v. 30, n. 1, p. 8–22, 2025
ISSN: 2238-3824 · DOI: 10.35699/2238-3824.2025.58837
Submetido em: 28/04/2025 · Aprovado em: 28/04/2025

caligrama
revista de estudos românicos

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: Multimodalide; 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' – independendentemente de como ou onde isso seja feito".[1]

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

[1] 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).[2]

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"[3], termo que usa a autora[4]) 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).[5]

---

[2] 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 com- position 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".

[3] Em inglês, *channels*.

[4] 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).

[5] 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,[6] 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

[6] Disponível em: https://www2.ufjf.br/framenetbr/reinventa/. 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 Sistêmico-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 anotaçã 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[7] (Belcavello *et al*., 2024):7 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 Multi30k (Viridiano *et al*., 2024):8 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):9 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

[7] Disponível em: https://huggingface.co/datasets/FrameNetBrasil/Frame2. Acesso em: 21 abr. 2025.

[8] Disponível em: https://huggingface.co/datasets/FrameNetBrasil/FM30K. Acesso em: 21 abril 2025.

[9] 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 Multi30k. 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 elicitação de legendas de imagens, interpretando-o no seu sentido da linguística documentária, como uma tarefa de elicitação 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* Frame², 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 dêitico dessas narrativas decorrentes de diferentes modos comunicativos devem ser interpretadas como colaborando para a progressão da narrativa. Do ponto de vista metodológico, Sigiliano delineia 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 dêiticas: centro dêitico, operações dêiticas, 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 dêiticas e categorias da FrameNet, culminando no mapeamento dos procedimentos dêiticos 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 sociossemiótica. 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 315749/2021-0).

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.; WILDEFEUER, 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*. Frame²: 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 Lingu@gem*, 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 Multi30k: A Frame-based Multimodal Multilingual Dataset. *In:* THE 2024 JOINT INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, LANGUAGE RESOURCES AND EVALUATION (LREC-COLING 2024). *Proceedings* [...]. [*S. l.*] p. 7438-7449, 2024.