230 minutes of video in total. Pedro Andrade, the presenter, visited a different country in each episode, focusing on locations that had recently undergone a social transformation that could be economic, cultural, governmental, and so forth. During each episode, Pedro Andrade introduced the viewer to beautiful locations, tried local food and drinks while describing his experience, interacted with local people, and interviewed people who had been living there – especially, although not exclusively, Brazilian people. For this reason, several of the interviews in the series were conducted in Portuguese. Other interviews, however, were conducted in English and a few in Spanish. The interviews occurred in different places, including restaurants and parks, and some of them occurred while Pedro Andrade and his interviewee were walking on the street, on their way to a specific location. This variety of settings proved to be an asset for our research.

This audiovisual resource was used to create a dataset composed of annotated data collected from all the verbal language in the series (using both the transcribed audio portion of the series and the subtitles present on screen when the interviews were not conducted in Portuguese), and also from the visual portion of the series. The transcribed audio portion of the ten episodes, as well as all the subtitles, considered together as text, have been lexically annotated for frames and their corresponding frame elements, generating 16,458 annotation sets. The visual portion of the series has been annotated for objects recognizable by computational vision and for the frames evoked by them, generating 7,520 bounding boxes.

As explained by Belcavello *et al.* (2020), images and text may work together in the production of meaning. Taking this into consideration, the video annotators for Frame<sup>2</sup> dataset used a multimodal approach in completing their task (Belcavello *et al.*, 2024): the annotation of images started only after the text annotation was completed; when the visual annotators were working on creating the image dataset, they were instructed to first watch the video (with the audio on); after that, they were required to read each annotated sentence and then watch the video sequence related to that sentence, looking either for the objects described in it or for objects that were somehow related to these sentences. For example, Belcavello *et al.* (2024) describe how, during the annotation process, a graffiti on a wall was considered the visual manifestation of Behavior, a frame element in the frame Mental\_property, evoked by the lexical unit *criativo* ('creative') in the sentence *o povo voltou a ser criativo* ('the people became

creative once more'). In this way, it was demonstrated that this multimodal approach adds layers of meaning to FrameNet capabilities.

For its multimodal capabilities, Frame<sup>2</sup> dataset proved to be invaluable for our project.

## 1. 2 Pragmatic Frames

In the article "Beyond Lexical Semantics", Czulo, Ziem and Torrent (2020) discuss the usefulness of creating a way for pragmatic frames to be annotated and become part of the FrameNet dataset. As the authors point out, Berkeley FrameNet, founded by Fillmore, was created to deal with lexicographical problems (Fillmore; Baker, 2009, for example), and as other FrameNets for other languages were created, their aim remained the same. For this reason, they do not have the necessary apparatus for annotation of the pragmatic aspects of terms and expressions. However, Czulo, Ziem and Torrent (2020) demonstrate the necessity of recognizing and working with the pragmatic nature of certain frames.

Czulo, Ziem and Torrent (2020) make it clear that their proposal is not to modify any existing theory. Both Frame Semantics (Fillmore, 1982) and Frame Analysis (Goffman, 1974) allow for the possibility of the existence of pragmatic frames. Moreover, the proposal is not to cover the entire field of pragmatics, but only the study of linguistic expressions that have a conventionalized pragmatic use.

The authors begin their discussion with an analysis of tag questions, demonstrating that pragmatic meaning does not lie in each separate lexeme that is part of such expressions. Instead, it is the construction itself that performs the different pragmatic functions of tag questions, depending on the context in which they are used and fulfilling an interactive role between interlocutors (Czulo; Ziem; Torrent, 2020, p. 2). "Tag questions are multi-word expressions that evoke frames in a holistic rather than a compositional fashion, in that the building blocks of the expressions cannot be considered units carrying frame-semantic information on their own" (Czulo; Ziem; Torrent, 2020, p. 2).

The authors use, as a basis for their arguments, observations of the analysis of the results obtained by the Global FrameNet Shared Annotation Task. Annotation of the text of the TED talk "Do Schools Kill Creativity?", presented by Sir Ken Robinson (2006), was made using the translations (provided with the video) in Brazilian Portuguese, German and English, using Berkeley FrameNet 1.7. While undertaking this task, the annotators were allowed to include new lexical units to the dataset; however, they were not allowed to include new frames or to modify the existing ones. The first thirty sentences annotated in this way were, then, analyzed by an evaluation system (Czulo; Ziem; Torrent, 2020).

Comparing the annotations made for these three languages, Czulo, Ziem and Torrent (2020) demonstrate that some differences in the three annotation sets were caused by lack of an appropriate form to annotate pragmatic aspects of terms and expressions such as question tags and greetings. 'Good morning', for example, was not annotated in either Portuguese or English, as annotators recognized the pragmatic nature of the expression and did not have the means to indicate it. German annotators, however, chose to annotate the German term *Guten Morgen* ('Good morning'), as two separate lexical units – *Guten*, evoking the Desirability frame, and *Morgen*, evoking the Calendric\_unit frame – which caused the expression to lose the pragmatic meaning of a greeting. In fact, 'good morning' (and its equivalent expressions

in other languages) is a multi-word unit (like tag questions) that, to work as a greeting, has to meet certain conditions: the fact that the speaker is involved in an interaction with their interlocutor, the fact that it is morning, and the fact that it is the first time that the interlocutors meet that morning (Czulo; Ziem; Torrent, 2020, p. 3).

We consider it interesting to notice that the equivalent expression in Brazilian Portuguese, *bom dia*, means literally ‘good day’, as in (1):

(1) Brazilian Portuguese

Bom	dia.
Good	day
‘Good morning.’	

*Dia* (‘day’), in Portuguese, either refers to the whole period of twenty-four hours, or to the part of this period during which there is sunlight, similarly to the word ‘day’ in English. Nevertheless, in Portuguese, the use of the greeting *bom dia* is restricted to the morning period, exactly the way ‘good morning’ is used in English and *Guten Morgen* is used in German. This fact reinforces the pragmatic value of the expression *bom dia*.

Even though German annotators made the choice of annotating *Guten Morgen* separately during the Global FrameNet Shared Annotation Task, German FrameNet and Constructicon was the first FrameNet to start working on pragmatic frame definition (Ziem; Willich; Triesch, 2023). Taking the analysis and observations of Czulo, Ziem and Torrent (2020), as well as the work of German FrameNet and Constructicon as departing points, FrameNet Brasil has started the development of creation and annotation of pragmatic frames, which are now part of our database.

### 1.3 Interactive Gestures

As part of our work within ReINVenTA, it seemed logical to develop pragmatic frame annotation as a multimodal task. The first step, then, was to choose a specific topic of investigation. The topic chosen was the strategies used by communicators, during face-to-face conversation, to manage conversational turns. More specifically, we chose the gestures used by these communicators to indicate different choices such as keeping the turn, passing it, asking for it, and so forth. As Sacks, Schegloff and Jefferson (1974) and Bavelas (2022) observed, face-to-face conversation is the most basic form of communication. Nevertheless, research in turn management is far from abundant.

Janet Bavelas (Bavelas *et al.*, 1995; Bavelas, 2021, 2022) is probably the most prolific author in the study of gestures used by communicators for face-to-face conversation. Her research involves different aspects of this form of communication, and one of these aspects is the types of gestures used by communicators during conversation. She separates these gestures, which she calls “co-speech gestures” (Bavelas, 2022, p. 13), into two different categories: topic gestures and interactive gestures. Topic gestures are those used to illustrate the topic of conversation – for example, when a person indicates with a hand a height level while saying something like, “the bookshelf was this high”. Interactive gestures, however, refer to a person’s understander. These gestures are used to refer, for example, to something one’s communica-



tor mentioned previously. Bavelas *et al.* (1992, p. 469) give as an example a situation in which the utterer says, “and Customs is DEFINITELY is career oriented”. When the utterer mentioned “Customs”, they made a brief movement towards their understander, which the authors, through observation and analysis of the video, considered to mean something like “which YOU suggested” (Bavelas *et al.*, 1992, p. 471).

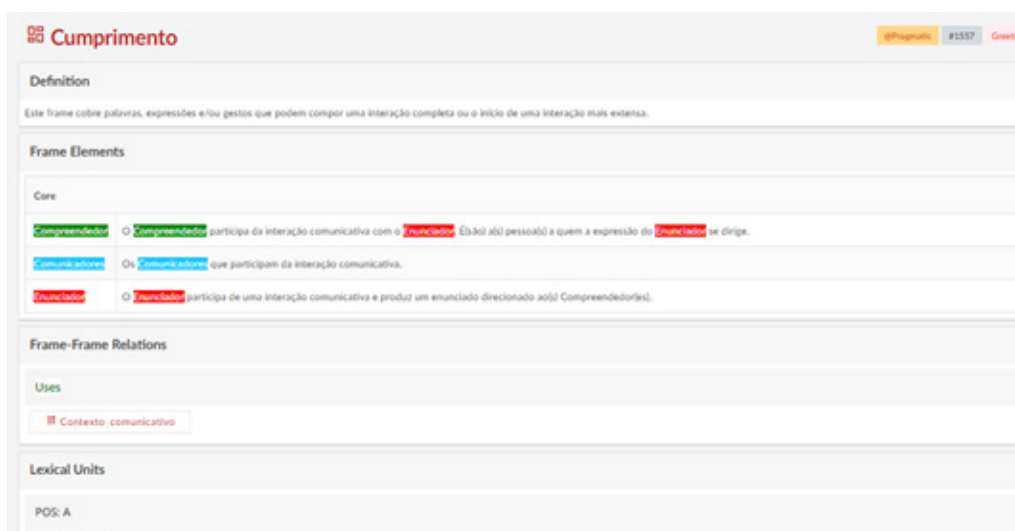
Interactive gestures are also used for the management of conversational turns (Bavelas *et al.*, 1995; Bavelas, 2021, 2022), proving to be an important tool that helps communicators indicate their intentions of keeping the conversational turn, passing it to a chosen person, asking permission to take it and so forth, while the topic of conversation continues to flow. When an utterer “passes” the turn to another person, for instance, a typical turn-passing gesture is one that is similar to the gesture of passing a real object to someone. This gesture may be used when only two people are participating in conversation, but it can be especially useful when three or more people are involved. It is important to note that the message conveyed through these gestures is not mentioned in the conversation. The gestures work as a second layer of communication, which makes face-to-face conversations multimodal.

The following sections will present how our dataset was developed, a discussion of our observations and our conclusions.

## 2 Dataset Development

The first step in our work was to develop the annotation of pragmatic frames within Webtool. The first two pragmatic frames to be inserted into Webtool were *Contexto comunicativo* (‘Communicative\_context’) and *Cumprimento* (‘Greeting’), which is demonstrated by Figure 1.

Figure 1- Greeting



Source: FrameNet Brasil Lab (c2014-2025d).

The decision to start with these frames was based on the conclusions reached by Czulo, Ziem and Torrent (2020) about expressions such as ‘good morning’ having pragmatic mea-



ning, as presented in section 1, and the fact that human conceptualization also makes use of the communicative context frame together with the greeting frame. The relation between these two frames is indicated in the “Frame-to-Frame Relations” area, which can be observed in Figure 1 above. As indicated, the frame *Cumprimento* (“Greeting”) uses the frame *Contexto comunicativo* (“Communicative\_context”).

The frame elements (FEs from now on) in the frame Greeting are the people involved in the communicative scene: *enunciador* (‘utterer’, the person who talks or gestures), *compreendedor* (‘comprehender’, the person to whom the communicative act is directed) and *comunicadores* (‘communicators’, encompassing utterer and comprehender together, as a unit). On the upper right-hand corner of the screen, the orange label indicates the type of frame it is – in this case, a pragmatic one.

As we had in mind the type of pragmatic frames we would be looking for in the videos, based on the paper by Czulo, Ziem and Torrent (2020) and on the research on interactive gestures presented in section 1, the next frames to be inserted into Webtool were: *Organization\_of\_conversation* (of which the next frames would be sub-frames); *Turn\_passing* (in which the utterer indicates the passage of conversational turn to the comprehender); *Turn\_taking* (in which a communicator takes the communicative turn); *Turn\_confirmation* (in which the comprehender indicates the utterer should keep the turn); and *Turn\_keeping* (in which the utterer indicates the wish to continue communicating). Figure 2 shows the frame *Organiza  o\_da\_conversa* (‘Organization\_of\_conversation’). In its Frame-Frame Relations panel we can observe its subframes and the fact that it uses the frame *Contexto comunicativo* (‘Communicative\_context’). After these frames were ready for annotation, we proceeded to the next phase of our work.

Figure 2- *Organization\_of\_conversation*



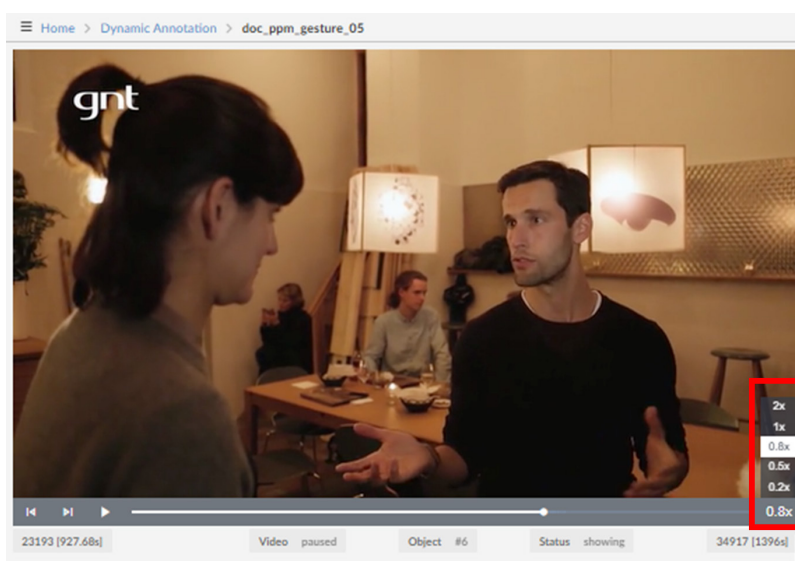
Source: FrameNet Brasil Lab (c2014-2025e).

Our corpus, *Pedro Pelo Mundo - Gestures*, was developed as a subdataset within Frame<sup>2</sup>. This subdataset is composed of the same episodes of the first season of the series *Pedro Pelo Mundo* that had already been fully annotated for lexical frames and the visual

objects related to them. Pedro Pelo Mundo - Gestures, however, was created within Webtool 4.0 separately, so that we could annotate each gesture that evokes a pragmatic frame.

The next step was to look for the interactive gestures that evoked each of the pragmatic frames previously mentioned, as they were used by the communicators during all the interviews. Webtool works in such a way that in order to look for the gestures we have the option of watching the video at normal speed, double the speed, or we can watch it in slower motion (0.8, 0.5 or 0.2 times in relation to the normal speed). The speed can be altered in the right-hand corner of the bar underneath the video panel on the screen, as shown by Figure 3, marked by a red box.

Figure 3: Video speed

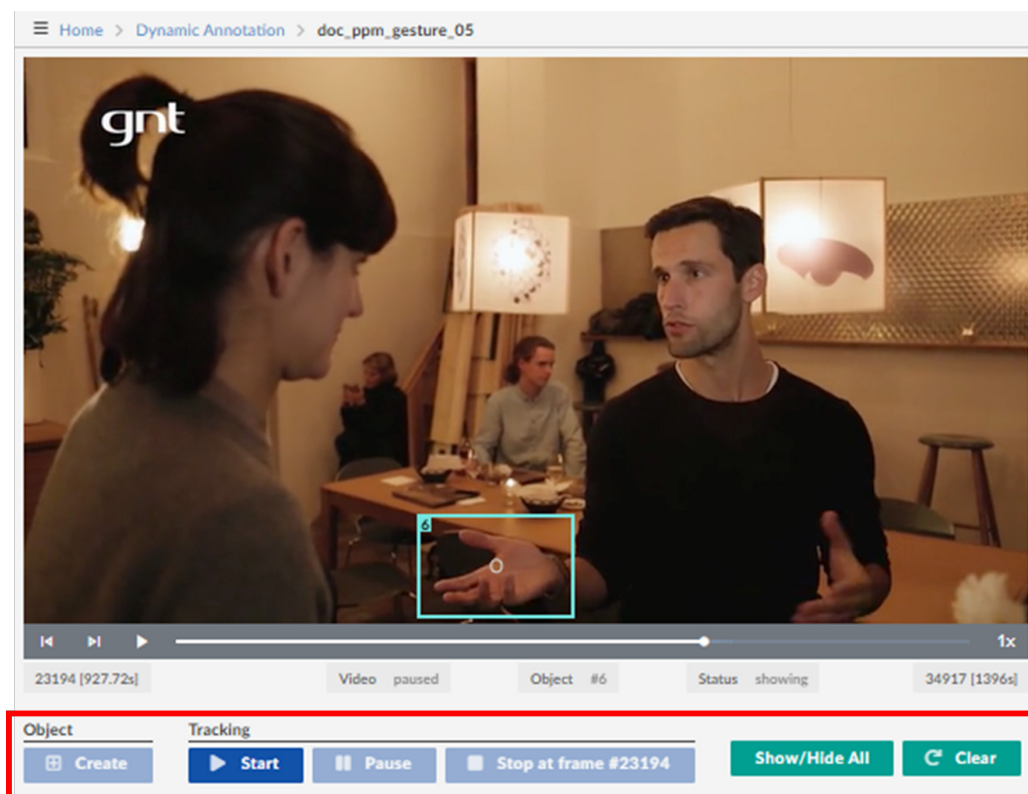


Source: FrameNet Brasil Lab (c2014-2025g).

We observed that the best speed for identifying the gestures is 0.8, which is almost normal speed, and yet slow enough for the gestures to be more easily recognized. Once the gesture is recognized, slower speeds help locate the precise moment of the beginning and the end of the movement. After pinpointing the moment the gesture begins, we start the process of creating a bounding box by pressing the button “create” on the left-hand side of the screen, under the video panel. After pressing this button, we can move the cursor and choose a place in the paused video image, click there and drag the cursor to create the bounding box. The box can then be moved, and its sides can be adjusted as necessary. Then we can press the button “start” and the bounding box will automatically follow the object as it moves within the video, until we press the button “pause” or “stop at frame #XXXX”.<sup>1</sup> We also have the option of using the button “move forward”, to the left of the “play” button, to move the video slowly, frame by frame, and make any adjustments to the bounding box that may be necessary. The buttons are shown in Figure 4, within a red box.

<sup>1</sup> It is important to notice that this button indicates the movie frame, which is a term used in the film industry and is unrelated to frames as in the frame theory. “Frame” as used in the film industry refers to each “picture” that, together with all the others, and in motion, compose a video.

Figure 4: Bounding box creation



Source: FrameNet Brasil Lab (c2014-2025b).

Once the bounding box is created, we proceed to the top right-hand side of the screen, as depicted in Figure 5, to manually insert the chosen frame, the FE and the CVName (Computer-vision Name). In the case of gestures, the CVNames are the lexical identification of the visual object marked within the bounding box. In our example in Figure 4, it is the utterer's hand, marked by the bounding box number 6.

Figure 5: Completing information

The screenshot shows the 'Completing information' form in the FrameNet Dynamic Annotation interface. The form is titled 'Pedro Pelo Mundo - 1ª Temporada EP 05' and has a 'Build Sentences' button. The 'Current Object' is set to #6, with a range of 23194/23210 and a CVName of #9809. The 'FE' (Frame Element) is set to 'Utterer' and the 'CVName' (Computer-vision Name) is set to 'Partes\_do\_corpo.mão.n'. The 'Start frame' is 23194 and the 'End frame' is 23210. There is a checkbox for 'Is blocked?' which is currently unchecked. The form has 'Save' and 'Clone' buttons at the bottom.

Source: FrameNet Brasil Lab (c2014-2025c).

Webtool 4.0 can be seen in Brazilian Portuguese or in English. In the example shown by Figure 5, in English, the chosen frame is Turn\_passing; the FE is 'utterer', as it is the utterer's hand that is marked by the bounding box; and the LU is *mão* ('hand'), within the frame *Partes\_do\_corpo* ('Body\_parts'). The start frame and the end frame numbers are provided automatically; however, they can be manually adjusted if necessary. Once these options are chosen, we press the button "save" and the object appears within a panel in the bottom right quadrant of the screen. We can see the individual object more clearly in Figure 6:

Figure 6: Annotated object information

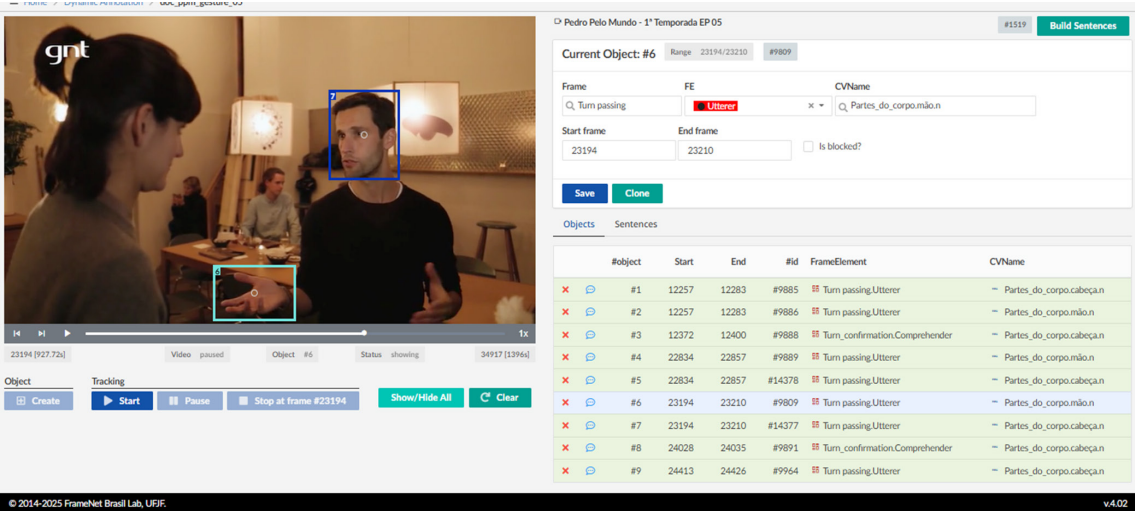
✕	🗨	#5	22834	22857	#14378	Turn passing.Utterer	Partes_do_corpo.cabeça.n
✕	🗨	#6	23194	23210	#9809	Turn passing.Utterer	Partes_do_corpo.mão.n
✕	🗨	#7	23194	23210	#14377	Turn passing.Utterer	Partes_do_corpo.cabeça.n

Source: FrameNet Brasil Lab (c2014-2025a).

This panel contains such information as the number of the object within the episode (in this case, number 6), the start and end movie frames of the bounding box (23194 and 23210, respectively), the object number in relation to the whole dataset (object number 9809), the frame (Turn\_passing), the FE (Utterer) and the CVName (*mão*, 'hand', which is part of the frame Body\_parts).

It is important to mention that turn passing gestures may involve not only the utterer's hand or hands, but also the person's head. Sometimes it is only with the head that the utterer indicates passage of turn. In the example depicted in Figure 4, Pedro, the utterer, used his right hand and his head to make the gesture. However, we can only annotate each body part separately. In this case, his hand is object 6, and his head is object 7. By clicking on the button "show/hide all", we can make all bounding boxes visible at the same time. Figure 7 depicts the whole screen, with the video paused at a moment in which we can see the passing of turn, the two parts of the utterer's body involved in the gesture (head and hand), as well as all the other objects (parts of the body involved in interactive gestures) annotated in this whole episode. It also shows how the whole screen is visualized on Webtool 4.0.

Figure 7: Whole screen



Source: FrameNet Brasil Lab (c2014-2025h).

### 3 Discussion

This corpus was selected for its advantages, as presented in section 1; however, it has some limitations that should be considered. First of all, the series is not composed only of interviews. There are long sequences in which Pedro Andrade is showing the viewers different places, trying different types of food, and so forth, as mentioned before. The interviews are only one of the multiple parts of each episode. As a result, of the 23 minutes that compose one episode, there may be less than 10 minutes of actual interview time. Secondly, because it is a television program, the camera angles chosen are not always the best for simultaneously observing both communicators. Sometimes only one of them is visible on the screen, and it is not possible for us to observe all the expected gestures or the responses to them which would be necessary to confirm these gestures as interactive ones. At other times, the camera is focused on the communicator's face and the gestures they might be making are hidden from the viewer. Thirdly, during the interviews, the video sequence at times shows not the communicators themselves but objects, people, places or scenes that are related to the subject being discussed.

Although this presentation format is a great resource for multimodal annotation of frames in general (Belcavello *et al.*, 2024), it was impossible for us to observe what gestures the communicators were using at those times. For all these reasons, in each episode there are only a few minutes of fully observable interactions between communicators. Nevertheless, the corpus is significant enough in a way that we were able not only to observe and confirm the existence of gestures that had been previously recorded by other



researchers, but also to make our own observations which were added to the body of knowledge about interactive gestures.

One of these observations was the fact that the gestures are not always directed to the utterer's comprehender. It is our understanding that the context in which interactions were recorded in previous work (interviews and conversations filmed in laboratories) did not allow for the observation of what communicators do in other settings. For example, some of the interviews conducted by Pedro Andrade happened while he and his communicator were walking on the street. In these situations, while both communicators were walking side by side, seemingly paying partial attention to where they were walking, some of the turn passing gestures were not pointed at the utterer's communicator. Instead, these gestures were made with the utterer's hand extended to his front and still they could be easily recognizable. Figure 8 shows one of these interactions.

Figure 8 : Turn passing gesture



Source: FrameNet Brasil Lab (c2014-2025f).

## 4 Conclusions

This paper aimed to present pragmatic frames as they are now being developed and annotated by FrameNet Brasil. Czulo, Ziem and Torrent (2020) had already demonstrated the opportunities for annotating pragmatic meaning and the advantages that it would entail. This paper has shown that FrameNet Brasil, through its Webtool, has been able to produce



such annotations and confirmed the views of Czulo, Ziem and Torrent (2020). It has also shown that pragmatic frames may be evoked lexically or visually.

Moreover, our research has demonstrated that, through annotation of pragmatic frames, we were able to confirm that the management of conversational turns is multimodal. This fact proves that the annotation of these frames is a useful tool for research in linguistics as well as other areas involved with language, such as, but not limited to, conversation analysis and Natural Language Processing.

Further developments of our work will contribute to the expansion of FrameNet multimodal analysis to include turn organization pragmatic frames, which is fundamental for computational representation of different multimodal text genres, such as interviews and debates. Also, further developments of our work will contribute to pragmatic investigations, through the introduction of a computational implemented tool for analysis of pragmatic phenomena in a multimodal context.

Such contributions will increase both the analytical and theoretical impact of the multimodal turn in FrameNet development. From the analytical side, including the annotation of pragmatic frames evoked by gestures opens up the possibility to include in the resulting datasets a myriad of meaning making interrelations between communicative modes yet to be fully explored by the FrameNet Brasil model.

From the theoretical side, they bring an additional opportunity to extend the application of the theory of Frame Semantics to a domain that, despite being included in Fillmore's original proposals for a Semantics of Understanding (Fillmore, 1985), is yet to be fully explored by practitioners in the field.

## Acknowledgments

This paper is a project within ReINVenTA – Research and Innovation Network for Vision and Text Analysis of Multimodal Objects. ReINVenTA is funded by FAPEMIG grant RED 00106/21, and CNPq grants 408269/2021-9 and 420945/2022-9. Helen de Andrade Abreu's research is funded by FAPEMIG grant RED-00106-21.

## References

- BAVELAS, J. B. *Face-to-Face Dialogue: Theory, Research, and Applications*. Oxford: Oxford University Press, 2022.
- BAVELAS, J. B.; CHOVIL, N.; COATES, L.; ROE, L. Gestures Specialized for Dialogue. *Personality and Social Psychology Bulletin*, v. 21, n. 4, p. 394-405, 1995.
- BAVELAS, J. B.; CHOVIL, N.; LAWRIE, D. A.; WADE, A. Interactive Gestures. *Discourse Processes*, v. 15, n. 4, p. 469-489, 1992.
- BAVELAS, J. Pragmatics of Human Communication 50 Years Later. *Journal of Systemic Therapies*, v. 40, n. 2, p. 3-25, 2021. DOI: 10.1521/jsyt.2021.40.2.3.

- BELCAVELLO, F. *et al.* Frame<sup>2</sup>: A FrameNet-Based Multimodal Dataset for Tackling Text-image Interactions in Video. *In: JOINT INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, LANGUAGE RESOURCES AND EVALUATION (LREC-COLING 2024)*, 2024, Torino. *Proceedings [...]*. Torino: European Language Resources Association (ELRA)/ ICCL, 2024. p. 7429-7437.
- BELCAVELLO, F.; VIRIDIANO, M.; COSTA, A. D. da; MATOS, E. E.; TORRENT, T. T. Frame-Based Annotation of Multimodal Corpora: Tracking (A)Synchronies in Meaning Construction. *In: LREC INTERNATIONAL FRAMENET WORKSHOP*, 2020, Marseille. *Proceedings [...]*. Marseille: ELRA, 2020. p. 23-30.
- BELCAVELLO, F.; VIRIDIANO, M.; MATOS, E.; TORRENT, T. T. Charon: A FrameNet Annotation Tool for Multimodal Corpora. *In: LINGUISTIC ANNOTATION WORKSHOP (LAW)*, 16., 2022, Marseille. *Proceedings [...]*. Workshop presented at LREC 2022. Marseille: ELRA, 2022. p. 91-96.
- BOAS, H. C.; ZIEM, A. Constructing a Constructicon for German. *In: LYNGBELT, B.; BORIN, L.; OHARA, K.; TORRENT, T. (ed.). Constructicography: Constructicon Development Across Languages*. Amsterdam: John Benjamins Publishing Company, 2018. p. 183-228.
- CZULO, O.; TORRENT, T. T.; MATOS, E.; COSTA, A. D. da; KAR, D. Designing a Frame-Semantic Machine Translation Evaluation Metric. *In: WORKSHOP ON HUMAN-INFORMED TRANSLATION AND INTERPRETING TECHNOLOGY (HiT-IT)*, 2., 2019, Varna. *Proceedings [...]*. Varna: Incoma, 2019. p. 28-35.
- CZULO, O.; ZIEM, A.; TORRENT, T. T. Beyond Lexical Semantics: Notes on Pragmatic Frames. *In: LREC INTERNATIONAL FRAMENET WORKSHOP*, 2020, Marseille. *Proceedings [...]*. Marseille: ELRA, 2020. p. 1-7.
- DORNELAS, L.; GAMONAL, M. A.; PAGANO, A. S. Análise semântica de audiodescrição em curta-metragem: uma abordagem multimodal a partir da Semântica de Frames. *Domínios de Linguagem*, Uberlândia, v. 18, e1866, 2024.
- FILLMORE, C. J. An Alternative to Checklist Theories of Meaning. *In: COGEN, C. THOMPSON; H. THURNGOOD, G.; WHISTLER, K. (ed.). Proceedings of the Berkeley Linguistic Society*. Berkeley: Berkeley Linguistics Society, 1975. p. 123-131.
- FILLMORE, C. J. Frame Semantics. *In: LINGUISTIC SOCIETY OF KOREA (ed.). Linguistics in the Morning Calm*. Seoul: Hanshin Publishing, 1982. p. 111-137.
- FILLMORE, C. J. Frames and the Semantics of Understanding. *Quaderni di Semantica*, v. 6, p. 222-254, 1985.
- FILLMORE, C. J. The Merging of Frames. *In: FAVRETTI, R. R. (ed.). Frames, Corpora, and Knowledge Representation*. Bologna: Bononia University Press, 2008. p. 1-12.
- FILLMORE, C. J.; BAKER, C. A Frames Approach to Semantic Analysis. *In: HEINE, B. et al. (ed.). The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, 2009. p. 791-816.
- FILLMORE, C. J.; JOHNSON, C. R.; PETRUCK, M. R. L. Background to FrameNet. *International Journal of Lexicography*, v. 16, n. 3, p. 235-250, 2003a.
- FILLMORE, C. J.; LEE-GOLDMAN, R.; RHOMIEUX, R. A. FrameNet in Action: The Case of Attaching. *International Journal of Lexicography*, v. 16, n. 3, p. 297-332, 2003b.
- FRAMENET BRASIL LAB. *Annotated Object Information*. [S. l.], c2014-2025a. Available at: <https://webtool.frame.net.br/annotation/dynamicMode/1519>. Accessed on: 10 Jan. 2025.

- FRAMENET BRASIL LAB. *Bounding Box Creation*. [S. l.], c2014-2025b. Available at: <https://webtool.frame.net.br/annotation/dynamicMode/1519>. Accessed on: 10 Jan. 2025.
- FRAMENET BRASIL LAB. *Completing Information*. [S. l.], c2014-2025c. Available at: <https://webtool.frame.net.br/annotation/dynamicMode/1519>. Accessed on: 10 Jan. 2025.
- FRAMENET BRASIL LAB. *Greeting*. [S. l.], c2014-2025d. Available at: <https://webtool.frame.net.br/report/frame/1557>. Accessed on: 10 Jan. 2025.
- FRAMENET BRASIL LAB. *Organization\_of\_conversation*. [S. l.], c2014-2025e. Available at: <https://webtool.frame.net.br/report/frame/1561>. Accessed on: 10 Jan. 2025.
- FRAMENET BRASIL LAB. *Turn Passing Gesture*. [S. l.], c2014-2025f. Available at: <https://webtool.frame.net.br/annotation/dynamicMode/1516>. Accessed on: 10 Jan. 2025.
- FRAMENET BRASIL LAB. *Video Speed*. [S. l.], c2014-2025g. Available at: <https://webtool.frame.net.br/annotation/dynamicMode/1515>. Accessed on: 10 Jan. 2025.
- FRAMENET BRASIL LAB. *Whole Screen*. [S. l.], c2014-2025h. Available at: <https://webtool.frame.net.br/annotation/dynamicMode/1519>. Accessed on: 10 Jan. 2025.
- GRUZITIS, N.; NESPORE-BERZKALNE, G.; SAULITE, B. Creation of Latvian FrameNet Based on Universal Dependencies. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 11., 2018, Miyazaki. *Proceedings* [...]. Edited by T. T. Torrent, L. Borin and C. F. Baker. Miyazaki: ELRA, 2018. p. 23-27.
- HAHM, Y.; NOH, Y.; HAN, J. Y.; OH, T. H.; CHOE, H.; KIM, H.; CHOI, K. S. Crowdsourcing in the Development of a Multilingual FrameNet: A Case Study of Korean FrameNet. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE (LREC), 12., 2020, Marseille. *Proceedings* [...]. Marseille: ELRA, 2020. p. 236-244.
- LUZ, A. C. L.; BRAZ, G.; RUIZ, L. P.; PINTO, M. C.; BELCAVELLO, F.; SIGILIANO, N. S.; TORRENT, T. T. Anotação do Dataset Multimodal da ReINVenTA. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 14., 2023, Porto Alegre. *Anais* [...]. Porto Alegre: SBC, 2023. p. 360-364.
- OHARA, K.; KAWAHARA, D.; SEKINE, S.; INUI, K. Linking Japanese FrameNet with Kyoto University Case Frames Using Crowdsourcing. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 11., 2018, Miyazaki. *Proceedings* [...]. Edited by T. T. Torrent, L. Borin and C. F. Baker. Miyazaki: ELRA, 2018. p. 56-61.
- SACKS, H.; SCHEGLOFF, E.; JEFFERSON, G. A Simplest Systematics for the Organization of Turn Taking for Conversation. *Language*, v. 50, n. 4, p. 696-735, 1974.
- SUBIRATS-RÜGGERBERG, C.; PETRUCK, M. R. Surprise: Spanish FrameNet! In: WORKSHOP ON FRAME SEMANTICS, 2003, Prague. Workshop presented at the XVII International Congress of Linguists (CIL). *Proceedings* [...]. Edited by E. Hajivcová, A. Kotévsocová, and J. Mirovsky. Prague: Matfyzpress, 2003.
- TORRENT, T. T. *et al.* A Flexible Tool for a Qualia-Enriched FrameNet: The FrameNet Brasil WebTool. *Lang Resources & Evaluation*, 2024. DOI: <https://doi.org/10.1007/s10579-023-09714-8>.
- TORRENT, T. T.; ELLSWORTH, M. Behind the Labels: Criteria for Defining Analytical Categories in FrameNet Brasil. *Veredas*, v. 17, n. 1, p. 44-65, 2013.

VIRIDIANO, M. *et al.* Framed Mult30K: A Frame-Based Multimodal-Multilingual Dataset. *In: JOINT INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, LANGUAGE RESOURCES AND EVALUATION (LREC-COLING)*, 2024, Torino. *Proceedings* [...]. Torino: ELRA/ICCL, 2024. p. 7438-7449.

VIRIDIANO, M.; TORRENT, T. T.; CZULO, O.; LORENZI, A.; MATOS, E.; BELCAVELLO, F. The Case for Perspective in Multimodal Datasets. *In: WORKSHOP ON PERSPECTIVIST APPROACHES TO NLP*, 1., 2022, Marseille. *Proceedings* [...]. Workshop presented at LREC 2022. Marseille: ELRA, 2022. p. 108-116.

YOU, L.; LIU, K. Building Chinese FrameNet Database. *In: INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE PROCESSING AND KNOWLEDGE ENGINEERING*, 2005, Wuhan. *Proceedings* [...]. Wuhan: IEEE, 2005. p. 301-306.

ZIEM, A.; WILlich, A.; TRIESCH, S: Expanding the German FrameNet: Pragmatic Frames Across Lexicon and Grammar. *In: INTERNATIONAL COGNITIVE LINGUISTICS CONFERENCE*, 16., 2023, Düsseldorf. *Proceedings* [...]. Düsseldorf: ICLC, 2023. Available at: [https://iclc16.github.io/abstracts/ICLC16\\_BoA.pdf](https://iclc16.github.io/abstracts/ICLC16_BoA.pdf). Accessed on: 13 Feb. 2025.

# Multimodal Frame Semantics: Expanding the Analytical Categories of FrameNet Brasil Multimodal Datasets

## *Semântica de Frames Multimodal: expandindo as categorias de análise dos datasets multimodais da FrameNet Brasil*

**Natália Sathler Sigiliano**

Universidade Federal de Juiz de Fora (UFJF)

| Juiz de Fora | MG | BR

natalia.sigiliano@ufjf.br

<https://orcid.org/0000-0002-8460-5546>

**Abstract:** In this paper, I propose an extension of the joint attention scenes models to include mediated multimodal communicative settings. From this extension, I discuss how the FrameNet analytical model can be used for the annotation of multimodal genres so as to include categories capable of accounting for invited shifts in joint attention in visual narratives. I revisit the literature on deixis and claim that, despite the fact that much of the foundational research on this topic proposes correlations between deixis in verbal language and other semiotic modes, the focus of research lies on the linguistic material in which the expression of deixis is grounded. By bringing together contributions from human communication studies, from the analysis of deixis in narratives and from Frame Semantics, I claim that semiotic devices of different types can be mobilized in filmic narratives to invite viewers to promote shifts in the center of the joint attention scene. To illustrate the proposed typology, I provide example analyses of deictic center shifts in films and describe the annotation methodology incorporated to the FrameNet model to account for them.

**Keywords:** deixis; joint attention; multimodality; Frame Semantics; filmic narrative.

**Resumo:** Neste artigo, proponho uma extensão dos modelos de cena de atenção conjunta para incluir cenários comunicativos multimodais mediados. A partir dessa extensão, discuto como o modelo analítico da FrameNet pode ser usado para a anotação de gêneros multimodais, de modo a incluir categorias capazes de



explicar as mudanças convidadas na atenção conjunta em narrativas visuais. Revisito a literatura sobre dêixis e afirmo que, apesar de grande parte da pesquisa fundamental sobre esse tópico propor correlações entre dêixis na linguagem verbal e outros modos semióticos, o foco da pesquisa está no material linguístico no qual a expressão de dêixis está fundamentada. Ao reunir contribuições dos estudos de comunicação humana, da análise de dêixis em narrativas e da Semântica de Frames, proponho que dispositivos semióticos de diferentes tipos podem ser mobilizados em narrativas fílmicas para convidar os espectadores a promover mudanças no centro da cena de atenção conjunta. Para ilustrar a tipologia proposta, apresento exemplos de análises de deslocamentos do centro dêítico em filmes e descrevo a metodologia de anotação incorporada ao modelo FrameNet para explicar esses deslocamentos.

**Palavras-chave:** dêixis; atenção conjunta; multimodalidade; Semântica de Frames; narrativa fílmica.

## 1 Introduction

More than one century ago, Ferdinand de Saussure (1916) defined the object of Linguistics as a system composed of linguistic signs – *langue* – separating the field from other areas of study, such as Psychology and Semiotics. More recently, however, as the discussion on the relation between language use and multimodality advances, research has been emphasizing the need for a multidisciplinary approach to semantics, given that the analysis of verbal text as the sole system for meaning construction is not enough. This is not to say that phenomena beyond verbal text analysis have been ignored by linguists, rather, taking into consideration the field to which they are circumscribed, there has been a natural emphasis on the proposition of analytical categories based on verbal text. Nonetheless, the multimodal turn in Linguistics has been highlighting the need for rethinking the emphasis on verbal text. Such a communicative mode should be taken as one of the aspects of language, and the multiple semioses mobilized for using language for meaning construction should be taken into consideration.

The multimodal turn alluded to above is present in several subfields of Linguistics. The New London Group has pioneered in highlighting the notion of multiliteracy. This concept broadens the idea of literacy by looking at elements other than verbal language, while introducing metalanguage to describe and interpret the design elements of different modes for meaning construction (The New London Group, 1996). Analytical apparatus for multimodal phenomena built on contributions from diverse fields of research, such as Semiotics, Communication Studies and Film Analysis, has been proposed by the Bremen-Groningen group led by Bateman – see Bateman, Wildfeuer and Hiippala (2017) for an overview. Also,



building on the parallel architecture model proposed by Jackendoff (2002) for the analysis of verbal language, Cohn proposes a “grammar” for image comprehension, applying it first for comics (Cohn, 2013) and later extending it to film analysis (Cohn, 2016b).

In Cognitive Linguistics, research on co-speech gesture – see, among others, Sweetser (2007), Steen et al. (2018) and Cienki (2022) – has expanded the analytical coverage of models such as Mental Spaces and Blending Theory and Construction Grammar. Such an expansion is a natural path in Cognitive Linguistics, since, as pointed out by Turner (2018, p. 357):

Construction Grammar [in particular and Cognitive Linguistics in general] accepts responsibility to account for forms of creativity otherwise almost entirely ignored in linguistics. This commitment is wise, given that creativity is the engine that develops systems of communication.

As for Frame Semantics, since the original propositions of the theory, Fillmore (1976, 1982, 1985) has recruited the notion of visual scenes to explain the role of frames in organizing knowledge. Moreover, the proposed theory has highlighted the fact that frames play a role in understanding that goes beyond lexical semantics. At some point in the theory, distinctions between (lexical) semantic frames, interactional frames and cognitive frames (or scenes) have been proposed. However, those were later abandoned by Fillmore (2008) and merged into one single notion of frame. Nonetheless, such a diversity of possibilities for the application of frames to the analysis of meaning construction has been restricted – for the sake of feasibility – to lexical semantics, as the theory was implemented in FrameNet (Fillmore *et al.*, 2003).

Such an implementation restriction found a welcoming home in Computational Linguistics, which, back in the early 2000s, was exclusively centered on the analysis of strings of characters (see Dannélls et al., 2022). Therefore, the original FrameNet methodology had as its first step transforming text used in the analysis into those types of strings, stripping out of it all kinds of information that could not be encoded as UTF-8 characters. Such a method was – and still is to a large extent – standard in Computational Linguistics. Even considering that a multimodal turn has been also taking place in Computational Linguistics, the approach to multimodality is still very reductionist, to the extent that it mostly focuses on the compositional analysis of the elements depicted in the image. Such an analysis takes images as some sort of ground truth in reference to which the text can be analyzed (see Viridiano et al., 2022 for discussion).

More recent research has been extending the FrameNet methodology to the analysis of both textual genres (Dutra; Sigiliano, 2021) and static and dynamic images (Belcavello et al., 2020, 2022; Torrent et al., 2022; Viridiano et al., 2022; Belcavello, 2023). Those analyses have been uncovering correlations which are fundamental for the development of a Multimodal Frame Semantics. Nonetheless, they still take verbal language as a starting point, since they seek to associate entities depicted in the images to verbal language superimposed to them.

Considering the varied resources that the different communicative modes recruit for meaning construction, I claim that two additional extensions of the analytical framework of Frame Semantics – and FrameNet – are crucial: first, the very nature of the phenomena usually analyzed should be broadened; second, it is important to build multidisciplinary teams whose role will be that of defining analytical categories for multimodal texts.

In this context, I propose that Frame Semantics should be repositioned to account for key aspects of multimodal texts. Given the importance of deixis for meaning construction,

in this paper, I will be focusing on the proposition of improvements to the Frame Semantics model, more specifically to its implementation as FrameNet, so that it can account for key deictic phenomena in multimodal texts.

## 2 Language and Deixis

In this section, I discuss the centrality of the notions of shared intentionality, shared attention and cooperative motivation to the model of language assumed in this paper. I later point out the connections between this model and the study of deixis, emphasizing that such a connection is not limited to linguistic conventions, but is also present in other communicative semioses.

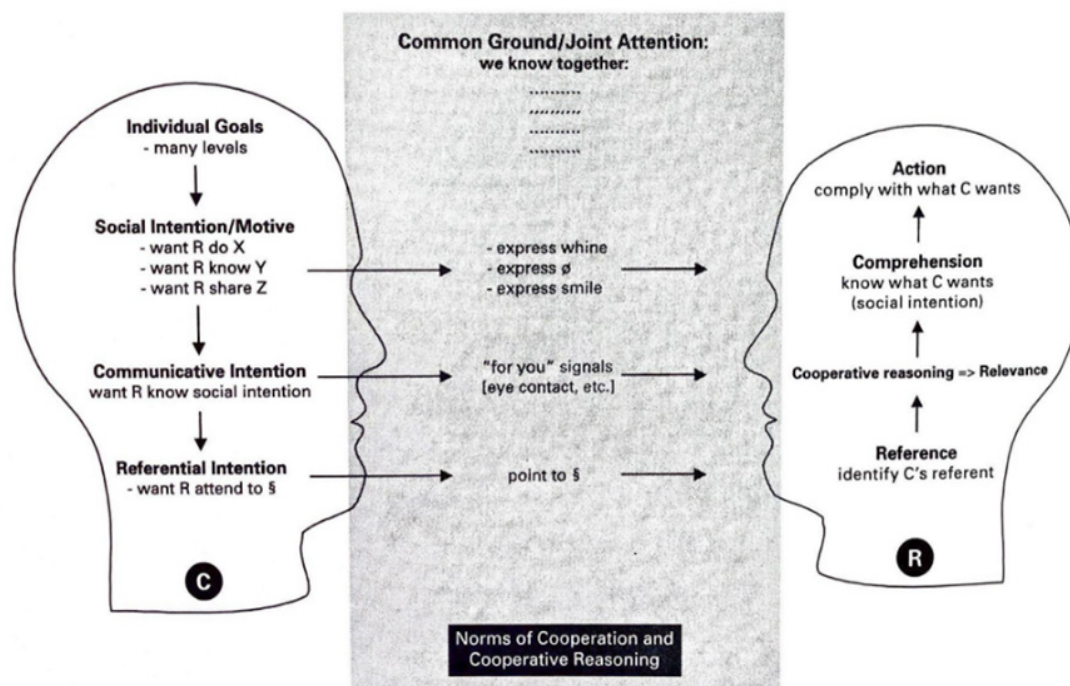
### 2.1 The Cooperative Foundations of Human Communication

Every linguistic use involves intentionality (Grice, 1957), and shared intentionality “presupposes a background sense of the other as a candidate for cooperative agency” (Searle, 1990, p. 414-415). As pointed out by Wittgenstein (1953), language is based on a non-linguistic structure comprising intentional knowledge and shared concepts. In this context, Tomasello (2008) proposes that the analysis of human communication must depart from the observation of non-conventionalized, non-coded communication, as well as from forms of connection other than language, such as natural gestures – e.g. pantomiming and nodding –, for example. Despite being simple and natural, such gestures reveal great communicative power.

Therefore, communication in general – be it linguistic or not – requires shared intentionality and a context of common knowledge. Tomasello (2018) proposes a model for human communication that involves different language manifestations. Although he frames one of the participants as the “recipient”, Tomasello argues for the existence of cooperation norms and cooperative reasoning throughout human communication. Moreover, he claims that social, communicative and referential intentionalities are the individual goals of communication, as seen in Figure 1.

From Figure 1, communicative acts involving different semiotic modes can be present in any regular communicative setting. Smiles, gesture and eye gaze, for example, can be used for reinforcing the demonstration of intentionalities, according to the model proposed.

Figure 1- Summary of cooperative model of human communication



Source: Tomasello (2018, p. 98).

Caption: C = communicator; R = recipient.

Another important aspect of communication, according to Tomasello, is the necessity of a joint attention scene for linguistic learning to take place. Tomasello (2003) defines language as a joint action whose fulfillment is tied to shared attention. When discussing the stages of child development, Tomasello (2003) situates joint attention in the period when children start to perceive themselves, the other and the world surrounding them. At this point, children also start to recognize their interlocutors as intentional agents. Both the child and their interlocutor may intentionally define external referents and activities that compose the joint attention frames. Also, children understand the role fulfilled by themselves, their caregiver and the external referent, as well as the interchangeability of such roles, allowing them to adopt an external perspective to build concepts about themselves or to understand the intentions of caregivers when using linguistic symbols to manipulate their attention (Tomasello, 2003). Together with the ability to build shared attention frames, children develop communicative behaviors using triadic deictics, manipulating the attention of their interlocutor towards external referents, changing them (imperative gesture) or indicating them (descriptive gesture). Pointing gestures can be regarded, then, as the foundational manifestation of deixis in human communication. I will turn to them next.

## 2.2 Shared Intentions, Shared Attention and Deixis

The use of deictic gestures constitutes the first pillar of reference processes underlying the development of intentional communication (Goldin-Meadow, 2007; Liszkowski; Brown;

Callaghan; Takada; de Vos, 2012; Tomasello; Carpenter; Liszkowski, 2007). Tomasello, Carpenter and Liszkowski (2007) claim that:

[...] these early pointing gestures are used not merely to direct attention either to the self or to an object, but to influence the mental states of others. [...] Pointing may thus represent a key transition, both phylogenetically and ontogenetically, from non-linguistic to linguistic forms of human communication (Tomasello; Carpenter; Liszkowski, 2007, p. 720).

The basic function of directing someone's attention to something through gesture is present in every known human society (Kita, 2003). Such an ability can still be used in support of language, in some contexts, after the latter is acquired (Tomasello, 2008). However, pointing gestures alone are not capable of expressing the whole range of intentionality. Tomasello (2018, p. 91) states that "[...] a pool of common ground between the pointing infant and her recipient, including shared assumptions about why she wants to establish joint attention in the first place" is required.

Rodríguez, Moreno-Núñez, Basilio and Sosa (2015) advocate in favor of the existence of ostensive gestures preceding and providing the foundation for the acquisition of deictic gestures. They argue that the gestures for showing – i.e. holding an object so that one can see it – and giving occur before that of pointing and are intentional. Wilkins (2003), when analyzing the gesture of pointing with the index finger as not being universal in sociocultural and semiotic terms, concludes that "[...] pointing (i.e., the use of some part of the body to make deictic gestural reference) appears to be universal. However, the use of the index finger for pointing does not appear to be universal" (Wilkins, 2003, p. 212). Therefore, regardless (a) of being used to refer to internal or external objects that are given to the interlocutor, (b) of the way those objects are presented to the interlocutor (with open hands or while pointing), or (c) of the motion associated with pointing, it is important to look at the pointing gesture as a signal of intentionality in the joint attention scene and as a foundation for human communication, coming before language acquisition.

The study of pointing translates into Linguistics as deixis. Levinson (2004) describes the issues revolving around how deixis is accounted for in Linguistics as follows:

For those who want to treat language as a generative system for objectively describing the world, deixis is one hell of a big black fly in the ointment. Deixis introduces subjective, attentional, intentional and of course context-dependent properties into natural languages. Further, it is a much more pervasive feature of languages than normally recognized, and is theoretically puzzling in many regards. All this makes difficult a tidy treatment within formal theories of semantics and pragmatics (Levinson, 2004, p. 97).

Levinson (2004) differentiates deixis and indexicality, explaining that, while the first comes from the Linguistic tradition, the latter belongs to Philosophy. He uses indexicality to label broader phenomena of contextual dependency, and deixis to refer to narrower linguistic aspects of indexicality, namely, deictic or indexical expressions in language such as *you*, *here*, *now*. He stresses the fact that indexicality cannot be reduced to such a study, since any referential

*expression can be used indexically. When discussing deixis, Levinson includes examples where this phenomenon includes elements beyond the limits of verbal language (Levinson, 2004, p. 106).*

Levinson's (2004) analyses indicate that a pragmatic resolution for deictics is crucial, because "there is a close relation between exophora and anaphora—in both cases we have contextual resolution of semantically general expressions, in the one case in the physical space-time context of the speech event, and in the other in the ongoing discourse" (Levinson 2000, p. 268 apud Levinson, 2004, p. 107). Moreover, he frames indexicality as both an intentional and attentional phenomenon, revolving around the spatio-temporal center of verbal interaction (cf. Bühler, 1934).

While discussing uses of linguistic expressions that are semantically insufficient for successfully producing reference without contextual support, Levinson (2004) explains that such a support is provided by interlocutors' mutual attention and by their ability to reconstruct the referential intentions indicated by the speaker through clues in context. Such clues include gesture or eye gaze, which may take part in the indexical sign. Nonetheless, such an analysis would not be enough to account for Bühler's *Deixis am Phantasma* ('deixis in the imagination'). In those cases, the deictic origo is shifted in a series of transpositions triggered by the process of imagining oneself at a different place. *Deixis am Phantasma* demonstrates how easily human cognition can change joint attention scenes (cognitively speaking) to, through the use of deictics, achieve intended goals.

Levinson (2004) lists yet other challenges for the analysis of deictics, which are related to the fact that deixis is much broader than what a description of purely verbal language aspects could capture, even when context is taken into consideration. This is because deixis is a cognitive tool for shifting attentional focus, which can be performed through language, gesture, gaze, body movement, among others, including even the simulation of spatial, temporal and eventive motion. Even analyses focusing on linguistic deixis recognize that other communicative modes and strategies can be recruited. When discussing the use of demonstratives in language, Levinson (2004) highlights that deictic expressions arise from a necessity to locate the intended reference, directing the attention of the interlocutor.

Similar discursive strategies are found in other authors discussing deixis in Linguistics. Bühler (1934) describes deixis as a mental imagery system, where an imaginary referential line along which elements are positioned is built in the mind of the interlocutor. Similarly, as per Hanks (2008), the deictic system of the language is composed of a field of spatial-temporal coordinates, which can be phenomenologically expanded to include the perceptual coordinates from constructive fantasy, namely from ontological memories of the past.

When differentiating deixis and non-deixis, Fillmore (1975) proposes an analogy where a sculpture is seen as non-deictic, since it can be observed in its entirety when standing on a square, for example. On the other hand, a photograph would be deictic, since a point of view was adopted when it was taken. The analogy proposed by Fillmore makes clear the possibility of extending the concept of deixis to other semioses. Moreover, while analyzing the linguistic expression of deixis, similarly to Levinson (2004) and the other authors discussed so far, Fillmore claims that certain lexical and grammatical items can only be interpreted when the sentences where they occur are understood against some social context, defined so as to



identify the participants in the communicative act and their location in space and time for the duration of said act.

Those contributions demonstrate that the study of deixis should not be restricted to the analysis of verbal language. Kita (2003, p. 325) corroborates this claim by proposing that “the coordination of gaze, torso movement, gesture, and speech is motivated by communicative factors but also by the interlinkage among various cognitive processes”. She proceeds by stating that “Both interactional and cognitive sides of the story need to be taken into account in order to reach a full understanding of various body movements and speech as a unified system of communication” (Kita, 2003, p. 327). Therefore, looking at deixis as a phenomenon involving multiple physical, linguistic and cognitive tasks provides clues for understanding its role beyond strictly linguistic processing. Deixis plays a key role in our ability to read the world and understand diverse interaction situations, during which we use deictics to indicate our intentions.

Theoretical constructs central to the cognitive models of human communication – such as intentionality and the ways to express it, joint attention and cooperation – and their relation to deixis are therefore also relevant for the analysis of multimodal text. This is so because in a text the choices reflected on its organization – be they linguistic, typographic, musical, or imagetical, among others – are intentional and rely on the cooperation between author and reader via text to be properly understood. In other words, deixis plays a key role in manipulating the reader’s attention so that they can understand the intentions motivating the text and reconstrue the joint attention scene needed for cooperative communication to take place.

Therefore, concerning multimodal text and, more specifically in this paper, filmic narratives, I claim that expanding our understanding of deixis can provide means to analyze how deictic operators trigger cognitive rearrangements that allow comprehenders to construe meaning from sequences of events. To investigate this claim, I will propose an analytical model built upon the existing work developed by FrameNet Brasil for the annotation of multimodal text for Frame Semantics. Multimodal texts abound with elements functioning as clues inviting the comprehender to shift the deictic center of the joint attention scene. I will turn to the relation between deixis and filmic narrative next.

### 3 Multimodal Filmic Narrative and Deixis

It is a fact that multimodal text has always existed and that we are immersed in multimodality in our everyday lives. It is also a fact that multimodal features in text have been enriched and made more available with the advancement of technology (Bateman, 2008). Multimodality, according to Bateman, Wildfeuer and Hiipala:

[...] is a way of characterizing communicative situations (considered very broadly) which rely upon combinations of different ‘forms’ of communication to be effective—the TV programme uses spoken language, pictures and texts; the book uses written language, pictures, diagrams, page composition and so on; talking in the cafeteria brings together spoken language with a host of bodily capabilities and postures; and the computer game might show representations of any of these things and include movement and actions as well (Bateman; Wildfeuer; Hiipala, 2017, p. 7).



Such a concept makes the complexity of multimodal phenomena explicit, especially if we consider the diversity of resources, textual genres and communicative settings involved in the analyses of the phenomena. Multimodal analysis is not about putting communicative modes side by side in a sum. Rather, it is a multiplication operation (Lemke, 1998). Analyzing such multiplying factors in context – aiming at describing them at work – is crucial for understanding multimodal phenomena, despite the associated challenge, which is “to find ways of characterising the nature of such interdependencies and to develop methodologies for investigating them empirically” (Bateman; Wildfeuer; Hiipala, 2017, p. 17).

Besides the multiplication of meaning construction possibilities emerging from the combination of communicative modes, research points to the sharing of meaningful properties across modes – see Jackendoff and Lerdahl (2006) for language and music, Steedman (2002) for language and body action, and Cohn (2016a) for language and visual sequences in comic strips. In this context, investigating, as proposed by Belcavello *et al.* (2020), *how different modes interact during human comprehension of multimodal genres is key.*

I subscribe to their claim that Frame Semantics is an adequate model for representing such interactions. In this theory, “meaning is relativized to scenes” (Fillmore, 1977, p. 59), and, although this maxim has been largely explored only in relation to language, it does not mean that it cannot be extended to other communicative modes. Sound, image and language, representative of distinct modes, bear similarities in their organization and in how they are perceived cognitively. If we consider sound, for example, it allows humans to gather information on space, distance, directions, emotional and physical states, among other features. Our cognitive capacity for the perception of sound sequences also allows humans to remember musically arranged sequences and predict harmonic progression (Bateman; Wildfeuer; Hiipala, 2017, p. 28). The knowledge humans acquire by experiencing textual genres throughout life also allows them to make predictions about language use patterns (Bakhtin, 1984). Therefore, it is only natural that the confluence of communicative modes in a multimodal genre allows humans to combine different types of formal clues in the process of re-constructing intended meaning.

Frames work precisely in this process. They provide schematic structures that can be accessed and brought into play by comprehenders for making sense of varied formal clues found in (multimodal) text. When we isolate communicative modes, we are surely still able to analyze how each type of formal pattern – linguistic, musical, visual – can serve as a clue for activating the frames needed for interpretation. However, important aspects of the meaning construction process will be lost. Hence, approaching multimodal genres requires trying to look at all the diversity of communicative modes involved, and such an effort, although challenging, is crucial for the adequacy of the analyses.

Therefore, even phenomena that have been traditionally associated with their linguistic manifestations, such as deixis, must be considered under the perspective of the combination of communicative modes for meaning construction. When it comes to studying filmic narratives, this proposition can be pervasively illustrated. Fillmore (1975) already recognized deictic centering to be both more general and more abstract than the linguistic material pointing to shifts in deixis, such as here, now and you. While buy and sell and loan and borrow can demonstrate the perspectival nature of linguistic symbols (Fillmore, 1975; Tomasello, 1999), go and come can do so for the abstractness of the deictic center in language (Fillmore, 1975). In all those cases, verb alternations encode different vantage points adopted while describing

the relations between the participants in the scene, or whether motion happens toward or away from the speaker. Similarly, shot composition, camera positions and editing choices in a filmic narrative can serve as clues for meaning construction.

It is thus necessary to recognize that deixis goes beyond the linguistic expression of categories such as person, space and time, for example. It is a foundational phenomenon in human perception and in the organization of semioses contributing to our ability to make sense of the world from constantly rebuilding (joint) attention scenes. This is so because we engage in meaning construction by individually observing the world and sharing our perspective on it through joint attention scenes, which involve not only individual, but shared intentionalities. Through pointing – which can manifest via language, gesture, gaze, body motion – we can intentionally guide our interlocutor to new joint attention scenes. Therefore, motion and change are also foundational to approaches to meaning construction involving multiple semioses.

How does pointing occur in filmic narratives? In this case, would it be limited to the analysis of gesture and language? How are intended attention foci defined and how is the need for viewers to reframe their attention focus signaled? Those are the questions I aim to investigate.

### 3.1 Deixis in Narrative

Segal (1995), when approaching the relation between narrative and deixis, proposes that:

when one reads a narrative as it is meant to be read, he or she is often required to take a cognitive stance within the world of the narrative. A location within the world of the narrative serves as the center from which the sentences are to be interpreted (Segal, 1995, p. 15).

In this scenario, deixis is then seen, as per Zubin and Hewitt (1995), as the structuring framework around which narrative emerges. These authors proceed with the following claim:

stories are made possible because readers can import knowledge of the everyday world and of other possible worlds into the current story world; this provides the listener/reader with the illusion of mentally inhabiting a fully specified and coherent world (Zubin; Hewitt, 1995, p. 130).

Zubin and Hewitt (1995) situate deixis in narrative from the perspective of the Deictic Center (DC) Theory, which attempts to model the consequences of deictic shifts beyond the here and now of real-world interaction and into fictional text. Such a theory is founded on Deictic Shift Theory (DST), which, in turn, claims that the DC frequently changes from the here and now to a locus in the mind model representing the discourse world (Segal, 1995). According to this later theory, in fictional narrative, readers and authors change their deictic centers and cognitively project themselves to a place inside the world of the story, taking over other deictic centers. According to Segal's (1995) view, the DC is a structure providing cohe-

rence to a text, even when it is not directly represented in the lexicon or syntax, and which changes according to the progression of the story.

Zubin and Hewitt (1995) illustrate deictic change in narrative through the observation of typical oral storytelling styles, stating that:

The story is not addressed to the audience in the way conversation or a lecture is; rather, it opens a conceptual window through which the story world can be glimpsed. The story is self-enclosed. Its deictic structure presupposes its own story world, and not the current interactional context of the teller and audience. In fact, the listener's deictic perspective becomes the one chosen for him or her by the teller. In a successful story, we have the illusion of experiencing the fictional world directly, because we unconsciously adopt the deixis of the DC as our own (Banfield, 1982; Wiebe, this volume) (Zubin; Hewitt, 1995, p. 131).

The illustration provided by Zubin and Hewitt aligns with Turner's (2017) idea of Blended Classic Joint Attention (BCJA) Scenes. According to Turner, the everyday experience of watching the news on TV requires a complex blend:

The news anchor is not actually in a scene of classic joint attention with the viewer; [...] *here* for the participants in the news interaction is not actually a single shared space ("It's good to have you here," says the news anchor, but where is "here"?); *now* for the participants in the news interaction need not be a particular moment ("Now we have a special announcement coming up for you here," says the news announcer, but perhaps it was recorded, perhaps it is meant to be viewed at many different times, perhaps the announcer did not even know what the special announcement would be; in addition, who is "we," and again, where is "here"?). But we can blend all these elements into a scene of *blended classic joint attention*, which is tractable and familiar because it draws on our understanding of classic joint attention. Most of the language that is available for running a scene of classic joint attention can be projected, adapted, and used for BCJA. BCJA is a generic integration template—that is, a well-known general pattern of blending that can guide the mental construction of indefinitely many specific networks (Turner, 2017, p. 3).

Human ability to take over different DCs in narrative helps narrative progression, and the DC window provides the listener-reader with two shifting foci, defined by Zubin and Hewitt (1995) as "focalizing perspective" – origin – and "focalized perspective" – content. The first refers to the reader's point of view, to what is being shown to them. The latter to what can be seen inside the deictic window in terms of motion in space, time and people inside the story world, considering the purpose of the focalizing perspective.

To analyze how DCs are organized in written narratives, Zubin and Hewitt (1995) propose four basic concepts: WHO, WHERE, WHEN and WHAT, which are subject to several operations allowing the narrative to be introduced, maintained, changed or suspended during text. The authors map DC-devices in written text and define them as instruments for signaling DC stability or change, while maintaining textual cohesion. DC-devices manifest as either morphemic or syntactic structures building the DC and guiding the reader in reconstructing it (Zubin; Hewitt, 1995, p. 141). They also propose that deictic operations in the DC are carried out by the reader-listener during the process of interpreting a narrative text, including: (a) introducing people, objects, places and time; (b) maintaining the stability of

the DC; (c) changing the WHO, WHAT, WHEN or WHERE to another one; (d) voiding one or more of the components of the DC, which become no longer relevant at that point of the narrative. Finally, Zubin and Hewitt (1995) highlight that listeners-readers build DCs not only from devices present in the text, but also based on shared world knowledge and on cooperative principles.

Analyses, propositions and descriptions such as the ones summarized in the previous paragraphs are crucial for understanding narrative structure. Nonetheless, their scope is limited to language. In a context where multimodal filmic narratives are in focus, there is still a lack of studies and models aimed at accounting for the correlations between linguistic and non-linguistic aspects of deixis. The next section presents a proposal for that area.

### 3.2 A Model for Analyzing Multimodal Deixis

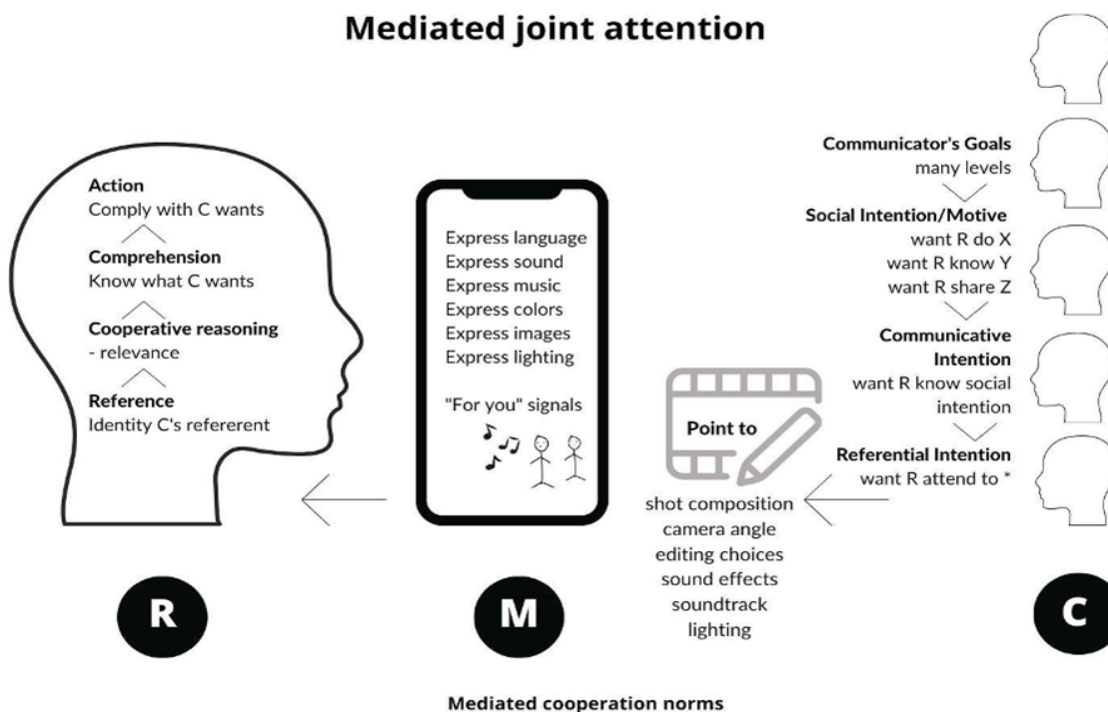
Based on the cooperative human communication model by Tomasello (2018) – Figure 1 – and drawing on Turner's (2017) notion of BCJA, I propose a new model – Figure 2 – where multiple communicators (authors) choose a screen – the medium – (M) to occupy the position of the communicator (C). Together with the comprehender (R), and sharing intentionality with them, the joint attention scene is built and the comprehender of the multimodal text is expected to interact with it, deploying their knowledge about the world. The comprehender is supposed to construe their comprehension of the text by following the clues provided in the text by the multiple communicators using clues from different semioses, which are broadcast via the screen.<sup>1</sup>

The cooperation principle still holds between the communicator and the comprehender. However, it may be the case – and it usually is – that the comprehender is not aware of the communicators involved. There is a prominence of the text – broadcast via the screen – as an indicator of the communicative intentions. The situation is similar to that of reading a written novel, where the reader interacts directly with the narrative, where intentionalities have been indicated by linguistic material. In multimodal text, however, many other elements, besides the linguistic ones, are at play and operate on the DC, especially in filmic narratives. If, on the one hand, in written narratives, one has to pay attention to morphological, lexical and syntactic aspects to analyze deixis in narrative progression, in multimodal narratives, visual and sound aspects must be also analyzed.

---

<sup>1</sup> It may be the case that R's meaning construction ignores C's intentions completely. However, it does not mean that C has no communicative goals that they want R to recognize.

Figure 2 - Summary of model for mediated communication focusing on the role of DC shifts in meaning construction



Source: The author. Based on Tomasello (2018).

In this context, the comprehender engages in a mediated cognitive interaction with the communicators in the narrative world. In a multimodal joint attention scene mediated by devices, joint attention depends on the intentionality of the comprehender to be established. They must turn on the device and pay attention to it. Next, it is expected that the comprehender aligns their own deictic perspective to the one that was intentionally chosen by the communicators, who, throughout the narrative, establish joint attention micro scenes guiding textual understanding.<sup>2</sup>

While building these micro scenes, communicators use deictic devices which may or may not comprise language, and go beyond it, becoming multiplying agents in meaning construction. Such devices collaborate to the maintenance or alteration of the joint attention micro scenes. The alteration of the micro scenes prompts alterations in the frames invoked in the mind of the comprehender and promotes narrative progression.

As previously indicated, pointing – or deixis – is an important element in everyday joint attention scenes. We use pointing by means of several communicative modes related to

<sup>2</sup> Note that, in the case of filmic narratives, it is usually the case that the narrator is not present in the medium used to broadcast the story, which makes the process of building the blended joint attention scene more complex, to the extent that the projection of the dialogue between C and R requires the construction of an abstract or idealized C.

cognitive processes, to refer to an element external to the interaction, directing the attention of our interlocutor and marking intentionality. The coordination for a new attention focus of the interaction arises from deictic pointing actions. One of the participant's attention in the interaction is redirected to some specific point in space and in time. In filmic narratives, similarly, new joint attention scenes are previously established by communicators every time shifts in the DC are intended. Such shifts can be motivated by the WHO, WHAT, WHERE and WHEN features (cf. Zubin; Hewitt, 1995), and the forms they manifest linguistically have been described. Nonetheless, their manifestation in other semioses has not been the focus of descriptions and discussions. Considering the prominent role of deictics in the various semioses, in the means used for calling attention and indicating intentions, analyzing their manifestation in multimodal narratives is certainly relevant.

In the following section, I will further explain the model summarized in Figure 2, providing elicited and didactic examples of multimodal deictic devices in filmic narratives, especially in what concerns the establishment of joint attention micro scenes. I will also explain why the incorporation of such a model into FrameNet brings both new challenges and new possibilities.

## 4 Incorporating Deixis into Multimodal FrameNet Analysis: Challenges and Possibilities

Linguistics was defined as a science focusing on the analysis of verbal language. To do so, it considers the surrounding context of the element being analyzed to provide descriptions and explanations on several aspects of such elements: how they developed in time, how they are acquired by infants, how they interact with other elements in language, how they contribute to meaning construction, how they function in textual genres and communicative settings, and so on.

Genres take communicative intentionality as a foundational aspect of their organization, of what one can expect from them. Moreover, genres occur in communicative situations where they are relevant. Therefore, the genre a communicator chooses to express their intentionality is not fortuitous. In this context, genres can be regarded as large deictic elements signaling expected intentionalities to the interlocutor. In what specifically concerns filmic narratives, it is expected that one or more stories will be told, forming a plot, and that elements such as character, time, objects and space will interact to build the narrative. Any movement or change involving any of those elements may result in a shift in the narrative's deictic center. Such shifts are expected in a narrative, since the plot comprises constant reframing allowing for textual progression.

In Frame Semantics, as already pointed out in this text and elsewhere (Dannélls *et al.*, 2022), the analysis is mostly circumscribed to the lexicon and to grammatical constructions. Only recently has the model been extended to include analyses going beyond sentence level, allowing for the annotation of genre-relevant structure and multimodal text. Such an extension has brought new challenges to the model. This is because signaling meaning correla-



tions between what is expressed in the text and what is expressed by the image is only part of the analysis of texts whose organization presupposes multimodality.

If, in a text purely composed by means of verbal language, the use of some specific lexical item may lead to the activation of a frame, in a multimodal text, such as a filmic narrative, the mere act of a character closing their eyes can (a) evoke a frame, (b) serve as a clue for the comprehender to invoke a frame, and (c) invite the comprehender to reframe their interpretation of the text, requiring them to cognitively move towards a new joint attention micro scene. Those three functions may also be associated with a lexical item, but in a much more conventionalized fashion. One of the parade examples of frame invocation by Fillmore (1985, p. 232) demonstrates (a) and (b). I reproduce the example in (1).

(1) We never open our presents until morning.

If this sentence were to be annotated in FrameNet, it would result in the evocation of the following frames: *Negation*, by *never.adv*; *Closure*, by *open.v*; *Giving*, by *present.n*; *Time\_vector*, by *until.prep*; and *Calendric\_unit*, by *morning.n*.<sup>3</sup> Nonetheless, anyone familiarized enough with the western culture would say that this is a sentence about Christmas, even though there is no LU evoking a frame for modeling the meaning of this festivity. This phenomenon is what Fillmore called frame invocation. In other words, while the LUs listed above are conventionally associated with the frames mentioned, there is an additional frame that plays a key role in the interpretation of (1) which is reconstrued by the comprehender without being grounded in any specific LU. I will claim that frame invocation is even more common in multimodal text. This is because the act of closing one's eyes, in a multimodal context, may evoke, for example, the *Body\_movement* frame, but may also be a clue for the invocation of the *Death* or the *Fall\_asleep* frames. Moreover, it can function as a device signaling the starting point for a new segment of the narrative, where the deictic center is moved towards another setting. However, two key aspects of semiotic modes other than language are at play in this elicited example.

First, it is important to consider that the eye closing per se is not as conventionally associated with the *Death* frame as the verb *to die* is. Language is a more conventionalized mode than the other ones involved in a filmic narrative. The character in question could be going to sleep, or passing out, or, if the shot is in slow motion, blinking. For the invocation of the *Death* frame to take place in this case, other clues are needed in a filmic narrative, such as where the character is located, what the comprehender knows about the character's previous health conditions, the soundtrack and lighting composition choices for the scene, among others. Therefore, the first important question to ask when it comes to the association of visual clues to frames is: which elements in the scene are responsible for that frame to be invoked in the comprehender's mind?

Second, while annotating multimodal text, it is also important to investigate the DC-devices that point to either the constitution of a new joint attention micro scene or to their maintenance in different textual genres. Such devices may be linguistic or not and must

---

<sup>3</sup> All frames mentioned in *Courier* font in this paper can be found at FrameNet Brasil Lab (c2014-2025). The use of such a font is a convention adopted in most FrameNet-related publications.

be accounted for in terms of their central contribution to the progression of the meaning construction process beyond a sentence or a sequence of shots in a film.

Therefore, while annotating multimodal text, mapping the correlations between the frames evoked in the image and those evoked in the text only covers part of the meaning construction process. In a film, the act of closing the eyes may invoke the death of a character or just a kid going to sleep after listening to a bedtime story. Regardless of which of the two is invoked, what comes next in the sequence will rely on the activation of new joint attention micro scenes to be interpreted so that the narrative can advance. In this context, several pragmatic elements and shared knowledge between the comprehender and the communicators would be brought into play to guide expectations regarding a new joint attention micro scene.

The act of closing the eyes, if performed by a character in the narrative, could have its invoked meaning multiplied by other elements in the multimodal composition, such as the slow rhythm in the background soundtrack, a blueish or noir color palette in the image, a tear-drop on the face of another character, a dog lowering its ears. In that case, the act of closing the eyes would trigger the invocation of the Death frame and the sadness atmosphere would both be explained by such invocation and reinforce it.

It is also important to point out that, as much as the act of closing one's eyes may be associated with the invocation of different frames, diverse image sequences may also lead to the invocation of the Death frame. Suppose the filmic narrative, at some point, shows one of the characters taking care of a garden in what, at that point, appears to be a shot in an opening sequence that is not key for the narrative. At a later point of the narrative, a shot of an abandoned garden may serve as a deictic device situating the comprehender at a new attention focus, guiding them towards a new joint attention micro scene. Note that no linguistic or imagetic element would need to be used to explicitly evoke the Death frame in this case. The invocation of such a frame would derive from a reframing of the role of that initial gardening shot in contrast with the new aspect of the garden. Therefore, non-verbal deictic devices would have been used to signal to the comprehender the need to seek an explanation for the change in the state of the garden, identifying the possible intentionality behind the communicators' choice. If, from the opening sequence, the comprehender has learnt that the character takes good care of the garden, and now the garden is no longer in good shape, some element in the frame referring to the gardening process must be no longer present. In this case, the death of the gardener could be one of the hypotheses for the current state of the garden and the progression of the narrative would confirm or not the possibility of invoking the Death frame.

Of course, there is an individual dimension to the interaction between the comprehender and the text, meaning that the effects intended by the communicators may not be successfully achieved, while other ones which were not predicted may occur. However, this possibility is not exclusive to multimodal text and the analytical focus should rely on the intentionality behind the elements chosen to enable the progression of the narrative. Focusing on that will of course bring new challenges to FrameNet annotation, namely that of looking beyond the immediate co-text of the element evoking a frame and looking more closely into frame invocation, an aspect of Frame Semantics not properly explored in its computational implementation (Torrent *et al.*, 2022). Nonetheless, it could open a series of new

possibilities in using FrameNet annotation for building semantic representations of multimodal genres which capture key aspects of textual progression.

In the final section of this paper, I will provide a sample analysis of real data based on the model sketched so far.

## 5 Sample Analysis

The incorporation of the mediated joint attention scene model and the consequent multimodal approach to deictic center shifts to FrameNet multimodal annotation requires new analytical categories to be included in the model. To demonstrate the feasibility of this proposal, in this section I analyze *The Neighbors Window*, an Academy winning short film by Marshall Curry.<sup>4</sup> I start by showing the aspects of multimodal composition currently accounted for by the FrameNet Brasil multimodal annotation model (Belcavello *et al.*, 2020, 2022; Belcavello, 2023), and then, present the additional analytical layers made possible by the proposal devised in this paper.

### 5.1 FrameNet Multimodal Annotation of Dynamic Images

For exemplifying the methodology proposed by Belcavello *et al.* (2022), I will consider the sequence of the short film in Figure 3. In the sequence, which is the first of the film, we see the main character Alli cleaning up the floor amidst kids toys. She bangs her head on the table when standing up. Next, her husband enters the room, after successfully putting the kids to sleep. The couple sits to have dinner, and Alli's attention turns to something in front of them. We realize next that it is a scene of her neighbors having sex.

---

<sup>4</sup> Available at *The Neighbors* (2019).

Figure 3: Sequence of stills for The Neighbors Window

 <p>&lt;Background music fades&gt;</p>	 <p>Alli: Ow! Oh my God...</p>	 <p>Alli: Wow, are they down? Nice job. That was fast. Husband: Yeah, They're whooped.</p>
 <p>Alli: I thought I was gonna lose it when they wouldn't quit with that Captain Underpants thing. Husband: They have no idea what it was like to grow up back when kids got spanked.</p>	 <p>Alli: Oh God ... That looks good! Husband: Have a little. Alli: No, I shouldn't.</p>	 <p>Alli: What the...? Husband: What?</p>
 <p>Husband: Wow.</p>	 <p>Alli: Oh my god! Oh, I can't watch. Husband: As she keeps watching.</p>	 <p>Alli: But seriously, they need to order some drapes.</p>

Source: The Neighbors (2019).

Following Belcavello's (2023) methodology, the first step in the analysis is that of annotating the audio transcription for frames and frame elements (FEs). For instance, the sentence spoken by Alli when the eighth video frame in Figure 3 is shown could be annotated for the *Perception\_active* frame, shown in Figure 4. This frame is evoked by *watch.v*. The resulting annotation is seen in Figure 5. Note that the *Perceiver\_agentive* FE is assigned to I, while the *Phenomenon* is noted as a definite null instantiation, meaning that it can be inferred or found in the context of the annotation, but not in the syntactic locality of the lexical unit that is the target of the annotation. In this specific case, because this is a multimodal annotation, and annotators are instructed to analyze the text while watching the video, they know that both the *Perceiver\_agentive* and the *Phenomenon* FEs will be shown in the video.

Figure 4: The Perception\_active frame in FrameNet Brasil

Perception\_active

@Psychology #60 Perception\_active [en] PDF

Definition

This frame contains perception words whose perceivers intentionally direct their attention to some entity or phenomenon in order to have a perceptual experience. For this reason we call the perceiver role in this frame **Perceiver\_agentive**.

She|per **gazed** **upon him** **fondly**.

Comparing the Perception\_active frame to the Perception\_experience frame, we note that for some modalities there are different lexical items in each frame. For instance, whereas Perception\_active contains the verb phrase look at, Perception\_experience contains see. For other sense modalities, we find the same lexical item in both frames. To illustrate, consider the verb smell.

This first sentence exemplifies the Perception\_active use of the verb smell: **Smell this to see if it's fresh.** **CNI**

This second sentence exemplifies its Perception\_experience sense:  
I **smell** something rotten.

Frame Elements

Core			
<b>Perceiver_agentive</b>	The <b>Perceiver_agentive</b> performs some action in order to have a perceptual experience.	CoreSet: Body_part	@sentient
<b>Phenomenon</b>	The <b>Phenomenon</b> indicates the entity or phenomenon to which the <b>Perceiver_agentive</b> directs his or her attention in order to have a perceptual experience.	CoreSet: Direction	

Source: FrameNet Brasil Lab (c2014-2025).

Figure 5: Sentence annotated for the Perception\_active frame

Details multimodal The\_neighbors\_window #603233 Previous Next

Oh my god! Oh, I can't watch.

LU: Perception\_active.watch.v Alternative LUs #181634

CNI Oh my god! Oh, I can't watch.

DNI

Phenomenon

INI

INC

Perceiver\_agentive

Phenomenon

Duration

Ground

Manner

Means

Obscuring\_medium

Place

Purpose

State

Time

Depictive

Expected\_entity

Location\_of\_protagonist

Body\_part

Direction

Comment

Delete this AnnotationSet

Source: FrameNet Brasil Lab (c2014-2025).

Once the transcribed audio is annotated for frames and FEs, the methodology states, as the next step, that the video sequences are annotated. The video annotation is shown in Figures 6 and 7. In Figure 6, the annotation represents the moment where Alli notes the neighbors having sex in the other building and focuses her attention on that scene. Note that, on top of recording that this portion of the image represents a FE in the Perception\_active frame, the annotation also indicates that the object being annotated is a *woman.n*, a lexical unit of the People frame.

Next, in Figure 7, Objects 2 and 3 are annotated at the same portion of the image. Object 2 indicates the instantiation of the Sex frame and marks the Participants FE for this frame, which is labeled using the *couple.n* lexical unit evoking the Personal\_rela-

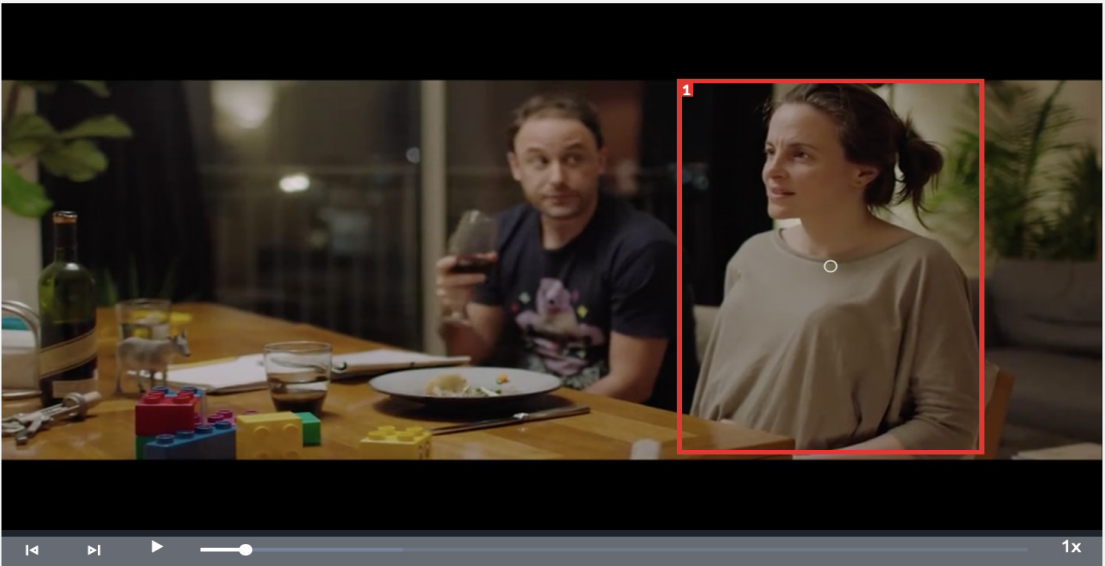


tionship frame. Object 3 is a clone of Object 2 in terms of spatial and temporal properties in the video, but registers the fact that the sexual act happening in the other apartment is the Phenomenon FE of the Perception\_active frame which was a definite null instantiation (DNI) in the text annotation.

As the example shows, the methodology proposed by Belcavello (2023) already captures several complexities of the meaning-making process involved in the interpretation of the filmic narrative presented to the comprehender. Next, I propose an augmentation of the analytical categories FrameNet Brasil uses for the annotation of multimodal genres to capture the role of deictic center shifts in the process of interpreting filmic narratives.

Figure 6: Annotation of the Perceiver\_agentive FE as Object #1

Home > Dynamic Annotation > The\_neighbors\_window



1725 [68.96s] Video paused Object #1 Status showing 30936 [1237s]

Object

Tracking

Create Start Pause Stop at frame #1725 Show/Hide All Clear

The neighbors window #1490 Build Sentences

Current Object: #1 Range 1725/1725 #17047

Frame FE CVName

Perception\_active Perceiver\_agentive People.woman.n

Start frame End frame Is blocked?

1725 1725

Save Clone

Objects Sentences


#object	Start	End	#id	FrameElement	CVName
#1	1725	1725	#17047	Perception_active.Perceiver_agentive	People.woman.n

Source: FrameNet Brasil Lab (c2014-2025).



Figure 7: Annotation of the Participants FE in the Sex frame and of the Phenomenon FE in the Perception\_active frame as Objects #2 and #3

Home > Dynamic Annotation > The\_neighbors\_window



1789 [71.52s] Video paused Object #2 Status showing 30936 [1237s]

Object

Tracking

Create Start Pause Stop at frame #1789 Show/Hide All Clear

The neighbors window #1490 Build Sentences

Current Object: #2 Range 1789/1915 #17038

Frame FE CVName

Sex Participants Personal\_relationship.couple.n

Start frame End frame Is blocked?

1789 1915

Save Clone

Objects Sentences

	#object	Start	End	#id	FrameElement	CVName
✖	#1	1725	1725	#17047	Perception_active.Perceiver_agentive	People.woman.n
✖	#2	1789	1915	#17038	Sex.Participants	Personal_relationship.couple.n
✖	#3	1789	1915	#17044	Perception_active.Phenomenon	Sex.sex.n

Source: FrameNet Brasil Lab (c2014-2025).

## 5.2 Annotating for Deictic Center Shifts in FrameNet

The methodology for the annotation of deictic center shifts I present in this paper was based on the categories devised by Zubin and Hewitt (1995). Such categories, in turn, were based on studies focusing on written narratives and the role of deictic centers in their progression. They served as the basis for the definition of the types of deixis to be annotated. Next, the idea was to find ways through which changes in the deictic center could be annotated for different communicative modes. The notion of communicative modes presented by Kalantzis, Cope and Pinheiro (2020) proved to be a good foundation for the establishment of categories for annotation.

Nonetheless, the methodology still lacked granularity. For the visual mode, for instance, video editing choices and resources needed better definition. I used the work by Bordwell and Thompson (2013) as the basis for the definition of filmic categories for the visual mode, such as framing, cut, among others. The categories proposed are presented in Table 1.

Table 1: Categories for the annotation of deictic center shifts in filmic narratives

Macro category	Types	Trigger/ Object	Specification	How to annotate
Deictic center	Reader's DC			Checkbox
	Character's DC			
	Reader's and character's DC			
Deictic operations	DC Introduction			Checkbox
	DC Maintenance			
	DC Change			
	DC Annulment			
Types of deixis	Place			Checkbox
	Person			
	Time			

Meaning-making mode	Gestural		Bounding box
	Tactile		Bounding box
	Spatial	Scene	Bounding box
		Region	
	Visual	Character	Bounding box
		Artifact	
		Camera angle	Checkbox
		Shot	Checkbox
		Camera motion	Checkbox
	Cut	Cut	Checkbox
		Fade in	
		Fade out	
		Fusion	
		Jump cut	
		Cut in	
		Final cut	

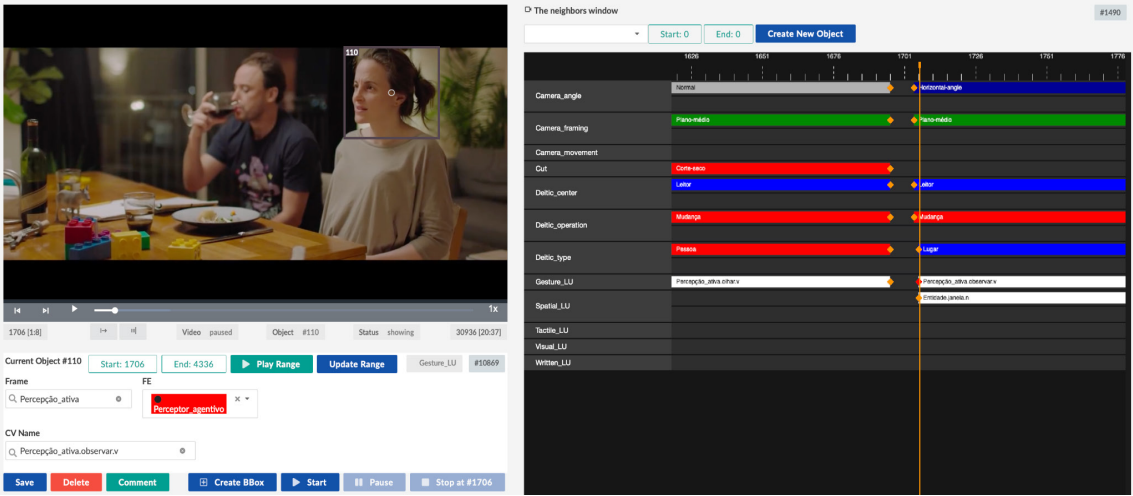
	Written	In scene	Bounding box
		Overlay	Bounding box
	Oral		Text annotation
	Audio	Sound effect	Checkbox
		Soundtrack	

Source: elaborated by the author.

Caption: blank cells indicate the lack of intermediate categories for a given level of analysis.

Once categories were defined, a first round of annotation of the short film *The Neighbors Window* was carried out using a spreadsheet. This experiment showed that the timespan was the main axis along which annotation needed to be performed, since, in the demarcation of changes to the deictic center, one needs to not only indicate when the DC changes, but also how long it takes until another change happens. The need for indicating specific elements in the video which could trigger the changes in the DC was also clear. The resulting interface is shown in Figure 8.

Figure 8: Annotation of the deictic center shift marked by the change in Alli's focus of attention towards the neighbors' window



Source: FrameNet Brasil Lab (c2014-2025).

Figure 8 shows the annotation for the moment the DC changes from the conversation between the couple over dinner to the event at the neighbor's apartment. Note that the change in this case refers to the reader's DC, that is, the intention of the shot composition is to invite the reader to pay attention to the fact that something in the dynamics of the story will change. Alli's focus is no longer on her husband, or on the dinner, but on something else. The horizontal camera angle in a medium shot of the living room helps build the setting for the change. The meaning-making mode used as the main trigger for the DC shift is gestural, meaning that it is the change in Alli's gaze towards the window that provides the main clue

for the reader to notice the DC shift. Note that the annotation system allows for the specification of the gesture in terms of frame and FE, since it can be annotated as a bounding box.

The annotation exemplified in Figure 8 adds to the meaning representation of multimodal objects currently used in FrameNet Brasil multimodal annotation, to the extent to which it allows for frame semantic information to be correlated with shifts in the narrative progression. In the near future, the main goal is that of augmenting and refining the categories related to meaning-making modes such as audio, for example.

## 6 Conclusions and Outlook

In a scenario where text analysis must go beyond the analysis of sequences of characters, linguists agree on the need to broaden the scope of the science they practice. More specifically, in areas such as Computational Linguistics, where the demand for multimodal datasets is growing, research initiatives have been undertaken as a means of grounding language analyses and what machines can learn from them. Even if the kind of multimodal analysis currently implemented by FrameNet Brasil represents an advance in the area, limiting the annotation to what is said and directly seen in multimodal genres leaves out foundational aspects of textual progression.

In this context, I have argued in favor of repositioning Frame Semantics so that its computational implementation – FrameNet – can embrace the analysis of deictic devices and their role in textual progression. Considering the foundational role of such devices in human communication (cf. Tomasello, 2003), analyzing their behavior in multimodal texts is key in the process of representing how such texts create meaning.

To illustrate this claim, I proposed exemplar analyses of filmic narratives in which I showed how shifts in the deictic center triggered by different communicative modes in interaction allow for the progression of the narrative. These analyses were performed within a model of Mediated Joint Attention, based on Tomasello's (2018) model of in presence communication and drawing from Turner's (2018) notion of blended joint attention scenes. Within this new model, in a filmic narrative, for example, the joint attention scene is established through some medium in which the intentionalities of one or more communicators are projected. Those communicators intentionally act to produce a multimodal text aimed at potential comprehenders. In this process, communicators make use of several communicative modes appealing to human sensory perception, mostly vision and hearing. All those modes collaborate for meaning making in multimodal texts and can be used for promoting shifts in the deictic centers, creating new joint attention micro scenes. Looking at only one of these semioses would be ignoring important aspects of these texts and their meaning.

The analytical model proposed reinforces the need for FrameNet multimodal analyses to consider (a) the inclusion of other communicative modes in the annotation methodology, (b) the need to build multidisciplinary teams including experts in Music, Film Analysis, Semiotics and Linguistics, at least, and (c) the need to consider other annotation levels that make room for categories lying closer to the pragmatic end of the Semantics-Pragmatics continuum.

I hope the proposed model can contribute to broadening the Frame Semantics perspective on the field of Multimodal Analysis, with replications for both Computational

Linguistics and Language Pedagogy. I believe that more comprehensive representations of how texts invite readers to construe meaning and of how such meaning-making process progresses as they read may help students in thinking critically about multimodal texts and further analyzing them.

## Acknowledgements

The analytical proposal devised in this paper was developed during the author's sabbatical period as a visiting researcher at the Department of Swedish, Multilingualism and Language Technology of the University of Gothenburg. The author acknowledges the contributions of the undergraduate student annotators involved in the pilot analyses aimed at testing the analytical model: Rian Costa and Taís Reis, whose work was supported by scholarships provided by the Federal University of Juiz de Fora and FAPEMIG, respectively. The author is also thankful to Dr. Ely Matos for all the support in developing the module of the FrameNet Brasil WebTool used for the deixis annotation.

## References

- BAKHTIN, M. M. *Esthétique de la création verbale*. Paris: Gallimard, 1984.
- BATEMAN, J. *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. Berlin: Springer, 2008.
- BATEMAN, J.; WILDFEUER, J.; HIIPALA, T. *Multimodality: Foundations, Research and Analysis – A Problem-Oriented Introduction*. Berlin: Walter de Gruyter GmbH & Co, 2017.
- BELCAVELLO, F. *FrameNet Annotation for Multimodal Corpora: Devising a Methodology for the Semantic Representation of Text-Image Interactions in Audiovisual Productions*. 2023. 134 f. Thesis (Doctorate in Linguistics) – Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora, 2023.
- BELCAVELLO, F.; VIRIDIANO, M.; COSTA, A. D. da; MATOS, E. E. da S.; TORRENT, T. T. Frame-Based Annotation of Multimodal Corpora: Tracking (A) Synchronies in Meaning Construction. In: INTERNATIONAL FRAMENET WORKSHOP: TOWARDS A GLOBAL, MULTILINGUAL FRAMENET, 2020, Marseille. *Proceedings* [...]. Marseille: European Language Resources Association (ELRA), 2020. p. 23-30.
- BELCAVELLO, F.; VIRIDIANO, M.; MATOS, E.; TORRENT, T. T. Charon: A FrameNet Annotation Tool for Multimodal Corpora. In: LINGUISTIC ANNOTATION WORKSHOP (LAW-XVI), 16. Marseille. *Proceedings* [...]. Workshop presented at LREC2022. Marseille: ELRA, 2022. p. 91-96.
- BORDWELL, D.; THOMPSON, K. *A arte do cinema: uma introdução*. Tradução: Roberta Gregoli. Campinas, SP: Editora da Unicamp; São Paulo: Editora da USP, 2013.
- BÜHLER, K. *Sprachtheorie*. Jena: Fischer, 1934.
- CIENKI, A. The Study of Gesture in Cognitive Linguistics: How It Could Inform and Inspire Other Research in Cognitive Science. *Cognitive Science*, v. 13, n. 6, e1623, 2022.



- COHN, N. A Multimodal Parallel Architecture: A Cognitive Framework for Multimodal Interactions. *Cognition*, v. 146, p. 304-323, 2016a.
- COHN, N. From Visual Narrative Grammar to Filmic Narrative Grammar: The Narrative Structure of Static and Moving Images. In: WILDFEUER, J.; BATEMAN, J. (ed.). *Film Text Analysis: New Perspectives on the Analysis of Filmic Meaning*. London: Routledge, 2016b. p. 94-117.
- COHN, N. Visual Narrative Structure. *Cognitive Science*, v. 37, n. 3, p. 413-452, 2013.
- DANNÉLLS, D.; TORRENT, T. T.; SIGILIANO, N. S.; DOBNIK, S. Beyond Strings of Characters: Resources Meet NLP – Again. In: VOLODINA, E.; DANNÉLLS, D.; BERDICEVSKIS, A.; FORSBERG, M.; VIRK, S. (ed.). *Live and Learn: Festschrift in Honor of Lars Borin*. Gothenburg: Institutionen för Svenska, Flerspråkighet och Språkteknologi – Göteborgs Universitet, 2022. p. 29-36.
- DUTRA, L. V.; SIGILIANO, N. S. Ferramenta linguístico-computacional como facilitadora para o ensino de gramática na escola. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 2021, [S. l.]. *Anais [...]. [S. l.]*: SBC, 2021. p. 432-436.
- FILLMORE, C. J. Frame Semantics and the Nature of Languages. In: ANNALS OF THE NEW YORK ACADEMY OF SCIENCES. *Conference on the Origin and Development of Language and Speech*, v. 280, p. 20-32, 1976. Work presented at the Conference on the Origin and Development of Language and Speech, 1976, New York.
- FILLMORE, C. J. Frame Semantics. In: THE LINGUISTIC SOCIETY OF KOREA (ed.). *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Company, 1982. p. 111-137.
- FILLMORE, C. J. Frames and the Semantics of Understanding. *Quaderni di Semantica*, v. 6, n. 2, p. 222-254, 1985.
- FILLMORE, C. J. *Santa Cruz Lectures on Deixis, 1971*. Bloomington, IN: Indiana University Linguistics Club, 1975.
- FILLMORE, C. J. The Case for Case Reopened. In: COLE, P.; SADOCK, J. M. (ed.). *Syntax and Semantics: Grammatical Relations*. New York: Academic Press, 1977. v. 8, p. 59-81.
- FILLMORE, C. J. The Merging of Frames. In: ROSSINI FAVRETTI, R. (ed.). *Frames, Corpora, and Knowledge Representation*. Bologna: Bononia University Press, 2008. p. 1-12.
- FILLMORE, C. J.; PETRUCK, M. R.; RUPPENHOFER, J.; WRIGHT, A. FrameNet in Action: *The Case of Attaching*. *International Journal of Lexicography*, v. 16, n. 3, p. 297-332, 2003.
- FRAMENET BRASIL LAB. *WebTool*. c2014-2025. Available at: <http://webtool.frame.net.br/>. Accessed on: 31 Jan. 2025.
- GOLDIN-MEADOW, S. Pointing Sets the Stage for Learning Language – and Creating Language. *Child Development*, v. 78, n. 3, p. 741-745, 2007.
- GRICE, H. P. Meaning. *The Philosophical Review*, v. 66, n. 3, p. 377-388, 1957.
- HANKS, W. F. *Língua como prática social: das relações entre língua, cultura e sociedade a partir de Bourdieu e Bakhtin*. Textos selecionados por Hanks, traduzidos por Anna Christina Bentes, Marco Antônio Rosa Machado, Marcos Rogério Cintra e Renato C. Rezende. São Paulo: Cortez, 2008.
- JACKENDOFF, R. *Foundations of Language*. Oxford: Oxford University Press, 2002.

- JACKENDOFF, R.; LERDAHL, F. The Capacity for Music: What is It, and What's Special About It? *Cognition*, v. 100, p. 33-72, 2006.
- KALANTZIS, M.; COPE, B.; PINHEIRO, P. *Letramentos*. Campinas: Unicamp, 2020.
- KITA, S. *Pointing: Where Language, Culture and Cognition Meet*. Mahwah, NJ: Lawrence Erlbaum, 2003.
- LEMKE, J. Multiplying Meaning. In: MARTIN, J. R.; VEEL, R. (ed.). *Reading Science: Critical and Functional Perspectives on Discourses of Science*. London: Routledge, 1998. p. 87-114.
- LEVINSON, S. C. Deixis. In: HORN, L. R.; WARD, G. (ed.). *The Handbook of Pragmatics*. Malden, MA: Blackwell Publishing, 2004. p. 97-121.
- LISZKOWSKI, U.; BROWN, P.; CALLAGHAN, T.; TAKADA, A.; DE VOS, C. A. Prelinguistic Gestural Universal of Human Communication. *Cognitive Science*, v. 36, p. 698-713, 2012.
- RODRÍGUEZ, C.; MORENO-NÚÑEZ, A.; BASILIO, M.; SOSA, N. Ostensive Gestures Come First: Their Role in the Beginning of Shared Reference. *Cognitive Development*, v. 36, p. 142-149, 2015.
- SAUSSURE, F de. *Cours de linguistique générale*. Paris: Payot, 1916.
- SEARLE, J. R. Collective Intentions and Actions. In: COHEN, O.; MORGAN, J.; POLLACK, M. (ed.). *Intentions in Communication*. Cambridge, MA: MIT Press, 1990.
- SEGAL, E. M. Narrative Comprehension and the Role of Deictic Shift Theory. In: DUCHAN, J.; BRUDER, G. A.; HEWITT, L. E. (ed.). *Deixis in Narrative: A Cognitive Science Perspective*. London: Psychology Press, 1995. p. 3-18.
- STEEDMAN, M. J. Plans, Affordances, and Combinatory Grammar. *Linguistics and Philosophy*, v. 25, n. 5-6, p. 723-753, 2002.
- STEEN, F.; HOUGAARD, A.; JOO, J.; OLZA, I.; CÁNOVAS, C.; PLESHAKOVA, A.; RAY, S.; UHRIG, P.; VALENZUELA, J.; WOŹNY, J.; TURNER, M. Toward an Infrastructure for Data-Driven Multimodal Communication Research. *Linguistics Vanguard*, v. 4, n. 1, e20170041, 2018.
- SWEETSER, E. Looking at Space to Study Mental Spaces: Co-Speech Gesture as a Crucial Data Source in Cognitive Linguistics. In: GONZALEZ-MARQUEZ, M.; SPIVEY, M. J.; COULSON, S.; MITTELBERG, I. (ed.). *Methods in Cognitive Linguistics*. Amsterdam: John Benjamins, 2007. p. 201-224.
- THE NEIGHBORS Window. Written and directed by Marshall Curry. New York: Marshall Curry Productions, 2019. Available at: <http://www.theneighborswindow.com/>. Accessed on: 31 Jan. 2025.
- THE NEW LONDON GROUP. A Pedagogy of Multiliteracies: Designing Social Futures. *Harvard Educational Review*, v. 66, n. 1, p. 60-93, 1996.
- TOMASELLO, M. *Becoming Human: A Theory of Ontogeny*. Cambridge, MA: The Belknap Press of Harvard University Press, 2018.
- TOMASELLO, M. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press, 2003.
- TOMASELLO, M. *Origins of Human Communication*. Cambridge, MA: MIT Press, 2008.
- TOMASELLO, M. *The Cultural Origins of Human Cognition*. Cambridge, MA: MIT Press, 1999.

TOMASELLO, M.; CARPENTER, M.; LISZKOWSKI, U. A New Look at Infant Pointing. *Child Development*, v. 78, n. 3, p. 705-722, 2007.

TORRENT, T. T.; MATOS, E. E. 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, e838441, 2022.

TURNER, M. Multimodal Form-Meaning Pairs for Blended Classic Joint Attention. *Linguistics Vanguard*, v. 3, n. s1, e20160043, 2017.

TURNER, M. The Role of Creativity in Multimodal Construction Grammar. *Zeitschrift für Anglistik und Amerikanistik*, v. 66, n. 3, p. 357-370, 2018.

VIRIDIANO, M.; TORRENT, T. T.; CZULO, O.; LORENZI, A.; MATOS, E.; BELCAVELLO, F. The Case for Perspective in Multimodal Datasets. In: WORKSHOP ON PERSPECTIVIST APPROACHES TO NLP, 1., 2022, Marseille. *Proceedings [...]*. Workshop presented at LREC2022. Marseille: European Language Resources Association, 2022. p. 108-116.

WILKINS, D. Why Pointing with the Index Finger is not a Universal (in Sociocultural and Semiotic Terms). In: KITA, S. (ed.). *Pointing: Where Language, Culture and Cognition Meet*. Mahwah, NJ: Lawrence Erlbaum, 2003. p. 171-216.

WITTGENSTEIN, L. *Philosophische Untersuchungen*. London: Kegan Paul, 1953.

ZUBIN, D. A.; HEWITT, L. E. The Deictic Center: A Theory of Deixis in Narrative. In: DUCHAN J. F.; BRUDER G. A.; HEWITT, L. E. (ed.). *Deixis in Narrative: A Cognitive Science Perspective*. New York: Routledge: Taylor & Francis Group, 1995. p. 129-155.

Varia

# Uma casa não é um lar: o afeto e o espaço físico em *Madame Bovary* (1856)

## *A House is Not a Home: The Affection and the Physical Space in Madame Bovary (1856)*

**Laís Marx Umpierre**

Universidade Federal do Rio Grande do Sul (UFRGS) | Porto Alegre | RS | BR  
lais.marx@ufrgs.br  
<http://orcid.org/0000-0003-4687-7601>

**Luísa Freire**

Universidade de Lille | Lille | FR  
luisagfreire@gmail.com  
<http://orcid.org/0000-0002-2415-6320>

**Resumo:** O presente trabalho almejou examinar quatro elementos do espaço físico de *Madame Bovary* pelo viés do afeto e do espaço psicológico das personagens, analisando o espaço a partir de Bachelard e Weisgerber e os coloridos do discurso do narrador e de Emma Bovary com base em Bakhtin. Dessa forma, demonstrou-se como a visão dos personagens influencia o discurso do narrador na descrição dos locais estudados, são eles: (i) a casa de Tostes, (ii) o castelo de *La Vaubyessard*, (iii) a casa de Yonville; e, por fim, (iv) o quarto no *Hôtel de Boulogne*. Percebeu-se essas interferências pelas emoções da protagonista refletidas nas escolhas lexicais do narrador, dentre as quais se focou, sobretudo, na dicotomia casa/lar (*sa/ma maison* e *chez elle/chez nous*). Assim, concluiu-se que a forma através da qual os ambientes são descritos no romance é permeada pelos sentimentos e percepções das personagens, especialmente de Emma Bovary.

**Palavras-chave:** Madame Bovary; Flaubert; espaço físico; afeto.

**Abstract:** The present paper aimed to examine four elements of the physical space in *Madame Bovary* through the perspective of affection and the characters' psychological space, analysing the space based on Bachelard and Weisgerber and the tones of the narrator and Emma's speech, from a bakhtinian approach. Thus, it was demonstrated that the characters' vision influences the description of the studied spaces through the narrator's speech. Four spaces from the novel were chosen: (i) the house in Tostes; (ii) the castle of *La Vaubyessard*; (iii) the house in Yonville; and, finally, (iv) the room in *Hôtel de Boulogne*. These interferences manifest themselves



because of the protagonist's emotions, which were reflected in the narrator's lexical choices, among which the analysis focused mainly on the house/home dichotomy (*sa/ma maison* e *chez elle/chez nous*). Therefore, it was concluded that the manner in which the spaces are described is permeated by the characters' feelings and perceptions, especially those of Emma Bovary.

**Keywords:** Madame Bovary; Flaubert; physical space; affection.

## 1 Considerações iniciais

O presente artigo propõe-se a discutir as relações entre o afeto e o espaço físico em *Madame Bovary*, romance de Gustave Flaubert, publicado em quatro folhetins pela *Revue de Paris* em 1856 e, posteriormente, em forma de livro, em dois volumes, pelo selo de Michel Lévy em 1857.<sup>1</sup> Considerado como marco inicial do realismo francês pela sua originalidade, suas características inspiraram uma escola literária e uma geração de artistas e escritores.

Verificamos a aproximação da narrativa com o real, sobretudo, pela inspiração da trama ser oriunda de um *fait divers*,<sup>2</sup> sua essência está calcada em um acontecimento concreto. No entanto, apesar da autenticidade inerente à procedência do romance analisado, frisamos, desde já, que problematizamos as supostas objetividade e neutralidade do narrador distanciado de Flaubert. Como expõe Bakhtin (2011, p. 309), “um enunciado absolutamente neutro é impossível. A relação valorativa com o objeto do discurso [...] também determina a escolha dos recursos lexicais, gramaticais e composicionais do enunciado”. O anseio pela impessoalidade da narrativa era um projeto do realismo, porém, sua inviabilidade se evidencia no desenovelar da narrativa.

Com essa breve explicação contextual do romance, passamos ao tema que guia nossa reflexão. No gênero narrativo, em oposição ao cinema e ao teatro, por exemplo, os espaços físicos presentes nas obras são arquitetados apenas por palavras, por descrições dos elementos que os compõem. A apreensão dessas criações autênticas não ocorre por meio da visão, de imagens ou de objetos concretos, precisamos concebê-las por meio do texto. Nesse sentido, a despeito da focalização do narrador, entrevemos os cenários dos eventos através dos olhos de outrem, portanto, trata-se de uma visão parcial, suscetível a emoções. Podemos, assim,

---

<sup>1</sup> A primeira tradução para o português a circular no Brasil data de 1881, feita pelo tipógrafo português Francisco Ferreira da Silva Vieira (Müller, 2013).

<sup>2</sup> “Notícia cujo interesse reside naquilo que tem de insólito, extraordinário, surpreendente” (*Houaiss eletrônico*, 2009). Este breve artigo jornalístico não é estranho à literatura, originando outros romances famosos, como *Le Rouge et le noir* (1830) de Stendhal.



encontrar traços das sensações daquele que descortina os espaços – ou, no caso de um narrador que se aproxima de uma personagem, vislumbramos algumas sensações desta.

Após um breve resumo de *Madame Bovary*, dedicamo-nos a aprofundar a complexidade do narrador do romance, concomitantemente realista e tendencioso e de significância para a edificação dos elementos espaciais. Além disso, observamos também alguns dos aspectos realistas que fazem de *Madame Bovary* um célebre romance do século XIX. Iniciamos nossa análise, então, abordando quatro espaços significativos para a protagonista da trama, são eles (i) a casa de Tostes, (ii) o castelo de *La Vaubyessard*, (iii) a casa de Yonville; e, por fim, (iv) o quarto no *Hôtel de Boulogne*. Entremeamos o estudo desses com apontamentos acerca do espaço psicológico de Emma, demonstrando a influência de suas emoções nas apresentações dos espaços examinados. Em seguida, concentramo-nos em apresentar o cerne deste estudo, argumentando a evidente diferença entre “casa” e “lar” no romance. Para tanto, embasamo-nos principalmente nos estudos sobre o espaço na narrativa de Gaston Bachelard (1979) e Jean Weisgerber (1971), a fim de discorrer sobre como a afetividade influencia e molda a percepção do ambiente das personagens que o vivenciam em *Madame Bovary*.

## 2 *Madame Bovary*, o romance realista por excelência

No romance de Gustave Flaubert, conhecemos Emma Rouault, a jovem filha de um fazendeiro, amante da literatura, que sonha viver um amor como aqueles dos livros e por quem Charles Bovary se apaixona. Após se casarem, mudam-se para Tostes, onde Charles exerce a medicina. No entanto, Emma começa a sentir as frustrações da vida pacata de uma esposa em uma pequena comunidade. Para deixar sua nova esposa feliz, Bovary procura uma posição em Yonville, para onde se mudam ao descobrirem que Emma está grávida de Berthe. Nesse vilarejo, Emma conhece Léon Dupuis, ambos sentem uma atração imediata, porém, antes de expressarem-na, o escrivão muda-se para Rouen. Em seguida, Rodolphe Boulanger corteja madame Bovary e ambos deixam-se seduzir; assim começa o primeiro caso amoroso de Emma. O relacionamento encerra-se de maneira abrupta: Rodolphe parte para evitá-la e Emma adoece, aproximando-se da morte. Após muitos meses, os Bovary reencontram Léon em uma ópera em Rouen; Emma e Léon declaram seu antigo afeto e começam um relacionamento amoroso, encontrando-se toda semana no *Hôtel de Boulogne*. Com problemas econômicos agravando-se durante vários meses, os Bovary tomam decisões irresponsáveis, Emma se desespera e recorre a todos que poderiam ajudá-la. Desamparada, madame Bovary toma uma decisão trágica.

### 2.1 O realismo e o narrador em *Madame Bovary*

*Madame Bovary* se caracteriza como uma das obras mais significativas para o realismo francês. Alguns dos elementos mais relevantes dessa escola são a busca pelo verossímil e pela objetividade, ou seja, visava a relatar uma história próxima à realidade, por meio de um narrador imparcial. Em um movimento contrário aos exageros do romantismo, o realismo planeja reproduzir a realidade “nua e crua”. Considerado “o romance dos romances”, de acordo com Thierry Laget, ensaísta e crítico literário francês:

*Madame Bovary* é sobretudo filha de uma época e de uma civilização que encontram sua glória em criar a fundição, locomotivas, tecidos de algodão e poetas desamparados pela vulgaridade e brutalidade dos tempos, impotentes face ao reino da matéria positiva que substitui as épocas meditativas (Laget, 2020, p. 14, tradução nossa).<sup>3</sup>

No seu romance mais famoso, Flaubert engloba a vida como ela era na França do século XIX, os valores relevantes à sociedade da época estão presentes na trágica narrativa de Emma Bovary. Desde o casamento, a agricultura, o comércio até o Estado, a Igreja e o romance; enfim, a vida em sua totalidade (Laget, 2020). A verossimilhança prezada pela escola realista é manifesta na trama, bem como na voz que a relata. Contudo, talvez mais importante seja a sua influência nas temáticas que atravessam e guiam as ações das personagens.

Em coerência com o projeto de impessoalidade discursiva do realismo, o narrador em *Madame Bovary* é heterodiegético (Genette, 1972), ou seja, a voz narrativa não integra o rol das personagens, não vivencia os acontecimentos, mas os relata. Embora onisciente, o narrador do romance em análise elege (intencionalmente ou não) Emma Bovary como a principal personagem a ser acompanhada e refletida em sua narração – podemos pensar em uma “onisciência seletiva” (Friedman, 2002). Esse é um aspecto relevante, uma vez que a descrição do quadro espacial do romance – que imaginamos objetiva e isenta de julgamentos – é maculada pelas emoções de Emma, ecoadas pela voz narrativa.

Entretanto, como veremos adiante, diferente da descrição das casas de Tostes e Yonville, o retrato do quarto no *Hôtel de Boulogne* repercute sutilmente os sentimentos de Emma e Léon. A título de exemplo, apresentamos os seguintes trechos:<sup>4</sup> “[l’appartement] semblait tout commode pour les intimités de la passion” e “comme ils aimaient cette bonne chambre pleine de gaieté, malgré sa splendeur un peu fanée!” (Flaubert, 2020, p. 349).<sup>5</sup> Nessas passagens, destacamos que o narrador retrata o quarto como “próprio para as intimidades da paixão” ou “belo quarto cheio de alegria”; no momento em que emprega tais atributos em seu discurso, sua percepção mescla-se com a daqueles emocionalmente envolvidos com o espaço descrito no texto. Adentramos o palco dos encontros amorosos, não por uma caracterização imparcial, mas pelo ponto de vista dos amantes, tingido de emoção a ponto de influenciar o narrador.

A busca por uma escrita (ou, no âmbito da diegese, por uma narração) impessoal diverge do valor simbólico que permeia a descrição das personagens e dos espaços, e em razão do qual encontramos indícios do espaço psicológico das personagens. No que toca à construção do cenário do romance, compreendemos ser impossível a neutralidade do narrador, pois o espaço é “uma realidade psicológica em estado puro: criada e habitada pelo ser que escreve e que, por assim dizer, confunde-se com ela” (Weisgerber, 1971, p. 161, tradução

<sup>3</sup> No original: “*Madame Bovary* est surtout fille d’une époque et d’une civilisation qui trouvent leur gloire à produire de la fonte, des locomotives, des toiles de cotons, et des poètes désemparés par la vulgarité et la brutalité des temps, impuissants face au règne de la matière positive qui se substitue aux âges méditatifs” (Laget, 2020, p. 14).

<sup>4</sup> Em virtude de as escolhas lexicais em língua francesa serem significativas para nossa análise, optamos por manter no corpo do texto as citações originais do romance e, em notas de rodapé, as citações em língua portuguesa.

<sup>5</sup> Em português: “[o cômodo] parecia mesmo próprio para as intimidades da paixão” e “como eles gostavam daquele belo quarto cheio de alegria, apesar de seu esplendor um tanto gasto!” (Flaubert, 1981, p. 198, tradução de Araújo Nabuco).

nossa).<sup>6</sup> Desse modo, apesar de ser heterodiegético, ao vivenciar os espaços através do olhar das personagens, a voz narrativa confunde suas percepções com as dos sujeitos que compõem os eventos e mesmo os ambientes.

### 3 O quadro espacial afetivo em *Madame Bovary*

Por tratar-se de uma narrativa longa, no seu decorrer, somos transportados por diversos recintos caracterizados por valores ora positivos, ora negativos e ora inclinando-se para um distanciamento. Em vista de nosso objetivo, realizamos um recorte e enfocamos quatro espaços do romance: (i) a primeira residência em que Emma morou com Charles após o casamento em Tostes; (ii) o castelo de *La Vaubyessard*, local que marca a fuga da vida provincial pacata de Emma; (iii) o segundo domicílio dos Bovary em Yonville; e (iv) alguns locais de Rouen, principalmente o quarto em que Emma e Léon se encontravam no *Hôtel de Boulogne*.

Este exame parte da premissa de Weisgerber (1971, p. 153, tradução nossa) ao afirmar que uma pessoa que “‘vive’ o espaço em todas as suas fibras, constrói-o literalmente sob os nossos olhos, isso significa que o espaço sugerido pelas palavras do discurso vê-se determinado, em primeira instância, pela pessoa e pela situação do narrador”.<sup>7</sup> Assim, estabelece que o quadro espacial da narrativa é exposto com a intervenção dos sentidos e do ponto de vista parcial do narrador, corroborando o pensamento de Bakhtin (2011) no que toca à inviabilidade de um discurso neutro.

Nesse sentido, um adendo importante a se fazer é o fato de que os locais escolhidos para o estudo envolvem, de forma majoritária, três personagens: Emma, Charles e Léon. Contudo, nossa referência sempre será Emma, uma vez que vivencia todos os ambientes e que seu envolvimento com as outras duas personagens ocasionam a valoração dos espaços. O protagonismo da jovem e entediada esposa direciona o olhar e a voz narrativa do romance, permitindo-nos analisar trechos tanto do discurso descritivo do narrador quanto do discurso direto de Emma. Ao levarmos em consideração que o espaço se constitui enquanto uma projeção daquele que o retrata, as impressões do observador permeiam suas escolhas lexicais ao fazê-lo e isso, por conseguinte, molda a maneira como adentramos e passamos a conhecer esses ambientes. Por isso, notamos a diferença entre as descrições espaciais de acordo com o sentimento, as relações que os envolvem e quem o contempla, como veremos durante esta análise.

#### 3.1 A primeira casa em Tostes

No quinto capítulo da primeira parte do livro, após o casamento com o médico Charles Bovary, Emma muda-se para a casa de seu marido em Tostes e, logo na sua chegada, enquanto a per-

<sup>6</sup> No original: “une réalité psychologique à l'état pur : créée et habitée par l'être qui écrit et qui, pour ainsi dire, se confond avec elle” (Weisgerber, 1971, p. 161).

<sup>7</sup> No original: “[l'homme] ‘vit’ l'espace dans toutes ses fibres le bâtit ici littéralement sous nos yeux, ce qui signifie que l'espace suggéré par le mot du récit se voit déterminé au premier chef par la personne et la situation du narrateur” (Weisgerber, 1971, p. 153).

sonagem explora sua nova residência, há uma descrição extensa e detalhada dos espaços internos e externos desta. Ao longo dela, não constatamos marcas evidentes valorativas das impressões da protagonista, mas percebemos que o local não está em perfeitas condições: “*Elle s’occupe, les premiers jours, à méditer des changements dans sa maison*” (Flaubert, 2020, p. 82, grifos nossos).<sup>8</sup> Conquanto, de maneira geral, o retrato desse primeiro domicílio do casal seja bastante objetivo, no trecho citado, identificamos um indício de sentimento (ou falta deste). O narrador prefere referenciar a casa de Tostes como *sa maison* (sua casa), ou *la maison d’Emma*; em outras palavras, há um certo distanciamento se compararmos essa escolha lexical com a expressão francesa *chez quelqu’un* (nesse caso, *chez elle*), cujo sentido é carregado de afeto — como discutiremos detalhadamente. Em contrapartida, se compararmos *sa maison* com a fórmula ainda mais desapaixonada *la maison* (simplesmente, a casa), a voz narrativa não parece estar despida de toda emoção. Interpretamos a partir dessa alusão que, embora recém chegada à nova moradia, Emma inicialmente se empenha em torná-la *sua*, em tomar para si esse espaço que dividirá com o marido.

Na continuação do capítulo, acompanhamos a felicidade de Charles com sua nova esposa em sua casa. Contudo, esse sentimento não é mútuo, Emma começa a experimentar as frustrações e melancolias da vida conjugal em uma pequena província. Nesse ponto do romance, introduz-se o bovarismo da personagem, isto é, sua insatisfação com a realidade. Evidencia-se também como o mundo imaginado por Emma provinha de suas leituras de romances.

“Avant qu’elle se mariât, elle avait cru avoir de l’amour; mais le bonheur qui aurait dû résulter de cet amour n’étant pas venu, il fallait qu’elle se fût trompée, songeait-elle. Et Emma cherchait à savoir ce que l’on entendait au juste dans la vie par les mots de félicité, de passion et d’ivresse, qui lui avaient paru si beaux dans les livres” (Flaubert, 2020, p. 84, grifos nossos).<sup>9</sup>

O trecho acima expõe a completa desilusão de Emma em relação ao seu casamento e a Charles; a sensação de que fora enganada transforma-se na sua insatisfação com a vida real. A primeira moradia do casal representa, portanto, um ambiente em que Emma nutria uma esperança inicial pelo casamento ideal, pelo amor, pela vida romântica a dois e, em conformidade com a protagonista, transforma-se em um local tedioso e estagnado, que conseguimos ocupar recorrendo à fuga mental/emocional. Tal como quando lemos um romance, já diria Emma. Esses sentimentos negativos manifestam-se nos próximos capítulos por meio da ausência de traços de afetividade na descrição do espaço da casa, algo que se mantém na segunda moradia dos Bovary.

<sup>8</sup> Em português: “Logo nos primeiros dias, começou a pensar em operar modificações *na [sua] casa*” (Flaubert, 1981, p. 29, grifos nossos).

<sup>9</sup> Em português: Antes de se casar, julgava sentir amor; mas, como a ventura resultante desse amor não aparecia, com certeza se enganara, pensava ela. E procurava saber qual era, afinal, o significado certo nesta vida, das palavras ‘felicidade’, ‘paixão’ e ‘embriaguez’, que nos livros pareciam tão belas (Flaubert, 2020, p. 84, grifos nossos).

### 3.2 A vida idealizada no castelo de *La Vaubyessard*

Antes de debruçarmo-nos sobre a casa dos Bovary em Yonville, julgamos pertinente abordar o *château de La Vaubyessard*, enquanto espaço simbólico representativo da vida e do lar desejados por Emma. No apogeu de seu desencantamento, Emma e Charles são convidados pelo marquês de Andervilliers para um baile em seu castelo; esse evento é mencionado pelo narrador como algo extraordinário em meio à melancolia da jovem madame Bovary. Na chegada ao *château de La Vaubyessard*, seu detalhamento personifica o desejo de Emma pela vida romântica e luxuosa, passados anos na fazenda de seu pai ou os meses casada em um pequeno vilarejo.

*Le château, de construction **moderne**, à l'Italienne, avec deux ailes avançant et trois per-  
rons, se déployait au bas d'une **immense** pelouse où paissaient quelques vaches, entre **des  
bouquets de grands** arbres espacés, tandis que des bannettes d'arbustes, rhododendrons,  
seringas et boules-de-neige bombaient leurs touffes de **verdure inégales** sur la ligne  
courbe du chemin sablé. Une rivière passait sous un pont ; à travers la brume, on distin-  
guait des bâtiments à toit de chaume, éparpillés dans la prairie, que bordaient en pente  
douce deux coteaux couverts de bois, et par-derrière, dans les massifs, se tenaient, sur deux  
lignes parallèles, les remises et les écuries, restes conservés de l'ancien château démoli  
(Flaubert, 2020, p. 98, grifos nossos).<sup>10</sup>*

Notamos como todo o retrato do castelo se constrói em um campo semântico relacionado à riqueza: uma construção moderna cercada por um verdor desigual, um imenso gramado, grandes árvores e variadas espécies flores e arbustos. Desde então, estabelece-se o cenário que Emma recorda em seus momentos de maior desilusão, nessa primeira metade do romance, o castelo de *La Vaubyessard* é como um refúgio distante onde estaria a vida ideal almejada. Esse valor simbólico é reforçado, uma vez que os Bovary entram no vestíbulo: “Il était pavé de dalles en marbre, très haut, et le bruit des pas, avec celui des voix, y retentissait comme dans une église” (Flaubert, 2020, p. 99).<sup>11</sup> A comparação desse pequeno cômodo com uma igreja contribui para a reprodução do espaço enquanto sagrado e paradisíaco, algo que ficará para sempre nas lembranças de Emma. Além da própria arquitetura, a atmosfera é descrita pesadamente, sobrecarregando os sentidos com o ar quente e perfumado, a fumaça que exalava do jantar, o brilho trêmulo das velas e seu reflexo nos cristais. Segundo o narrador, mesmo o açúcar, naquele salão deslumbrante, parecia mais fino e branco para a madame Bovary (Flaubert, 2020).

A noite passada no castelo concretiza o deslumbramento pela opulência e pela moda nobre, que se tornam centrais para Emma. Em outros momentos da narrativa, mesmo ao serem confrontados por dificuldades financeiras profundas, a protagonista parece incapaz de renunciar seus interesses extravagantes. O nome da filha dos Bovary, Berthe, também

<sup>10</sup> Em português: “O palácio, de construção *moderna*, à italiana, com duas alas salientes e três escadarias, surgia do meio de *um extenso prado*, onde pastava algum gado, entre *grupos de grandes árvores* espaçadas, ao mesmo tempo que canteiros de arbustos, rododendros e bolas-de-neve arredondavam as suas massas de *verdura desigual* na linha curva dos caminhos areados. Sob a ponte passava um ribeiro; *através da bruma* distinguiam-se casinhas colmadas, espalhadas pelo prado e rodeadas em suave declive por duas encostas cobertas de arvoredo; por detrás, nas moitas, erguiam-se, em dois renques paralelos, as cocheiras e cavalariças, reminiscências do antigo palácio demolido” (Flaubert, 1981, p. 40, grifos nossos).

<sup>11</sup> Em português: “Este era lajeado de mármore, muito alto, onde o ruído dos passos e das vozes ecoava como numa igreja” (Flaubert, 1981, p. 40).



encontra sua origem nesse baile, pois Emma escuta a marquesa chamar uma jovem moça por esse nome — meses mais tarde, na gravidez, ainda lembra desse momento e parece querer recriá-lo na sua realidade cotidiana por meio da recém-nascida.

Dessa maneira, apesar dos Bovary não habitarem no *château de la Vaubyessard*, julgamo-lo um ambiente fundamental na construção do quadro espacial do romance, posto que é retratado fortemente a partir do olhar encantado de Emma. O local corporifica as fantasias românticas da leitora assídua e cristaliza-se como a existência da vida ideal, do lar perfeito.

### 3.3 A segunda casa em Yonville

Faz-se impossível afirmar que a província de Yonville-l'Abbaye é uma repetição do cenário solitário e apático vivenciado por Emma em Tostes. Nesse vilarejo, por exemplo, a protagonista estabelece-se na comunidade, dando à luz sua filha, que será afilhada do farmacêutico, cuja esposa se tornará uma amiga próxima. Além disso, é nesse espaço que madame Bovary conhecerá seus dois futuros amantes: Léon Dupuis, entusiasta de música e literatura, e Rodolphe Boulanger, um charmoso burguês, com muitas amantes. Em Yonville, também, Emma será enterrada.

No entanto, esta segunda morada, descrita no capítulo dois da segunda parte, desde o primeiro momento, provoca-lhe desconforto: “*dès le vestibule, sentit tomber sur ses épaules, comme un linge humide, le froid du plâtre*” (Flaubert, 2020, p. 142).<sup>12</sup> A residência estava desorganizada e os móveis foram deixados no meio do apartamento, aumentando o estranhamento. Semelhante à casa em Tostes, Emma e o narrador se referem a este espaço como *ma/sa maison* (minha/sua casa). Segundo Bachelard (1961), através do sonho e da imaginação, o passado se faz presente em uma nova casa. Sejam sonhos ou pesadelos, Emma não consegue escapar de sua infelicidade e não controla as fugas da realidade que a acometem. Na segunda residência dos Bovary, sentirá a mesma melancolia vivenciada em Tostes e desejará novamente algo extraordinário no seu cotidiano pacato.

Destacamos a ocorrência de *ma maison*, quando Léon (ainda apenas um vizinho) pergunta à jovem madame Bovary se ela teria abandonado a música, ao que ela responde: “*Ah! mon Dieu, oui! n'ai-je pas ma maison à tenir, mon mari à soigner, mille choses enfin, bien des devoirs qui passent auparavant!*” (Flaubert, 2020, p. 166, grifos nossos).<sup>13</sup> Constatamos que Emma elenca dois deveres enquanto esposa e que agem também como justificativas para renunciar à música: cuidar da casa e do marido. Nesse contexto, compreendemos que a referência à propriedade da casa ocorre em razão do contrato do casamento, uma vez que no decurso da narrativa, evidencia-se que Emma não se sente acolhida ou protegida em sua residência. A casa de Yonville nada mais é do que um bem compartilhado por dois cônjuges, não se trata de um lar ou um refúgio, mas sim uma de suas obrigações enquanto esposa.

Nestes capítulos da segunda parte, a protagonista revela-se otimista e esperançosa sobre o futuro, acredita que a mudança de ambiente representa uma nova fase em sua vida: “Ela não podia acreditar que as coisas pudessem surgir sempre iguais em lugares diferen-

<sup>12</sup> Em português: “no vestíbulo, sentiu cair-lhe nos ombros alguma coisa como se fosse um pano úmido; era o frio do estuque” (Flaubert, 1981, p. 67).

<sup>13</sup> Em português: “Mas é claro! Não tenho *minha casa de que me ocupar, meu marido para tratar*, mil coisas, enfim, mil deveres que estão em primeiro lugar?” (Flaubert, 1981, p. 82, grifos nossos).



tes; e, uma vez que a parte já vivida fora má, tinha esperanças de que a que lhe restava viver havia de ser melhor” (Flaubert, 2020, p. 143).<sup>14</sup> Mesmo que outras experiências tenham sido negativas e apenas o ambiente habitado tenha mudado, este que desde o começo a incomoda, Emma acreditava que a situação melhoraria. Em certo grau, Yonville será o cenário da transformação de sua personalidade e caráter: após retorno da dolorosa desilusão de *Vaubysard* estar fora de alcance, sua fuga emocional ocorre por meio de casos amorosos e impulsiona sua obsessão pela riqueza.

### 3.4 O lar no *Hôtel de Boulogne*

Nas casas e nos quartos, cada elemento é adaptado às necessidades e representativo das preferências e personalidades daqueles que nele habitam, todavia, um quarto de hotel é um cômodo genérico, arquitetado não para acolher alguém específico, mas sim uma diversidade de pessoas que por ali passam. Não se espera um ambiente íntimo e caloroso, visto que se trata de um local provisório, de pernoite, cuja estadia não seria suficiente para estabelecer laços afetivos. Em *Madame Bovary*, esta regra encontra sua exceção.

Três anos após a partida de Léon para Rouen, os Bovary encontram-no em uma noite de ópera no teatro e, no dia seguinte, um novo espaço ganha relevância: um quarto no *Hôtel de Boulogne*. Neste, os amantes passam a encontrar-se todas as quintas-feiras, a fim de viverem seu amor. Por conseguinte, esse local passa a ser colorido por sentimentos e, semana a semana, sua imagem transforma-se aos olhos de Emma (e do narrador). No quinto capítulo da terceira parte, decorrido determinado período dos encontros secretos de Emma e Léon, o quarto é descrito de forma detalhada pelo narrador:

*Le tiède appartement, avec son tapis discret, ses ornements folâtres et sa lumière tranquille, semblait tout commode pour les intimités de la passion. Les bâtons se terminant en flèche, les patères de cuivre et les grosses boules de chenets reluisaient tout à coup, si le soleil entrait. Il y avait sur la cheminée, entre les candélabres, deux de ces grandes coquilles roses où l'on entend le bruit de la mer quand on les applique à son oreille.*

**Comme ils aimaient cette bonne chambre pleine de gaieté, malgré sa splendeur un peu fanée!** Ils retrouvaient toujours les meubles à leur place, et parfois des épingles à cheveux qu'elle avait oubliées, l'autre jeudi, sous le socle de la pendule [...]. Ils étaient si complètement perdus en la possession d'eux-mêmes, **qu'ils se croyaient là dans leur maison particulière**, et devant y vivre jusqu'à la mort, comme deux éternels jeunes époux. Ils disaient **notre chambre, notre tapis, nos fauteuils, même elle disait mes pantoufles**, un cadeau de Léon, une fantaisie qu'elle avait eue. C'étaient des pantoufles en satin rose, bordées de cygne (Flaubert, 2020, p. 349-350, grifos nossos).<sup>15</sup>

No trecho apresentado acima, destacamos algumas passagens que nos levam ao cerne do nosso exame: o retrato do cômodo é transpassado por um matiz afetuosos, nas palavras do narrador, mas pelo olhar da protagonista. Ao inaugurar a descrição designando o apar-

<sup>14</sup> Em português: “Elle ne croyait pas que les choses pussent se représenter les mêmes à des places différentes, et, puisque la portion vécue avait été mauvaise, sans doute ce qui restait à consommer serait meilleur” (Flaubert, 1981, p. 67).

<sup>15</sup> Em português: “O *tépido* aposento, com seu tapete discreto, os seus ornatos alegres e a sua luz tranquila, parecia mesmo próprio para as intimidades da paixão. As molduras terminavam em flechas, as salvas de cobre e as grandes

tamento como “tépido” e de “esplendor um pouco desbotado”, imaginamos referir-se à sua simplicidade, à ausência de luxos (cobiçados por Emma durante tanto tempo). Em seguida, a voz narrativa denuncia sua proximidade com a protagonista, detalhando o quarto do hotel como próprio às intimidades da paixão e como um espaço sempre alegre.

A presença, então, dos pronomes possessivos da primeira pessoa do singular (*mes*, *minhas*) e da primeira e terceira pessoa do plural (*notre/nos*, *nosso/s*; *leur*, *deles*) acresce um tom evidentemente sensível e apaixonado ao simples quarto de hotel. Até este ponto do romance, o pronome possessivo possui poucas ocorrências, como a referência à casa de Tostes (pelo narrador) em um momento de otimismo ou à residência em Yonville enquanto parte dos deveres de uma esposa.

No espaço do *Hôtel de Boulogne*, a voz narrativa aparenta esquecer a prudência com que estas estruturas lexicais foram usadas até então – ou sente tão intensamente a paixão e liberdade experimentadas pela protagonista, que não é mais necessário poupar e pesar palavras. Emma e Léon começam a construir seu lar, enxergam-se nos objetos comuns da hospedagem e passam a imaginá-los como seus. A *fantasia* se fortalece e Emma parece finalmente adentrar as páginas de uma obra do romantismo.

Conforme Bachelard (1961), a casa, tomada por nós enquanto lar, é o porto seguro do sonhador e permite-lhe sonhar em paz. Observamos que a metamorfose do tépido quarto de hotel em lar ocorre a partir de sua capacidade de isolar os amantes, de apresentá-los uma nova realidade, uma vida *sonhada*. Nesse espaço, Emma reimagina sua existência, esquece as dores e insatisfações, sonha em paz com uma outra verdade. Léon, por sua vez, faz o mesmo movimento, concebe sua amada como uma mulher inocente, isenta de pecados e livre de amarras com a realidade (leia-se, seu casamento, sua filha, sua dependência financeira). No meio desse sonho interno ao lar do casal, ambos fogem de suas vidas insatisfatórias e encontram acolhimento juntos, um no outro.

## 4 A afetividade na visão das personagens

Expostos quatro espaços que julgamos fundamentais para a concepção de “casa” e “lar” em *Madame Bovary*, detemo-nos na influência da afetividade na percepção de espaços físicos. De acordo com Weisgerber (1971, p. 152, tradução nossa), “o espaço da narrativa englobará, então, representações mentais, o meio psíquico [...] os sonhos, o desejo e a especulação, e, de outra parte, as sensações e as percepções ocasionadas por um mundo físico”.<sup>16</sup> Afirmo, portanto, que

---

bolas do fogão reluziam repentinamente, se entrava o sol. Em cima do fogão, entre os castiçais, havia duas grandes conchas rosadas, em que se ouvia o ruído do mar quando se encostavam ao ouvido.

*Como eles gostavam daquele belo quarto cheio de alegria, apesar do seu esplendor um tanto gasto!* Achavam sempre os móveis no seu lugar e às vezes os grampos dos cabelos, de que se esquecera na outra quinta-feira, debaixo do pedestal do relógio [...]. Estavam tão completamente perdidos na posse de si mesmos, *que se julgavam já na sua própria casa*, onde deveriam viver até a morte, como dois eternos noivos. Diziam ‘*o nosso quarto, o nosso tapete, as nossas poltronas*’ e *Ema chegava mesmo a dizer ‘os meus chinelos*’, uma prenda de Léon, uma *fantasia* que ela tivera. Eram uns chinelos de cetim cor-de-rosa, bordados (Flaubert, 2020, p. 349-350, grifos nossos)” (Flaubert, 1981, p. 198, grifos nossos).

<sup>16</sup> No original: “l’espace du récit englobera donc les représentations mentales, le milieu psychique [...] des songes, du désir et de la spéculation, et, d’autre part, des sensations et perceptions occasionnées par un monde

o psicológico da personagem mantém um vínculo direto com sua forma de conceber, e talvez vivenciar, o mundo ao seu redor.

A título de exemplo, citamos uma passagem (presente no capítulo três da terceira parte), em que Emma e Léon estão desfrutando de um momento romântico em uma ilha do rio Sena, em Rouen. Esse ambiente aprofunda o retrato do relacionamento dos amantes – que apresentamos ao abordar o quarto no *Hôtel de Boulogne* –, uma vez que estabelecem seu lar onde quer que estejam.

*Ils se couchaient sur l'herbe; ils s'embrassaient à l'écart sous les peupliers; et ils auraient voulu, comme deux Robinsons, vivre perpétuellement dans ce petit endroit, **qui leur semblait, en leur béatitude, le plus magnifique de la terre**. Ce n'était pas la première fois qu'ils apercevaient des arbres, du ciel bleu, du gazon, qu'ils entendaient l'eau couler et la brise soufflant dans le feuillage; **mais ils n'avaient sans doute jamais admiré tout cela, comme si la nature n'existait pas auparavant, ou qu'elle n'eût commencé à être belle que depuis l'assouvissement de leurs désirs** (Flaubert, 2020, p. 340, grifos nossos).<sup>17</sup>*

A narração é atravessada pela intensidade das emoções sentidas pelas personagens: esse pequeno espaço é magnífico. Assim como no *château de La Vaubyessard*, o açúcar parecia diferente a Emma, nessa ilha, as árvores, o céu, a grama, o som da água do rio e mesmo a brisa do ar pareciam-lhes diferentes, mais admiráveis. A percepção da beleza da natureza depende da satisfação de seus desejos, detalham o espaço como um reflexo de seus sentimentos. Nesse sentido, para Emma, as casas em Tostes e Yonville representavam a desilusão, e o castelo, o deslumbramento por outra realidade. Na ilha e no quarto de hotel, o espaço ao redor do casal se altera com a intervenção de seus olhares que, por sua vez, guiam a voz narrativa.

Segundo Dominique Fernandez (2007), o isolamento absoluto sentido em uma ilha (dissociada da vida cotidiana em terra firme), constrói a impressão do recomeço do mundo. A “lua de mel” de Emma e Léon nesta pequena ilha próxima ao Sena é simbólica: a vida recomeçava, o amor perfeito era possível e o coração poderia estar livre de preocupações. Durante três dias, os amantes ignoram suas responsabilidades, seus compromissos e suas famílias; acreditam viajar no tempo para uma época em que escolhas melhores possam ser feitas. Neste recomeço, o lar de ambos se estabelece em um e no outro, mesmo as paredes simples do hotel não são capazes de comportá-lo.

#### 4.1 Uma casa não é um lar

Inúmeras são as referências nas artes à imagem da casa (do lar, do abrigo, do ninho), pois esta tem um valor simbólico de grande expressão devido à sua importância para o ser humano. Para Bachelard (1979, p. 201), “ela mantém o [sujeito] através das tempestades do céu e das tempestades da vida. Ela é corpo e alma. É o primeiro mundo do ser humano”. Em outros termos,

---

physique” (Weisgerber, 1971, p. 152).

<sup>17</sup> Em português: “Deitavam-se na relva; abraçavam-se debaixo dos choupos; e que riam, como dois Robinsons, viver perpetuamente naquele pequeno sítio que *lhes parecia, no meio da sua beatitude, o mais belo da terra*. Não era a primeira vez que viam árvores, céu azul e relva, que ouviam a água corrente e a brisa ramalhando a folhagem: *mas nunca decerto tinham admirado tudo isso, como se anteriormente a natureza não existisse, ou como se não tivesse começado a ser bela senão depois de eles terem saciado os seus desejos*”.

concebemos a casa enquanto materialização do eu-interior, da intimidade, consagrando-se enquanto refúgio; e, por isso, somos tão conectados emocionalmente com esse ambiente.

Pelo nosso ensejo de discorrer sobre a influência do afeto nas descrições, escolhemos esses espaços da casa e/ou do lar, no qual é relevante a presença de um forte laço sentimental. Entretanto, nem sempre o afeto relaciona-se à casa em si, à estrutura física onde habitamos e descansamos, por vezes outros fatores são importantes para que consideremos um lugar como nosso porto seguro. Além disso, pode não ser um lugar, pode ser alguém e este é um fator notável em *Madame Bovary*.

No que tange à semântica, essa concepção de um “refúgio” pode ser feita a partir da distinção entre “casa” e “lar” – introduzida no decurso de nossa análise da caracterização espacial no romance de Flaubert. Ao pesquisarmos nos dicionários da língua portuguesa *Houaiss*, *Aurélio* e *Michaelis* não observamos, no geral, diferenças expressivas entre as acepções dos dois vocábulos, ambos compartilham valores de sentido relacionados à residência, ao local físico onde habita a família e à estrutura familiar em si.

Na língua inglesa, há as palavras *house* e *home* (equivalentes, respectivamente, à casa e lar). De acordo com o dicionário *Cambridge* de língua inglesa, a primeira possui os mesmos valores de seu correspondente em português, mas a segunda manifesta uma diferença expressiva ao trazer o sentido de “o local de origem de alguém ou algo, ou o local ao qual uma pessoa sente pertencer” (Cambridge, 2021, tradução nossa).<sup>18</sup> No português do Brasil, há o vocábulo “pátria”, que não exprime necessariamente a noção de um local de pertencimento, um lugar onde “nos sentimos em casa” (como popularmente expressamos esse sentimento). Esse é o ponto de partida para a visão poética e afetiva que caracteriza um lar, pois, como apontamos, os espaços nos quais Emma morou com seu marido e aqueles em que compartilhava momentos de júbilo com seu amante, como o quarto de hotel e a ilha do Sena, eram vivenciados de formas distintas pela protagonista.

No tocante à língua francesa, destacamos, nas seções anteriores, algumas questões lexicais simbólicas para nossa interpretação: a diferença entre referir-se ao espaço como *la maison* (a casa), *ma/sa maison* (minha/sua casa), o possível emprego da expressão *chez moi* (variação de “na minha casa”), bem como a utilização de pronomes possessivos com respeito aos objetos do hotel. Em toda a extensão de *Madame Bovary*, Emma jamais designa suas casas, em Tostes e Yonville, pela expressão *chez moi* ou *chez nous*. Ao contrário disso, a personagem e o narrador optam pelo vocábulo *maison*, que não detém a mesma carga sentimental. De acordo com o dicionário de língua francesa *Le Petit Robert*, a expressão *chez* possui a acepção “domicílio pessoal (com valor afetivo)” (Robert, 2014, tradução nossa).<sup>19</sup> Podemos, então, considerar tal construção, *chez* seguido de um pronome oblíquo (neste caso, interessa-nos *moi*, *elle* e *nous*), como um sinônimo em francês da nossa concepção de lar, devido ao valor emotivo somado à noção de moradia.

Em nosso exame do quadro espacial do romance, essa distinção verifica-se tanto pelas descrições dos ambientes onde os acontecimentos ocorrem, como também por estas escolhas lexicais na voz narrativa e no discurso de Emma. Constatamos duas incidências do termo *chez* no romance. A primeira ocorre pelo narrador, ao relatar a volta de Emma para casa de seu

<sup>18</sup> No original: “someone’s or something’s place of origin, or the place where a person feels they belong” (Cambridge, 2021).

<sup>19</sup> No original: “domicile personnel (avec valeur affective)” (Robert, 2014).

pai (a fazenda dos Bertaux) depois de partir da pensão do convento em que estudava: “*Emma, rentrée chez elle, se plut d’abord au commandement des domestiques, prit ensuite la campagne en dégoût et regretta son couvent*” (Flaubert, 2020, p. 90, grifos nossos).<sup>20</sup> O bovarismo característico da personagem se faz patente nesse trecho. O espaço inicialmente descrito como seu lar (a casa de seu pai, a fazenda na qual ela cresceu) é, em seguida, representado pelo hiperônimo “campo” que a desagrada. Notamos, contudo, que apesar de uma única frase conter os dois retratos, há um intervalo de tempo (mesmo que curto) entre eles. *Chez elle* era o local que Emma buscava ao estar confinada e infeliz no convento, acreditava pertencer junto de seu pai na vida campesina; porém, ao retomar essa realidade, decepçiona-se e passa a experimentar uma melancolia sem solução.

A segunda ocorrência produz-se no capítulo sete da terceira parte; desesperada em consequência das dívidas que contraiu e perto de perder sua casa para o banco, Emma se dirige à residência de Léon em Rouen a fim de convencê-lo a ajudá-la. Tão logo o escrivão abre a porta para que entre, a protagonista responde “— *Oh! non, là-bas, chez nous. Et ils allèrent dans leur chambre, à l’hôtel de Boulogne*” (Flaubert, 2020, p. 386, grifos nossos).<sup>21</sup> Este é o clímax produzido pelo romance, uma vez que acompanhamos os amantes desde a primeira atração em Yonville, os interesses em comum, as ambições sonhadas, chegamos ao quarto *deles* no hotel, presenciamos o amor *deles* na ilha do Sena. Essa conexão emocional que Emma demonstra ter não somente com o quarto do *Hôtel de Boulogne*, mas com qualquer espaço que tenha ocupado com Léon, explica a forma como o próprio narrador, influenciado por ela, retrata-os.

Emma encontra-se, mais do que nunca, em um contexto desolador, sente-se em uma ilha desabitada, isolada e sozinha, como todo/a verdadeiro/a herói/heroína de romance (Fernandez, 2007), a solidão a preenche. No entanto, nessa única exclamação, a jovem protagonista luta contra esse sentimento que a acompanhou durante a vida: no convento, na fazenda do pai e nas casas compartilhadas com o marido. Isolamento que sente mesmo com seu primeiro amante, Rodolphe, e no castelo de *La Vaubyessard*. Nestes, poderia se encontrar luxos e riquezas, mas é na pobre simplicidade de um quarto de hotel, junto a seu amante, que descobre um lar. Emma se afirma digna de amor, digna de um lar, de *chez elle*, com ou em outra pessoa, em Léon.

## 5 Considerações finais

Ao longo deste trabalho, comprovamos indiretamente que a impessoalidade idealizada pela escola realista fora apenas um projeto e não uma possibilidade realizável pelo discurso. Embora nos debruçemos sobre uma pequena parcela do quadro espacial compreendido em *Madame Bovary*, os traços valorativos presentes na voz narrativa são palpáveis, estando presentes em outros aspectos do romance, como nos retratos físicos e psicológicos das personagens. Toda a narrativa é influenciada pela focalização do narrador heterodiegético, neste

<sup>20</sup> Em português: “Ema, voltando para [sua] casa, regozijava-se com dar ordens aos criados; depois se desgostou do campo, e teve saudade do convento” (Flaubert, 1981, p. 34, grifos nossos).

<sup>21</sup> Em português: “— Não, não, no nosso quarto. E foram para o seu quarto no Hotel de Boulogne” (Flaubert, 1981, p. 221, grifos nossos). Interpretamos o *chez nous*, traduzido por “nosso quarto” como uma referência à “casa” compartilhada por ambos: “na nossa casa”.



caso, sendo guiado pela visão e pelas sensações de Emma Bovary. Essa relação existente entre protagonista e narrador proporciona-nos um ponto de vista tendencioso, mas também permite-nos vivenciar os acontecimentos com Emma, experienciando o mundo e aqueles ao seu redor segundo sua visão.

Em nossa análise, engajamo-nos em expor esses coloridos afetivos, por meio de enunciados do narrador e da protagonista, no tocante aos ambientes essenciais para conhecermos o espaço psicológico desta. Demonstramos, assim, como a realidade descrita é uma construção que passa pelo filtro psicológico de quem a vivencia, gerando a influência emotiva. A casa de Tostes será o cenário do otimismo inicial e da desilusão conjugal; na residência de Yonville, insatisfeita desde sua chegada, Emma sofrerá o desespero e a morte. O castelo de *La Vaubyessard* permanecerá como um sonho distante, uma noite de opulência que retorna à mente de Emma nos seus momentos de maior melancolia, representando uma vida inalcançável.

Em contrapartida, o quarto no *Hôtel de Boulogne* e a ilha no rio Sena são gradativamente acrescidos de carinho e amor, neles Emma encontra o espaço para “sonhar em paz” (Bachelard, 1979, p. 201) que não descobrira na casa que dividia com sua família. Aquele local nunca protegeu a sonhadora protagonista, ao contrário disso, foi o seu leito de morte. No hotel, “*elle était l’amoureuse de tous les romans, l’héroïne de tous les drames, le vague Elle de tous les volumes de vers*” (Flaubert, 2020, p. 350).<sup>22</sup> Em seu lar, junto a Léon, Emma alcança as personagens que sempre leu, realizando a fuga total da dura realidade.

Desse modo, justifica-se a frieza da narração ao dissertar sobre as casas dos Bovary que, para a protagonista, não pareciam ser de tamanha conexão emocional, eram apenas espaços em que Emma deveria cumprir as funções que eram atribuídas às esposas na época, onde ela dormia, mas não sonhava. Os valores afetivos mesclados à descrição do quarto do *Hôtel de Boulogne*, caracterizam-no não somente como o local da vida em casal com Léon, por quem ela nutria sentimentos românticos, mas como o único espaço em que se sentia livre para sonhar.

## Referências

BACHELARD, Gaston. *La poétique de l’espace*. 3. ed. Paris: Les Presses universitaires de France, 1961.

BACHELARD, Gaston. *Os Pensadores*. Tradução de Antônio da Costa Leal e Lúcia do Valle Santos Leal. São Paulo: Abril Cultura, 1979.

BAKHTIN, Mikhail. *Estética da criação verbal*. 6. ed. São Paulo: WMF Martins Fontes, 2011.

CHEZ. In: ROBERT, Paul. *Le Petit Robert électronique*. CD-ROM, Paris, Dictionnaires Le Robert, 2014.

FERREIRA, Aurélio Buarque de Holanda. *Dicionário da língua portuguesa*. 5. ed. Curitiba: Positivo, 2010.

FERNANDEZ, Dominique. *L’art de raconter*. Paris: Éditions Grasset, 2007.

FLAUBERT, Gustave. *Madame Bovary*. Tradução de Araújo Nabuco. São Paulo: Abril Cultural, 1981.

FLAUBERT, Gustave. *Madame Bovary*. Paris: Gallimard, 2020. (Coleção Folio classique).

---

<sup>22</sup> Em português: “Ema era a apaixonada de todos os romances, a heroína de todos os dramas, a vaga ‘ela’ de todos os volumes de versos” (Flaubert, 1981, p. 298).



FRIEDMAN, Norman. O ponto de vista na ficção: o desenvolvimento de um conceito crítico. *Revista USP*, São Paulo, v. 53, p. 166-182, 2002.

GENETTE, Gérard. *Figures III*. Paris: Éditions du Seuil, 1972.

HOME. In: CAMBRIDGE. *Cambridge English Dictionary*. Cambridge: Cambridge University Press, 2021. Disponível em: <https://dictionary.cambridge.org/dictionary/english/home>. Acesso em: 20 jun. 2021.

HOUAISS, Antônio. *Dicionário Houaiss da Língua Portuguesa*. Rio de Janeiro: Objetiva, 2009. CD-ROM.

LAGET, Thierry. Préface. In: FLAUBERT, Gustave. *Madame Bovary*. Paris: Gallimard, 2020. p. 7-42. (Coleção Folio classique).

MICHAELIS. *Moderno dicionário da língua portuguesa*. São Paulo: Melhoramentos, 2021. Disponível em: <http://michaelis.uol.com.br/moderno/portugues/index.php>. Acesso em: 20 jun. 2021.

MÜLLER, Andréa. Os percursos de “Madame Bovary” no Brasil. *Revista Letras*, Santa Maria, v. 23, n. 47, p. 157-174, 2013.

WEISGERBER, Jean. Notes sur la représentation de l'espace dans le roman. *Revue de L'Université de Bruxelles*, 2-3, p. 149-165, 1971.