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

Mariane de Carvalho Pinto

Anotação Multimodal para Copilotos de Produção de Tecnologias Assistivas: uma proposta para a audiodescrição

Juiz de Fora
Dezembro 2025

Mariane de Carvalho Pinto

Anotação Multimodal para Copilotos de Produção de Tecnologias Assistivas: uma proposta para a audiodescrição

Dissertação de mestrado submetida ao Programa de Pós-Graduação em Linguística da Faculdade de Letras da Universidade Federal de Juiz de Fora, como requisito parcial à obtenção do título de Mestre em Linguística.

Orientador: Prof. Dr. Tiago Torrent

Juiz de Fora
Dezembro 2025

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

de Carvalho Pinto, Mariane .
Anotação Multimodal para Copilotos de Produção de Tecnologias Assistivas: uma proposta para a audiodescrição / Mariane de Carvalho Pinto. -- 2025.
119 f.

Orientador: Tiago Timponi Torrent
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Faculdade de Letras. Programa de Pós-Graduação em Linguística, 2025.

1. Semântica de Frames. 2. FrameNet. 3. Anotação multimodal. 4. Audiodescrição. 5. Copilotos de IA. I. Timponi Torrent, Tiago, orient. II. Título.

Mariane de Carvalho Pinto

Anotação Multimodal para Copilotos de Produção de Tecnologias Assistivas: uma proposta para a audiodescrição

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

Aprovada em 17 de dezembro de 2025.

BANCA EXAMINADORA

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

Prof.(ª)Dr(ª). Maucha Andrade Gamonal
Universidade Federal de Juiz de Fora

Prof.(ª)Dr(ª). Flávia Affonso Mayer
Universidade Federal da Paraíba

Juiz de Fora, 08/12/2025.

Documento assinado eletronicamente por **Tiago Timponi Torrent**, **Coordenador(a)**, em 17/12/2025, às 10:54, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **maucha andrade gamonal**, **Usuário Externo**, em 17/12/2025, às 11:07, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Flávia Affonso Mayer**, **Usuário Externo**, em 29/12/2025, às 10:43, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2782823** e o código CRC **4B33F38F**.

AGRADECIMENTOS

Nenhum trabalho é feito sozinho, e este, mais do que qualquer outro, só foi possível graças à generosidade e à presença de muitas pessoas.

Aos meus pais, Maria Madalena e Paulo, agradeço por acreditarem em mim antes mesmo que eu soubesse no que acreditar. Espero, um dia, poder retribuir tudo o que fazem por mim.

Aos meus irmãos, Melissa e Carlos, por dividirem comigo os pequenos e grandes acontecimentos do cotidiano: os novos planos, as estripulias das crianças, as plantas que vingaram (ou não), os dias bons e os difíceis. A distância nem sempre é simples, mas vocês tornam tudo mais leve.

Aos meus sobrinhos, Otávio e Heitor, por me lembrarem, com suas descobertas diárias, da importância de seguir em frente.

Aos amigos de sempre e aos de agora — especialmente Karolyne, Fernando, Caio e Ana Letícia —, por estarem comigo em todas as versões de mim mesma que apareceram na vida acadêmica. Pela escuta atenta, pelas horas de estudo, pelos inúmeros cafezinhos e pelos silêncios partilhados. Vocês são parte fundamental desse caminho.

À equipe da FrameNet Brasil e, em particular, aos colegas de laboratório, por fazerem da pesquisa um lugar de encontro e compartilharem comigo sonhos, dúvidas e ideias (brilhantes!) para a nossa grande pilha de trabalhos futuros.

Ao meu orientador, Prof. Dr. Tiago Timponi Torrent, pela paciência, sinceridade e confiança com que guiou cada etapa deste trabalho. Muito do que aprendi sobre pesquisa acadêmica tem o seu nome.

À Prof.ª Dra. Maucha Andrade Gamonal e à Prof.ª Dra. Flávia Affonso Mayer pela leitura cuidadosa e pelas valiosas contribuições que enriqueceram esta dissertação.

À Capes, pelo apoio financeiro que viabilizou este trabalho (bolsas 88887.963011/2024-00 e 8887.236362/2025-00), e à Fapemig (RED00106/21), pelo fomento às iniciativas de pesquisa da FrameNet Brasil, das quais este trabalho faz parte.

RESUMO

Esta pesquisa investiga de que forma a anotação multimodal de eventos na FrameNet Brasil (FN-Br), desenvolvida com base na Semântica de Frames (Fillmore, 1982), pode contribuir para a geração de roteiros de audiodescrição por sistemas de inteligência artificial atuando como copilotos para audiodescritores humanos. Parte-se da hipótese de que esse tipo de anotação, ao mapear semanticamente os eventos e seus participantes na produção audiovisual, pode favorecer a criação de roteiros de audiodescrição mais precisos, rápidos e adequados ao contexto narrativo. O objetivo do estudo é observar a performance de sistemas de IA como auxiliares na produção de roteiros de audiodescrição, comparando aqueles baseadas na anotação de *frames* com as que não utilizam essa abordagem, além de refletir sobre o potencial dessa ferramenta como apoio ao trabalho do audiodescritor. Para isso, cenas dos episódios 1 e 7 da série de viagens *Pedro Pelo Mundo*, da GNT, foram anotadas com base no modelo da FN-Br, a partir das quais foram gerados roteiros preliminares de audiodescrição por IA. Os roteiros produzidos foram, então, comparados e analisados quanto à sua aproximação com os critérios do *Guia para Produções Audiovisuais Acessíveis* (Naves et al., 2016). Os resultados indicam que as anotações de frames atuam como um modulador do desempenho do modelo, favorecendo, com maior frequência, maior alinhamento entre a descrição gerada e a organização visual das cenas, sem comprometer a qualidade dos roteiros quando não há ganhos imediatos. Do ponto de vista qualitativo, a presença de metadados tende a orientar as descrições pela dinâmica visual, enquanto sua ausência favorece descrições mais estativas e enumerativas. Esses resultados reforçam o potencial da anotação multimodal baseada na Semântica de Frames como apoio à produção de roteiros de audiodescrição por IA.

Palavras-chave: Semântica de Frames, FrameNet, Anotação multimodal, Audiodescrição, Copilotos de IA.

ABSTRACT

This research investigates how the multimodal event annotation in FrameNet Brazil (FN-Br), developed based on Frame Semantics (Fillmore, 1982), can contribute to the generation of audio description scripts by artificial intelligence systems acting as copilots for human audio describers. The study is based on the hypothesis that this type of annotation, by semantically mapping events and their participants in audiovisual productions, can support the creation of more accurate and contextually appropriate audio description scripts. The objective is to observe the performance of AI systems as assistants in the production of audio description scripts, comparing those based on frame annotation with those that do not use this approach, while also reflecting on the potential of this tool to support the work of audio describers. To this end, scenes from episodes 1 and 7 of the travel series *Pedro Pelo Mundo*, produced by GNT, were annotated using the FN-Br model, and preliminary audio description scripts were generated by AI based on these annotations. The resulting scripts were then compared and analyzed in terms of their alignment with the criteria outlined in the *Guide for Accessible Audiovisual Productions* (Naves et al., 2016). The results indicate that frame annotations act as a modulator of the model's performance, more frequently favoring stronger alignment between the generated descriptions and the visual organization of the scenes, without compromising the quality of the scripts when no immediate gains are observed. From a qualitative perspective, the presence of metadata tends to guide descriptions according to the visual dynamics of the scenes, whereas their absence favors more stative and enumerative descriptions. These results reinforce the potential of multimodal annotation based on Frame Semantics as support for the production of audio description scripts by AI systems.

Keywords: Frame Semantics, FrameNet, Multimodal annotation, Audio description, IA copilot.

SUMÁRIO

1 INTRODUÇÃO

A audiodescrição (AD) é uma prática de interação entre videntes e não videntes que tem como objetivo ampliar o acesso de pessoas com deficiência visual às informações transmitidas visualmente (Mayer, 2016). Mais do que descrever o que se vê ou ouve, essa prática exige escolhas interpretativas sensíveis ao contexto narrativo, aos efeitos de sentido e à interação com outros modos semióticos, como som, fala e música. Por isso, o trabalho do audiodescritor é complexo: envolve diferentes etapas e demanda atenção tanto aos aspectos visuais quanto à forma como eles se articulam à narrativa e aos demais recursos expressivos da obra.

Embora as políticas públicas de acessibilidade voltadas para o audiovisual tenham avançado gradualmente no Brasil — como o Projeto de Lei nº 4.248/2012, que passou a exigir a presença de audiodescrição em filmes exibidos nos cinemas e na televisão, e o Programa de Apoio à Distribuição de Conteúdo Acessível da Ancine, lançado em 2017, que incentiva a inclusão de audiodescrição, Libras e LSE em produções nacionais —, a implementação da audiodescrição em produções audiovisuais ainda enfrenta entraves, como o tempo necessário para sua elaboração e os custos envolvidos. Segundo Campos (2019), esse cenário impulsiona a busca por soluções que funcionem como suporte ao trabalho do audiodescritor e agilizem o processo. Diante disso, iniciativas baseadas em tecnologias assistivas e inteligência artificial têm ganhado destaque, ao propor formas semiautomatizadas de geração de audiodescrição. Essas propostas visam a otimizar e acelerar etapas da produção, como a elaboração de roteiros de AD, sem comprometer a qualidade do resultado final.

A incorporação dessas tecnologias, contudo, torna premente a necessidade de se examinar com cuidado aquilo que é produzido, bem como de se compreender como os sistemas de IA estão interpretando e organizando as informações visuais. Atualmente, ainda são poucos os estudos que se dedicam a analisar, de forma sistemática, os textos gerados por IA no contexto da audiodescrição. Nesse sentido, investigar o uso de anotações semânticas estruturadas, como as orientadas por *frames*, no processo de geração automática desses textos permite observar se e como a organização conceitual prévia do conteúdo visual pode contribuir para descrições mais precisas e adequadas ao contexto narrativo. Do ponto de vista

social, considerando que a audiodescrição é um recurso fundamental para o acesso de pessoas com deficiência visual ao audiovisual, é essencial que a expansão de soluções automatizadas venha acompanhada de reflexões sobre qualidade, e não apenas sobre velocidade de produção.

Inserida nesse contexto, esta dissertação tem como objetivo avaliar a contribuição da anotação multimodal de eventos na FrameNet Brasil, com base na Semântica de Frames (Fillmore, 1982), para a geração de roteiros de audiodescrição a partir de um copiloto[1] de inteligência artificial. Especificamente, busca-se: (i) realizar a anotação multimodal de cenas dos episódios 1 e 7 da série *Pedro pelo Mundo*, da GNT, à luz da Semântica de Frames e do modelo da FrameNet Brasil; (ii) gerar, com o auxílio de um copiloto de inteligência artificial, dois conjuntos de roteiros de audiodescrição, sendo um orientado por essas anotações e outro produzido sem esse embasamento semântico; (iii) comparar os dois conjuntos quanto à precisão, à coerência e à adequação à dinâmica narrativa audiovisual; e (iv) analisar em que medida a estruturação semântica de eventos contribui para a redução de lacunas interpretativas e para a produção de descrições mais contextualizadas.

Nossa hipótese é a de que a anotação multimodal de eventos, ao mapear semanticamente quais são os eventos em curso na produção audiovisual, quais participantes deles fazem parte e quais papéis desempenham, fornece ao modelo de IA uma estrutura semântica capaz de orientar a geração de descrições mais alinhadas à dinâmica narrativa. Parte-se do pressuposto de que essa modelagem prévia do conteúdo visual, por organizar a cena em termos de ações, participantes e relações, reduz lacunas interpretativas, evita descrições excessivamente estáticas e favorece a produção de roteiros de audiodescrição mais precisos, coesos e sensíveis às demandas da experiência audiovisual.

Para perseguir o objetivo e testar a hipótese indicados acima, esta dissertação organiza-se nos capítulos a seguir definidos. O segundo capítulo deste trabalho apresenta a Semântica de Frames (Fillmore, 1982) e a FrameNet Brasil (FN-Br). Em seguida, o terceiro capítulo aborda os fundamentos dos estudos em multimodalidade (Lemke, 1998; Bateman, 2017; Gualberto, Santos, 2019) e destaca

[1] Copiloto, aqui, designa ferramentas baseadas em modelos de língua de grande escala (LLMs), que atuam como assistentes interativos na realização de tarefas como a geração de textos, a organização de informações e a interpretação de comandos complexos.

os esforços da FN-Br de, no processo de anotação, abarcar diferentes modos comunicativos para além da língua verbal falada ou escrita, haja vista a necessidade de que eles também sejam contemplados (Belcavello et al., 2022; Belcavello, 2023; Luz et al., 2023; Viridiano, 2024; Abreu, Torrent e Matos, 2025). O quarto capítulo, por sua vez, discorre sobre a audiodescrição dentro do contexto dos estudos da tradução, além de apresentar um breve panorama histórico da audiodescrição no Brasil, bem como trabalhos anteriores que estabelecem a conexão entre a Semântica de Frames e a audiodescrição. O quinto capítulo apresenta os materiais e métodos usados no âmbito deste trabalho, ao passo que o sexto capítulo apresenta a análise realizada durante a pesquisa. O capítulo de conclusão sumariza os achados da pesquisa e aponta para caminhos futuros de investigação.

2 A SEMÂNTICA DE FRAMES E A FRAMENET BRASIL

Esta seção tem como objetivo apresentar a Semântica de Frames, modelo teórico que fundamenta as análises propostas nesta dissertação, e a FrameNet Brasil, que aplica esse modelo no contexto do português brasileiro. Nesse sentido, será abordada a perspectiva da Semântica de Frames sobre a compreensão do significado das palavras e a forma como a FrameNet Brasil organiza e descreve as relações semânticas no português.

2.1 A SEMÂNTICA DE FRAMES

Embora já fosse empregado na área da linguística nos anos 1960, o termo *linguística cognitiva* passou a ser utilizado, no fim da década de 70, para designar uma abordagem que se afastava das premissas do gerativismo. Essa abordagem enfatizava a interação entre linguagem e experiência humana, destacando a importância do contexto e da perspectiva dos falantes no uso da linguagem. Na época, a expressão foi adotada por autores como Langacker, Talmy, Fillmore, Fauconnier e Lakoff, os quais, estudiosos da Linguística Gerativa, mostravam-se insatisfeitos com a maneira como a teoria tratava os processos sintáticos e semânticos de forma independente (Ferrari, 2011), sem considerar como o contexto e as experiências cognitivas dos falantes afetavam a construção de significado e a interpretação linguística em situações de uso real da língua.

Esses estudiosos, então, ao buscar compreender como se dava a relação entre cognição e linguagem no uso efetivo da comunicação, afastaram-se da visão modular da mente, defendida pelo gerativismo. Para Chomsky (1965), precursor da Linguística Gerativa, a mente humana seria composta por módulos distintos e especializados, cada um responsável por uma função cognitiva específica. A linguagem, por sua vez, seria um desses módulos, o que garantiria aos seres humanos uma predisposição biológica para a aquisição linguística. Sendo assim, o foco dos estudos gerativistas era compreender a competência linguística do falante, ou seja, o conhecimento implícito e internalizado das regras e estruturas da língua. Esse conhecimento seria distinto do desempenho linguístico, que diz respeito ao

uso da língua em situações concretas, visto que o desempenho seria influenciado por fatores externos e não afetaria a análise do sistema linguístico em si.

Nesse contexto, Chomsky (1965) introduziu a idealização do falante ideal em uma comunidade de fala homogênea como uma abstração teórica para investigar a competência linguística. No modelo, considera-se um falante plenamente competente, que domina a gramática de forma integral e não é afetado por quaisquer limitações de memória ou atenção, como referência teórica para estudar a estrutura e o funcionamento da língua. Por sua vez, para construir o significado das sentenças, o falante ideal seria guiado pela transparência semântica, somando o sentido das partes para compreender o todo.

Essa visão, no entanto, é problematizada por Fillmore (1979), o qual propõe uma segunda idealização: a do falante/ouvinte inocente. Nessa perspectiva, o falante/ouvinte proposto pelo paradigma gerativista de então seria caracterizado pela capacidade de reconhecer os morfemas da língua e os seus significados, bem como as estruturas gramaticais e os processos que envolvessem esses morfemas, mas apresentaria uma série de limitações linguísticas por ter um conhecimento semântico totalmente literal e, assim, não ser capaz de estabelecer inferências. Entre essas limitações, Fillmore destaca a incapacidade do falante/ouvinte inocente de compreender expressões idiomáticas e metáforas, por exemplo, e a sua inabilidade para lidar com situações que exigem contexto ou interpretação de comunicação indireta.

Ao problematizar a idealização feita pela linguística gerativista, Fillmore buscava demonstrar que a hipótese composicional, embora útil para explicar certos aspectos da linguagem, não é suficiente para dar conta da complexidade dos significados que emergem das interações linguísticas em contextos reais. Com isso, o foco das pesquisas do autor e de outros estudiosos da linguística cognitiva desloca-se para a análise de usos contextualizados da língua, nos quais o significado não se constrói apenas de forma literal ou composicional.

Nesse cenário, emergem diversos estudos que aprofundam a compreensão da interação entre a experiência humana e sua relação com o uso da linguagem. Destacam-se, por exemplo, as pesquisas de Lakoff e Johnson (1980;1999) sobre a função dos esquemas sensório-motores como domínios-fonte das metáforas conceptuais e os estudos de Talmy (2000) sobre o papel dos esquemas imagéticos na estruturação do léxico e da gramática. A diversidade das investigações na área

reflete, conforme explicitado por Salomão (2010), a heterogeneidade da Linguística Cognitiva, um campo teórico amplo que reúne diferentes perspectivas sobre a relação entre cognição e linguagem.

Assim como esses trabalhos, a Semântica de Frames, teoria linguística fundada por Fillmore em 1982, busca, também, apresentar uma visão particular de se olhar para o significado das palavras (Fillmore, 1982, p.111), que considera as experiências cognitivas e culturais que moldam a construção de sentidos na língua. Para Fillmore (1982), compreender uma palavra é acessar uma estrutura conceitual mais ampla, ativando um conjunto de expectativas sobre os elementos envolvidos na cena que ela evoca. Nesse sentido, o significado lexical não reside apenas na relação entre palavra e referente, mas é construído a partir de um pano de fundo de conhecimentos culturais, sociais e enciclopédicos compartilhados pelos falantes.

Fillmore (1985) caracteriza, então, a Semântica de Frames como uma *semântica da compreensão*, em oposição à abordagem da semântica das condições de verdade. Enquanto esta última parte do pressuposto de que o significado de uma sentença decorre exclusivamente das condições sob as quais ela pode ser considerada verdadeira ou falsa, a proposta de Fillmore enfatiza o papel do contexto experiencial e cultural dos falantes na atribuição do significado.

Como ressalta o autor, compreender uma expressão linguística requer acesso à moldura (*frame*) conceitual na qual ela está inserida — ou seja, a um conjunto de informações que tornam a interpretação possível, como, por exemplo, quais são os participantes daquela cena, as suas intenções e os papéis desempenhados por eles naquela interação. A Semântica de Frames enfatiza que o entendimento linguístico está intrinsecamente ligado à experiência humana e ao modo como os falantes categorizam eventos e situações com base em conhecimentos compartilhados. A linguagem, portanto, não é apenas um reflexo do mundo, mas uma ferramenta de construção e negociação de sentidos, ancorada na cognição e nas práticas culturais.

Nessa perspectiva, na Semântica de Frames, considera-se que o sentido é construído com base em conhecimentos prévios organizados em *frames*, ou seja, estruturas cognitivas que relacionam conceitos e experiências de forma contextualizada. Na teoria, o conceito *frame* é usado por Fillmore como uma espécie de construto fundador para outros conceitos que fazem parte da área da Linguística,

como, por exemplo, os conceitos de esquema e script. Segundo o autor, o termo *frame* pode ser definido como

> [...] qualquer sistema de conceitos relacionados de tal forma que, para compreender qualquer um deles, é preciso compreender toda a estrutura à qual ele pertence; quando um dos conceitos dessa estrutura é introduzido em um texto ou em uma conversa, todos os outros são automaticamente disponibilizados. (Fillmore, 1982, p.111)[2]

Sob essa ótica, um *frame* pode ser entendido como uma representação esquemática que abrange tanto os sentidos linguísticos das palavras quanto as interpretações oriundas das experiências dos falantes no mundo. Na perspectiva do autor, o significado de uma palavra é relativo à cena que ela evoca e é compreendido a partir do conhecimento dos falantes sobre essa cena e sobre os participantes que dela fazem parte (os quais são denominados Elementos de Frame – EFs). Um *frame* corresponde, portanto, a um conjunto de conhecimentos internalizados sobre o mundo, os quais possibilitam a construção de significado durante o processo comunicativo (Fillmore, Baker, 2009).

Para elucidar a questão, Fillmore (1982) utiliza o exemplo da transação comercial. Segundo o autor, o significado da palavra "vender" é compreendido porque se conhece a existência de uma cena associada a esse verbo, que envolve um "vendedor" e uma "mercadoria", além de outros elementos que podem ou não estar instanciados na sentença, dependendo da estrutura utilizada, como o "comprador" e o “dinheiro”. Todos esses conceitos, mencionados ou não em uma sentença com a palavra “vender”, são disponibilizados quando o falante acessa, em sua mente, a cena de transação comercial, uma vez que ele faz parte de uma cultura em que as práticas comerciais são cotidianas. Dessa forma, nas análises semânticas baseadas em *frames*, argumenta-se que o significado de um termo decorre da relação entre o próprio termo e o seu pano de fundo (Fillmore, 1982).

Nesse sentido, a Semântica de Frames considera que um mesmo evento pode ser analisado sob diferentes pontos de vista. Segundo Fillmore (1982), os *frames* podem adotar perspectivas variadas em relação a uma mesma cena. Isso significa que um mesmo evento pode ser representado de maneiras diferentes,

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

dependendo de qual participante da cena é tomado como foco. Nas palavras do autor,

> A maneira como o intérprete concebe o mundo do texto atribui a ele uma perspectiva e um ponto de vista específico. Um relato sobre alguém comprando algo ativa o frame de um evento comercial, mas o apresenta, ao menos naquele momento, sob o ponto de vista de um dos participantes da cena. (Fillmore, 1982, p.122)[3]

Para ilustrar essa ideia, pode-se considerar uma situação em que uma pessoa adquire uma mercadoria. Esse evento pode ser descrito sob diferentes pontos de vista: o do vendedor, o do comprador e o da mercadoria, como mostram, respectivamente, as sentenças (1), (2) e (3).

(1) A livraria **vendeu** um livro por cem reais.
(2) Pedro **comprou** um livro por cem reais.
(3) O livro **custava** cem reais.

Embora todas as sentenças descrevam o mesmo evento — o de uma transação comercial —, cada uma adota uma perspectiva distinta sobre ele. Em (1), o foco recai sobre o vendedor, que realiza a ação de vender; em (2), destaca-se o comprador, que adquire o item; e em (3), a mercadoria ocupa o papel central, com ênfase em seu valor. Os verbos utilizados — vender, comprar e custar — refletem essas diferentes formas de enquadrar a cena, destacando certos elementos dela em detrimento de outros. Cada verbo, então, evoca um *frame* específico do domínio do comércio, que reflete o ponto de vista assumido pelo falante ao representar aquela cena na linguagem.

Além disso, as diferentes perspectivas observadas em (1), (2) e (3) não se manifestam apenas em termos conceituais, mas também têm implicações diretas na estrutura dessas sentenças. A escolha de um verbo, ao evocar um determinado *frame*, impõe um conjunto específico de participantes esperados — ou seja, define um padrão de valência associado à cena. Esses padrões registram as potencialidades sintático-semânticas de cada verbo, refletindo o papel que cada elemento da cena desempenha conforme o ponto de vista adotado. Ao sistematizar essas relações entre léxico, sintaxe e significado, a Semântica de Frames amplia

[3] The interpreter's envisionment of the text world assigns that world both a perspective and a history. A report of somebody buying something evokes the frame of the commercial event, but sees that event for the moment at least, from the point of view of one of its participants. [Tradução nossa]

seu escopo analítico e abre caminho para outro de seus objetivos centrais: o mapeamento das estruturas de valência associadas aos *frames*.

Acerca disso, em 1992, Fillmore e Atkins desenharam o plano do que seria um modelo de léxico eletrônico. Ao investigarem a polissemia da palavra “risk”, no inglês, os autores analisaram as diferentes valências semânticas e sintáticas que ela apresentava em contextos diversos e como elas afetavam a sua definição. Ou seja, o sentido do verbo variava conforme os papéis semânticos assumidos pelos participantes envolvidos na cena — quem corre o risco, o que está em risco, em que circunstâncias —, evidenciando limitações dos dicionários tradicionais em representar essas configurações.

Na proposta dos autores, o léxico seria estruturado por meio de *frames* e não mais por classes gramaticais, como nos modelos tradicionais, permitindo ao usuário acessar diferentes níveis de informação ao consultar uma palavra. Além de sua definição, o usuário teria acesso, por exemplo, ao *background* conceitual do termo, como os contextos em que ela costuma ocorrer, os elementos com os quais frequentemente se associa e as relações que estabelece com outros termos do léxico (Fillmore, Atkins, 1992). Essa proposta serviu de base, anos depois, para o desenvolvimento da FrameNet, projeto lexicográfico que organiza os significados das palavras com base nos *frames* que elas evocam.

2.2 A FRAMENET BRASIL

A FrameNet é um projeto lexicográfico que foi fundado por Fillmore, em 1997, no International Computer Science Institute (ICSI), em Berkeley, na Califórnia. Desde sua criação, a iniciativa adota a Semântica de Frames (Fillmore, 1982, 1985) como referencial teórico e concentra-se na investigação da atribuição de sentido nas línguas naturais. De acordo com Ruppenhoffer et al. (2016), o objetivo do projeto é

> [...] documentar a variedade de possibilidades combinatórias semânticas e sintáticas — valências — de cada palavra em cada um de seus sentidos, por meio de anotação de sentenças assistida por computador e da tabulação automática e exibição dos resultados da anotação. (Ruppenhoffer et al., 2016, p.7)[4]

[4] The aim is to document the range of semantic and syntactic combinatory possibilities valences of each word in each of its senses, through computer-assisted annotation of example sentences and automatic tabulation and display of the annotation results. [Tradução nossa]

Dessa forma, a FrameNet oferece um mapeamento detalhado das estruturas de valência das palavras a partir da anotação e análise de textos autênticos, compilados em corpora. A escolha por dados reais de uso de língua natural para as anotações reitera um dos objetivos da base teórica da FrameNet, a Semântica de Frames, que é oferecer um programa de pesquisa em semântica empírica para descrição de língua (Fillmore, 1982). Por meio da anotação assistida por computador, o projeto viabiliza a organização sistemática dos dados reais anotados, possibilitando avanços tanto no campo da linguística e da lexicografia quanto em aplicações computacionais, como o processamento de língua natural.

Inicialmente voltada para a língua inglesa, a FrameNet vem sendo expandida ao longo dos anos para contemplar outras línguas. Um exemplo dessa expansão é a FrameNet Brasil (FN-Br), laboratório de linguística computacional sediado na Universidade Federal de Juiz de Fora (UFJF), em Minas Gerais, que desenvolve anotações em português brasileiro e também propõe extensões à base de *frames* e à estrutura de dados originalmente concebidas pela Berkeley FrameNet. Seguindo os princípios do projeto original, a FN-Br utiliza dados reais extraídos de textos autênticos para construir um banco de dados de *frames* semânticos do português brasileiro (Salomão, 2009). No projeto, os corpora são anotados manualmente por bolsistas e voluntários em uma ferramenta própria do laboratório, sendo posteriormente utilizados em tarefas de processamento e compreensão de língua natural.

Para que a metodologia de anotação da FrameNet e da FN-Br seja melhor entendida, é necessário que algumas nomenclaturas usadas no processo de anotação sejam apresentadas. O ponto de partida de toda anotação é a unidade lexical (LU – *lexical unit*), ou seja, uma palavra ou expressão capaz de evocar um *frame*. Esse *frame*, por sua vez, pode ser composto por diferentes elementos a depender da cultura e dos costumes dos falantes. Na Figura 1, pode-se ver como o *frame* Comércio_vender[5] é definido na FN-Br.

[5] Seguindo a metodologia adotada pelas FrameNets, a menção aos nomes dos *frames* é sempre apresentada em fonte `Courier`.

Figura 1 — Frame Comércio_vender
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/report/frame/156>, último acesso em 30 jan. 2025.

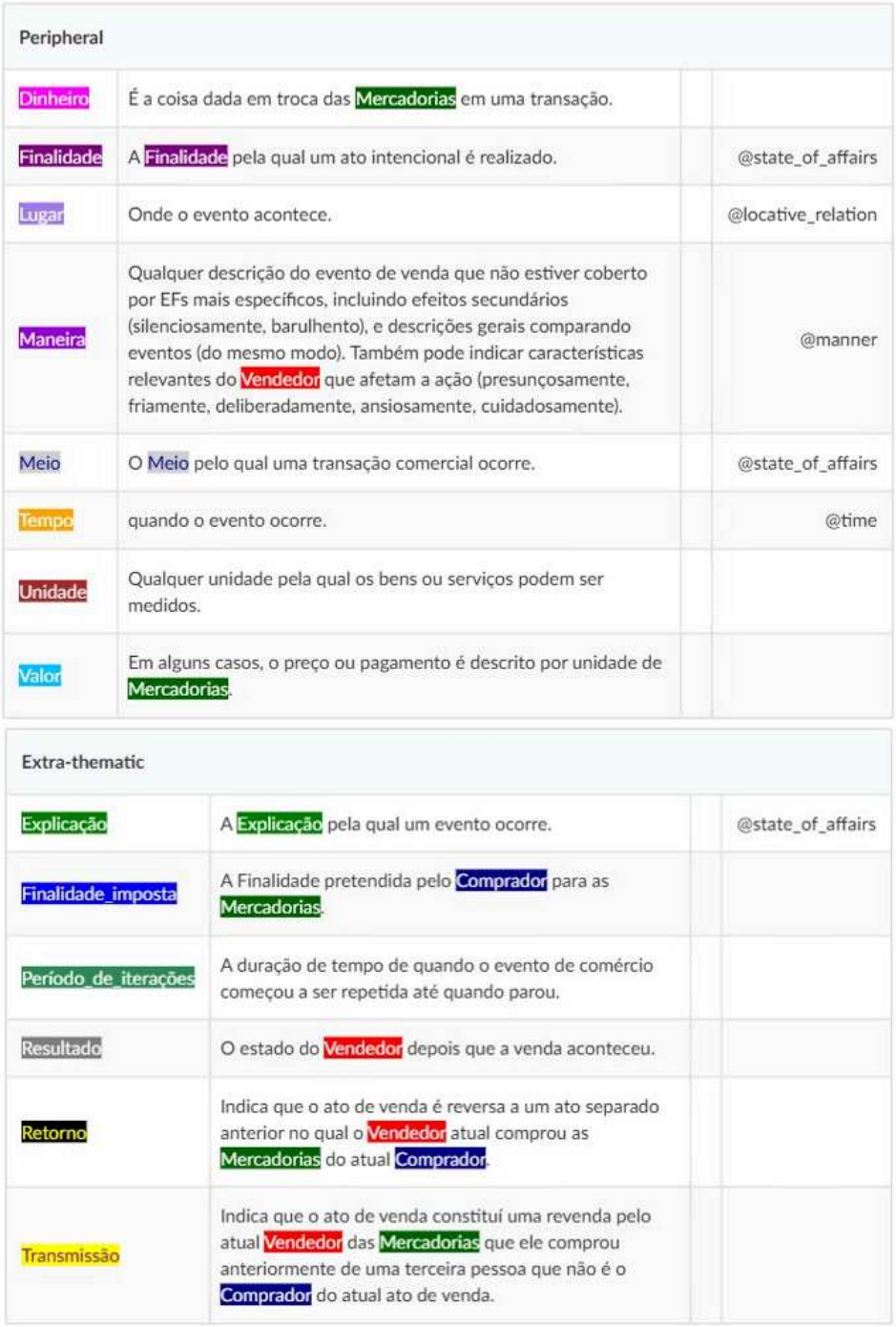
Na Figura 1, portanto, vê-se o nome do *frame* no topo da página e, em seguida, a sua definição. No caso do *frame* `Comércio_vender`, sabe-se que se trata de um evento comercial em que o vendedor vende uma mercadoria para um comprador em troca de dinheiro. Apresentada a definição do *frame*, são listados os Elementos de Frame (EFs) em seguida. Na FrameNet, os Elementos de Frame funcionam como micropapéis temáticos (Salomão, 2009), sendo definidos em relação a uma cena específica, o que busca garantir análises mais fiéis das relações semânticas entre participantes em uma sentença.

No caso da Figura 1, o *frame* `Comércio_vender` apresenta três EFs nucleares, ou seja, elementos que devem aparecer obrigatoriamente na sentença quando o *frame* `Comércio_vender` é instanciado. São eles: COMPRADOR, MERCADORIA e VENDEDOR.[6] A classificação desses elementos como nucleares indica que a ocorrência do evento de venda depende, essencialmente, da presença desses participantes. No relatório do *frame*, cada um deles é apresentado, também, com a sua definição.

De forma similar, os EFs não nucleares, que não precisam, necessariamente, fazer parte da cena, também estão listados e definidos no *report* (vide Figura 2). Diferentemente dos EFs nucleares, esses elementos desempenham funções circunstanciais, identificando apenas informações auxiliares ao evento principal,

[6] Também seguindo a metodologia das FrameNets, nomes de EFs são apresentados em VERSALETE.

como onde e quando ele ocorre, por exemplo. Na FrameNet, os EFs não nucleares são categorizados em dois tipos: periféricos, que adicionam características adicionais sobre a circunstância em que a cena descrita no *frame* acontece, e extra-temáticos, que incorporam informações que vão além do escopo do *frame* em questão, introduzindo novos *frames* na sentença.



Peripheral			
Dinheiro	É a coisa dada em troca das Mercadorias em uma transação.		
Finalidade	A Finalidade pela qual um ato intencional é realizado.		@state_of_affairs
Lugar	Onde o evento acontece.		@locative_relation
Maneira	Qualquer descrição do evento de venda que não estiver coberto por EFs mais específicos, incluindo efeitos secundários (silenciosamente, barulhento), e descrições gerais comparando eventos (do mesmo modo). Também pode indicar características relevantes do Vendedor que afetam a ação (presunçosamente, friamente, deliberadamente, ansiosamente, cuidadosamente).		@manner
Meio	O Meio pelo qual uma transação comercial ocorre.		@state_of_affairs
Tempo	quando o evento ocorre.		@time
Unidade	Qualquer unidade pela qual os bens ou serviços podem ser medidos.		
Valor	Em alguns casos, o preço ou pagamento é descrito por unidade de Mercadorias.		

Extra-thematic			
Explicação	A Explicação pela qual um evento ocorre.		@state_of_affairs
Finalidade_imposta	A Finalidade pretendida pelo Comprador para as Mercadorias.		
Período_de_iterações	A duração de tempo de quando o evento de comércio começou a ser repetida até quando parou.		
Resultado	O estado do Vendedor depois que a venda aconteceu.		
Retorno	Indica que o ato de venda é reversa a um ato separado anterior no qual o Vendedor atual comprou as Mercadorias do atual Comprador.		
Transmissão	Indica que o ato de venda constituí uma revenda pelo atual Vendedor das Mercadorias que ele comprou anteriormente de uma terceira pessoa que não é o Comprador do atual ato de venda.		

Figura 2 — Elementos não-nucleares no *frame* Comércio_vender
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/report/frame/156>, último acesso em 30 jan. 2025.

Em algumas situações de anotação, há sentenças em que EFs nucleares não são instanciados. Casos como esses são chamados de instanciações nulas e são divididos em três tipos: Instanciação Nula Definida, Instanciação Nula Indefinida e Instanciação Nula Construcional. Ainda que não sejam instanciados por algum material linguístico, tais EFs continuam acessíveis na sentença, seja como informação recuperável pelo contexto, seja como informação inferível.

A Instanciação Nula Definida (DNI) diz respeito à omissão de um EF que pode ser identificado com base no contexto linguístico. Na sentença (4), por exemplo, o verbo comer está anotado no *frame* `Ingestão`, que tem dois EFs nucleares: quem ingere, o INGESTOR, e o que é ingerido, o INGERÍVEL. Observa-se que, embora o agente da ação, o INGESTOR, esteja expresso na sentença, o EF INGERÍVEL não aparece nela, sendo considerado, portanto, uma instanciação nula. O fato de o conteúdo dessa instanciação poder facilmente ser recuperado pelo contexto — mais especificamente pela sentença (5), que a antecede no corpus — faz com que ela seja classificada como uma DNI.

(4) Quero começar, [a gente$_{\text{INGESTOR}}$] COME$^{\text{INGESTÃO}}$ assim… (Pedro pelo Mundo)

(5) Pombo com sopa. (Pedro pelo Mundo)

Assim, o EF INGERÍVEL será classificado como uma DNI durante o processo de anotação multimodal, uma vez que aquilo que é ingerido pelo INGESTOR está explícito em uma sentença anterior do corpus: pombo com sopa.

A Instanciação Nula Indefinida (INI), por sua vez, diz respeito a casos em que um EF nuclear não está lexicalmente expresso na sentença nem pode ser recuperado a partir do contexto. Nesses casos, o referente tende a ser interpretado de forma genérica. Na sentença (6), por exemplo, o verbo comer também foi anotado no *frame* Ingestão, no entanto, diferentemente do que aconteceu na sentença (4), o EF INGERÍVEL não é mencionado de forma explícita nem pode ser inferido pelo contexto, o que configura uma ocorrência de Instanciação Nula Indefinida. Algo semelhante ocorre na sentença (7). Nela, o *frame* evocado por cozinhar.v, `Criação_culinária`, é composto pelos elementos de *frame* (EFs)

nucleares COZINHEIRO e COMIDA_PRODUZIDA. Nesse exemplo, o EF COMIDA_PRODUZIDA não está lexicalmente realizado nem pode ser inferido a partir do contexto, sendo, por isso, marcado como INI durante a anotação.

(6) Fora isso, [eles$_{INGESTOR}$] COMEM^{INGESTÃO} em casa. (Pedro pelo Mundo)
(7) [Você$_{CRIADOR}$] está COZINHANDO^{CRIAÇÃO_CULINÁRIA} e aí toma uma bebida quente como esta. (Pedro pelo Mundo)

Já a Instanciação Nula Construcional (CNI) refere-se aos casos em que a estrutura da língua licencia a omissão de um EF nuclear, o que, no português brasileiro, comumente acontece nas sentenças com sujeito desinencial ou nas construções imperativas, por exemplo.

A sentença (8) é uma oração imperativa, portanto, a expressão do agente da ação não é obrigatória no português. Nesse exemplo, o EF INGESTOR não é realizado lexicalmente e é classificado como uma CNI, uma vez que essa omissão é licenciada pela própria estrutura da língua. A sentença (9), por sua vez, apresenta um sujeito desinencial. Nela, o verbo *fazer.v* está anotado no *frame* `Agir_intencionalmente`, que tem como EF nuclear o AGENTE, responsável por indicar quem realiza a ação. Embora esse elemento não esteja expresso na sentença, seu referente pode ser recuperado pela forma verbal *fiz*, que codifica a primeira pessoa do singular. A omissão do sujeito, nesse caso, é gramaticalmente permitida pelo português brasileiro.

(8) COMA^{INGESTÃO} toda a comida. (Pedro pelo Mundo)
(9) FIZ^{AGIR_INTENCIONALMENTE} ampliações dessa foto por muitos e muitos anos. (Pedro pelo Mundo)

Existem momentos, ainda, em que a própria UL que evoca o *frame* já incorpora um dos EFs em seu radical, o que é chamado de Incorporação (INC). Na sentença (10), por exemplo, a unidade lexical roupa está associada ao *frame* `Vestuário`, cujo EF nuclear é TRAJE, que identifica a peça de vestuário utilizada. Nesse caso, não há um constituinte separado na sentença que realize o EF TRAJE,

pois ele já está presente na própria UL *roupa.n*, o que caracteriza uma ocorrência de INC.

(10) Esta ROUPA$^{\text{VESTUÁRIO}}$ é dos anos vinte. (Pedro pelo Mundo)

A partir dessa organização interna dos *frames* e seus elementos, a FrameNet Brasil classifica os *frames* em cinco categorias principais: entidade, estado, evento, atributo e relação (Torrent et al., 2022). Os *frames* de entidade referem-se a objetos ou conceitos com existência independente. O *frame* `Pessoas`, por exemplo, pertence a essa categoria, por ser evocado por ULs gerais para indivíduos, como *gente.n*, *povo.n* e *humano.n*. Outro *frame* que também se encaixa nessa categoria é o de Utensílios, evocado por ULs relacionadas a recipientes ou ferramentas de uso doméstico, como *garfo.n*, *panela.n* e *tigela.n*.

Os *frames* de estado, por sua vez, descrevem condições ou situações estáveis, como estados emocionais ou físicos. São exemplos de *frames* dessa categoria o *frame* `Emoção_com_foco_no_experienciador`, evocado por ULs como *amor.n* e *raiva.n*, e o `Condições_em_saúde`, formado por ULs como *cardiopata.a* e *diabético.a*.

Já os *frames* de evento envolvem ações ou processos que ocorrem ao longo do tempo, como atividades ou acontecimentos. Um exemplo de *frame* de evento é o `Criação_culinária`, que descreve situações em que um cozinheiro prepara um alimento. Esse *frame* é evocado por ULs como *cozinhar.v*, *preparar.v* e *temperar.v*. O *frame* `Comércio_vender`, apresentado anteriormente, também pertence a essa categoria, pois representa uma cena em que um vendedor vende uma mercadoria para um comprador, sendo evocado por ULs como *vender.v* e *negociar.v*.

Os *frames* de atributo estão relacionados às características ou propriedades de entidades. Um exemplo é o *frame* `Cor`, formado por ULs associadas a cores, como *vermelho.a* e *acizentado.a*. Os *frames* `Forma` e `Tamanho` também se enquadram nessa categoria, pois qualificam entidades, sendo evocados por ULs como *retangular.a* e *grande.a*, respectivamente.

Por fim, os *frames* de relação são aqueles que estabelecem vínculos entre duas ou mais entidades, eventos ou estados. O *frame* `Parentesco`, por exemplo, agrupa ULs que expressam relações familiares, como *irmão.n*, *filho.n* e *pai.n*.

Essa distinção entre as categorias dos *frames* contribui para a organização deles na base de dados da FrameNet Brasil. Para compreender como a rede da FN-Br é estruturada, contudo, é necessário ir além da análise isolada dos *frames*. Na subseção seguinte, serão abordadas as relações semânticas que se estabelecem entre eles, responsáveis por estruturar a rede de forma dinâmica e interconectada.

2.2.1 As relações entre *frames*

Ao modelar um *frame* em uma rede como a FrameNet, é possível estabelecer conexões entre os *frames* criados e os já existentes. As relações frame-a-frame, como são chamadas, são essenciais no projeto, já que os *frames* não são unidades isoladas, mas sim parte de um sistema interconectado de conhecimento. As relações entre *frames* são categorizadas em: Herança, Subframe, Uso, Precedência, Perspectiva, Causativo_de, Incoativo_de e Veja_também. Na FrameNet Brasil, é possível observar as relações frame-a-frame através do Grapher, disponível na WebTool. Nele, cada relação possui uma cor específica, como pode ser observado na Figura 3.

Figura 3 — Relações frame-a-frame

Fonte: Adaptado de Berkeley FrameNet (2025)

A Figura 4, por exemplo, mostra as relações do *frame* `Comércio_vender` com outros *frames* da base. Percebe-se, pelas cores das setas, que esse *frame* estabelece relação de Herança (seta vermelha) e relação de Uso (seta verde) com outros *frames* da FrameNet Brasil. Essas duas categorias de relações, assim como as demais presentes no modelo, serão detalhadas a seguir.

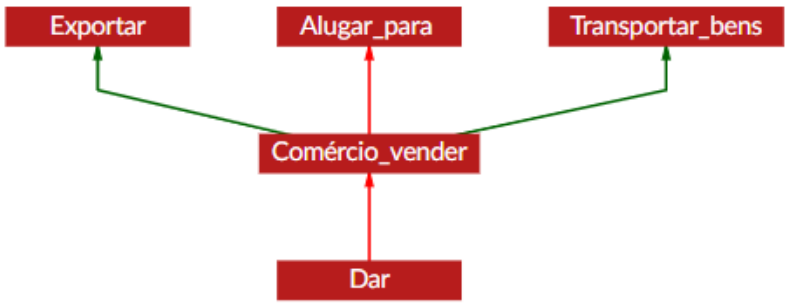


Figura 4 — Relações do *frame* Comércio_vender
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/grapher/frame>, último acesso em 30 mar. 2025

A primeira delas, a relação de Herança (*Inheritance*), é estabelecida quando um *frame* filho herda as características de um *frame* mãe. Diz-se, então, que o *frame* filho é um tipo mais específico do *frame* mãe, já que ele elabora um ou mais de seus elementos. O *frame* `Comércio_vender`, por exemplo, herda características do *frame* `Dar`, apresentado na Figura 5.

Figura 5 — *Frame* Dar
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/report/frame/127>, último acesso em 30 mar. 2025.

Logo, além de herdar os seus EFs, o novo *frame* os especifica. Enquanto em `Dar`, o EF DOADOR é mais genérico, sendo definido como uma entidade que transfere a posse de um tema para outra, em `Comércio_vender`, o EF VENDEDOR é especificado como aquele que transfere a posse de uma mercadoria para um comprador em troca de pagamento. De forma semelhante, os EFs TEMA e RECIPIENTE também são especificados ao serem herdados de `Dar`, tornando-se, respectivamente, a MERCADORIA vendida e o COMPRADOR (vide Figura 1).

A relação de Subframe, por sua vez, ocorre entre um *frame* que codifica um evento complexo e dois ou mais *frames* que representam subeventos delimitáveis que o compõem. Esses subframes descrevem etapas ou partes constituintes do evento mais amplo, permitindo uma representação mais detalhada de sua estrutura interna. O *frame* `Dar`, por exemplo, é considerado um subframe do *frame* `Cenário_de_doação`, que descreve a situação completa em que uma doação ocorre, já que o ato de dar algo a alguém é uma parte essencial desse cenário mais amplo. Também são subframes de `Cenário_de_doação` os *frames* `Pré-doação` e `Pós-doação`, os quais estabelecem um outro tipo de relação entre si, a de Precedência (Figura 6).

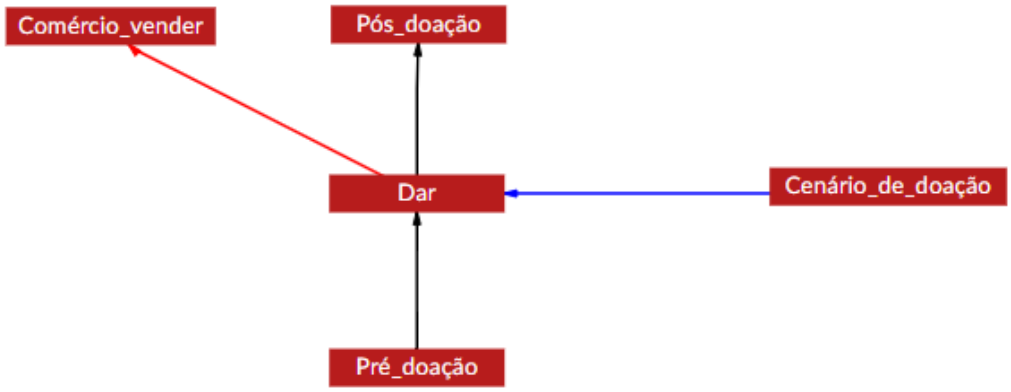


Figura 6 — Relações de Herança, Subframe e Precedência no *frame* Dar
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/report/frame/127>, último acesso em 30 mar. 2025.

A relação de Precedência ocorre quando *frames* estão dispostos em uma sequência temporal, ou seja, acontecem em ordem cronológica. Nesse caso, o *frame* Dar é precedido pelo *frame* `Pré-doação` e precede o *frame* `Pós-doação`, estabelecendo uma sequenciação de eventos, já que um ocorre após o outro.

Já a relação de Uso acontece quando um *frame* pressupõe a existência de um ou mais *frames* como o seu pano de fundo. O *frame* `Comércio_vender`, por exemplo, é usado pelo *frame* `Transportar_bens`. Isso acontece porque o ato de transportar bens frequentemente envolve um contexto comercial, no qual os itens transportados resultam de uma transação de venda. Assim, o *frame* `Comércio_vender` fornece a base conceitual necessária para a interpretação do *frame* `Transportar_bens`, garantindo que o transporte de mercadorias seja compreendido dentro de um cenário de venda.

A relação de Perspectiva, por sua vez, conecta *frames* que oferecem diferentes pontos de vista sobre uma mesma cena. Na definição do *frame* `Comércio_vender` (Figura 1), por exemplo, percebe-se que esse *frame* toma como perspectiva da cena o vendedor, diferentemente de outro *frame* da FN-Br, o `Comércio_comprar`, que enfatiza a visão do comprador sobre a cena de transação comercial. Ambos os *frames*, portanto, estabelecem uma relação de perspectiva com o *frame* `Comércio_transferência_de_mercadorias`, que

representa a cena de forma abrangente, sem assumir o ponto de vista de um dos participantes específicos da transação, conforme Figura 7.

Por fim, as relações de `Causativo_de` e `Incoativo_de` indicam, respectivamente, relações de causalidade e mudança de estado entre dois *frames*. Essas relações não podem ser demonstradas a partir do *frame* `Comércio_vender`, logo, para exemplificá-las, é necessário recorrer a outros *frames*. O *frame* `Estar_seco`, por exemplo, tem como causativo o *frame* `Causar_ficar_seco`, que indica uma situação em que um agente seca algo ou alguém. Por outro lado, `Estar_seco` estabelece uma relação de incoativo com um outro *frame*, o `Tornar-se_seco`, que representa a transição de uma entidade para o estado seco sem que haja, necessariamente, um agente responsável mencionado. A Figura 8 mostra as relações entre esses três *frames* no Grapher.

Figura 7 — *Frame* Cenário_transferência_de_mercadorias
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/report/frame/189>, último acesso em 30 mar. 2025.

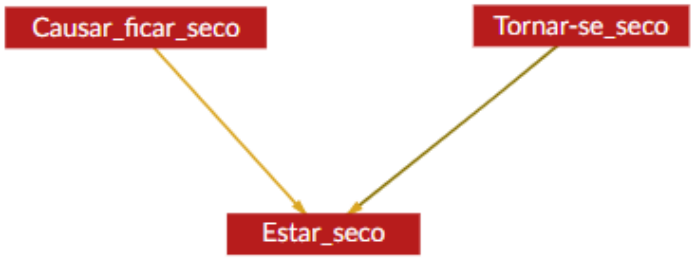


Figura 8 — Relações Causativo_de e Incoativo_de no *frame* Estar_seco
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/grapher/frame>, último acesso em 05 abr. 2025.

As relações frame-a-frame descritas até aqui demonstram como os *frames* se organizam e se conectam dentro da rede semântica da FrameNet Brasil, refletindo diferentes formas de associação entre eles. No entanto, nem todas as conexões relevantes na FrameNet ocorrem no nível dos *frames*. Há relações mais específicas que se estabelecem diretamente entre unidades lexicais de diferentes *frames* e que são fundamentais para capturar nuances semânticas importantes, especialmente em domínios mais especializados.

É nesse contexto que se inserem as relações qualia, cuja modelagem na FN-Br se inspira na teoria de Pustejovsky (1995). Embora não façam parte da Berkeley FrameNet, essas relações foram incorporadas ao projeto brasileiro para expandir a sua capacidade descritiva, resolvendo, também, a ausência de conexões diretas entre ULs dentro da base de dados da FrameNet.

2.2.2 As relações qualia ternárias

Além das relações frame-a-frame, originalmente propostas pela Berkeley FrameNet, a FrameNet Brasil incorporou à sua base um novo tipo de relação semântica: as relações qualia, implementadas com base no estudo de Pustejovsky (1995). A noção de papéis qualia parte da ideia de que a compreensão do significado das palavras pode ser influenciada por quatro fatores geradores, que representam diferentes maneiras pelas quais os humanos interpretam as relações entre objetos no mundo. Segundo Pustejovsky (1995), existem quatro categorias diferentes de relações desse tipo, que são subdivididas em:

> Formal — descreve a categoria básica do item e fornece a informação que distingue uma entidade dentro de um conjunto maior dentro de seu domínio semântico.

> Constitutivo — expressa uma variedade de relações sobre a constituição interna de uma entidade.
> Télico — diz respeito à função ou propósito típico de uma entidade, ou seja, para que a entidade serve.
> Agentivo — refere-se à origem de uma entidade, seu criador ou seu processo de surgimento.[7] (Torrent et al., 2022, p.8)

Essas quatro dimensões — formal, constitutiva, télica e agentiva — oferecem uma estrutura para explicar como os falantes compreendem o significado de itens lexicais com base em diferentes aspectos das entidades que eles designam, como sua natureza, composição, finalidade ou origem. Com base nessa perspectiva, a FrameNet Brasil adotou uma abordagem específica, a partir de *frames*, para representar essas relações no banco de dados.

Em vez de simplesmente listar as categorias de qualia como relações diretas entre ULs, o projeto propõe uma modelagem mais detalhada, que articula cada tipo de qualia por meio de um *frame* mediador. Essa proposta, denominada qualia ternária, foi desenvolvida por Costa (2020) e permite representar semanticamente os vínculos entre duas ULs de forma mais precisa. Nesse modelo, a relação qualia passa a ser ancorada em um *frame* específico, que serve de pano de fundo conceitual para a relação.

Segundo Belcavello et al. (2020), nessa modelagem, duas ULs são conectadas por meio de um *frame* que representa semanticamente o tipo de relação qualia em questão. Cada UL, então, é associada a um Elemento de Frame dentro do mesmo *frame*. Assim, o *frame* estabelece pontes semânticas entre as ULs na base de dados, atuando como uma estrutura intermediária que especifica a natureza da relação entre elas e os papéis desempenhados por cada termo envolvido. De acordo com os autores, a direção da relação também é relevante: ela é sempre unidirecional, ou seja, parte de uma UL e se direciona a outra.

A aplicação prática das relações qualia pode ser exemplificada por meio da unidade lexical *pizza.n*, que, no banco de dados da FrameNet Brasil, relaciona-se a outras cinco ULs por meio de diferentes tipos de relações qualia, cada uma mediada por um *frame* distinto. A relação agentiva, por exemplo, conecta *pizza.n* às ULs

---

[7] Formal — describes the basic category for the item and provides the information that distinguishes an entity within a larger set inside its semantic domain.
Constitutive — expresses a variety of relations concerning the internal constitution of an entity.
Telic — concerns the typical function or purpose of an entity, i.e., what the entity is for.
Agentive — concerns the origin of an entity, its creator, or its coming into being. [Tradução nossa]

pizzaria.n e *cozinheiro.n* por meio do *frame* `Criação_culinária`. Nesse caso, a pizza é interpretada como o alimento produzido (EF COMIDA_PRODUZIDA), enquanto a pizzaria e o cozinheiro são vinculados ao papel de agente (EF COZINHEIRO). A relação constitutiva, por sua vez, ocorre entre *pizza.n* e *farinha.n*, mediada pelo *frame* `Ingredientes`, que associa *pizza.n* ao EF PRODUTO e *farinha.n* ao EF MATERIAL. Já a relação formal é representada no *frame* `Exemplar`, conectando *pizza.n* ao EF EXEMPLAR e *comida.n* ao EF TIPO, ao passo que a relação télica vincula *pizza.n* ao EF UTENSÍLIO no *frame* `Finalidade_do_utensílio` — que representa objetos ou processos criados para alcançar um propósito específico — e relaciona a Unidade Lexical *comer.v* ao EF FINALIDADE (vide Figura 9).

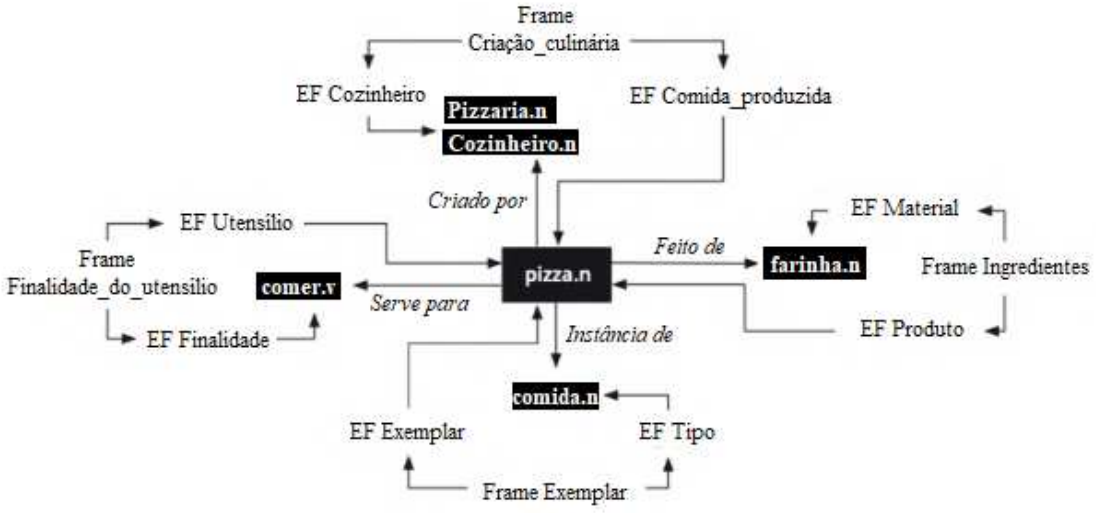


Figura 9 — As relações qualia a partir da UL pizza.n
Fonte: Adaptado de Belcavello et al., 2020.

A Figura 9 ilustra a estrutura das relações qualia no banco de dados da FrameNet Brasil, com base nas conexões entre diferentes unidades lexicais mediadas por *frames* distintos. A organização das relações na FN-Br, tanto entre *frames* quanto entre unidades lexicais, serve como base para as anotações do projeto, as quais são feitas por meio da WebTool, plataforma desenvolvida para sistematizar esse processo. A seguir, são detalhados os procedimentos de anotação adotados.

2.2.3 O passo a passo da anotação

Atualmente, todas as anotações realizadas na FrameNet Brasil são feitas por meio da WebTool, plataforma responsável pelo gerenciamento do banco de dados da FN-Br. Dentro da ferramenta, os anotadores podem acessar e interagir com as unidades lexicais (ULs) e os *frames* presentes no banco de dados, permitindo uma análise detalhada de cada item. Nesta seção, será descrita uma das principais modalidades de anotação realizadas na FrameNet Brasil, a anotação de texto corrido, e como, ao longo do tempo, foi percebida a necessidade de contemplar outras modalidades semióticas da língua no processo de anotação.

A anotação de texto corrido acompanha a metodologia da FrameNet desde suas origens. Inicialmente, os dados trabalhados no projeto, eram, assim como o padrão na área de processamento de língua natural, compostos majoritariamente por textos escritos, os quais eram reduzidos a sequências de caracteres para o tratamento computacional (Dánnells et al., 2022). Nesse cenário, as anotações estavam restritas ao plano verbal, com foco exclusivo na análise textual. Na anotação de texto corrido, a metodologia adotada visa a identificar as Unidades Lexicais (ULs) e os *frames* evocados por elas, atribuindo-lhes categorias com base nas relações semânticas presentes no texto escrito.

Nesse tipo de anotação, diferentes lotes de sentenças são distribuídos entre os anotadores do laboratório. Ao analisar essas sentenças, o anotador deve identificar e atribuir um *frame* a cada Unidade Lexical (UL) presente nelas. Para isso, na interface gráfica da WebTool, ele seleciona a palavra que deseja anotar, o que automaticamente carrega um quadro com os *frames* associados àquela UL disponíveis na FN-Br. O anotador deve, então, ler o relatório de cada *frame* para definir qual deles é o mais apropriado para a UL no contexto da sentença. Uma vez que o anotador o escolhe, é gerada uma camada de anotação do Elemento de Frame (EF), permitindo que os demais itens da sentença sejam categorizados de acordo com o *frame* escolhido.

Na Figura 10, visualiza-se a tela de anotação de uma sentença na WebTool. O processo de anotação inicia-se, em geral, com a seleção de uma UL na sentença por parte do anotador. As ULs exibidas em retângulos estão aptas a serem anotadas por integrarem a base lexical da FrameNet. Suponhamos que o anotador opte por começar pelo verbo, selecionando o lema *ver.v*.

Figura 10 — Anotação de texto corrido
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/fullText/sentence/74772>, último acesso em 11 mai. 2025.

Na tela de anotação, então, aparecem as opções de *frame* evocados por ULs cuja contraparte formal contém o referido lema (Figura 11) e o anotador deve ler os relatórios de cada *frame* na WebTool e escolher aquele que melhor se aplica ao contexto da sentença. No caso apresentado, Requerer_entidade foi o escolhido.

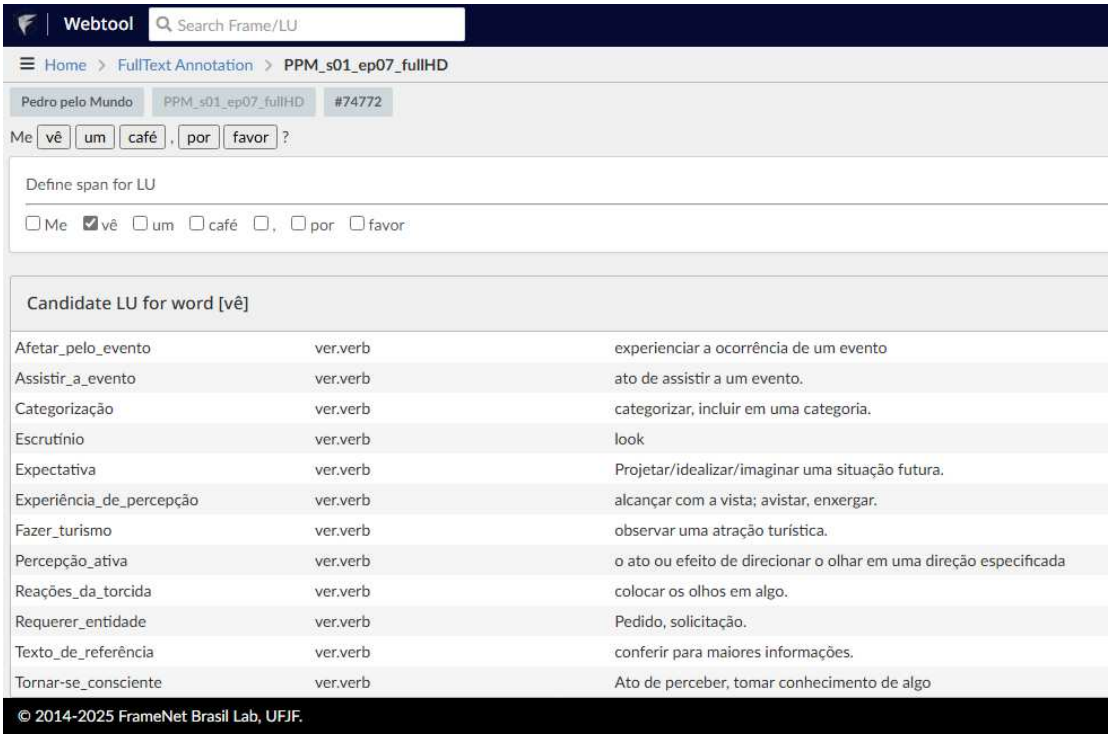


Figura 11 — *Frames* evocados por *ver.v* na FrameNet Brasil
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/fullText/sentence/74772>, último acesso em 11 mai. 2025.

Após a escolha do *frame*, o anotador passa a marcar os Elementos de Frame (EFs) correspondentes na sentença. Essa etapa consiste em identificar quais expressões linguísticas realizam os papéis semânticos previstos no *frame* selecionado. Na interface da WebTool, essa marcação é feita manualmente: o

anotador seleciona trechos da sentença que correspondem a cada EF e os associa à respectiva categoria semântica definida pelo *frame*. No caso do *frame* Requerer_entidade, os EFS nucleares são CLIENTE, ENTIDADE e FORNECEDOR. Quando um dos elementos nucleares esperados pelo *frame* não está expresso na sentença, o anotador deve indicar esse fato utilizando as categorias de instanciação nula fornecidas pela FN-Br, que aparecem do lado esquerdo da Figura 12.

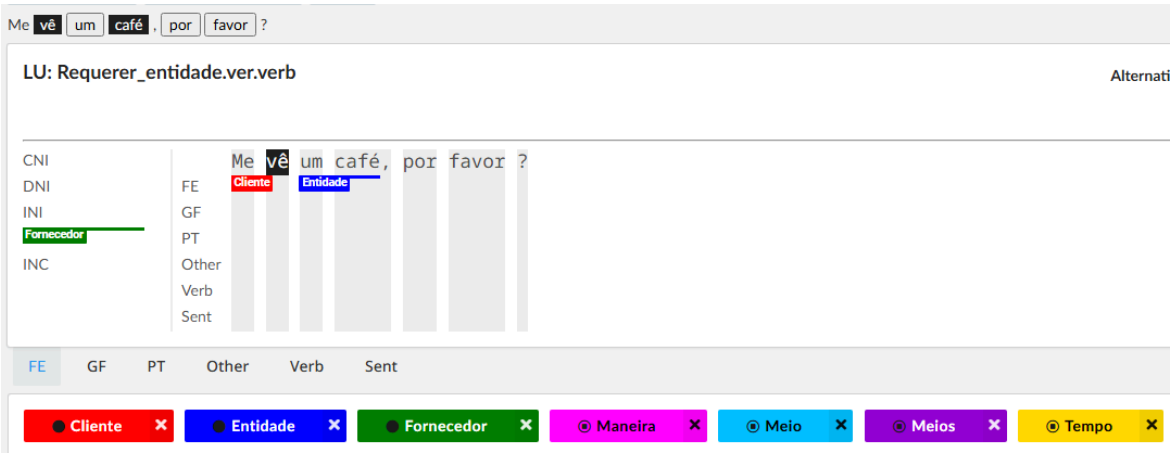


Figura 12 — Anotação dos Elementos de Frame
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/fullText/sentence/74772>, último acesso em 11 mai. 2025.

Observa-se na Figura 12 que, embora os EFs CLIENTE e ENTIDADE estejam instanciados na sentença, o EF FORNECEDOR não aparece nela e, por não poder ser retomado pelo contexto, é anotado como uma Instanciação Nula Indefinida (INI). Uma vez marcados todos os EFs nucleares presentes na sentença, o anotador pode prosseguir com a anotação das demais ULs disponíveis nela e concluir o processo de anotação.

A análise de uma sentença, contudo, pode mudar quando se considera o caráter multimodal do material anotado. É o que acontece nesse exemplo. Ao assistir ao vídeo de onde a sentença acima foi extraída, percebe-se que o elemento marcado como uma INI, o FORNECEDOR, está presente de forma explícita na imagem (Figura 13), delimitado pela *bounding box* verde. Embora ausente no plano verbal, portanto, esse EF é acessível por meio da modalidade visual.

Casos como esse revelam uma limitação importante da anotação restrita ao texto: ela desconsidera informações que, embora não verbalizadas, estão disponíveis ao interlocutor em outras modalidades semióticas. A partir de situações

recorrentes como essa, as práticas de anotação realizadas no laboratório passaram a evidenciar que, em contextos multimodais, como o audiovisual, muitos dos elementos considerados ausentes no texto estavam, na verdade, representados na imagem.

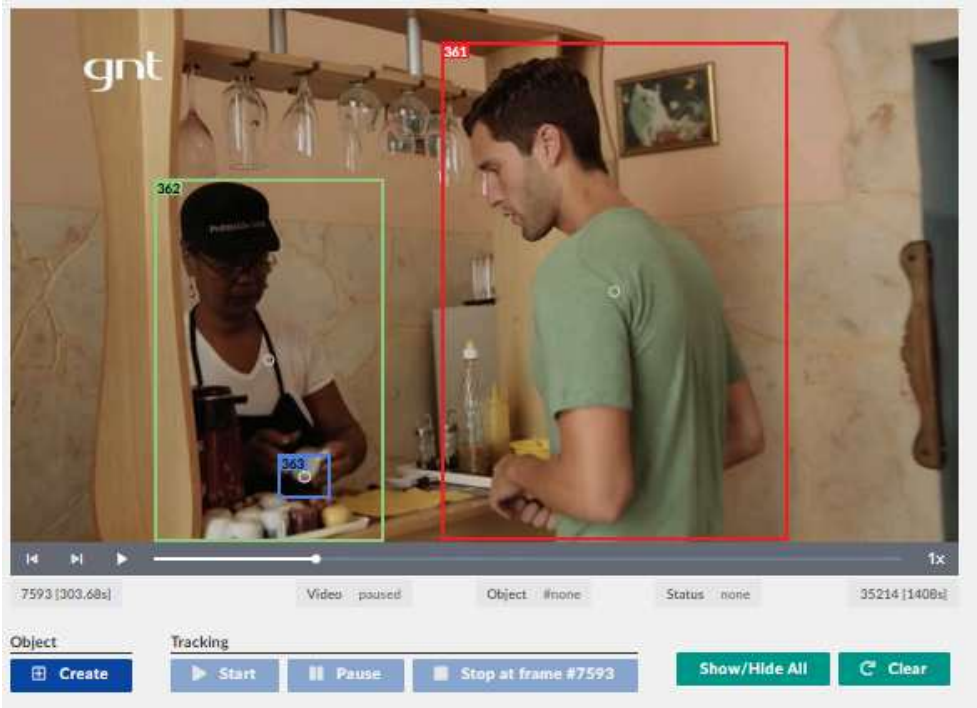


Figura 13 — Anotação multimodal
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/grapher/frame>, último acesso em 11 mai. 2025.

Esse tipo de observação reforçou a percepção de que a anotação exclusivamente textual não era suficiente para dar conta da complexidade dos eventos representados em gêneros multimodais, como os audiovisuais. Em tais gêneros, parte muito importante do sentido não é construída pela linguagem verbal, mas pelo entrelaçamento de diferentes modos de significação — e ignorar isso é perder parte essencial da cena analisada. A partir dessa constatação, a FrameNet Brasil passou a considerar a importância de incorporar outras modalidades semióticas às suas práticas de anotação, com o objetivo de representar com mais precisão os *frames* evocados em gêneros como esse. É nesse contexto que surge a proposta de anotação multimodal, tema da próxima seção.

3 A FRAMENET BRASIL ENCONTRA A MULTIMODALIDADE

A comunicação humana envolve múltiplos modos semióticos. Na visão de Bateman et. al (2017), esses modos são formados por três estratos: a materialidade, que diz respeito ao suporte físico no qual um modo opera; os recursos semióticos, que se referem à organização interna do modo e suas possibilidades combinatórias; e a semântica do discurso, que determina como esses recursos são interpretados em contextos específicos (Bateman et. al, 2017, p. 117). Como mostra a Figura 14, esses três níveis operam em conjunto.

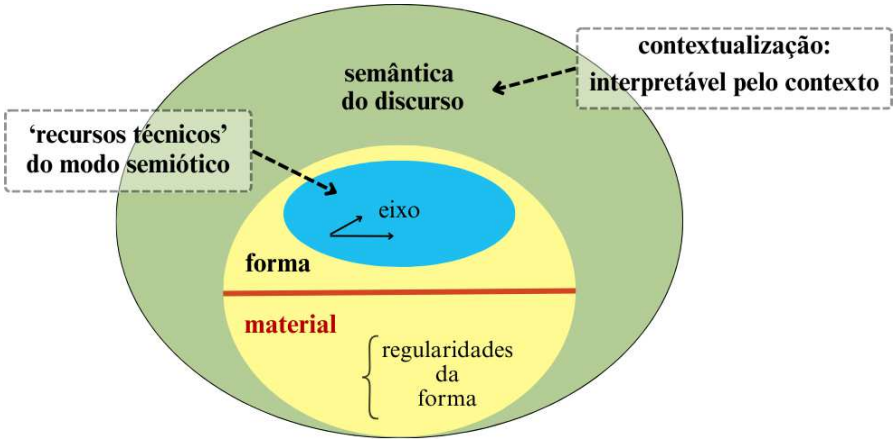


Figura 14 — Os estratos de um modo semiótico.
Fonte: Adaptado de Bateman et al., 2017, p. 117

Segundo os autores, o entendimento da estrutura individual de um modo nos permite analisar como ele contribui para a construção do significado. No entanto, embora cada modo tenha sua própria organização, o significado de um texto multimodal não é gerado apenas pela interação dos elementos internos de um único modo, mas também pela relação entre diferentes modos semióticos. Segundo Hodge e Kress (1988, p. 61), "os modos oferecem diferentes potenciais para criar significados" e, quando combinados, podem gerar novas camadas de sentido que não estariam presentes em sua análise isolada, complementando, reforçando ou até mesmo contrapondo interpretações.

Para Lemke (1998), a combinação entre modos em um texto multimodal faz com que os sentidos dele se multipliquem. De acordo com o autor,

> Em gêneros multimídia, os significados produzidos com cada recurso funcional em cada modalidade semiótica podem modular os significados de

> cada tipo em outras modalidades semióticas, multiplicando assim o conjunto de significados possíveis que podem ser produzidos (e, portanto, também a especificidade de qualquer significado particular produzido contra o pano de fundo desse conjunto maior de possibilidades). (Lemke, 1998, p.92).[8]

Dessa forma, na visão do autor, os significados gerados por um modo semiótico podem transformar ou ampliar os produzidos por outro. Assim, o sentido de um elemento verbal, por exemplo, pode ser influenciado por um elemento visual, e vice-versa. É justamente essa interdependência que gera um leque mais amplo de interpretações possíveis e torna o significado final do texto multimodal algo mais complexo do que a simples soma das partes (Lemke, 1998, p. 285). Essa perspectiva reforça a necessidade de analisar não apenas os modos isoladamente, mas também as dinâmicas de interação entre eles na produção de sentido.

Entender as relações que os modos estabelecem entre si, contudo, é um desafio. Na visão de Gualberto e Santos (2019), a multimodalidade é uma característica intrínseca a todos os textos, pois, em maior ou menor grau, todos combinam diferentes modos de comunicação na construção do significado (Gualberto; Santos, 2019, p. 6). Consoante as autoras, as análises multimodais, atualmente, buscam tanto compreender o papel de cada modo na construção de sentido, reconhecendo suas particularidades, quanto entender as possibilidades de interação entre os diferentes modos semióticos e os efeitos que essa relação gera no processo comunicativo (Gualberto; Santos, 2021).

As autoras também destacam que o aumento da produção e circulação de significados por meio de diversos modos e mídias levou à necessidade de revisar teorias que tradicionalmente focavam exclusivamente na construção de sentido pelo modo verbal. Nesse sentido, ao citarem Hodge e Kress (1988), Gualberto e Santos (2019) apontam que a incorporação da multimodalidade não se trata apenas de adicionar outros modos à análise, mas de desenvolver novas ferramentas metodológicas capazes de lidar com a complexidade da interação entre os diferentes modos semióticos.

Nessa mesma linha, Bateman et al. (2017) enfatizam que a pesquisa

[8] In multimedia genres, meanings made with each functional resource in each semiotic modality can modulate meanings of each kind in each other semiotic modality, thus multiplying the set of possible meanings that can be made (and so also the specificity of any particular meaning made against the background of this larger set of possibilities). [Tradução nossa]

multimodal não deve considerar os modos como categorias fixas e isoladas (Bateman et al., 2017, p. 117). Na obra, os autores destacam a importância de pesquisas que explorem os efeitos combinatórios que emergem da interação entre os modos, enfatizando que "o principal desafio da pesquisa multimodal é encontrar formas de caracterizar a natureza dessas interdependências e desenvolver metodologias para investigá-las empiricamente" (Bateman et al., 2017, p. 17).

Essa abordagem rompe com a visão segmentada tradicional, que analisava a linguagem verbal separadamente de outros modos, como imagens, gestos e sons, ao reconhecer que a interação entre os modos é essencial para compreender o significado como um todo. Isso implica, contudo, a necessidade de abordagens que permitam descrever e analisar não apenas os elementos individuais de um texto multimodal, mas também os mecanismos pelos quais eles se influenciam mutuamente para produzir novos sentidos.

Até recentemente, as análises realizadas pela FN-Br estavam concentradas na modalidade textual da linguagem. No entanto, diante do caráter essencialmente multimodal da comunicação humana, viu-se a necessidade de incluir outras modalidades nos estudos do projeto, a fim de que fossem realizadas análises semânticas mais completas, que não se limitassem a uma única modalidade comunicativa (Belcavello et al., 2022; Belcavello, 2023; Luz et al., 2023).

Assim, com o objetivo de viabilizar estudos desse tipo no laboratório, a FN-Br passou a buscar métodos de anotação multimodal, impulsionados pela criação da — Research and Innovation Network for Vision and Text Analysis of Multimodal Objects. Essa iniciativa tem como um de seus focos principais o desenvolvimento de um modelo semântico-computacional capaz de processar informações multimodais, incluindo imagens estáticas e dinâmicas. Atualmente, a FN-Br possui propostas de anotação multimodal utilizando ferramentas próprias, o que permite investigar a interação entre diferentes modalidades e seu impacto na construção de sentidos (Torrent et al., 2022).

Por meio dessas propostas, foi possível associar imagens e vídeos a dados sobre *frames* e Elementos de Frame evocados por entidades visuais, ampliando as informações vinculadas às Unidades Lexicais já registradas na base de dados da FrameNet Brasil (Viridiano, 2024). Entre os últimos estudos realizados no laboratório que contavam com análises multimodais, destaca-se a pesquisa de Belcavello

(2023), que propõe uma metodologia para a anotação multimodal baseada na FrameNet, com o objetivo de representar semanticamente as interações entre texto e imagem em produções audiovisuais. A hipótese central do estudo é que, assim como palavras evocam *frames* e organizam seus elementos no contexto sintático, elementos visuais também podem evocar *frames* e estruturar seus significados na tela.

Foi com a aplicação dessa metodologia que surgiu, no âmbito da FN-Br, a proposta de anotação de imagens dinâmicas. Essa nova modalidade ampliou as possibilidades de análise multimodal ao incorporar vídeos como objetos anotáveis dentro do projeto. Nessa modalidade, os anotadores trabalham com o vídeo e o áudio transcrito de cada episódio simultaneamente, permitindo uma análise integrada das duas modalidades. A anotação das imagens dinâmicas acontece em uma parte específica da WebTool, chamada *Dynamic Mode*, desenvolvida para lidar com vídeos. Para acessá-la, o anotador deve clicar em *Dynamic Mode* na página inicial da ferramenta, como mostra a Figura 15.

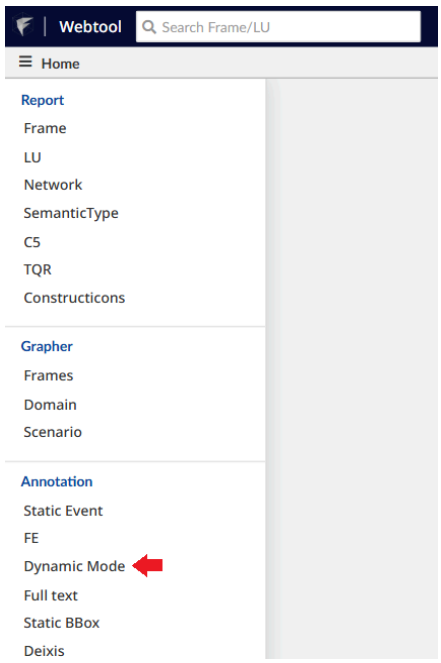


Figura 15 — Página inicial da Webtool
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/>, último acesso em 23 fev. 2025.

Depois, ele deve selecionar a pasta em que o corpus está alocado e escolher o material audiovisual que será anotado. A tela que aparecerá em seguida é a tela de anotação, que conta com três painéis e um arquivo de vídeo, o qual detém

objetos a serem anotados. Como pode ser observado na Figura 16, do lado esquerdo da interface está localizado o vídeo para a anotação, ao passo que, do lado direito, estão listados os objetos visuais já anotados. Ao clicar em um desses objetos, o anotador é direcionado para o momento em que ele aparece no vídeo.

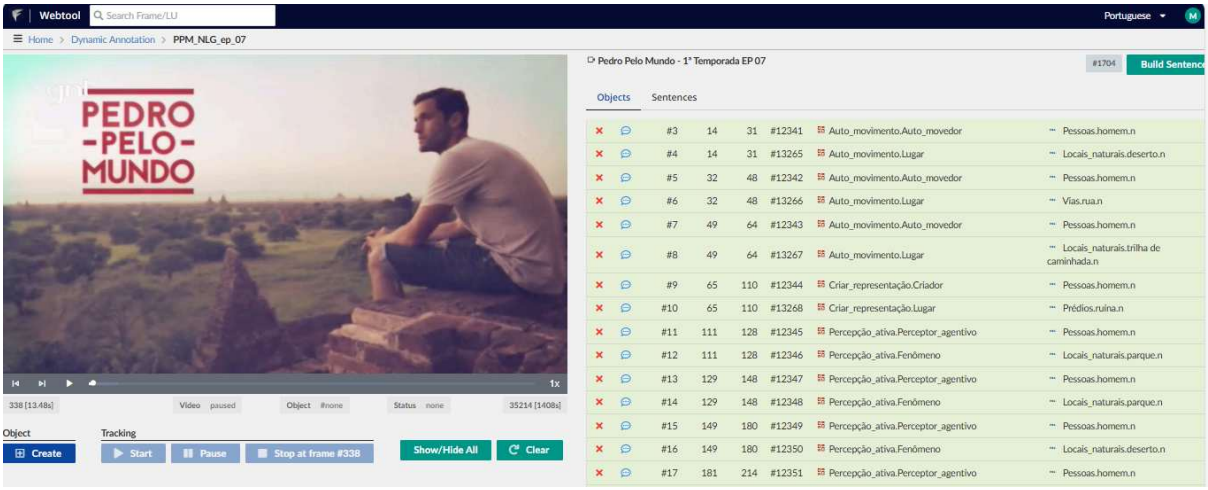


Figura 16 — Interface de anotação
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/1704>, último acesso em 23 fev. 2025.

O processo de anotação de imagens dinâmicas parte da anotação verbal já realizada: os *frames* evocados no texto servem como guia inicial — embora não sejam uma limitação — para orientar a análise do vídeo. A partir disso, o anotador delimita manualmente, com o uso de *bounding boxes*, os objetos visuais relevantes da cena. Cada objeto demarcado é, então, associado a um *frame* da base da FN-Br, conforme sua função semântica na cena. Em seguida, o anotador identifica o Elemento de Frame correspondente e atribui um *Computer Vision Name* (CV Name) ao objeto, ou seja, uma unidade lexical que represente, de forma objetiva, aquilo que está visivelmente demarcado na cena. Para ilustrar esse procedimento, considera-se a sentença a seguir, previamente anotada no modo de texto corrido, conforme mostra a Figura 17.

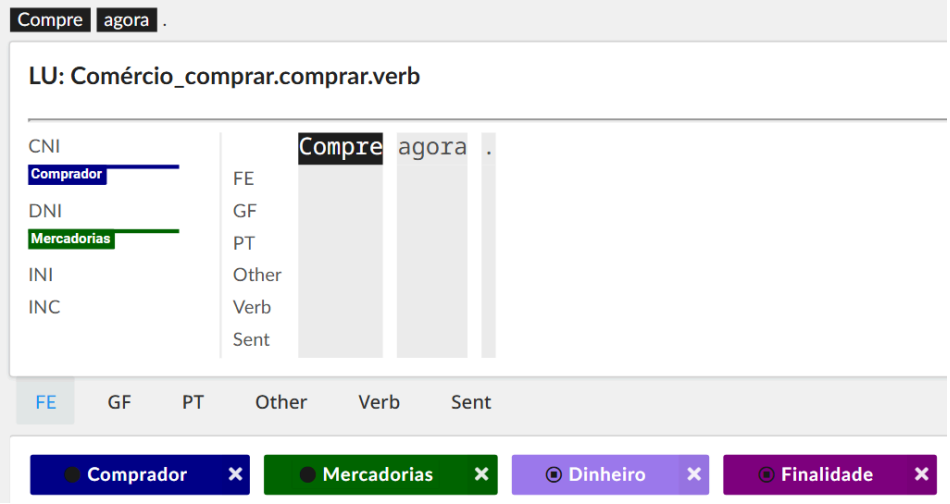


Figura 17 — Anotação de texto corrido da sentença *Compre agora*
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/fullText/sentence/73366>, último acesso em 11 mai. 2025.

Nesse caso, o verbo comprar.v foi anotado no *frame* Comércio_comprar, que tem como EFs nucleares o COMPRADOR e a MERCADORIA. Na sentença, ambos os elementos são marcados como instanciações nulas, por não aparecerem explicitamente na sentença anotada. Percebe-se, na Figura 17, que o COMPRADOR foi marcado como uma Instanciação Nula Construcional, uma vez que, por se tratar de uma sentença imperativa, o português brasileiro licencia a omissão do sujeito da sentença. Já a MERCADORIA foi marcada como uma Instanciação Nula Definida, por poder ser recuperada pelo contexto.

Sendo assim, com base nessa anotação textual, o anotador realiza a análise da cena correspondente no vídeo, delimitando os objetos visuais relevantes com o uso de *bounding boxes*. Como mostra a interface apresentada na Figura 18, o vídeo anotável aparece no lado esquerdo da tela, onde estão visíveis duas *bounding boxes*: uma azul e uma verde.

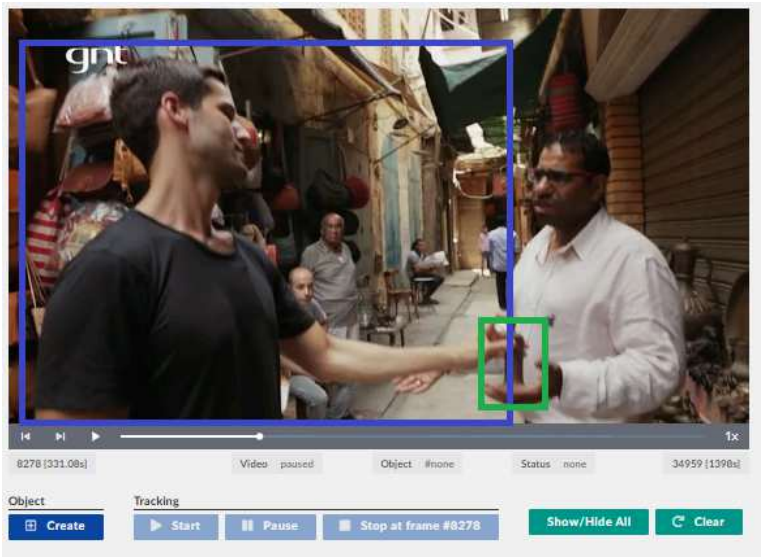


Figura 18 — Anotação de imagem dinâmica da sentença *Compre agora*
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/502>, último acesso em 11 mai. 2025.

A *bounding box* azul delimita o apresentador do programa, Pedro, enquanto a verde delimita um objeto segurado por ele. Orientando-se pela anotação de texto corrido, o anotador atribuiu a Pedro o EF nuclear de COMPRADOR na cena, ao passo que, no CV Name, associou a *bounding box* na tela à UL pessoa, já que era o que ele estava vendo objetivamente na cena (um homem). A UL pessoa faz parte, na FrameNet, do *frame* Pessoas. Em seguida, o anotador atribuiu ao objeto em sua mão o EF nuclear de MERCADORIA, enquanto, no CV Name, associou a sua *bounding box* à UL moedor, que faz parte do *frame* Utensílios.

Percebe-se, dessa forma, que, mais uma vez, os dois EFs que aparecem como instanciações nulas na anotação de texto corrido estão, na verdade, presentes visualmente na cena anotada. Como aponta Belcavello (2023), isso mostra que os elementos visuais de uma produção são capazes de instanciar elementos que fazem parte do *frame* escolhido e não foram mencionados ao longo da narração, complementando o que está sendo narrado.

Além desse tipo de correlação, outras formas de interação entre texto e vídeo foram identificadas nos estudos de Belcavello (2023). O autor registra, por exemplo, momentos em que diferentes *frames* são associados a um mesmo elemento da cena, bem como casos em que múltiplos elementos visuais se relacionam a uma única UL. Durante a pesquisa, o autor também observou que os *frames* evocados pelas imagens podem oferecer novas perspectivas sobre as cenas apresentadas.

No segundo episódio da série (Figura 19), por exemplo, foi atribuído o *frame* Turismo_de_atração à expressão “atrações turísticas”. Para anotar a cena que faz parte desse momento da narração, no entanto, o anotador selecionou o *frame* Atrair_turistas ao marcar a catedral que aparece no vídeo. A escolha de frames ativa formas distintas de estruturar a cena: enquanto o frame Turismo_de_atração não assume uma perspectiva específica dentro da cena, descrevendo uma situação em que o local tem potencial de atrair visitantes, o frame Atrair_turistas estrutura a cena a partir da experiência do turista, que se desloca em direção à atração.

Figura 19 — Catedral anotada no *frame* Atrair_turistas
Fonte: Adaptado de Belcavello (2023)

Tais observações reforçam que restringir a análise semântica de objetos multimodais, como o programa de televisão analisado, apenas ao texto verbal pode ser prejudicial à compreensão dos efeitos de sentido, já que as imagens anotadas enriquecem de forma muito significativa a anotação. Por meio da análise de *frames* na anotação multimodal, foi possível realizar comparações entre texto e vídeo e observar como os elementos de ambos os modos se correlacionam na construção de sentidos.

Na pesquisa de Belcavello (2023), todos os episódios da primeira temporada da série de viagens *Pedro pelo mundo* tiveram a narração e o vídeo anotados. Os

dados reunidos resultaram no dataset Frame[29], um conjunto de anotações semânticas detalhadas que relacionam a interação entre texto e imagem no contexto audiovisual.

A incorporação da multimodalidade pela FrameNet reflete a necessidade de ampliar a análise semântica para além do nível estritamente verbal, considerando como diferentes modos interagem na construção do significado. Essa abordagem tem possibilitado novas investigações sobre a relação entre linguagem e outros recursos semióticos, incluindo sua aplicação em contextos específicos, como a audiodescrição.

[9] Disponível em: https://huggingface.co/datasets/FrameNetBrasil/Frame2. Acesso em: 03 mai. 2025.

4 A TRADUÇÃO AUDIOVISUAL NA MODALIDADE DE AUDIODESCRIÇÃO

A Tradução Audiovisual (TAV) é um ramo da tradução especializada voltado para a mediação linguística e cultural de produtos audiovisuais. Seu processo tradutório exige atenção à articulação simultânea de múltiplos modos semióticos — como imagem, som, linguagem verbal e elementos gráficos —, que interagem entre si na construção de sentido.

Como apontam Franco e Araújo (2011), a terminologia utilizada para nomear esse campo passou por diversas transformações ao longo das últimas décadas, acompanhando a crescente diversidade de gêneros multimodais e de suportes midiáticos. Nesse cenário, expressões como “tradução fílmica” ou “tradução cinematográfica” foram progressivamente substituídas por uma nomenclatura mais abrangente, capaz de contemplar não apenas o cinema e a televisão, mas também o teatro, os jogos eletrônicos e os vídeos produzidos e veiculados em plataformas digitais e redes sociais.

Entre as modalidades de TAV, a dublagem e a legendagem são as mais conhecidas. No entanto, o campo é muito mais amplo e inclui outras práticas que respondem a diferentes demandas comunicacionais e contextos de acessibilidade. Franco e Araújo (2011) destacam que fazem parte das modalidades de TAV a legendagem para ouvintes, a legendagem para surdos e ensurdecidos (LSE), a legendagem eletrônica, a dublagem, o voice-over, a narração (ou *voice-off*) e a audiodescrição (AD), como mostra o Quadro 1.

Quadro 1 — Modalidades de TAV

Modalidade de TAV	Definição
Legendagem para ouvintes	Consiste na transcrição do conteúdo verbal da trilha sonora em forma escrita, normalmente na parte inferior da tela, de modo sincrônico com os diálogos. Essa modalidade não inclui informações sonoras não verbais, pois é voltada a espectadores ouvintes.
Legendagem para surdos e ensurdecidos (LSE)	Variante da legendagem que inclui não apenas os diálogos, mas também informações sonoras relevantes (como ruídos, músicas e efeitos sonoros) para garantir a compreensão total da mensagem por espectadores surdos ou com deficiência auditiva.

Legendagem eletrônica	Sistema em que as legendas são inseridas eletronicamente no momento da exibição, permitindo a escolha da língua ou da modalidade de legenda.
Dublagem	Substituição da trilha sonora original (falada) por uma nova trilha com vozes em outro idioma, sincronizadas com os movimentos labiais dos personagens. A voz original é completamente apagada.
Voice-over	Técnica em que a voz original é mantida com volume mais baixo, enquanto a tradução é lida por cima, geralmente sem preocupação com a sincronização labial.
Narração (ou voice-off)	Voz que é inserida na trilha sonora, mas não está associada a um personagem visível na tela. Pode ser utilizada para comentários, reflexões ou explicações adicionais.
Audiodescrição (AD)	Tradução intersemiótica que traduz imagens em palavras. Trata-se da inserção de uma faixa narrativa que descreve elementos visuais importantes (ações, cenários, expressões faciais, etc.) para que pessoas com deficiência visual possam compreender o conteúdo audiovisual.

Fonte: Adaptado de Franco e Araújo (2011, p.5-19).

Segundo as autoras, tanto a LSE quanto a AD enfrentaram, inicialmente, resistência por parte da comunidade acadêmica para serem reconhecidas como modalidades de TAV por não se caracterizarem como uma tradução entre línguas, mas entre modos semióticos. Díaz Cintas (2005) defende a inclusão dessas modalidades, reiterando que, independentemente do tipo de barreira enfrentada — seja ela linguística, como no caso da dublagem e da legendagem, ou sensorial, como ocorre na LSE e na AD —, o propósito central da TAV permanece o mesmo: tornar o conteúdo audiovisual acessível a diferentes públicos. Nas palavras do autor,

> Dublar, legendar ou traduzir em voice-over um programa é compartilhar com a ideia de acessibilidade da mesma forma que a LSE e a AD. Apenas os públicos-alvo é que são diferentes. Se o desafio é uma barreira linguística ou sensorial, o objetivo do processo tradutório é exatamente o mesmo: facilitar o acesso a uma fonte de informação e entretenimento anteriormente hermética. Nesse sentido, a acessibilidade se torna um denominador comum que permeia essas práticas. (Díaz Cintas, 2005, p. 4).[10]

[10] [...] to lip-sync, to subtitle or to voice-over a programme shares as much the idea of accessibility as SDH or AD. Only the intended audiences are different. Whether the hurdle is a language or a sensorial barrier, the aim of the translation process is exactly the same: to facilitate the access to an otherwise hermetic source of information and entertainment. In this way, accessibility becomes a common denominator that underpins these practices. [Tradução nossa]

Essa discussão sobre os contornos da Tradução Audiovisual ganha ainda mais densidade quando se considera a proposta de Jakobson (1959), que amplia a noção de tradução para além da tradução interlingual. Ao propor uma tipologia que inclui a tradução intralingual (reformulação dentro da mesma língua) e a intersemiótica (transposição entre sistemas de signos distintos), o autor oferece uma base teórica para compreender modalidades como a LSE e a AD não como exceções à tradução, mas como expressões legítimas de práticas tradutórias que operam entre diferentes modos de significação. Partindo dessa perspectiva, Naves et al. (2016) consideram também a Janela de Libras (Língua Brasileira de Sinais) como uma modalidade de tradução audiovisual acessível, haja vista que, nela, a informação sonora e verbal é transposta para o visual-gestual, permitindo o acesso de pessoas surdas ao conteúdo audiovisual.

Nesse contexto, a acessibilidade deixa de ser um elemento periférico e passa a ocupar um lugar central na TAV, especialmente no que diz respeito à inclusão de pessoas com deficiência visual e surdos e ensurdecidos. Conforme destaca Díaz Cintas (2005), a acessibilidade no âmbito audiovisual deve ser encarada como uma questão fundamental, visando garantir que todos os indivíduos possam usufruir plenamente dos produtos midiáticos, independentemente de suas capacidades sensoriais.

Para Pablo Romero-Fresco (2013), a acessibilidade na TAV deve ser considerada em todo o processo de produção audiovisual, para que as necessidades de acessibilidade não sejam tratadas como um acréscimo posterior, mas como parte integrante da criação do conteúdo. Segundo o autor, é fundamental que os produtos midiáticos sejam planejados com o propósito de serem acessíveis ao público mais variado possível, prevendo, desde a etapa de criação, a inclusão de ferramentas de acessibilidade correspondentes às diferentes modalidades de TAV.

Assim, observa-se que a acessibilidade vem assumindo um papel cada vez mais relevante na TAV, refletindo a crescente valorização de práticas tradutórias voltadas à inclusão de públicos com diferentes necessidades. Entre essas práticas, destaca-se a audiodescrição, cuja trajetória e regulamentação no contexto brasileiro merecem atenção especial.

4.1 A AUDIODESCRIÇÃO NO BRASIL

A audiodescrição é uma modalidade de tradução intersemiótica que consiste na tradução de informações visuais para a linguagem verbal, com o objetivo de tornar conteúdos audiovisuais acessíveis a pessoas cegas ou com baixa visão. Segundo Mayer (2016), a audiodescrição

> Constitui-se como uma atividade de interação entre videntes[11] e não videntes, com objetivo de contribuir para que pessoas com deficiência visual tenham um maior acesso às informações visuais oculares. Na atividade de audiodescrição, ocorre a descrição de detalhes visuais importantes como cenários, figurinos, indicação de tempo e espaço, movimentos, características físicas de pessoas/personagens e expressões faciais. (Mayer, 2016, p.4)

Como destaca a autora, a audiodescrição tem como finalidade possibilitar uma compreensão mais completa do conteúdo audiovisual, por meio da descrição de elementos relevantes para a construção do sentido, como ações, cenários, expressões faciais e marcas temporais ou espaciais. Essa modalidade de tradução audiovisual é inserida por meio de uma narração entre diálogos, sonoplastias e trilhas relevantes da produção, permitindo que o espectador com deficiência visual acesse informações visuais essenciais à narrativa. No entanto, mais do que relatar o que se vê, a audiodescrição envolve escolhas interpretativas e comunicativas que levam em conta o contexto narrativo, os efeitos de sentido e a interação com outros modos semióticos presentes na obra, como som, fala e música.

A partir dos estudos de Gregory Frazier (1975), a audiodescrição começou a ser desenvolvida formalmente em 1965, nos Estados Unidos, e se expandiu para outros países por volta da década de 1980. No Brasil, a audiodescrição chegou em 2003, durante o Festival *Assim Vivemos – Festival Internacional de Filmes sobre Deficiência*[12], realizado no Centro Cultural Banco do Brasil (CCBB). Na ocasião, foram exibidos documentários com recursos de acessibilidade, como a audiodescrição e as legendas para surdos e ensurdecidos (LSE). A narração das informações visuais foi feita ao vivo por dois atores, enquanto o público com deficiência visual acompanhava o conteúdo por meio de fones de ouvido (Franco e Silva, 2010).

---

[11] Termo usado para denominar pessoas com acuidade visual considerada “regular”.

[12] O festival *Assim Vivemos* reúne filmes de diferentes países e oferece uma experiência acessível ao público, com recursos como audiodescrição, catálogos em Braille, LSE e interpretação em Libras. As produções exibidas têm em comum o protagonismo de pessoas com deficiência, colocando suas histórias e vivências no centro da narrativa.

Em 2005, o lançamento do filme *Irmãos de Fé*, dirigido por Moacyr Góes, foi considerado o primeiro longa-metragem brasileiro a contar com audiodescrição. A ampliação dessa funcionalidade em outras mídias também avançou nos anos seguintes: em 2007, a peça *Andaime*, apresentada em São Paulo, foi a primeira produção teatral a utilizá-la; e, em 2008, a marca Natura lançou a primeira campanha publicitária brasileira com audiodescrição. Um exemplo mais recente de uso da audiodescrição em outras produções audiovisuais é a ópera *Mata Teu Pai, Ópera Balada*, apresentada no Centro Cultural Banco do Brasil em Belo Horizonte em 2025. O espetáculo é uma adaptação livre do mito Medeia e contou com sessões com audiodescrição e interpretação em Libras.

Desde 2010, as políticas públicas voltadas à acessibilidade para pessoas com deficiência visual têm avançado, pouco a pouco, no Brasil. A Portaria nº 188/2010, do Ministério das Comunicações, determinou a inclusão de, no mínimo, duas horas semanais de programação com audiodescrição nas emissoras de TV aberta com sinal digital. Pouco depois, em 2012, o Projeto de Lei nº 4.248/2012 ampliou essas iniciativas ao exigir a presença de audiodescrição em filmes exibidos nos cinemas e na televisão.

Apenas em 2016 foi lançado o Guia para Produções Audiovisuais Acessíveis (Naves et al., 2016), da Secretaria do Audiovisual do Ministério da Cultura. No material, há orientações para a produção de três modalidades de tradução audiovisual acessível: audiodescrição, legendagem para surdos e ensurdecidos e janela de língua de sinais. Segundo os autores, o objetivo do Guia é apresentar parâmetros para a elaboração dessas modalidades, a fim de que se estabeleçam diretrizes que orientem a implementação de recursos de acessibilidade nas produções audiovisuais (Naves et al., 2016).

As diretrizes relacionadas à audiodescrição têm como objetivo auxiliar a produção de roteiros de audiodescrição, pensando no que deve ser inserido na AD e nas especificidades de seu público-alvo. No documento, essas orientações estão divididas em três categorias: (1) questões técnicas, (2) questões linguísticas e (3) questões tradutórias.

As questões técnicas tratam da estrutura dos roteiros de AD. Elas incluem orientações sobre a marcação de tempo e a inserção das descrições entre os diálogos, além de destacar a importância de evitar sobreposições entre a AD e a narração, para que não interfiram em diálogos ou efeitos sonoros essenciais para a

narrativa. No entanto, segundo o guia, em casos em que os elementos visuais sejam mais relevantes do que o áudio para a construção de sentido da obra, a sobreposição é permitida. As diretrizes relacionadas às questões técnicas estão organizadas no Quadro 2.

Quadro 2 — Questões técnicas

Questões técnicas	
Tempo de inserção	“Cada uma das inserções de audiodescrição dentro de uma marcação de tempo, é colocada preferencialmente entre os diálogos e não interfere nos efeitos musicais e sonoros.”
Narração e sobreposição	“Uma boa narração deve ser fluida e não monótona, sem vida.”
	“Não é aconselhável que se sobreponha aos diálogos ou a sons importantes para o enredo, a menos que uma ação relevante para a narrativa aconteça concomitantemente a um diálogo.”
	“Apesar da sobreposição da audiodescrição em filmes e programas de televisão não ser recomendada, poderá acontecer toda vez que a informação visual for mais relevante que a informação verbal para o desenvolvimento do enredo.”

Fonte: Adaptado de Naves et al. (2016, p.11-22)

As questões linguísticas, por sua vez, focam em como a língua deve ser utilizada nas audiodescrições, garantindo clareza e adequação ao público-alvo. Conforme o guia, a linguagem deve ser objetiva e descritiva, priorizando a escolha de adjetivos e advérbios que expressem emoções de forma precisa. Além disso, são apresentadas orientações sobre o uso do tempo verbal, a estruturação das frases e a adaptação da linguagem para diferentes públicos, como crianças. O Quadro 3 apresenta as diretrizes presentes no guia relacionadas às questões linguísticas.

Quadro 3 — Questões linguísticas

Questões linguísticas	
Uso geral da linguagem	“[Deve ser] objetiva, simples, sucinta, porém vívida e imaginativa, ou seja, priorizando o uso de léxico variado e se adequando à poética e à estética do produto audiovisual.”
	“Quanto à complexidade sintática, recomenda-se o uso de orações coordenadas, sem muita complexidade; ou períodos simples, principalmente devido ao pouco espaço entre as falas dos personagens.”
Uso de adjetivos	“Os adjetivos descritivos são muito importantes na AD, pois tornam cenas, ações, características dos personagens e ambientes mais claros para o espectador.”

	“Os adjetivos devem expressar estados de humor e de emoções condizentes com os construtos universais sem valoração subjetiva por parte do audiodescritor.”
Uso de verbos	“Usar verbos específicos que indiquem a maneira de realização das ações ex: pular, saltar, saltitar.”
	“O uso do presente do indicativo é recomendado, pois torna o texto fluido e expressa o fato no momento em que acontece.”
Uso de advérbios	“Os advérbios e locuções adverbiais ajudam na descrição de uma ação, tornando-a mais clara e aproximada possível.”
	“Assim como os adjetivos, devem expressar estados de humor e de emoções condizentes com os construtos universais sem valoração subjetiva por parte do audiodescritor.”

Fonte: Adaptado de Naves et al. (2016, p.11-22)

Já as questões tradutórias, como mostra o Quadro 4, envolvem a necessidade de descrever com precisão as características físicas dos personagens, seus figurinos e estados emocionais, além de localizar os ambientes das cenas e ler os elementos visuais verbais presentes no vídeo, como títulos, legendas e créditos, por exemplo.

Quadro 4 — Questões tradutórias

Questões tradutórias	
Personagens	“Na descrição dos atributos físicos de um personagem é recomendável a seguinte sequência: gênero, faixa etária, etnia, cor da pele, estatura, compleição física, olhos, cabelos e demais características marcantes.”
	“Não é necessário descrever em detalhes as características dos personagens que não têm relevância para a trama.”
	“Os personagens são nomeados na AD quando são nomeados na narrativa. Enquanto isso não acontece, são identificados por suas características físicas. O mesmo acontece para profissões ou funções.”
Tempo e espaço	“Da mesma forma que a mudança de cenário/ambiente, a mudança de tempo é anunciada logo que aconteça para o melhor entendimento da cena. Exemplos: “é dia”, “é fim de tarde...”,“de madrugada...”
	“Sugere-se audiodescrever os elementos importantes para caracterização dos ambientes de acordo com sua importância para a compreensão da obra.”
	“É necessário localizar sempre os ambientes, dizer que o personagem volta a um determinado ambiente em que já esteve; deixar claro caso um mesmo ambiente tenha sofrido mudanças e descrever quais.”
	“Além do ambiente, outra informação importante para o entendimento da cena é dizer quantos estão em cena e quem são.”

Figurino	“Começar pelas peças maiores e pela parte superior para depois passar para as menores e acessórios.”
	“Não é necessário descrever o figurino de todos os personagens em todas as cenas, pois o excesso de informação torna a audiodescrição cansativa e tira o foco do ponto principal”
Elementos textuais	“Recomenda-se que elementos visuais verbais, tais como créditos, textos, títulos, legendas e intertítulos, sejam lidos. [...] Sua leitura deve ser feita em momento que não se sobreponha à audiodescrição de cenas.”
Sons	“É preciso referenciar a fonte sonora, isto é, a identificação da origem do som.”

Fonte: Adaptado de Naves et al. (2016, p.11-22)

Essa parte do texto também reforça a importância do conhecimento sobre enquadramentos, pontos de vista e outros elementos do cinema para a elaboração da AD. Esses elementos são sistematizados no Quadro 5.

Quadro 5 — Enquadramentos e pontos de vista

Questões tradutórias — Enquadramentos e pontos de vista	
Grande Plano Geral (GPG)	“Enquadra uma grande área de ação, na qual o ambiente é mostrado de maneira ampla e é captado a longa distância, o que apresenta o local onde a história ocorrerá naquele momento e situa os personagens da trama. Por meio desse plano, o audiodescritor descreverá o ambiente, a fim de situar o espectador com relação ao espaço que é apresentado no filme.”
Plano Geral (PG)	“Possui um ângulo de visão menor do que o GPG. Por meio dele, o local é apresentado de forma mais precisa e é mostrada a posição do personagem em cena. Com esse plano, o audiodescritor poderá descrever locais mais específicos em que os personagens se encontram.”
Plano Médio (PM)	“Tem uma função descritiva e, para isso, os personagens são enquadrados da cintura para cima, dando destaque para a figura humana. Nesse momento, o audiodescritor pode fazer uma descrição mais precisa sobre as características físicas dos personagens e de suas vestimentas.”
Primeiro Plano (PP)	“Enquadra o personagem do busto para cima. Seu objetivo é mostrar os diálogos entre os personagens e suas expressões faciais, que podem ser mais bem detalhadas pelo audiodescritor.”
Primeiríssimo Plano (PPP)	“Enquadra somente a cabeça dos personagens. É utilizado para ressaltar as expressões dos personagens, a fim de revelar suas emoções. É utilizado para ressaltar as expressões dos personagens, a fim de revelar suas emoções.”
Close-up ou Plano Detalhe	“Enquadra apenas o que é essencial para a compreensão do que está sendo apresentado, destacando-o do resto da cena.”
Plongée e Contraplongée	“Também são muito significativos, aumentando ou diminuindo o tamanho dos personagens ou objetos, não só física mas também simbolicamente, o que deverá ser enfatizado na audiodescrição.”

Planos-ponto-de vista	“Mostram diferentes pontos-de-vista, que podem ser o do autor, o do narrador, ou o de um personagem e podem também ser explicitados na audiodescrição.”

Fonte: Adaptado de Naves et al. (2016, p.11-22)

A partir desse referencial, que contribuiu para sistematizar práticas já adotadas por profissionais da audiodescrição e das demais modalidades contempladas, surgiram iniciativas destinadas a incentivar a incorporação desses recursos no circuito audiovisual nacional. Em 2017, por exemplo, a Ancine lançou o Programa de Apoio à Distribuição de Conteúdo Acessível no Segmento de Exibição Cinematográfica. O principal objetivo do programa foi incentivar a presença de AD, Libras e LSE nos lançamentos nacionais de pequeno porte, ampliando o acesso de pessoas com deficiência visual e auditiva às produções audiovisuais.

Embora em meio a avanços, a implementação da audiodescrição em produções audiovisuais ainda enfrenta desafios. Segundo Campos (2019), a AD ainda não é um recurso amplamente utilizado no Brasil, o que pode estar associado, especialmente, ao tempo e aos eventuais custos envolvidos em sua elaboração, que muitas vezes dificultam a sua implementação em larga escala. De acordo com a autora, “este contexto motiva a busca de soluções que possam reduzir as barreiras de acesso à informação visual de pessoas cegas em plataformas de vídeo digital, especialmente quando profissionais não estiverem disponíveis.” (Campos, 2019, p.2).

Diante disso, tem crescido o interesse em soluções automatizadas baseadas em inteligência artificial, que prometem reduzir os custos e agilizar o processo. Estudos como o de Campos (2019), por exemplo, analisam a eficácia dessas tecnologias no contexto educacional explorando como a IA pode ser integrada para melhorar a acessibilidade de materiais educacionais. A pesquisa aponta, contudo, que a complexidade do trabalho do audiodescritor impõe desafios consideráveis para essas tecnologias, já que esse profissional não apenas narra o que é visto na tela, mas interpreta, seleciona e organiza informações de acordo com o contexto narrativo, ajustando a linguagem ao tom da obra. Esses aspectos exigem uma compreensão fina da narrativa, da cultura e dos efeitos de sentido — capacidades que, embora em expansão, ainda representam um obstáculo para sistemas automatizados.

Faz-se necessário, dessa forma, trabalhos que avaliem a qualidade do que é produzido, já que a automatização nem sempre garante a sensibilidade interpretativa e a adequação contextual exigidas pela audiodescrição, especialmente quando se considera a complexa interação entre os modos semióticos presentes em produções audiovisuais.

4.2 A SEMÂNTICA DE FRAMES APLICADA AO ESTUDO DA AUDIODESCRIÇÃO: TRABALHOS ANTERIORES

A relação entre a Semântica de Frames e a audiodescrição tem sido explorada em estudos recentes que buscam compreender como os frames são evocados tanto no texto da audiodescrição quanto nos elementos visuais das produções multimodais. Esses estudos contribuem para os campos da Tradução Audiovisual Acessível, da Linguística Computacional e da Multimodalidade, demonstrando como diferentes modos semióticos interagem na construção do significado.

Acerca disso, em 2022, foi publicado o estudo intitulado *A audiodescrição sob a perspectiva da Semântica de Frames: um estudo exploratório*, de Débora Soares de Souza, Adriana Silvina Pagano e Maucha Andrade Gamonal (2022). A pesquisa analisa as relações entre os *frames* identificados no texto transcrito da audiodescrição e no áudio original de um trecho do curta-metragem *Eu não quero voltar sozinho,* produzido pela Lacuna Filmes em 2010. O curta integra o *Audition* – Gamonal et al. (2025a); Gamonal et al. (2025b); Souza et al. (2025); Dornelas et al. (2022) –, um dataset multimodal composto por curtas-metragens brasileiros anotados semanticamente com base na Semântica de Frames e no modelo da FrameNet Brasil, voltado à integração dos modos sonoro e visual com foco na acessibilidade audiovisual. No estudo, adota-se a metodologia de anotação de texto corrido da FrameNet Brasil, utilizando a WebTool para categorização das ULs e suas respectivas evocações de *frames.*

Na pesquisa, observou-se que os *frames* evocados pelo áudio original e pela audiodescrição, de modo geral, se complementam, dado que os mesmos tipos de

Frames de Topo[13] são evocados em ambos. Notou-se, também, que a audiodescrição parece apresentar uma linguagem mais neutra, evitando a evocação de *frames* relacionados às emoções dos personagens. Esse resultado sugere que a AD tende a privilegiar descrições objetivas, reduzindo interpretações subjetivas sobre os sentimentos dos personagens, como, de fato, parecem orientar as diretrizes da área (Naves et al., 2016). Além disso, verificou-se que os *frames* mais frequentemente evocados na AD do curta-metragem expressam eventos, o que evidencia o foco da audiodescrição na narração de ações das cenas, assegurando a compreensão dos acontecimentos pelo público-alvo.

Outro estudo, desenvolvido por Dornelas (2023) e intitulado *A audiodescrição sob a perspectiva da semântica de frames: análise dos frames evocados pelo texto da audiodescrição e pelas imagens dinâmicas num curta-metragem*, investiga a evocação de *frames* não apenas a partir do texto, mas também das imagens do curta-metragem *Eu não quero voltar sozinho* (2010). O trabalho se insere nos Estudos da Tradução e também adota a Semântica de Frames (Fillmore, 1982) como arcabouço teórico. A metodologia consiste na anotação dos *frames* presentes no texto da audiodescrição e nas imagens dinâmicas do curta-metragem, utilizando as ferramentas da FrameNet Brasil e a proposta de anotação multimodal de Belcavello et al. (2022) e Belcavello (2023). Feitas as anotações, a autora realizou um levantamento dos *frames* evocados em cada modalidade.

Os resultados da pesquisa demonstram que ambas as modalidades, verbal e visual, atuam conjuntamente na evocação de *frames*, corroborando a ideia de que todos os modos comunicativos contribuem para a construção do sentido em um texto multimodal. Assim como na pesquisa de Souza et al. (2022), o estudo observou que a AD do curta-metragem prioriza as atividades desenvolvidas pelos participantes das cenas, evocando mais *frames* de evento.

Além disso, Dornellas (2023) também avaliou a conformidade da audiodescrição com os critérios propostos pelo *Guia para Produções Audiovisuais Acessíveis* (Naves et al., 2016), concluindo que os *frames* identificados respeitam essas diretrizes e promovem uma experiência acessível ao público-alvo. A autora

[13] Os *frames* chamados de *frames* de topo ocupam os níveis mais altos da hierarquia da rede por serem mais abstratos e não herdarem características de outros *frames*. Esses *frames* agrupam categorias conceituais amplas, pois servem como ponto de partida para a herança de muitos outros f*rames* e funcionam como intermediários em diversas relações. As principais categorias de *frames* de topo são: Evento, Relação, Estado, Entidade, Localidade e Processo.

aponta, entretanto, a limitação das descrições em alguns momentos-chave do filme, que poderiam proporcionar mais detalhes sensoriais e emocionais, enriquecendo a experiência do público.

Esses estudos destacam a contribuição da Semântica de Frames para o campo da audiodescrição e sugerem abordagens metodológicas que podem ser exploradas em futuras pesquisas na área, inclusive no âmbito da geração automática de AD. Além disso, indicam a relevância de uma abordagem multimodal para a compreensão da tradução audiovisual acessível, indicando como diferentes modos semióticos podem interagir na construção de significado em produções audiovisuais.

Com base nessa relevância, este trabalho busca explorar como a estruturação de eventos em *frames* pode contribuir para a geração automática de audiodescrição, considerando a interação entre diferentes modos semióticos na construção do significado. Acerca disso, na próxima seção, serão apresentadas as etapas metodológicas adotadas nesta pesquisa, com foco na anotação de eventos em produções audiovisuais e na análise da audiodescrição.

5 MATERIAIS E MÉTODOS

O objetivo deste capítulo é apresentar a metodologia que foi utilizada nesta dissertação. Para isso, expõem-se detalhes sobre o corpus anotado, o passo a passo das anotações multimodais desenvolvidas a partir dele e as diretrizes para audiodescrição.

5.1 CORPUS

Para as análises desta dissertação, foram utilizados os episódios 1 e 7 da primeira temporada da série de viagens *Pedro pelo mundo*, exibida pelo canal GNT a partir de 2016. A temporada, como um todo, integra o conjunto de dados multimodal da ReINVenTA (Research and Innovation Network for Vision and Text Analysis of Multimodal Objects), uma rede de pesquisa dedicada ao estudo da representação semântica de objetos multimodais e às suas aplicações em Inteligência Artificial e Tecnologias Assistivas.

Na série, acompanhamos o apresentador Pedro Andrade, enquanto ele explora diferentes culturas, gastronomia e paisagens ao redor do mundo. A cada episódio, com duração média de 23 minutos, o apresentador visita um novo destino, explorando suas particularidades e promovendo uma visão imersiva e descontraída das experiências locais. Nos episódios 1 e 7, por exemplo, Pedro visita, respectivamente, Egito e Cuba, apresentando uma visão ampla dos dois países durante o programa. Em ambos os episódios, o apresentador faz entrevistas com os locais e descobre suas histórias, costumes e perspectivas sobre a vida nesses lugares, proporcionando ao público uma compreensão mais profunda da cultura e da realidade social de cada destino.

No âmbito da FrameNet, a análise da série *Pedro Pelo Mundo* integra um conjunto mais amplo de pesquisas voltadas à representação semântica de objetos audiovisuais. A escolha da série de viagens como corpus para a anotação de eventos se justifica pelo fato de que o programa já foi previamente anotado para entidades (Belcavello et al., 2022; Belcavello, 2023; Luz et al., 2023) e está sendo usado para novas pesquisas na FN-Br no âmbito da dêixis e da organização de turnos conversacionais através de gestos (Sigiliano, 2025; Abreu, Torrent e Matos,

2025), o que facilita a confluência e a continuidade dos estudos promovidos pelo laboratório e o aproveitamento das anotações anteriores. Além disso, a baixa dimensionalidade dos dados — entendida aqui como a recorrência de padrões visuais semelhantes entre os episódios, como o apresentador caminhando, interagindo com interlocutores ou explorando paisagens urbanas e naturais — representa uma vantagem para o treinamento de modelos de aprendizado de máquina. Essa uniformidade reduz a complexidade do espaço de busca, favorecendo a generalização e melhorando a eficiência dos modelos treinados a partir desses dados.

5.2 ANOTAÇÃO DE EVENTOS

Como desdobramento do trabalho de Belcavello (2023), que se concentrou na relação entre texto verbal e elementos visuais da cena, este estudo foca especificamente na anotação de eventos. O objetivo é analisar como ações e situações representadas multimodalmente podem ser descritas a partir da FrameNet Brasil, visando à aplicação em roteiros de audiodescrição elaborados por inteligência artificial.

Para isso, foram utilizadas como referência as anotações previamente realizadas no âmbito do trabalho de Belcavello (2023), tanto as de texto corrido quanto as de imagens dinâmicas, disponíveis na base da FrameNet Brasil. A partir da análise dessas anotações, foram comparados os *frames* evocados no texto com o que efetivamente aparecia no vídeo, identificando-se o que já havia sido expresso verbalmente nos episódios (e, sobretudo, o que não havia), para então ser decidido quais eventos visuais seriam destacados na nova anotação.

Diferentemente da pesquisa de Belcavello (2023), em que cada Unidade Lexical (UL) anotada no texto era utilizada como ponto de partida para buscar correspondência no vídeo (incluindo tanto entidades quanto eventos) e em que a delimitação dos eventos dependia da presença de expressão verbal, nesta pesquisa, o foco recaiu exclusivamente sobre os eventos visuais em si, ou seja, sobre aquilo que efetivamente acontecia nas cenas do episódio. A atenção se concentrou no que era de fato expresso pela imagem, mesmo na ausência de

acompanhamento verbal, como ocorre nas cenas de transição entre blocos narrativos, em que ações relevantes para a narrativa ainda ocorrem.

O processo de anotação de eventos foi conduzido de forma manual[14]. Para isso, primeiramente, os episódios completos foram assistidos, a fim de que fossem compreendidos de forma geral. Em seguida, cada cena foi analisada individualmente, identificando-se os eventos que ocorrem em cada momento. A partir dessa análise, cena por cena, os objetos multimodais que compunham cada evento foram delimitados, marcados com *bounding boxes* e associados a um *frame* da lista de *frames* de evento disponíveis na base da FN-Br. Quando havia anotações anteriores para os eventos identificados, elas foram aproveitadas; quando não, novas marcações foram feitas.

Uma vez definido o *frame*, foi identificado qual Elemento de Frame (EF) estava presente na imagem e foi associado a ele um Computer Vision Name (CV Name). Essa última etiqueta tem a função de categorizar visualmente o elemento exibido, vinculando-o a um *frame* da base de dados da FrameNet Brasil.

É importante destacar que a anotação na FrameNet é perspectivizada: diferentes *frames* podem ser selecionados para descrever um mesmo evento, dependendo da perspectiva do anotador. Essa dimensão da anotação dialoga com discussões mais amplas sobre a incorporação de diferentes perspectivas humanas em processos de anotação de dados, como as desenvolvidas por Cabitza et al. (2021), que defendem a importância de considerar múltiplos olhares na construção de *datasets* anotados. Como, nesta pesquisa, a geração do roteiro para audiodescrição foi baseada diretamente nas anotações de eventos que serão detalhadas nesta seção, acredita-se que a descrição produzida também é uma entre várias possíveis interpretações do conteúdo audiovisual — isto é, uma perspectiva possível, que é guiada pelos *frames* ativados e reflete as escolhas feitas na etapa de anotação.

A título de exemplificação do processo de anotação, a Figura 20 apresenta uma anotação de imagem dinâmica. Ao assistir a cena em questão, identificou-se que se tratava de um evento de transação comercial, em que dois vendedores vendem as suas mercadorias (nesse caso, frutas e legumes) em uma banca na calçada. Tendo reconhecido o evento que estrutura a cena, optou-se por anotar os

[14] Nesta pesquisa, todas as anotações foram realizadas pela própria autora.

elementos constituintes da cena no *frame* `Comércio_vender`, que tem como EFs nucleares o VENDEDOR, o COMPRADOR e a MERCADORIA — vide 2.2. Sendo assim, com *bounding boxes*, os participantes da cena foram delimitados no vídeo e associados ao *frame* escolhido, a partir do qual foram identificados os EFs que compunham o evento. Em seguida, cada *bounding box* foi relacionada a um CV Name diferente, indicando o que era visto objetivamente na cena.

A *bounding box* #89, por exemplo, delimita um dos vendedores no vídeo e foi vinculada ao EF VENDEDOR do *frame* `Comércio_vender`, conforme ilustrado no lado direito da Figura 20. Para essa *bounding box*, foi atribuída a UL homem.n, uma vez que esse era o elemento identificado objetivamente na cena. Na FrameNet Brasil, a UL homem.n está associada ao *frame* `Pessoas`.

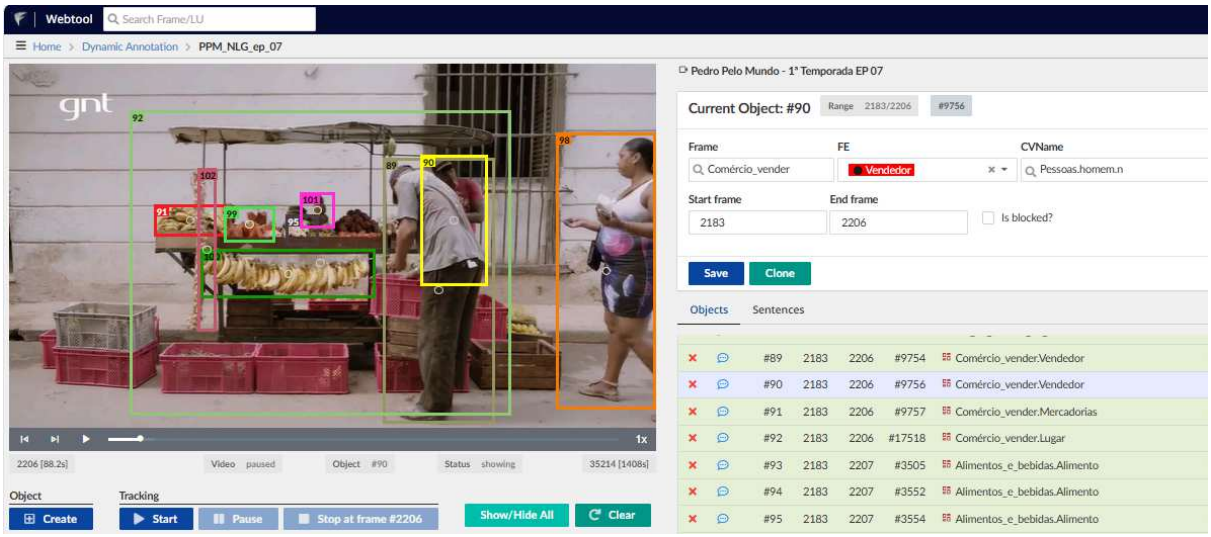


Figura 20 — Anotação de imagem dinâmica para eventos
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/1704>, último acesso em 23 fev. 2025

Na interface gráfica, também é possível ajustar o intervalo de tempo em que os objetos visuais aparecem no vídeo, rastreando-os durante todo o período de tempo em que o evento acontece. Durante os episódios, é comum que um mesmo objeto participe de diferentes eventos dentro de uma mesma cena. Nesses casos, optou-se por encerrar o rastreamento da *bounding box* e iniciar uma nova, de modo a garantir a segmentação dos eventos do episódio.

Além disso, houve situações em que mais de um evento acontecia em um mesmo momento do vídeo, como mostra a Figura 21. Na cena em questão,

enquanto o apresentador da série, Pedro Andrade, caminha por um calçadão e apresenta informações sobre o local, outras pessoas realizam diferentes ações no fundo do vídeo. Em momentos como esse, assim como caberia ao audiodescritor definir o que será priorizado no momento da narração, cabe ao anotador definir qual evento seria mais relevante para o entendimento do episódio.

Para além dos objetos multimodais que fazem parte de eventos em cada cena, também foram anotados os *letterings* presentes nos episódios, pois desempenham um papel fundamental na construção do significado. Nos vídeos, os *letterings* são trechos de texto inseridos na imagem (como legendas, títulos e nomes) que fornecem informações relevantes sobre os locais visitados pelo apresentador e identificam os falantes nas entrevistas, sendo essenciais para uma compreensão mais completa do material audiovisual.

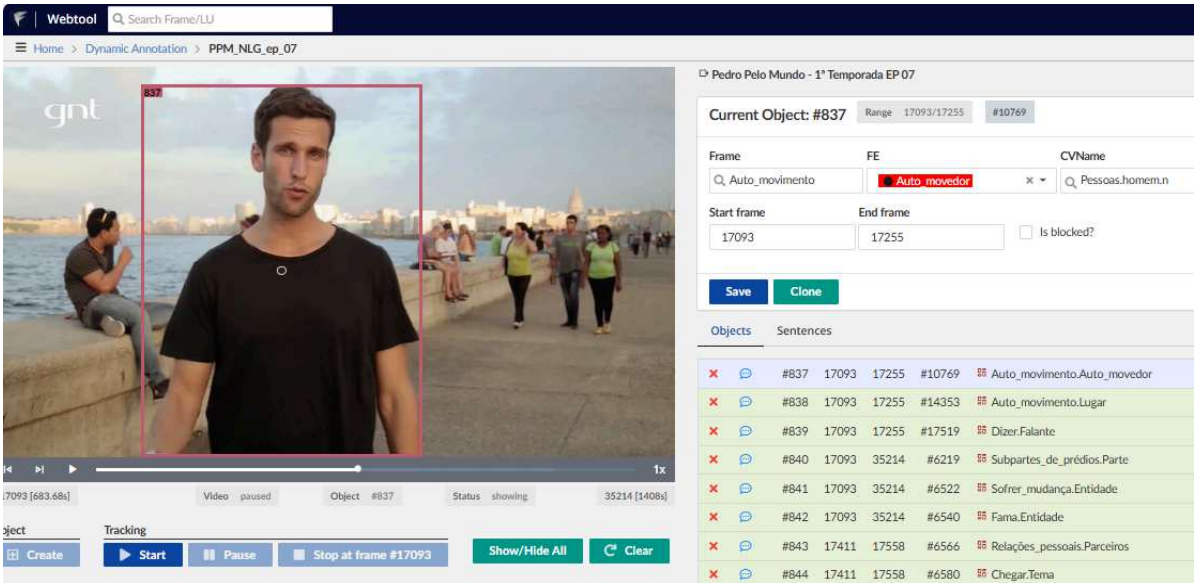


Figura 21 — Diferentes eventos em uma mesma cena
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/1704>, último acesso em 23 fev. 2025.

Com as anotações dos eventos dos episódios 1 e 7 da série *Pedro pelo mundo* concluídas, o próximo passo foi organizá-las e adaptá-las para a construção do *prompt*[15] utilizado no copiloto de IA. Essa etapa buscou garantir que as informações estivessem estruturadas de forma clara e adequada para a interação

[15] *Prompt* é o termo utilizado para designar o comando ou instrução textual fornecido pelo usuário a um modelo de linguagem, a partir do qual o sistema gera uma resposta em linguagem natural.

com o sistema. A seguir, será apresentado o processo de elaboração e aplicação desses *prompts* no contexto do uso do copiloto de inteligência artificial.

5.3 USO DE COPILOTOS DE IA

Durante o desenvolvimento da pesquisa, foram testados dois copilotos de inteligência artificial baseados em modelos de língua de grande escala (Large Language Models – LLMs), a fim de explorar seu potencial na geração de roteiros de audiodescrição a partir de anotações semânticas. Os copilotos testados foram o ChatGPT, da OpenAI, e o Gemini, da Google, utilizados em caráter exploratório. Ambos funcionam como assistentes virtuais capazes de receber comandos escritos e gerar respostas em língua natural, produzindo textos, resumindo informações e realizando tarefas variadas conforme as instruções fornecidas.

Com base nos testes realizados, contudo, optou-se pelo uso exclusivo do Gemini, acessado por meio do plano Gemini Plus, do Google DeepMind. Essa escolha foi motivada, principalmente, pela capacidade da ferramenta de processar vídeos, o que facilitou a análise detalhada das cenas dos episódios. Além disso, o Gemini também mostrou um bom desempenho ao interpretar instruções complexas e gerar textos com base em entradas estruturadas, como as anotações em *frames*. Ademais, pelo fato de a UFJF franquear a seus estudantes uma conta Google for Education, o acesso ao Gemini Plus foi gratuito.

A interação com o modelo foi feita por meio de *prompts* manuais específicos, elaborados de acordo com as diferentes condições experimentais da pesquisa: foram utilizados *prompts* distintos para os casos com e sem anotação de *frames*. Os *prompts* foram construídos de forma iterativa, com ajustes progressivos, até que se obtivesse um formato que favorecesse a geração de roteiros de AD mais adequados ao formato esperado.

Durante os testes iniciais, dois desafios principais foram identificados. O primeiro esteve relacionado ao processamento de vídeos longos: ao enviar o episódio completo ao modelo, surgiam muitas omissões de informações relevantes ou descrições pouco coerentes em grande parte dos trechos. Para contornar esse problema, o episódio foi dividido em segmentos menores, respeitando a sequência narrativa e preservando a coerência dos eventos. Cada segmento incluía todas as

informações visuais e os diálogos correspondentes, garantindo contexto suficiente para que o modelo gerasse uma audiodescrição precisa. Essa estratégia também permitiu controlar melhor o tempo das narrações, evitando sentenças longas demais ou perda de detalhes importantes.

O segundo desafio identificado esteve diretamente relacionado ao ritmo das narrações. Nas primeiras tentativas, o modelo gerava sentenças extensas, que não se ajustavam ao ritmo esperado para a audiodescrição nem respeitavam os intervalos de silêncio necessários entre as falas, como ilustra a Figura 22.

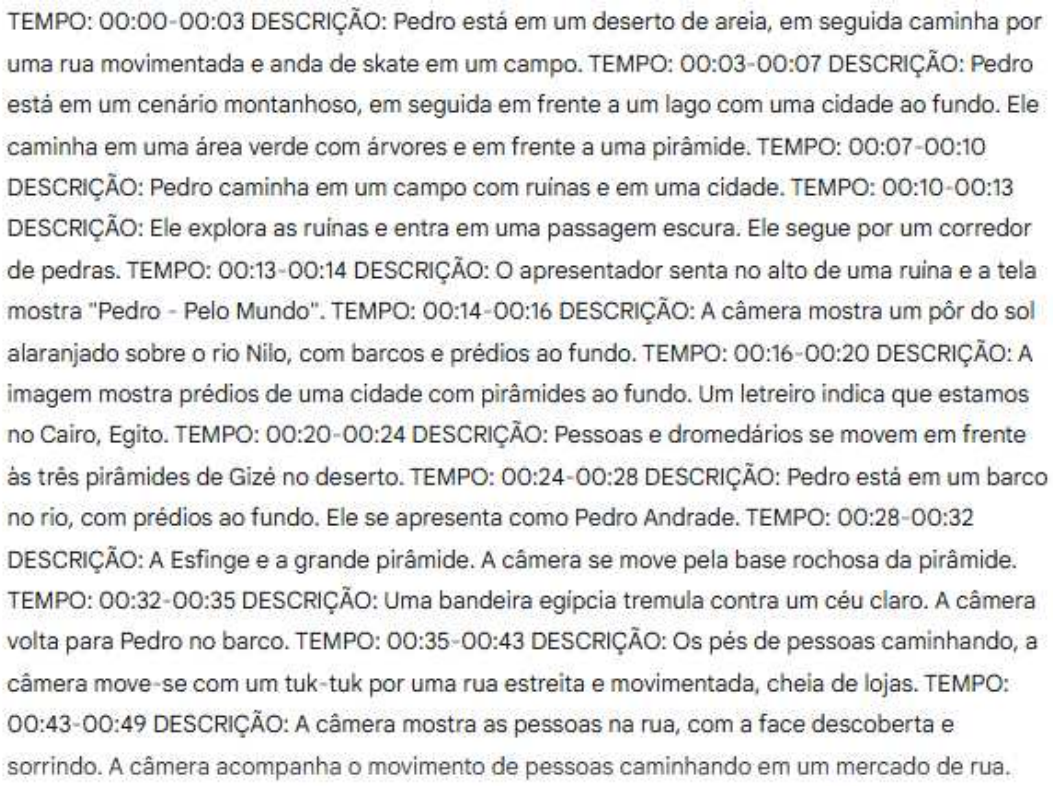

TEMPO: 00:00-00:03 DESCRIÇÃO: Pedro está em um deserto de areia, em seguida caminha por uma rua movimentada e anda de skate em um campo. TEMPO: 00:03-00:07 DESCRIÇÃO: Pedro está em um cenário montanhoso, em seguida em frente a um lago com uma cidade ao fundo. Ele caminha em uma área verde com árvores e em frente a uma pirâmide. TEMPO: 00:07-00:10 DESCRIÇÃO: Pedro caminha em um campo com ruínas e em uma cidade. TEMPO: 00:10-00:13 DESCRIÇÃO: Ele explora as ruínas e entra em uma passagem escura. Ele segue por um corredor de pedras. TEMPO: 00:13-00:14 DESCRIÇÃO: O apresentador senta no alto de uma ruína e a tela mostra "Pedro - Pelo Mundo". TEMPO: 00:14-00:16 DESCRIÇÃO: A câmera mostra um pôr do sol alaranjado sobre o rio Nilo, com barcos e prédios ao fundo. TEMPO: 00:16-00:20 DESCRIÇÃO: A imagem mostra prédios de uma cidade com pirâmides ao fundo. Um letreiro indica que estamos no Cairo, Egito. TEMPO: 00:20-00:24 DESCRIÇÃO: Pessoas e dromedários se movem em frente às três pirâmides de Gizé no deserto. TEMPO: 00:24-00:28 DESCRIÇÃO: Pedro está em um barco no rio, com prédios ao fundo. Ele se apresenta como Pedro Andrade. TEMPO: 00:28-00:32 DESCRIÇÃO: A Esfinge e a grande pirâmide. A câmera se move pela base rochosa da pirâmide. TEMPO: 00:32-00:35 DESCRIÇÃO: Uma bandeira egípcia tremula contra um céu claro. A câmera volta para Pedro no barco. TEMPO: 00:35-00:43 DESCRIÇÃO: Os pés de pessoas caminhando, a câmera move-se com um tuk-tuk por uma rua estreita e movimentada, cheia de lojas. TEMPO: 00:43-00:49 DESCRIÇÃO: A câmera mostra as pessoas na rua, com a face descoberta e sorrindo. A câmera acompanha o movimento de pessoas caminhando em um mercado de rua.

Figura 22 — Trecho do roteiro gerado sem o ajuste de densidade da locução.
Fonte: Elaborado pela autora a partir de interação com a ferramenta Gemini Plus.

Esse aspecto exigiu ajustes no formato dos *prompts*, que passaram a incluir uma instrução específica sobre a duração e o tempo das falas, destacadas na Figura 23.

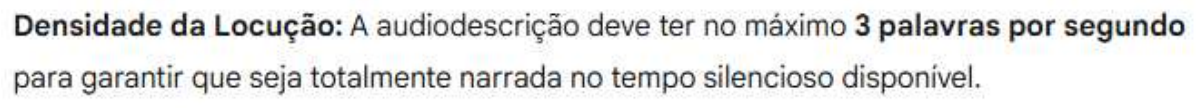
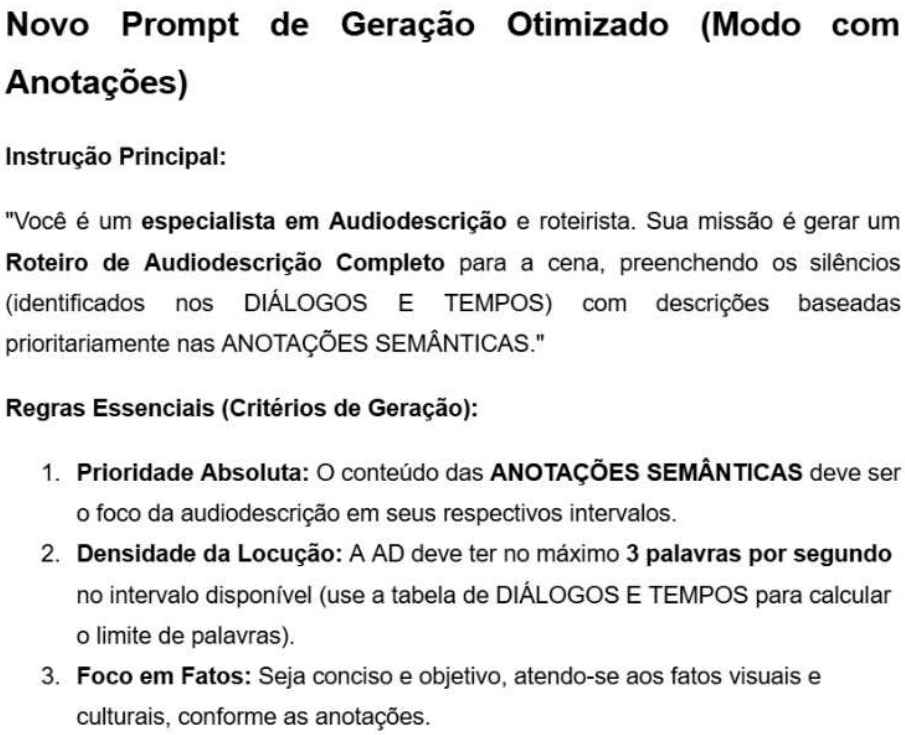

Densidade da Locução: A audiodescrição deve ter no máximo **3 palavras por segundo** para garantir que seja totalmente narrada no tempo silencioso disponível.

Figura 23 — Ajuste de densidade da locução.

Fonte: Elaborado pela autora.

Nas Figura 24 e 25, são apresentados os *prompts* já ajustados. A primeira figura é um exemplo do *prompt* utilizado para a condição com anotação de *frames*, ao passo que a segunda é um exemplo do *prompt* ajustado para a condição de controle, sem anotação semântica. É importante destacar que, embora cada imagem mostre apenas parte do diálogo da cena, o diálogo completo do trecho foi enviado ao modelo para a geração dos roteiros.



Novo Prompt de Geração Otimizado (Modo com Anotações)

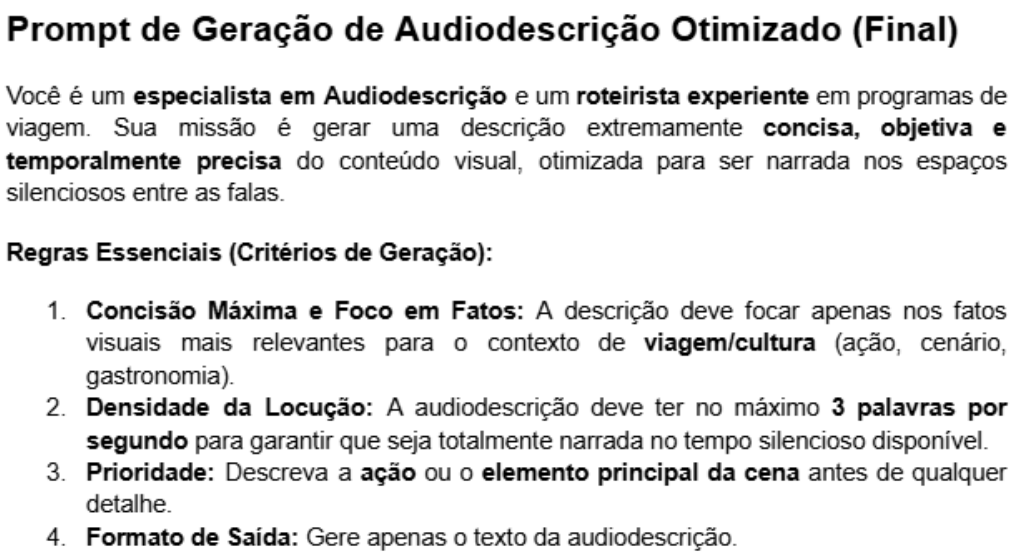
Instrução Principal:

"Você é um **especialista em Audiodescrição** e roteirista. Sua missão é gerar um **Roteiro de Audiodescrição Completo** para a cena, preenchendo os silêncios (identificados nos DIÁLOGOS E TEMPOS) com descrições baseadas prioritariamente nas ANOTAÇÕES SEMÂNTICAS."

Regras Essenciais (Critérios de Geração):

1. **Prioridade Absoluta:** O conteúdo das **ANOTAÇÕES SEMÂNTICAS** deve ser o foco da audiodescrição em seus respectivos intervalos.
2. **Densidade da Locução:** A AD deve ter no máximo **3 palavras por segundo** no intervalo disponível (use a tabela de DIÁLOGOS E TEMPOS para calcular o limite de palavras).
3. **Foco em Fatos:** Seja conciso e objetivo, atendo-se aos fatos visuais e culturais, conforme as anotações.

Figura 24 — *Prompt* para condição com anotação
Fonte: Elaborado pela autora.



Prompt de Geração de Audiodescrição Otimizado (Final)

Você é um **especialista em Audiodescrição** e um **roteirista experiente** em programas de viagem. Sua missão é gerar uma descrição extremamente **concisa, objetiva e temporalmente precisa** do conteúdo visual, otimizada para ser narrada nos espaços silenciosos entre as falas.

Regras Essenciais (Critérios de Geração):

1. **Concisão Máxima e Foco em Fatos:** A descrição deve focar apenas nos fatos visuais mais relevantes para o contexto de **viagem/cultura** (ação, cenário, gastronomia).
2. **Densidade da Locução:** A audiodescrição deve ter no máximo **3 palavras por segundo** para garantir que seja totalmente narrada no tempo silencioso disponível.
3. **Prioridade:** Descreva a **ação** ou o **elemento principal da cena** antes de qualquer detalhe.
4. **Formato de Saída:** Gere apenas o texto da audiodescrição.

Figura 25 — *Prompt* para condição sem anotação
Fonte: Elaborado pela autora.

Na condição com anotação, o modelo recebeu, para cada segmento do episódio, o trecho de vídeo correspondente, os diálogos daquele trecho e um arquivo PDF contendo somente as anotações semânticas daquele trecho específico — documento ilustrado na Figura 26. Em cada PDF, as anotações foram organizadas em tabelas que incluíam: o identificador da *bounding box* (ID), o momento de início e término dela no vídeo; e o *frame*, o EF e o CV Name aos quais ela estava associada.. Essa combinação permitiu que o Gemini se apoiasse nas informações estruturadas, acessando os eventos visuais e verbais de forma integrada e produzindo roteiros de audiodescrição mais precisos e coerentes.

EPISÓDIO 7 — TRECHO 01-03-03

IDDynamicObject	StartTime	EndTime	Frame	EF	CV Name
14806	258	259	Self_motion	Self_mover	homem.noun
14807	259	261	Self_motion	Self_mover	homem.noun
14808	264	271	Telling	Speaker	homem.noun
14809	271	273	Removing	Agent	homem.noun
14810	271	273	Removing	Theme	sapato.noun
14811	273	275	Activity	Agent	homem.noun
14812	273	275	Activity	Agent	homem.noun
14815	275	278	Chatting	Interlocutor_1	homem.noun
14816	275	278	Chatting	Interlocutor_2	homem.noun
14818	290	298	Telling	Speaker	homem.noun
14819	290	298	Self_motion	Self_mover	homem.noun
14820	290	298	Self_motion	Place	rua.noun

Figura 26 — Recorte do arquivo com as anotações semânticas de um trecho do Episódio 7
Fonte: Elaborado pela autora.

Já na condição sem anotação, o modelo recebeu apenas o trecho de vídeo e os diálogos correspondentes, sem o arquivo de anotações semânticas, precisando gerar os roteiros com base exclusivamente no material originalmente veiculado quando da exibição do programa de TV. Essa condição serviu como referência para avaliar o efeito do suporte das anotações sobre a qualidade e a precisão do texto produzido.

Com os roteiros de audiodescrição gerados para cada episódio, estes foram copiados integralmente, sem qualquer edição ou filtragem, garantindo a fidelidade ao output produzido pelo modelo. A partir desse material, tornou-se possível avançar para a avaliação do desempenho do copiloto na tarefa. Para tanto, foi aplicada a métrica de similaridade de cosseno, que permite comparar quantitativamente a proximidade semântica entre textos. Essa abordagem fornece uma medida objetiva do quanto os roteiros gerados pelo Gemini, em cada condição experimental, aproximam-se das anotações multimodais e entre si, oferecendo um panorama inicial sobre o efeito das anotações na qualidade e coerência das narrativas.

5.4 MÉTRICA DE SIMILARIDADE SEMÂNTICA

A fim de comparar os roteiros de audiodescrição gerados pelo sistema de IA Gemini e verificar o efeito das anotações multimodais sobre o texto produzido, foi aplicada a métrica de similaridade de cosseno. Esse procedimento permite quantificar o grau de proximidade semântica entre dois textos, atribuindo valores numéricos que indicam o quanto eles compartilham de significado.

Diferentemente de abordagens baseadas em *embeddings* derivados de modelos de língua, neste estudo, as representações vetoriais foram obtidas por meio de um algoritmo de *Spread Activation* aplicado sobre a rede de *frames* da FrameNet Brasil. Segundo Viridiano (2024), nesse modelo, cada texto (ou imagem) é representado pelos *frames* anotados, que funcionam como nós de ativação em um grafo semântico. Inicialmente, esses *frames* recebem o valor máximo de energia e essa energia é propagada ao longo das relações entre *frames*.

A cada iteração, a energia se espalha para os nós conectados e diminui gradualmente, até que o sistema se estabilize. O resultado desse processo é um vetor ponderado de *frames*, em que cada elemento reflete o grau de ativação alcançado por determinado *frame* dentro da rede. Esses vetores servem, então, como base para o cálculo da similaridade de cosseno.

A similaridade de cosseno, portanto, mede o ângulo entre os vetores correspondentes a dois textos. A relação entre esses vetores pode ser expressa pela seguinte fórmula (Figura 27):

$$\text{similaridade de cosseno} = \frac{A \cdot B}{\|A\| \times \|B\|}$$

Figura 27 — Fórmula da similaridade de cosseno
Fonte: Elaborado pela autora

Nela, A·B representa o produto escalar entre os vetores A e B, e ‖A‖ e ‖B‖ correspondem ao comprimento de cada vetor. O valor obtido varia entre –1 e 1, indicando o grau de proximidade entre os textos comparados.

Quanto menor o ângulo, maior é a semelhança semântica entre eles. O valor resultante é normalizado entre –1 e 1. Valores próximos de 1 indicam alta

similaridade, o que significa que os textos são semanticamente próximos; valores em torno de 0 apontam para baixa similaridade, indicando que os textos compartilham poucos elementos de sentido; e valores negativos revelam divergência semântica, sugerindo que os textos tratam de temas ou conteúdos opostos.

No presente estudo, a similaridade de cosseno foi empregada em três comparações principais: entre o roteiro gerado com anotações e o roteiro sem anotações, a fim de verificar o impacto do suporte multimodal sobre a formulação textual; entre o roteiro com anotações e as próprias anotações do vídeo, para avaliar o alinhamento entre o texto gerado e o conteúdo audiovisual original; e entre o roteiro sem anotações e as anotações do vídeo, de modo a observar o desempenho do modelo na ausência desse suporte. Os valores foram calculados para cada segmento dos episódios analisados, permitindo observar variações de desempenho conforme a extensão dos trechos e a densidade informacional das cenas. A interpretação dos resultados quantitativos é apresentada na seção seguinte, seguida de uma análise qualitativa que discute, em maior detalhe, as diferenças observadas entre os roteiros.

Os valores de similaridade obtidos servem como suporte para a interpretação dos resultados, mas não esgotam a compreensão sobre a adequação dos roteiros à audiodescrição. Por essa razão, a análise é complementada por um exame qualitativo, que considera a correspondência dos textos com as diretrizes do *Guia para Produções Audiovisuais Acessíveis* (Naves et al., 2016), avaliando a precisão, a coerência e a fluidez das narrativas geradas em cada condição experimental. Dessa forma, a combinação de métricas quantitativas e qualitativas possibilita uma avaliação mais completa do desempenho do modelo e do efeito das anotações semânticas sobre a geração da audiodescrição.

6 ANÁLISE: AVALIANDO O DESEMPENHO DE COPILOTOS DE IA PARA A GERAÇÃO DE ROTEIROS DE AUDIODESCRIÇÃO

Com os roteiros de audiodescrição gerados pelo Gemini, tanto na condição com anotação de *frames* quanto na condição sem anotação, tornou-se possível realizar uma análise detalhada de seu desempenho[16]. O objetivo desta etapa foi avaliar, de forma quantitativa e qualitativa, como as diferentes condições experimentais impactaram a qualidade, a coerência e a fidelidade das narrativas geradas. Para isso, utilizou-se uma abordagem mista: a métrica de similaridade de cosseno forneceu uma medida objetiva da proximidade semântica entre os textos, enquanto a análise qualitativa permitiu examinar a aderência dos roteiros às diretrizes do *Guia para Produções Audiovisuais Acessíveis* (Naves et al., 2016) e à sequência narrativa das cenas.

6.1 ANÁLISE QUANTITATIVA

A avaliação do desempenho dos roteiros gerados pelo Gemini foi realizada a partir dos valores de similaridade de cosseno, calculados para cada trecho dos episódios analisados. Cada episódio foi segmentado em múltiplos vídeos curtos, respeitando a sequência narrativa, de modo a facilitar o processamento pelo modelo e reduzir omissões ou inconsistências nas descrições. A similaridade de cosseno foi calculada utilizando vetores derivados dos *frames* anotados para cada trecho[17], conforme o método de *Spread Activation* aplicado sobre a rede de *frames* da FrameNet Brasil. A divisão do primeiro episódio resultou em 39 vídeos, enquanto o episódio 7 foi segmentado em 23 vídeos devido à maior duração de cada cena.

No episódio 1, os resultados indicam que a presença de anotações semânticas favoreceu o alinhamento do roteiro com o conteúdo audiovisual. A média da similaridade de cosseno entre o roteiro com anotações e as próprias anotações do vídeo foi de 0,503, enquanto a média do roteiro sem anotações em relação às anotações do vídeo foi de 0,415. Já a comparação direta entre os

[16] Neste trabalho, *desempenho* refere-se à avaliação dos roteiros de audiodescrição gerados por IA em duas dimensões: uma qualitativa, relacionada à adequação descritiva e ao alinhamento com o conteúdo audiovisual, e uma quantitativa, mensurada por meio da similaridade de cosseno.

[17] A lista completa de todos os *frames* anotados e utilizados no cálculo da similaridade de cosseno está apresentada no Apêndice I.

roteiros com e sem anotação mostrou que 24 dos 39 trechos apresentaram maior similaridade na condição com anotação, enquanto 2 trechos resultaram em empate. Esses dados sugerem que o suporte multimodal contribuiu para uma aproximação semântica mais consistente entre as descrições textuais e o conteúdo audiovisual.

Para fins de visualização e comparação, a Tabela 1 apresenta os resultados de similaridade no Episódio 1. A primeira coluna mostra os identificadores usados para organizar o corpus; cada código indica, nessa ordem, o episódio, o segmento e o subsegmento analisado. As demais colunas apresentam, respectivamente, a similaridade entre o roteiro gerado com anotação multimodal e as anotações multimodais (COM/SEM), a similaridade entre esse roteiro e as anotações do vídeo (COM/VÍDEO) e, por fim, a similaridade entre o roteiro gerado sem anotação e as anotações do vídeo (SEM/VÍDEO).

Tabela 1 — Resultados de similaridade no Episódio 1
Fonte: Elaborado pela autora.

Trecho	COM/SEM	COM/VÍDEO	SEM/VÍDEO
01-01-01	0,53632	0,706004	0,525298
01-01-02	0,528381	0,458744	0,28259
01-01-03	0,580924	0,476613	0,405836
01-02-01	0,785912	0,471471	0,466233
01-02-02	0,536235	0,428909	0,67201
01-03-01	0,244231	0,287891	0,100989
01-03-02	0,696052	0,666046	0,617991
01-03-03	0,630764	0,672724	0,697631
01-03-04	0,771601	0,608378	0,49907
01-03-05	0,683587	0,64503	0,436157
01-03-06	0,74046	0,492336	0,498996
01-04-01	0,558007	0,786804	0,389245
01-04-02	0,559464	0,699674	0,194062
01-05-01	0,76356	0,63072	0,493071

01-05-02	0,721723	0,582953	0,620752
01-05-03	0,647761	0,594654	0,552559
01-05-04	0,22721	0,22717	0,328729
01-05-05	0,660012	0,43585	0,365813
01-05-06	0,344396	0,454908	0,447862
01-05-07	0,685909	0,60428	0,420079
01-06-01	0,780154	0,550275	0,668466
01-06-02	0,487023	0,548786	0,490794
01-07-01	0,676765	0,452124	0,417582
01-07-02	0,939207	0,649856	0,66024
01-07-03	0,882016	0,409072	0,434251
01-07-04	0,870861	0,59762	0,59243
01-08-01	1	0,376934	0,376934
01-08-02	1	0,373198	0,373198
01-09-01	0,550202	0,321184	0,32838
01-09-02	0,459194	0,451532	0,344618
01-09-03	0,552474	0,344231	0,261532
01-09-04	0,791251	0,290628	0,394199
01-09-05	-1	0,281215	-1
01-09-06	0,558155	0,524555	0,609304
01-10-01	1	0,46706	0,46706
01-10-02	0,782258	0,441394	0,393665
01-10-03	0,70541	0,36162	0,495428
01-10-04	0,784172	0,649913	0,511834
01-10-05	0,449476	0,612816	0,386722

No episódio 7, por sua vez, observou-se um comportamento mais irregular nas medidas de similaridade entre as versões geradas pelo modelo e o vídeo original, em comparação ao episódio 1. De modo geral, as pontuações oscilaram com maior amplitude, o que pode estar relacionado à maior complexidade narrativa e à densidade de eventos visuais presentes nesse episódio. Ao todo, foram analisados 23 trechos, nos quais o desempenho dos modelos se mostrou equilibrado: em 9 trechos, a versão com anotação apresentou valores superiores; em 10 vídeos trechos, a versão sem anotação obteve resultados ligeiramente melhores; e em 4 casos, houve empate. O equilíbrio entre o resultado das versões reforça que o efeito das anotações não é linear, mas dependente do tipo de informação presente no vídeo.

Assim como na Tabela anterior, a seguir são apresentados os resultados de similaridade no Episódio 7.

Tabela 2 — Resultados de similaridade no Episódio 7

Trecho	COM/SEM	COM/VÍDEO	SEM/VÍDEO
07-01-01	0,768695	0,688249	0,593828
07-01-02	0,633062	0,598702	0,658446
07-01-03	0,506286	0,387262	0,345024
07-01-04	0,787844	0,216621	0,495683
07-02-01	0,405081	0,607028	0,387592
07-02-02	0,804749	0,575103	0,610091
07-02-03	0,512892	0,792315	0,505557
07-02-04	0,848466	0,606696	0,606798
07-03-01	0,770663	0,495456	0,521891
07-04-01	1	0,406441	0,406441
07-04-02	0,756612	0,340980	0,284752
07-05-01	0,704537	-1	-1
07-05-02	0,413458	0,470755	0,424871

07-06-01	1	0,197351	0,197351
07-07-01	0,808321	0,297327	0,322358
07-08-01	0,853932	0,505280	0,554037
07-08-02	0,820978	0,655310	0,675979
07-09-01	0,444608	0,446247	0,267072
07-09-02	0,640918	0,612133	0,469910
07-10-01	0,861292	0,486398	0,566260
07-10-02	0,536331	0,535127	0,524218
07-10-03	1	0,401659	0,401659
07-11-01	0,637944	0,439574	0,462873

Fonte: Elaborado pela autora.

As médias gerais de similaridade foram próximas — cerca de 0,424 para a versão com anotações e 0,407 para a versão sem —, o que indica desempenhos comparáveis no conjunto do episódio. Observa-se, contudo, que a versão com metadados atingiu, em mais ocasiões, valores próximos de 0,7, enquanto a versão sem anotações tendia a permanecer abaixo desse patamar. Esse comportamento sugere que, embora ambas as versões oscilem em função das características contextuais das cenas, a versão com anotações tende a apresentar maior potencial de alinhamento com o conteúdo multimodal quando encontra condições favoráveis, como a presença de eventos mais claramente definidos ou relações semânticas mais estruturadas.

Sobre isso, a comparação entre os dois episódios revela um padrão consistente na distribuição das diferenças de similaridade entre as versões com e sem anotações. Em ambos os casos, observou-se que, quando a versão com anotações apresentou desempenho inferior, as diferenças negativas tenderam a ser menores, ao passo que, nas situações em que o desempenho foi superior, as diferenças positivas mostraram-se mais expressivas. No episódio 1, a média das diferenças favoráveis à versão com anotações foi de aproximadamente 0,16, enquanto a média das diferenças desfavoráveis foi de cerca de 0,07. De modo

semelhante, no episódio 7, as diferenças positivas tiveram média aproximada de 0,14, ao passo que as negativas se mantiveram próximas de 0,06.

Esse comportamento indica que a presença de informações estruturadas sobre eventos e papéis semânticos não apenas contribui para uma maior estabilidade do desempenho, como também amplia o potencial de alinhamento quando as condições contextuais são favoráveis. Em outras palavras, mesmo quando o uso de anotações não resulta em ganhos imediatos de similaridade, ele tende a não comprometer a qualidade da geração e, nas situações em que há convergência entre a estrutura anotada e o conteúdo visual, a contribuição é significativamente mais expressiva.

Além desses aspectos, diferentemente do que se poderia supor, a duração dos trechos não se mostrou um fator determinante para o desempenho. Trechos mais longos não apresentaram, de modo sistemático, menores índices de similaridade. Esse resultado reforça a necessidade de uma análise qualitativa para compreender em que contextos a versão com anotações se aproxima mais do conteúdo original. As diferenças pontuais entre as versões parecem estar mais associadas ao tipo de evento representado, à quantidade de participantes envolvidos ou ao grau de sobreposição entre os elementos visuais e linguísticos da cena.

Em síntese, os resultados quantitativos sugerem que o uso de anotações de *frames* atua como um modulador do desempenho do modelo, fortalecendo o alinhamento semântico quando há compatibilidade entre as estruturas anotadas e o conteúdo visual. As médias próximas entre as condições com e sem anotação indicam que o ganho não é uniforme, mas contextual, dependente das propriedades narrativas e da complexidade de cada cena. Dessa forma, a etapa seguinte da pesquisa dedica-se à análise qualitativa dos episódios, na qual se busca compreender em que situações específicas as anotações contribuem para o aumento da coerência multimodal e quando sua influência se torna residual ou neutra.

6.2 ANÁLISE QUALITATIVA

A etapa qualitativa desta pesquisa tem como objetivo aprofundar a compreensão dos resultados apresentados na análise quantitativa, examinando não apenas as diferenças de similaridade entre as versões com e sem anotação, mas também a qualidade dos roteiros produzidos pelo assistente IA. Busca-se, assim, compreender como as informações semânticas estruturadas, oriundas da anotação de *frames*, interferem na geração textual e em que medida as produções se aproximam das diretrizes estabelecidas no *Guia para Produções Audiovisuais Acessíveis* (Naves et al., 2016).

É importante destacar que tais diretrizes não foram disponibilizadas ao modelo durante a geração dos roteiros, o que permite avaliar de forma independente o grau de convergência espontânea entre as descrições produzidas e os princípios norteadores da prática de audiodescrição. Essa análise possibilita, portanto, observar não apenas a eficiência técnica das anotações na melhoria da similaridade semântica, mas também o seu impacto na adequação comunicativa e na coerência das descrições.

A análise qualitativa concentra-se em trechos representativos dos dois episódios (1 e 7), selecionados com base em critérios que permitem observar diferentes comportamentos do modelo: a presença de grandes divergências entre as medidas de similaridade das versões com e sem anotação, casos de empate ou desempenho equilibrado e a relevância narrativa do trecho dentro do episódio. Em cada um desses trechos, a análise considera três aspectos principais: o alinhamento da descrição com o conteúdo visual apresentado; a adequação às diretrizes do Guia; e a influência das anotações de *frames* na estruturação semântica das sentenças geradas. Esse conjunto de critérios permite identificar padrões de comportamento do modelo e compreender em que condições as anotações multimodais contribuem para o alinhamento entre língua e imagem.

Após o envio do material em formato de vídeo, a IA gerou as descrições correspondentes para cada trecho. A seguir, serão analisados três trechos do *corpus*, selecionados com base em diferenças de similaridade entre as versões com e sem anotação, bem como em momentos narrativos considerados representativos para a avaliação do alinhamento entre descrição e conteúdo visual.

6.2.1 Melhor desempenho da versão anotada

O primeiro trecho, 07-02-03, faz parte do Episódio 7 da série e tem duração de 38 segundos. O segmento foi selecionado por apresentar o maior valor de similaridade na versão com anotação e uma diferença expressiva em relação à versão sem anotação (0,79 e 0,50, respectivamente). Nesse momento da narrativa, Pedro chega a uma praça onde são vendidas obras de arte e souvenirs, comentando sobre o custo de vida em Cuba e o preço dos produtos em relação ao salário dos moradores locais. A Figura 28 mostra a sequência de imagens que representa a dinâmica da cena, sem as anotações visuais.

Figura 28 — Sequência de imagens do trecho 07-02-03 sem anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/1704>, último acesso em 9 nov. 2025.

O primeiro intervalo de silêncio desse trecho corresponde às três primeiras imagens da Figura 28. Para este momento, a audiodescrição gerada pela versão sem anotações está expressa na sentença (11):

(11) Pedro caminha. Esculturas de metal. Quadros de rua expostos.

Nessa versão, nota-se que a IA identifica o movimento do apresentador, que caminha pela calçada, e prioriza a enumeração de elementos visuais que são mais salientes, como esculturas e quadros. A descrição assume um tom informativo e sintético, centrado na listagem dos objetos visíveis, o que acontece em outros momentos do roteiro gerado pela versão sem anotações. Nesse sentido, a presença de estruturas nominais curtas sugere que o modelo concentra-se nos elementos estáticos da cena, enfatizando a ambientação sem estabelecer relações explícitas entre ela e a ação de Pedro na cena neste momento.

A Figura 29 mostra o mesmo trecho com anotações, acompanhado das delimitações visuais (*bounding boxes*) que indicam os elementos anotados semanticamente para *frames*.

Figura 29 — Primeiras três imagens do trecho 07-02-03 com anotações
Fonte: Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/1704>, último acesso em 9 nov. 2025.

No trecho em questão, composto por três momentos, diferentes elementos das cenas foram delimitados por *bounding boxes* e anotados no *frame* `Percepção_ativa`, que representa o evento de observar realizado pelo apresentador naquele momento. Os objetos visuais foram associados a distintos Elementos de Frame (EFs): Pedro foi anotado como PERCEPTOR_AGENTIVO, aquele que realiza a ação de observar, enquanto os itens expostos correspondem ao FENÔMENO,

isto é, aquilo que é percebido. Além disso, o apresentador foi também anotado no *frame* `Auto_movimento`, no EF AUTO_MOVEDOR, por estar em deslocamento na cena.

Todos os objetos visuais presentes na cena também tiveram o campo CV Name preenchido. As artes, por exemplo, foram anotadas como *quadro.n* no *frame* `Obra_de_arte_física`, ao passo que as estantes com os itens para venda foram associadas à UL *estante.n* no *frame* `Móveis`. Pedro, por sua vez, foi classificado como *pessoa.n* no *frame* Pessoas.

As outras pessoas que caminhavam pela praça (vide Figura 28) também foram anotadas na cena para o *frame* `Auto_movimento`, no EF AUTO_MOVEDOR. Além disso, elas, assim como Pedro, foram associadas ao *frame* `Pessoas` pela UL *pessoa.n* no CV Name.

A partir desse novo conjunto de dados, a descrição gerada difere da anterior. Com base nas anotações, o modelo passou a mencionar tanto a ação de Pedro observando os quadros quanto a presença e o movimento das pessoas na praça, ampliando o campo de observação da cena, como exemplificado na sentença (12):

(12) Pedro observa quadros. Várias pessoas caminham na praça.

Essa inclusão de informações reflete que o modelo utilizou as anotações para organizar o texto de forma a representar os eventos e interações da cena, evidenciando como as informações estruturadas são aplicadas na geração da audiodescrição. Nesse contexto, o modelo priorizou a ação de “observar” do *frame* `Percepção_ativa`, com Pedro como PERCEPTOR_AGENTIVO e os quadros como FENÔMENO, apesar de Pedro também estar anotado no *frame* `Auto_movimento` como AUTO_MOVEDOR. Essa escolha evidencia como o modelo seleciona quais aspectos salientar na descrição, direcionando, geralmente, o foco para a relação entre ação e objeto, preferência que também se manifestou em outras análises qualitativas realizadas ao longo da pesquisa, ainda que não descritas em detalhe neste capítulo.

Nesse sentido, comparando as duas versões, nota-se que ambas captam elementos centrais da cena, mas adotam perspectivas distintas. A versão sem anotações tende a organizar a descrição a partir de elementos visuais observáveis,

enquanto a versão com anotações evidencia relações dinâmicas entre personagens e ambiente, o que é indicativo de que a anotação baseada em eventos conduzida neste trabalho influenciou a geração do roteiro de audiodescrição, quando fornecida ao assistente de IA. Assim, cada descrição mobiliza um tipo diferente de foco: a primeira tende a se concentrar nos elementos individuais da cena (quem ou o que está presente, objetos, personagens e espaço) enquanto a segunda, motivada pelas anotações semânticas, enfatiza as relações entre esses elementos, destacando as ações, interações e a percepção do personagem em relação ao ambiente.

Embora com focos diferentes, em ambas as descrições é possível identificar aproximações com as diretrizes do *Guia para Produções Audiovisuais Acessíveis* (Naves et al., 2016). O Guia enfatiza, por exemplo, que as inserções descritivas devem respeitar os intervalos disponíveis no áudio, evitando sobreposições que comprometam a coerência da narrativa. Nesse sentido, ambas as versões seguem adequadamente o tempo de inserção, atendendo a um dos critérios estabelecidos nos *prompts* sobre adequação temporal. Além disso, as duas mencionam os quadros em destaque, elemento central do trecho analisado, mantendo coerência entre o conteúdo descrito e o ritmo do vídeo.

Além da adequação temporal, o Guia orienta que os ambientes devem ser claramente situados, indicando quando o personagem retorna a um espaço já visitado, informando alterações ocorridas e garantindo que o espectador compreenda a cena antes que a ação se desenrole. Nesse aspecto, observa-se uma diferença na forma como as versões analisadas inserem essas referências espaciais: a versão com anotações menciona a praça logo no início do trecho, posicionando o espectador no novo ambiente antes da transição visual, enquanto a versão sem anotações fornece essa localização apenas próximo ao final, quando Pedro já se desloca para fora da cena. Essa escolha de momento de inserção pode influenciar a experiência do público, afetando a forma como ele percebe e acompanha a cena.

Nas descrições geradas para a continuação do episódio, isso fica ainda mais evidente. No momento seguinte do vídeo, Pedro fala, de forma geral, sobre os preços em Cuba, gesticulando na frente de barracas de itens variados na praça, enquanto vendedores e pessoas circulam ao fundo. A Figura 30 ilustra esse momento, em que o apresentador está anotado como EF FALANTE, no *frame* `Dizer`.

Figura 30 — Quarta imagem do trecho 07-02-03 com anotações
Fonte: Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/1704>, último acesso em 9 nov. 2025.

No fim de sua fala, há um breve momento de silêncio, de menos de um segundo, seguido de uma cena de transição em que as camisetas de uma das barracas aparecem expostas. Na Figura 31, é possível visualizar este momento do vídeo com as anotações. Nessa imagem, os objetos marcados com *bounding boxes* foram anotados no *frame* `Comércio_vender`, no elemento de *frame* MERCADORIA, por expressarem aquilo que é oferecido à venda.

Figura 31 — Quinta imagem do trecho 07-02-03 com anotações
Fonte: Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/1704>, último acesso em 9 nov. 2025.

No total, considerando os dois trechos, o intervalo de silêncio disponível para a inserção da audiodescrição é de 3 segundos. As descrições geradas pelas versões sem anotações e com anotações para esse período são, respectivamente:

(13) Pedro gesticula na frente das barracas.

(14) Camisetas estampadas à venda.

Na versão sem anotações, a sentença (13) apresenta uma boa descrição, pois retoma a ação imediatamente anterior ao intervalo de silêncio (o gesto feito por Pedro). Além disso, ela localiza a cena ao mencionar as barracas, elemento que ainda não havia sido introduzido por essa versão, o que contribui para situar o espectador no espaço. No entanto, no momento em que a narração é inserida, o foco visual já recai sobre as camisetas expostas (vide Figura 31), que ocupam a maior parte desse intervalo de silêncio. Na versão com anotações, por sua vez, a descrição da sentença (14) coincide precisamente com o aparecimento desses objetos na tela, aproveitando o intervalo de silêncio disponível para a inserção da audiodescrição.

A partir da comparação entre as sentenças (13) e (14), observam-se diferenças na sincronicidade entre descrição e imagem. O roteiro com metadados, por gerar descrições baseadas em anotações multimodais associadas a segmentos temporais precisos do vídeo, alinha a narrativa aos elementos que surgem na tela, evidenciando maior correspondência entre descrição e foco visual. Já o roteiro sem anotações descreve a ação de forma mais geral, indicando a localização espacial apenas mais tardiamente; dessa forma, cumpre a função de situar o espectador no ambiente, mas sem acompanhar diretamente os elementos apresentados naquele instante.

Nesse caso, as camisetas aparecem justamente no intervalo disponível para a inserção da audiodescrição e, por estarem anotadas no *frame* `Comércio_vender`, foram destacadas na versão com anotações. Essa correspondência mais precisa entre descrição e elementos visuais, que aparece neste momento e na menção das pessoas andando na praça, contribuiu, então, para que a versão com anotações apresentasse uma nota expressivamente maior na similaridade de cosseno em relação à versão sem anotações, uma vez que essa métrica avalia a sobreposição entre conteúdo descrito e elementos do vídeo.

Esses resultados evidenciam como as informações estruturadas oriundas da anotação de *frames* permitem ao modelo direcionar o foco descritivo para os

elementos mais salientes no momento disponível para inserção da audiodescrição, favorecendo o alinhamento entre descrição e imagem. Dessa forma, a escolha do que e quando descrever impacta diretamente a percepção do público, influenciando a forma como ele acompanha a cena e interpreta as interações entre personagens e ambiente. Essa estratégia de temporização e destaque de elementos visuais está, de modo geral, alinhada a princípios de clareza e compreensão espacial recomendados para audiodescrição no Guia.

Na próxima subseção, será analisado o desempenho das versões em um trecho no qual a versão sem anotações apresentou resultados superiores à versão com anotações.

6.2.2 Melhor desempenho da versão sem anotações

O trecho 01-03-03 tem a duração de 1 minuto e 16 segundos e foi o segundo segmento selecionado para análise, por ter apresentado o melhor desempenho da versão sem anotações entre os dois episódios, conforme o cálculo de similaridade de cosseno (0,69 em relação ao vídeo, ante 0,67 da versão com anotações). Nessa parte do episódio, o apresentador Pedro caminha por um mercado, visita diferentes lojas e comenta sobre a importância do turismo para o Egito. A Figura 32 apresenta a sequência do segmento analisado sem as anotações visuais.

Figura 32 — Sequência de imagens do trecho 01-03-03 sem anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/2018>, último acesso em 9 nov. 2025.

No início do trecho, Pedro aparece caminhando pelo corredor do mercado, e a primeira descrição pela versão sem anotações e pela versão com anotações estão expressas, respectivamente, nas sentenças (15) e (16).

(15) O apresentador caminha nas ruelas do mercado, passando por lojas de artesanato e tecidos.

(16) Pedro caminha nas ruelas do mercado, entre as lojas de acessórios.

Nesse trecho, o apresentador está anotado como o EF AUTO_MOVEDOR no *frame* de `Auto_movimento`, uma vez que está andando pelas ruas. O local por onde ele anda também está anotado neste *frame*, no EF ÁREA. A Figura 33 mostra o segmento com as anotações, que orientam a formulação da descrição expressa na sentença (16).

Figura 33 — Primeiro segmento do trecho 01-03-03 com anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/2018>, último acesso em 9 nov. 2025.

As sentenças (15) e (16) reconhecem o movimento de Pedro pelo espaço e descrevem adequadamente o cenário, já que os produtos mencionados (artesanato, tecidos e acessórios) de fato aparecem nas lojas do mercado. Considerando que o tempo disponível para inserção de audiodescrição nesse trecho é de apenas 2 segundos, qualquer uma das versões tenderia a soar um pouco apressada na leitura. Ainda assim, a sentença (16) apresenta uma extensão menor, o que favorece sua fluidez. Por outro lado, a sentença (15) oferece maior especificidade ao detalhar os tipos de produtos visíveis, ainda que isso resulte em uma descrição mais longa.

Segundo o *Guia para Produções Audiovisuais Acessíveis* (Naves et al., 2016), é importante detalhar o ambiente sempre que possível, especialmente quando há tempo hábil para isso. Nesse caso, entende-se que o intervalo de silêncio em que a AD seria inserida é um pouco curto, fazendo com que parte da descrição da sentença (15) se sobrepusesse à narração do episódio, o que não é o ideal de acordo com as diretrizes que compõem o Guia.

Outro ponto relevante nas descrições geradas é a diferença na forma de referir-se ao protagonista: a versão sem anotações recorre com maior frequência ao termo “o apresentador” ao longo dos episódios, enquanto a versão com anotações utiliza diretamente “Pedro”. Essa diferença não decorre do conteúdo das anotações, que não incluem nomes próprios, mas parece estar relacionada ao modo como elas são temporalmente marcadas. Como cada elemento anotado aparece associado a um intervalo específico do vídeo, o modelo tende a ajustar a descrição para que caiba exatamente na janela temporal disponível para sua leitura.

Nesse contexto, o uso de “Pedro”, por ser uma forma mais curta, reduz a extensão verbal e facilita o encaixe da descrição no tempo marcado pelas anotações. Assim, a referência à função desempenhada por Pedro funciona bem na descrição, mas resulta em uma formulação mais longa; já o uso do nome próprio confere maior concisão e se ajusta melhor à restrição de tempo. De acordo com as diretrizes do Guia, a escolha pelo nome em vez da função só é adequada quando o participante já foi apresentado na narrativa. Nesse caso, isso não representa um problema, pois Pedro já foi introduzido anteriormente, no início do episódio.

A próxima imagem da sequência faz parte de um segmento com vários recortes de vídeo nos quais são mostrados os diferentes espaços do mercado, como lojas, cafés e corredores. Ambas as versões geradas pelo modelo de IA são descrições que resumem, de certa forma, o que aparece na tela, como pode ser visto nas sentenças (17) – sem anotações e (18) – com anotações.

(17) Pessoas sentadas em cafés na rua. Placa "Fechado para oração" pendurada.

(18) Homens jogam gamão em café, perto de uma mesquita.

A partir do contraste entre as sentenças, é possível observar que as duas descrições divergem na forma como enquadram a cena. No trecho analisado, ilustrado pela Figura 34, os participantes visíveis estão anotados como EF AGENTE no *frame* `Atividade`, enquanto o tabuleiro visível na cena está anotado como EF ATIVIDADE nesse mesmo *frame*, já que representa o objeto central da ação realizada. Essa estruturação orienta diretamente a descrição gerada pela versão com acesso aos metadados, concentrando-a na ação realizada pelos participantes no vídeo, como mostra a sentença (17).

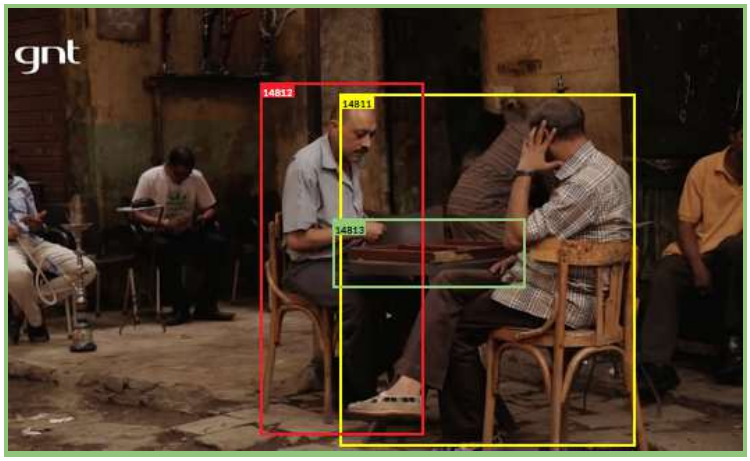


Figura 33 — Segundo segmento do trecho 01-03-03 com anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/2018>, último acesso em 9 nov. 2025.

Por outro lado, a versão sem anotações, ao não dispor dessas indicações, limita-se a enumerar o que aparece na tela — assim como na sentença (11) — e descreve o estado das entidades visíveis por meio de uma predicação estativa, isto é, uma sentença que apresenta apenas uma situação estática, não uma ação. Na sentença (17), por exemplo, o foco da descrição recai sobre o estado em que as pessoas se encontram (estarem sentadas) e a configuração espacial (em cafés, na rua), sem explicitar qualquer relação de ação ou engajamento em uma atividade. Trata-se, portanto, de uma caracterização descritiva e estática, que não organiza os participantes em papéis dentro de um evento.

Além disso, as descrições (17) e (18) também se distinguem na maneira de caracterizar os participantes e o espaço em que a cena ocorre. Sobre isso, as duas versões atendem às diretrizes do Guia que recomendam explicitar quem está em cena e onde a ação se desenvolve sempre que essas informações forem relevantes para a compreensão do contexto. Nesse sentido, ambas fazem um bom uso do trecho de silêncio disponível, aproveitando a pausa sonora para fornecer informações essenciais de identificação e ambientação.

Todavia, elas divergem na forma como caracterizam esses elementos. A versão sem anotações opta por mencionar apenas “pessoas”, enquanto a versão com anotações especifica que se trata de homens, alinhando-se às anotações realizadas, que também identificam os participantes da cena (vide Figura 34) como o EF PESSOA, no *frame* `Pessoas`, e os associam à UL *homem.n* no CV Name.

Ademais, a versão sem anotações caracteriza o local como uma rua, enquanto a versão com metadados situa a ação próximo a uma mesquita, provavelmente em razão da cena imediatamente anterior, que mostra um homem entrando no edifício religioso, anotado como EF PRÉDIO, no *frame* `Prédios`, e associado à UL *templo.n* no CV Name. Assim, cada versão apresenta o cenário na audiodescrição a partir de referências distintas.

Há também uma diferença quanto ao tipo de informação considerada pertinente. Sobre isso, a versão sem anotações antecipa a presença de uma placa de “fechado para oração”, elemento que só aparece posteriormente no vídeo. A versão com anotações, por outro lado, concentra-se exclusivamente no evento em curso, porque as anotações disponíveis naquele trecho, já mencionadas, coincidem com o ponto de silêncio e destacam justamente a ação dos participantes visíveis naquele momento. Observa-se, dessa forma, uma tendência do modelo a priorizar os elementos anotados e efetivamente presentes na cena quando eles coincidem com o momento de silêncio, em vez de mencionar informações que ainda não foram exibidas na imagem.

Na imagem seguinte da sequência, aparece a placa que indica que a loja está fechada para oração, exatamente como havia sido antecipado pela versão sem anotações. A sentença gerada para esse novo trecho de silêncio foi:

(19) Clientes compram em uma loja escura.
(20) Close em loja de decoração e artesanato.

A sentença sem anotações (19) inventa a presença de clientes, embora a loja esteja vazia, provavelmente como forma de dar continuidade ao que já havia sido mencionado anteriormente sobre o ambiente comercial, mas sem respaldo visual naquele momento. Já a sentença com anotações (20) descreve apenas o close da loja, mas não retoma a placa, embora ela seja claramente o foco da imagem. Nesse caso, a ausência da placa parece resultar do fato de que não havia anotações associadas a esse elemento no trecho correspondente, apenas à loja, o que faz com que o modelo deixe de mencioná-lo, mesmo sendo perceptível e relevante no contexto. A Figura 35 apresenta as anotações para este momento do vídeo.

Figura 35 — Terceiro segmento do trecho 01-03-03 com anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/2018>, último acesso em 9 nov. 2025.

É provável que seja justamente nesse ponto da sequência que a versão sem anotações tenha alcançado a pontuação superior na similaridade de cosseno. Mesmo que a sentença (19) seja inventada e não corresponda ao que aparece no vídeo naquele momento, ela ainda é computada na comparação, porque, em outro momento do mesmo trecho, ainda que distante, há de fato pessoas comprando mercadorias na rua. Como a métrica de similaridade avalia o segmento como um todo, essa menção deslocada acaba reforçando artificialmente a correspondência entre descrição e vídeo, contribuindo para o desempenho superior da versão sem anotações. Somado a isso, o uso recorrente do termo “apresentador” por essa versão — o qual evoca o *frame* `Pessoas_por_vocação`, frequentemente mobilizado para descrever os vendedores ao longo do trecho — também pode ter contribuído para aumentar a convergência entre o roteiro sem anotações e o conteúdo visual considerado.

Ao avançar para o par de sentenças seguinte, observa-se que as duas versões produziram descrições bastante próximas. A sentença (21) corresponde à geração sem anotações, enquanto a (22) àquela com anotações.

(21) O apresentador gesticula, com expressão séria.
(22) Pedro gesticula com a mão.

No caso do segundo método, as anotações presentes nesse segmento de vídeo identificaram o apresentador como EF AUTO_MOVEDOR no *frame* de `Auto_movimento` e como EF FALANTE no *frame* `Dizer`, como mostra a Figura 36.

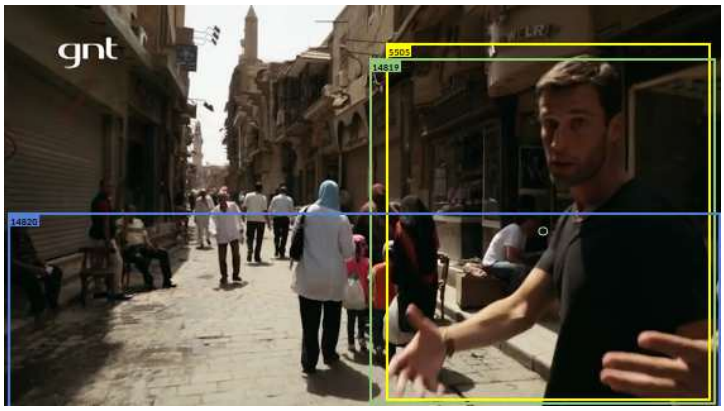


Figura 36 — Quarto segmento do trecho 01-03-03 com anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/2018>, último acesso em 9 nov. 2025.

Embora haja pequenas diferenças, como na forma de nomear o participante da cena, as sentenças (21) e (22) descrevem a mesma ação central: o gesto de Pedro enquanto fala. A versão sem anotações, contudo, inclui um traço interpretativo (“com expressão séria”) que não está marcado nas anotações visuais, mas aparece como inferência contextual. A versão com anotações, ao contrário, descreve apenas o movimento observado, de forma mais concisa. Nesse caso, embora não houvesse anotações para o gesto realizado por Pedro[18], ambas as versões recorrem à percepção visual, captando adequadamente o gesto da mão do participante, que se mostra bastante expressivo.

De maneira geral, a versão sem anotações tende a complementar a descrição com informações interpretativas sobre expressões ou sentimentos dos participantes da cena, enquanto a versão com anotações mantém-se mais próxima do que está objetivamente registrado, limitando-se ao que pode ser observado de forma direta na cena. Sobre isso, o *Guia para Produções Audiovisuais Acessíveis* (Naves et al., 2016) recomenda que a inclusão de emoções na AD seja feita em conformidade com os construtos universais de emoção. No entanto, o texto não fornece orientações detalhadas sobre como isso deve ser operacionalizado.

Logo em seguida, ainda dentro do mesmo intervalo de silêncio, a geração com anotações aproveita o espaço disponível para registrar exatamente o que está visível na imagem naquele momento e que havia sido anotado. Como mostra a

[18] Pesquisas recentes no âmbito da FrameNet Brasil investigam a representação de gestos como elementos semânticos que contribuem para a compreensão multimodal (Abreu, Torrent e Matos, 2025). Essas análises ainda não estavam disponíveis para os episódios analisados, de modo que os gestos não foram formalmente anotados.

Figura 37, as anotações identificam o participante da cena, que é delimitado como EF INGESTOR, no *frame* `Ingerir_substância`, e o narguilé, que, por sua vez, é anotado como EF INTOXICANTE no *frame* `Intoxicantes` e associado à UL *narguilé.n*.

Figura 37 — Quinto segmento do trecho 01-03-03 com anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/2018>, último acesso em 9 nov. 2025.

Essas marcações direcionam o modelo para esse evento, resultando na sentença (23). Como a ação do participante estava anotada e havia tempo suficiente, a inserção da sentença se encaixa naturalmente no intervalo, sem comprometer o ritmo da audiodescrição.

(23) Homem fuma narguilé.

A versão sem anotações também produz uma descrição nesse mesmo intervalo, mas com um comportamento distinto: ela antecipa a entrada de um vendedor que aparece apenas nesta transição específica, embora isso não seja estritamente o que se vê naquele instante. Essa antecipação, por sua vez, funciona como preparação para a ação presente no intervalo de silêncio seguinte, quando Pedro e o vendedor apertam as mãos.

(24) Um homem se aproxima do apresentador.
(25) Eles apertam as mãos.

É importante notar que nenhum dos dois personagens — o homem com o narguilé ou o vendedor — tem papel narrativo relevante; tratam-se apenas de

elementos pontuais das cenas, que se sucedem rapidamente. A versão com anotações não descreve o aperto de mãos — anotado no *frame* `Cumprimentar`, com ambos os participantes marcados como EF COMUNICADORES —, provavelmente devido à janela de silêncio reduzida, insuficiente para acomodar uma sentença mais longa. Já a versão sem anotações, por não ter mencionado o homem com o narguilé, mantém espaço textual para introduzir o vendedor e registrar o cumprimento, criando uma transição mais contínua entre as ações, ainda que nem todas sejam estritamente necessárias para a compreensão total do episódio.

Em seguida, ambas as versões geram descrições próximas. As duas indicam o movimento de Pedro, que está anotado como EF AUTO_MOVEDOR no *frame* de `Auto_movimento`, e destacam os itens do mercado. A Figura 38 mostra as *bounding boxes* anotadas para esse momento do vídeo.

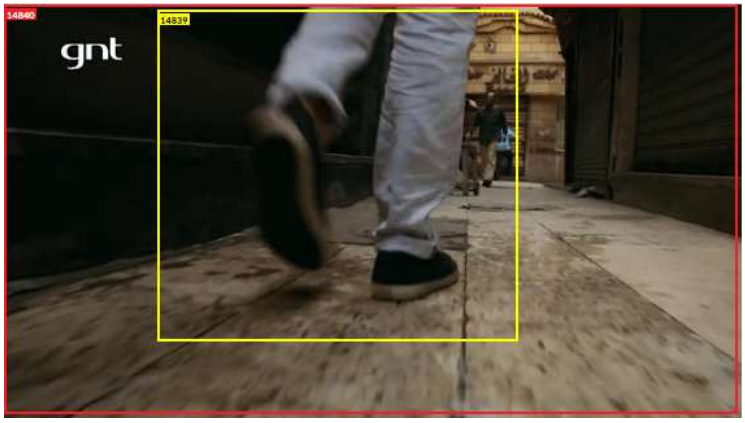


Figura 38 — Sexto segmento do trecho 01-03-03 com anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/2018>, último acesso em 9 nov. 2025.

A sentença (26) corresponde à descrição gerada sem anotações para esse segmento do episódio, enquanto a sentença (27) foi feita a partir do *prompt* que inclui as anotações.

(26) Apresentador volta a caminhar. Close em especiarias, tapetes e lenços coloridos do mercado.
(27) Pedro caminha. Close na venda de especiarias e lenços no mercado.

Nesse caso, na segunda parte da sentença (26), a versão sem anotações introduz novas entidades do cenário, ampliando a listagem dos elementos visuais

presentes no mercado. Já a versão com metadados, em contraste, introduz um evento: a palavra *"*venda*"* evoca o *frame* de transação comercial e reorganiza a cena a partir dessa ação. Assim, enquanto a primeira mantém, mais uma vez, uma perspectiva estática (enumerando objetos e características do espaço), a segunda assume como foco o evento, destacando a dinâmica da cena e não apenas os itens que a compõem.

Para além disso, a versão sem anotações é ligeiramente mais longa, pois especifica tudo aquilo que foi mostrado no mercado. Já a versão com anotações é mais concisa, concentrando-se no que é central da cena sem retomar todos os detalhes do espaço. Assim, o conteúdo central é o mesmo nas duas descrições mas a versão sem anotações elabora um pouco mais o cenário. Como há tempo suficiente no intervalo de silêncio, a expansão feita por ela enriquece a descrição do espaço sem comprometer o ritmo da audiodescrição, o que é sugerido pelo Guia.

No último conjunto de descrições do trecho, ambas as versões mencionam a ação central de Pedro examinando uma mercadoria no mercado. Na sentença (28), a versão sem anotações se mantém mais geral, referindo-se a "um objeto de metal", enquanto, na sentença (29), a versão com anotações especifica o item como "um moedor", seguindo a anotação do moedor como EF MERCADORIA no *frame* `Comércio_vender` e associado à UL *moedor.n* no CV Name, como apresenta a Figura 39. Neste momento, o apresentador também é identificado como EF COMPRADOR no *frame* `Comércio_vender`, enquanto o vendedor é anotado como EF VENDEDOR no mesmo *frame*.

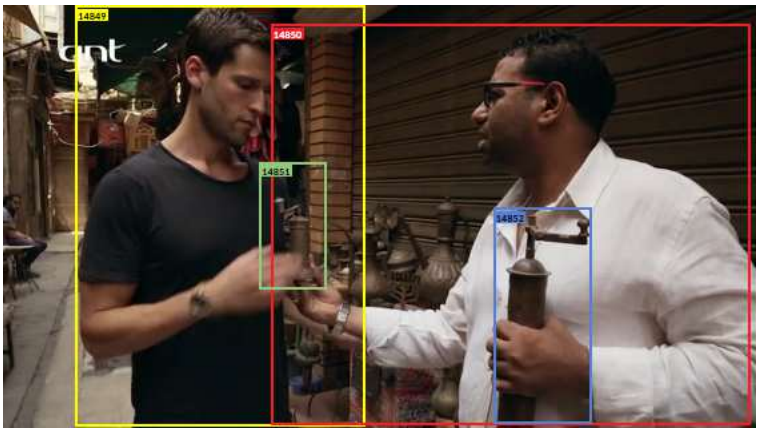


Figura 39 — Sétimo segmento do trecho 01-03-03 com anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/2018>, último acesso em 9 nov. 2025.

(28) O apresentador examina um objeto de metal.

(29) Pedro examina um moedor.

Além disso, a versão com anotações insere uma segunda descrição, a sentença (30), em um momento distinto da primeira, aproveitando outra janela de silêncio para registrar a movimentação de Pedro para longe do vendedor no fim do trecho. Nesse ponto, evidencia-se o ganho trazido pelas anotações semânticas com o uso do termo “vendedor” na descrição, que enquadra o participante dentro de uma interação comercial, ativando o *frame* `Comércio_vender` e organizando a cena como um evento de compra e venda — algo que a versão sem anotações não recupera.

(30) Pedro se afasta do vendedor.

Dessa forma, neste trecho, o conteúdo central é semelhante nas descrições de ambas as versões, mas a versão com anotações adiciona informações mais específicas e distribui a ação em duas sentenças, enquanto a versão sem anotações apresenta uma formulação única e mais genérica. Nenhuma das versões identifica o vendedor antes de Pedro se afastar, provavelmente devido à limitação de tempo disponível no intervalo de silêncio, que restringe a inserção de novas informações. A especificação do item como “moedor” na versão com anotações exemplifica o ganho proporcionado pelo uso do CV Name, permitindo que o modelo indique de forma precisa os elementos presentes na cena e enriqueça a descrição sem comprometer o ritmo da audiodescrição. De forma similar, o uso de “vendedor” na última descrição gerada pela versão com metadados evidencia como as anotações do *frame* `Comércio_vender` orientam o modelo a enquadrar corretamente o participante na função que desempenha no evento, organizando a cena como uma interação comercial e não apenas como a presença de mais uma pessoa no espaço.

Em síntese, neste trecho do episódio, observa-se uma diferença na sincronização das informações entre os os roteiros: o roteiro gerado a partir do *prompt* com anotações descreve predominantemente apenas os eventos visíveis em cada janela temporal, seguindo, na maioria das vezes, estritamente o que foi

registrado, enquanto o sem anotações inclui eventos ainda não exibidos, por não estar limitado ao conteúdo visual imediato.. Isso fez, por exemplo, com que a versão mencionasse clientes comprando — algo que só aparece em outra parte do trecho —, o que contribuiu para sua maior similaridade com o conteúdo visual considerado no cálculo de similaridade. Esse tipo de comportamento dá ao roteiro sem anotações uma certa flexibilidade para recuperar informações relevantes, mas também pode prejudicar a experiência do espectador ao inserir dados que não correspondem ao momento exibido no vídeo, criando um descompasso entre o que é narrado na AD e a sequência narrativa.

Além disso, a versão sem anotações, ao priorizar predicações estativas, tende a não enquadrar os elementos descritos conforme os papéis que desempenham em cada cena eventiva, o que faz com que parte da dinâmica da cena deixe de ser representada, especialmente quando comparadas às versões geradas com o apoio das anotações.

Ademais, a especificação de elementos da cena, como o moedor e vendedor, evidencia como as anotações podem orientar o modelo a produzir descrições mais precisas, mesmo em intervalos de silêncio curtos, respeitando o ritmo da audiodescrição. Na próxima subseção, será analisado um caso de empate entre as versões, permitindo observar como essas diferenças se manifestam em situações de similaridade mais equilibrada.

6.2.3 Empate entre as versões

O terceiro trecho analisado, identificado como 07-10-03, pertence ao Episódio 7 da série e tem duração de 1 minuto e 31 segundos. Diferentemente do anterior, este segmento apresentou valores idênticos de similaridade entre as versões com e sem anotação (0,40), configurando um caso de empate. Entre os trechos analisados para esta pesquisa, os empates nos valores de similaridade ocorreram, em sua maioria, nas aberturas ou nos encerramentos dos episódios, quando as descrições tendem a ser mais breves e estruturadas de forma semelhante. O trecho 07-10-03, entretanto, constitui uma exceção: trata-se do único caso de empate localizado no interior da narrativa e, sendo assim, foi selecionado justamente por esse motivo, a fim de se compreender por que as descrições geradas se mostraram tão semelhantes.

Nesse momento da narrativa, o apresentador entrevista Alicia Hechavarría, uma atriz cubana. A Figura 40 mostra a sequência dos principais momentos do trecho.

Figura 40 —Sequência de imagens do trecho 07-10-03 sem anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/1704>, último acesso em 9 nov. 2025.

Embora o final da cena seja descrito de modo semelhante por ambas as versões, o início apresenta diferenças relevantes. Parte dessa divergência decorre da própria escolha de momento para inserção da AD feita em cada caso. O roteiro baseado nas anotações, por exemplo, antecipa uma informação ao afirmar que Pedro e Alicia se cumprimentam antes do início da narração — ação que, no vídeo, ocorre apenas posteriormente, enquanto ainda há diálogo entre os personagens. A sentença (31) ilustra essa antecipação:

(31) Pedro e Alicia se cumprimentam.

A Figura 41 apresenta o momento exato em que o cumprimento realmente ocorre, bem como as *bounding boxes* que delimitam esse trecho no vídeo.

Figura 41 — Primeiro momento do trecho 07-10-03 com anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/1704>, último acesso em 9 nov. 2025.

Para este segmento do episódio, Pedro e Alicia estão anotados no *frame* `Cumprimento`, que reúne palavras, expressões e gestos associados a uma interação social, tanto quando ela se resume a uma saudação rápida quanto quando marca apenas o começo de uma interação mais extensa. Assim, Pedro e Alicia correspondem ao EF COMUNICADORES, que delimita aqueles que participam da interação comunicativa. Nesse sentido, a descrição gerada pelo modelo coincide com o conteúdo do *frame* anotado, o que sugere que as anotações influenciaram a seleção da ação verbalizada naquele ponto do roteiro.

A versão sem anotações, por sua vez, inseriu a descrição expressa na sentença (32) no momento em que os dois personagens sobem as escadas em direção ao local da entrevista (vide Figura 39). No entanto, descreve que Pedro gesticula em um instante em que ele permanece com as mãos paradas, uma delas segurando um copo, o que evidencia um descompasso entre a descrição verbal e a ação efetivamente observada na imagem.

(32) Pedro gesticula

No geral, o segmento em análise se mostra um trecho complexo para a inserção de descrições, uma vez que a fala do apresentador ocupa quase todo o tempo até o início da entrevista. Nesse contexto, as duas versões adotam estratégias diferentes para preencher os breves intervalos disponíveis.

O roteiro com anotações antecipa a ação de cumprimento, o que pode não ser o mais adequado do ponto de vista da sincronização e faz com que parte da descrição se sobreponha à música de transição entre cenas. Ainda assim, essa antecipação contribui para situar previamente os personagens e sinalizar o início da interação entre eles. O roteiro sem anotações, por sua vez, insere a descrição em um intervalo sem fala ou música, o que favorece a inserção da AD. A versão aproveita, portanto, uma pausa que não foi utilizada pela versão com anotações. No entanto, descreve que Pedro gesticula em um instante em que isso não acontece, cometendo um erro de correspondência visual ao atribuir uma ação inexistente ao apresentador.

De acordo com o *Guia para Produções Audiovisuais Acessíveis* (Naves et al., 2016), é importante que as descrições situem o espectador diante de mudanças de ambiente e da entrada de novos personagens em cena, de modo a garantir a compreensão espacial e narrativa. O Guia também recomenda, na seção destinada às questões tradutórias, que a AD seja sincronizada com os intervalos de silêncio e com as ações visuais, evitando sobreposição com trilhas sonoras ou falas, exceto quando a informação visual se mostrar de grande importância para a compreensão do trecho.

Sob essa perspectiva, o roteiro com anotações apresenta um ganho informativo por indicar os participantes da interação, mencionando a presença de Alicia na sequência seguinte. Embora a sobreposição com a música de transição não seja recomendada, entende-se que a informação acrescentada pelo modelo tem pertinência narrativa e contribui para situar o público quanto ao desenvolvimento do enredo, em consonância com as diretrizes do Guia.

Por outro lado, infelizmente, nenhuma das versões aproveita de maneira efetiva o momento de silêncio enquanto os participantes sobem as escadas. Por se tratar de um intervalo sem diálogo ou efeitos sonoros, esse trecho seria especialmente adequado para a inserção da AD, já que elimina o risco de sobreposição sonora. Ele poderia ter sido usado, por exemplo, para situar melhor o espectador e marcar com mais precisão as transições da cena, tal como indicado na seção do Guia sobre questões linguísticas.

Acerca disso, no caso do roteiro com anotações, vale destacar que havia informações adicionais que poderiam ter sido aproveitadas. Como mostra a Figura 42, ambos os participantes da cena, por estarem em movimento, estavam anotados

no EF AUTO_MOVEDOR, no *frame* `Auto_movimento`, assim como nos EFs FALANTE e DESTINATÁRIO do *frame* `Questionar`, por se tratar de uma entrevista.

Figura 42 — Segundo momento do trecho 07-10-03 com anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/1704>, último acesso em 9 nov. 2025.

Ainda assim, a versão baseada nas anotações não acionou esses elementos nesse momento do episódio, optando por manter o foco apenas no cumprimento no início do trecho. Nesse caso, como a ação dos participantes marcava uma mudança de ambiente, a utilização dessas anotações teria permitido uma descrição mais alinhada à progressão visual da cena. Isso reforçaria a noção de deslocamento dos personagens e proporcionaria ao público da AD uma percepção mais completa da transição de cenário e da continuidade narrativa entre os planos.

A partir desse ponto do vídeo, ambas as versões convergem na formulação das sentenças seguintes, descrevendo a cena de forma idêntica. Isso parece ocorrer porque, nos segmentos seguintes, as anotações não introduzem informações adicionais; assim, mesmo apenas um dos roteiros fazendo uso delas, tanto os acertos quanto as limitações se repetem nas duas versões.

Neste momento, Pedro continua a entrevista com Alicia e há poucas mudanças no enquadramento, o que faz com que, em todo o tempo, apenas os dois façam parte da cena. A Figura 43, com anotações, mostra o segundo intervalo de silêncio, propício à inserção de audiodescrição, do trecho 07-10-03. No segmento, Alicia segue anotada como DESTINATÁRIO no *frame* `Questionar`, e Pedro também permanece como o EF FALANTE neste mesmo *frame*.

Figura 43 — Terceiro momento do trecho 07-10-03 com anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/1704>, último acesso em 9 nov. 2025.

Nesse instante, a sentença (33) é narrada nos dois roteiros. A descrição se mostra adequada, pois coincide com o instante da Figura 32, em que a atriz realmente gesticula ao comentar sobre o medo de os Estados Unidos dominarem Cuba.

(33) Alicia gesticula.

Por outro lado, há também momentos em que ambas as versões revelam uma dificuldade comum: captar nuances temporais e sonoras fundamentais para a naturalidade da AD. No trecho em questão, Pedro e Alicia brindam, ação marcada não apenas pelo gesto, mas pelo som característico do tilintar das taças — elemento que funciona como um sinal claro do momento exato em que a ação ocorre. A Figura 44 apresenta esse momento, no qual Alicia permanece anotada como DESTINATÁRIO no *frame* `Questionar` e Pedro como o EF FALANTE neste mesmo *frame*.

Figura 44 — Quarto momento do trecho 07-10-03 com anotações
Fonte: FrameNet Brasil WebTool, disponível em: <https://webtool.frame.net.br/annotation/dynamicMode/1704>, último acesso em 9 nov. 2025.

Para esse momento do vídeo, a descrição gerada pelas duas versões está expressa na sentença (34) .

(34) Pedro e Alicia brindam.

Em ambos os casos, contudo, a descrição é antecipada, ocorrendo muito antes do som característico do brinde, o que compromete a correspondência temporal entre o áudio e a imagem. Nesse caso, ainda que não houvesse um intervalo de silêncio disponível exatamente no momento do brinde — o que impediria a sincronização ideal sem sobreposição —, a antecipação excessiva causa certo estranhamento no espectador. Isso se deve ao fato de que o som das taças funciona como um marcador auditivo evidente da ação, e sua inversão temporal tende a quebrar a naturalidade da sequência.

Segundo o *Guia para Produções Audiovisuais Acessíveis* (Naves et al., 2016), é preciso referenciar a fonte sonora, identificando a origem do som no momento em que ele ocorre. Assim, embora a AD não pudesse ser inserida exatamente no instante do brinde, caso se optasse por descrevê-lo, uma descrição mais próxima do evento sonoro preservaria melhor a coerência entre o que se ouve e o que se narra.

Essa observação ressalta, inclusive, o papel do audiodescritor como revisor especializado. Em materiais gerados automaticamente, nuances de temporalidade, como a articulação entre ações visuais e seus marcadores sonoros, podem passar despercebidas pelo sistema. Nesse sentido, a perspectiva de um profissional qualificado, o qual compreende as diretrizes e reconhece nuances da narrativa audiovisual, sobre o que foi gerado automaticamente permite corrigir antecipações desnecessárias e outras sutilezas na narrativa, evitar rupturas na fluidez e assegurar que a AD mantenha precisão e coerência com a experiência auditiva do público com deficiência visual.

No mais, entende-se que o trecho 07-10-03 é uma parte do vídeo relativamente estática, com poucos deslocamentos ou mudanças visuais significativas. As sentenças geradas por ambos os métodos (com e sem anotação)

refletem essa estabilidade da cena, concentrando-se em ações pontuais e recorrentes, como cumprimentar, gesticular e brindar. Devido ao tempo reduzido disponível para inserção da audiodescrição e à pouca variação visual, as descrições se limitam a registrar as ações mais evidentes, resultando em enunciados curtos e semelhantes, o que explica o equilíbrio nos valores de similaridade obtidos.

A diferença entre as versões aparece sobretudo no início do trecho: enquanto a versão com anotações tende a seguir o conteúdo marcado nos *frames*, a versão sem anotações tenta inferir um movimento do apresentador a partir das imagens, ainda que essa inferência nem sempre corresponda de fato ao que ocorre na cena.

Assim, o empate na similaridade entre as versões decorre menos do uso (ou não) das anotações multimodais e mais da natureza estática e previsível da interação filmada, que restringe o espaço para variações descritivas significativas.

7 CONSIDERAÇÕES FINAIS

Esta pesquisa teve como objetivo investigar como as anotações de *frames* influenciam a qualidade e o alinhamento das audiodescrições geradas por IA, partindo da hipótese de que a anotação multimodal de eventos, ao mapear semanticamente os eventos na produção audiovisual, fornece ao modelo uma estrutura capaz de orientar descrições mais alinhadas à dinâmica narrativa.

Para atender aos objetivos específicos do estudo, descritos na introdução desta pesquisa, foram realizadas anotações multimodais das cenas dos episódios 1 e 7 da série *Pedro pelo Mundo*, da GNT, à luz da Semântica de Frames e do modelo da FrameNet Brasil. Com base nessas anotações, foram gerados, com o auxílio do copiloto de inteligência artificial Gemini, dois roteiros de audiodescrição para cada episódio: um orientado pelas anotações semânticas para *frames* e outro produzido sem esse suporte. A comparação entre esses conjuntos, considerando precisão, coerência e adequação à dinâmica narrativa audiovisual, permitiu avaliar em que medida a estruturação semântica de eventos contribui para reduzir lacunas interpretativas e favorecer a produção de descrições mais contextualizadas.

Nesse sentido, os resultados quantitativos mostram que as anotações de *frames* atuam como um modulador do desempenho do modelo: quando a versão anotada apresenta desempenho inferior, as diferenças negativas tendem a ser pequenas; já quando apresenta desempenho superior, os ganhos são mais expressivos. Além disso, observou-se que a versão com metadados alcança níveis mais altos de alinhamento com o conteúdo visual com maior frequência e, mesmo quando não traz melhorias imediatas, não compromete a qualidade das descrições.

Os resultados qualitativos, por sua vez, revelam que as anotações também influenciam o tipo de descrição produzido pela IA. Na condição sem anotações, observa-se uma tendência a gerar predicações estativas, baseadas sobretudo na enumeração dos elementos visuais mais salientes dos episódios. Nesses casos, o texto assume um caráter mais descritivo e, em certos momentos, faz com que a dinâmica da cena seja pouco explorada, o que aparece em diferentes trechos analisados.

Já na condição com metadados, o modelo tende a produzir descrições orientadas pelos eventos registrados. As anotações funcionam como um eixo organizador: ao indicar quais eventos estão em curso, quem participa deles e como

esses participantes se relacionam, elas direcionam a IA a representar a cena como uma sequência de ações situadas, em vez de apenas como um conjunto de elementos visíveis. Nota-se, também, que o modelo especifica com mais precisão os itens presentes na cena, valendo-se das informações do CV Name e dos EFs para identificar objetos e participantes de forma mais detalhada.

Nesse contexto, os resultados indicam que o uso da Semântica de Frames e de anotações multimodais constitui um apoio relevante à produção de roteiros de audiodescrição por IA. A abordagem permite analisar como a organização semântica dos eventos influencia a coerência, a contextualização e o alinhamento das descrições com a dinâmica narrativa audiovisual, ao mesmo tempo em que abre espaço para refletir sobre as implicações da integração de sistemas de IA nessa prática.

Acerca disso, vale destacar que, embora Mayer (2016) defina a audiodescrição como uma interação entre videntes e não videntes, a incorporação de sistemas de inteligência artificial na produção de roteiros acrescenta um novo elemento a essa relação. Nesse sentido, a prática da AD continua a preservar seu caráter relacional e mediado, mas a IA passa a atuar como copiloto, acelerando a produção. Nesse contexto, o audiodescritor humano assume funções de supervisão e revisão especializada, ajustando escolhas lexicais, aprimorando a naturalidade das sentenças e garantindo coerência e fluidez nos roteiros, como evidenciado nos exemplos analisados ao longo deste estudo. Essa configuração evidencia a necessidade de repensar a noção tradicional de audiodescrição como interação, considerando cenários colaborativos que envolvem sujeitos humanos e sistemas automáticos.

Além disso, a produção de roteiros assistida por IA também aponta para a necessidade de se repensar o conceito de qualidade na audiodescrição. Diferentemente do que muitas vezes se supõe, a elaboração humana de roteiros não ocorre em condições ideais: grande parte é produzida para atender a exigências legais ou processos de judicialização, nem sempre permitindo tempo ou recursos para refinamento detalhado. A incorporação da IA como copiloto oferece a possibilidade de acelerar a produção e reduzir custos, ao mesmo tempo em que cria oportunidades para sistematizar critérios de qualidade, garantindo coerência e adequação narrativa. Dessa forma, o uso de IA evidencia a importância de investigar formas de aprimorar continuamente a audiodescrição, tanto em roteiros

existentes quanto em processos assistidos, promovendo práticas mais consistentes e reflexivas na área.

Como desdobramentos, a metodologia aqui empregada pode ser aplicada em outros gêneros audiovisuais, permitindo identificar padrões de desempenho da ferramenta em materiais com diferentes estruturas narrativas e demandas de audiodescrição. Ao mesmo tempo, esse movimento favorece a sistematização de parâmetros qualitativos de avaliação dos roteiros de audiodescrição gerados por IA, especialmente quando tal avaliação incorpora a participação de pessoas com deficiência visual no processo. Nesse caso, essa incorporação desloca o foco da análise de critérios exclusivamente técnicos para a consideração da experiência efetiva de recepção do roteiro, fortalecendo a discussão sobre as potencialidades e os limites do uso de sistemas automáticos na produção de audiodescrição.

Espera-se, por fim, que investigações futuras possam aprofundar e expandir os achados aqui apresentados, contribuindo para o aperfeiçoamento de sistemas automáticos de AD enquanto apoio ao audiodescritor, para o fortalecimento de pesquisas na área e, a longo prazo, para a otimização dos processos de produção de roteiros de AD, ampliando o alcance de conteúdos culturais por pessoas com deficiência visual.

REFERÊNCIAS

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

BATEMAN, J.; WILDFEUER, J.; HIIPPALA, T. **Multimodality: Foundations, research and analysis–A problem-oriented introduction**. Berlin: Walter de Gruyter GmbH & Co KG, 2017.

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

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

BELCAVELLO, F. **FrameNet Annotation for Multimodal Corpora: devising a methodology for the semantic representation of text-image interactions in audiovisual productions**. 135f. Tese (Doutorado em Linguística) — Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora, 2023.

CAMPOS, V. **Sistema de Geração Automática de Audiodescrição a Partir de Análise de Conteúdo de Vídeo**. 93f. Tese (Doutorado em Engenharia de Computação) — Faculdade de Engenharia Elétrica e de Computação, Universidade Federal do Rio Grande do Norte, Natal, 2019.

CHOMSKY, N. **Aspects of the Theory of Syntax**. Cambridge, MA: MIT Press, 1965.

DÁNNELLS, D.; TORRENT, T. T.; SIGILIANO, N. S.; DOBNIK, S. Beyond Strings of Characters: Resources meet NLP – Again. In: VOLODINA, E.; DÁNNELLS, D.; BERDICEVSKIS, A.; FORSBERG, M.; VIRK, S. (Org.). ***Live and Learn:*** *Festschrift in honor of Lars Borin*. Gothenburg: Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs Universitet, 2022.

DORNELAS, L. **A audiodescrição sob a perspectiva da semântica de frames: análise dos frames evocados pelo texto da audiodescrição e pelas imagens dinâmicas num curta-metragem**. 142f. Dissertação (Mestrado em Linguística) — Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2023.

DÍAZ CINTAS, J. Audiovisual Translation Today. A question of accessibility for all. **Translating Today**, v. 4, p. 3-5, 2005.

FRANCO, E.; ARAÚJO, V. Questões terminológico-conceituais no campo da tradução audiovisual (TAV). In: **Tradução em Revista**, n.11, 2011.

FERRARI, L. C. **Introdução à Linguística Cognitiva**. 1. ed. São Paulo: Contexto, 2011.

FILLMORE, C. J. Innocence: a second idealization for linguistics. In: ANNUAL MEETING OF THE BERKELEY LINGUISTICS SOCIETY, 1979, Berkeley, EUA. Anais... [S.l.: s.n.], 1979. v.5, p.63-76.

FILLMORE, C. J. Frame semantics. In.: **Linguistics in the morning calm.** Korea: Hanshin Publishing Company, 1982.

FILLMORE, C. J. Frames and the semantics of understanding. **Quaderni di Semantica**, v.5, n.2, p.222-254, 1985.

FILLMORE, C. J.; ATKINS, B. T. Toward a Frame-Based Lexicon: The Semantics of RISK and Its Neighbors. In A. Lehrer & E. F. Kittay (Eds.), **Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization**. Hillsdale, NJ: Lawrence Erlbaum Associates, 1992, p.75-102.

FILLMORE, C. J.; BAKER, C. A Frames Approach To Semantic Analysis. In: HEINE, B.; NARROG, H. (Orgs.). **The Oxford Handbook Of Linguistic Analysis.** (p.313–340). Oxford: Oxford University Press, 2009.

GUALBERTO, C. L.; SANTOS, Z. B. dos. Multimodalidade no contexto brasileiro: um estado de arte. **DELTA: Documentação E Estudos Em Linguística Teórica e Aplicada**, 35(2), 2019.

GUALBERTO, C. L.; SANTOS, Z. B. Multimodalidade e Hipertextualidade: Caminhos para pesquisa e ensino. **PERcursos Linguísticos**, v. 11, n. 29, p. 32–49, 2021. Disponível em: https://periodicos.ufes.br/percursos/article/view/36781. Acesso em: 24 mar. 2025.

HODGE, R.; KRESS, G. **Social Semiotics**. London: Polity Press, 1988.

JAKOBSON, R. On Linguistic Aspects of Translation. In: BROWER, R. A. (Ed.) **On Translation**. Cambridge, MA: Harvard University Press, 1959. p. 232-239.

LEMKE, J. **Multiplying meaning: Visual and verbal semiotics in scientific text**. In: J.R. Martin & R. Veel (Eds.), Reading Science. London: Routledge, 1998.

LUZ, A.; BRAZ, G.; RUIZ, L.; PINTO, M. C.; BELCAVELLO, F.; SIGILIANO, N. S.; TORRENT, T. Anotação do Dataset Multimodal da ReINVenTA. In: **Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana** (STIL), 14. Porto Alegre: Sociedade Brasileira de Computação, 2023, p. 352-356.

NAVES, S.; MAUCH, C.; ALVES, S.; ARAÚJO, V. L. S. (org.). **Guia para Produções Audiovisuais Acessíveis**. Brasília: Secretaria do Audiovisual do Ministério da Cultura, 2016.

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

ROMERO FRESCO, P. Accessible filmmaking: Joining the dots between audiovisual translation, accessibility and filmmaking. In: **The Journal of Specialised Translation**. 2013, v. 20. p.201-223.

RUPPENHOFER, J.; ELLSWORTH, M.; PETRUCK, M. R. L.; JOHNSON, C. R.; SCHEFFCZYK, J. **FrameNet II**: Extended theory and practice. Berkeley, International Computer Science Institute, 2016. Disponível em: http://framenet.icsi.berkeley.edu/. Acesso em: 23 fev. 2025.

SALOMÃO, M. FrameNet Brasil: um trabalho em progresso. **Calidoscópio**, 7(3), 2009.

SALOMÃO, M. Entrevista com Margarida Salomão. **Revista Investigações**, Recife, p.193-203, 2010.

SOUZA, D.; PAGANO, A.; GAMONAL, M. A audiodescrição sob a perspectiva da Semântica de Frames: um estudo exploratório. **Revista Gatilho**, Juiz de Fora, v. 23, p. 101-125, 2022.

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

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

VIRIDIANO, M. **Framed Multi30K: Um dataset multimodal-multilíngue baseado em semântica de frames**. 109f. Tese (Doutorado em Linguística) — Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora, 2024.

APÊNDICE I

Frames evocados na versão sem anotações — Episódio 1

ID Frame	Nome do Frame	Contagem
524	Pessoas_por_vocação	37
183	Vias	24
252	Alimentos_e_bebidas	24
278	Pessoas	22
157	Prédios	21
58	Auto_movimento	17
390	Artefato	17
30	Gesto	16
398	Veículo	13
60	Percepção_ativa	10
171	Locais_naturais	10
147	Expressão_facial	9
1375	Partes_do_corpo	9
337	Negócios	8
490	Peça_arquitetônica	8
571	Alvo	8
611	Substâncias	8
956	Causar_perceber	8
1184	Animais	8
3	Movimento	7
176	Locais_por_uso	7
196	Números_cardinais	7

492	Pessoas_por_idade	7
13	Postura	6
206	Unidade_calêndrica	6
237	Abundante_em	6
238	Ingestão	6
275	Rito	6
462	Fazer_cara	6
11	Movimento_corporal	5
491	Idade	5
621	Ingredientes	5
1414	Móveis	5
72	Recipientes	4
96	Localização_de_luz	4
123	Vestuário	4
131	Parte_orientacional	4
169	Subpartes_de_prédios	4
216	Transportar	4
226	Entidade	4
272	Texto	4
479	Conversar	4
486	Atributos_mensuráveis	4
623	Alcance	4
1253	Serviços_em_alimentação	4
49	Causar_movimento	3
148	Atividade_em_andamento	3

175	Locais_políticos	3
179	Relação_locativa	3
392	Cor	3
1427	Eletroeletrônicos	3
12	Manipulação	2
48	Chegar	2
51	Partir	2
56	Colocar	2
102	Causar_dano	2
154	Atividade_preparar	2
156	Comércio_vender	2
178	Agir_intencionalmente	2
193	Direção	2
233	Acompanhamento	2
234	Construir	2
374	Artes_performáticas	2
379	Pegar	2
505	Fazedores_de_barulho	2
756	Obras_de_arte_físicas	2
840	Colocação_temporal	2
849	Fazer_turismo	2
957	Representação	2
1000	Tamanho	2
1120	Representante	2
1187	Plantas	2

1190	Proximidade_não_graduável	2
1423	Serviço_em_alimentação	2
40	Emoção_direcionada	1
42	Emoção_com_foco_no_experienciador	1
57	Remover	1
59	Operar_veículo	1
63	Fazer_barulho	1
70	Posição_distribuída	1
75	Tempo_relativo	1
93	Relações_pessoais	1

Frames evocados na versão com anotações — Episódio 1

ID Frame	Nome do Frame	Contagem
278	Pessoas	36
252	Alimentos_e_bebidas	24
183	Vias	17
58	Auto_movimento	16
147	Expressão_facial	16
157	Prédios	14
171	Locais_naturais	14
524	Pessoas_por_vocação	12
1375	Partes_do_corpo	12
60	Percepção_ativa	11
238	Ingestão	11

3	Movimento	9
390	Artefato	8
398	Veículo	8
1184	Animais	8
30	Gesto	7
49	Causar_movimento	7
169	Subpartes_de_prédios	7
179	Relação_locativa	7
272	Texto	7
337	Negócios	7
956	Causar_perceber	6
206	Unidade_calêndrica	5
535	Origem	5
611	Substâncias	5
1253	Serviços_em_alimentação	5
1292	Meios_de_transporte	5
11	Movimento_corporal	4
72	Recipientes	4
176	Locais_por_uso	4
178	Agir_intencionalmente	4
216	Transportar	4
621	Ingredientes	4
757	Criar_representação	4
13	Postura	3
96	Localização_de_luz	3

175	Locais_políticos	3
275	Rito	3
356	Ingerir_substância	3
507	Pessoas_por_religião	3
623	Alcance	3
756	Obras_de_arte_físicas	3
846	Local_animado	3
1140	Nível_de_luz	3
1414	Móveis	3
1423	Serviço_em_alimentação	3
1427	Eletroeletrônicos	3
51	Partir	2
63	Fazer_barulho	2
172	Local	2
196	Números_cardinais	2
233	Acompanhamento	2
243	Criação_culinária	2
374	Artes_performáticas	2
379	Pegar	2
490	Peça_arquitetônica	2
505	Fazedores_de_barulho	2
571	Alvo	2
610	Atravessar	2
748	Deslocamento_intencional	2
825	Tornar-se_separado	2

1190	Proximidade_não_graduável	2
1231	Transporte	2
1404	Sofrer_transformação	2
12	Manipulação	1
19	Diferenciar	1
40	Emoção_direcionada	1
42	Emoção_com_foco_no_experienciador	1
57	Remover	1
70	Posição_distribuída	1
74	Frequência	1
93	Relações_pessoais	1
114	Evento_social	1
123	Vestuário	1
131	Parte_orientacional	1
146	Trajar	1
148	Atividade_em_andamento	1
237	Abundante_em	1
254	Ler	1
258	Fazer_fogo	1
276	Entidade_física	1

Frames evocados na versão sem anotações — Episódio 7

ID Frame	Nome do Frame	Contagem
756	Obras_de_arte_físicas	14
58	Auto_movimento	11

157	Prédios	9
175	Locais_políticos	9
183	Vias	9
491	Idade	9
30	Gesto	8
535	Origem	8
60	Percepção_ativa	7
278	Pessoas	7
374	Artes_performáticas	7
398	Veículo	7
94	Agregado	6
524	Pessoas_por_vocação	6
611	Substâncias	6
956	Causar_perceber	6
272	Texto	5
1375	Partes_do_corpo	5
169	Subpartes_de_prédios	4
171	Locais_naturais	4
176	Locais_por_uso	4
392	Cor	4
618	Campos	4
1190	Proximidade_não_graduável	4
123	Vestuário	3
359	Intoxicantes	3
490	Peça_arquitetônica	3

101	Cenário_da_educação	2
131	Parte_orientacional	2
178	Agir_intencionalmente	2
337	Negócios	2
382	Intérpretes_e_papéis	2
466	Causar_fazer_barulho	2
479	Conversar	2
486	Atributos_mensuráveis	2
957	Representação	2
1000	Tamanho	2
1029	Subpartes_de_artefato	2
1187	Plantas	2
1187	Plantas	2
3	Movimento	1
12	Manipulação	1
23	Julgar	1
48	Chegar	1
56	Colocar	1
67	Liderança	1
71	Causar_expansão	1
162	Comércio_pagar	1
179	Relação_locativa	1
196	Números_cardinais	1
237	Abundante_em	1
238	Ingestão	1

252	Alimentos_e_bebidas	1
277	Clima	1
295	Empregar	1
300	Formas	1
350	Costume	1
377	Fabricação	1
379	Pegar	1
390	Artefato	1
470	Sons	1
493	Pessoas_por_origem	1
505	Fazedores_de_barulho	1
621	Ingredientes	1
623	Alcance	1
716	Familiaridade	1
748	Deslocamento_intencional	1
791	Dinheiro	1
846	Local_animado	1
921	Confrontar_problema	1
996	Impressão	1
999	Coletivo_em_eventos_sociais	1
1046	Modo_de_viver	1
1063	Revolução	1
1120	Representante	1
1140	Nível_de_luz	1
1253	Serviços_em_alimentação	1

1270	Dançar	1
1423	Serviço_em_alimentação	1

***Frames* evocados na versão com anotações — Episódio 7**

ID Frame	Nome do Frame	Contagem
58	Auto_movimento	15
60	Percepção_ativa	12
157	Prédios	11
278	Pessoas	9
398	Veículo	9
171	Locais_naturais	8
183	Vias	8
374	Artes_performáticas	8
491	Idade	8
756	Obras_de_arte_físicas	8
272	Texto	7
169	Subpartes_de_prédios	6
175	Locais_políticos	6
176	Locais_por_uso	6
618	Campos	6
123	Vestuário	5
1190	Proximidade_não_graduável	5
30	Gesto	4
94	Agregado	4

390	Artefato	4
524	Pessoas_por_vocação	4
178	Agir_intencionalmente	3
234	Construir	3
252	Alimentos_e_bebidas	3
359	Intoxicantes	3
535	Origem	3
748	Deslocamento_intencional	3
791	Dinheiro	3
172	Local	2
179	Relação_locativa	2
356	Ingerir_substância	2
492	Pessoas_por_idade	2
623	Alcance	2
956	Causar_perceber	2
996	Impressão	2
1029	Subpartes_de_artefato	2
1187	Plantas	2
1253	Serviços_em_alimentação	2
1375	Partes_do_corpo	2
1414	Móveis	2
1423	Serviço_em_alimentação	2
3	Movimento	1
11	Movimento_corporal	1
13	Postura	1

48	Chegar	1
49	Causar_movimento	1
51	Partir	1
63	Fazer_barulho	1
76	Cenário_do_comércio	1
93	Relações_pessoais	1
95	Parentesco	1
101	Cenário_da_educação	1
113	Atividade	1
114	Evento_social	1
146	Trajar	1
156	Comércio_vender	1
238	Ingestão	1
255	Criar_intencionalmente	1
288	Vir_a_existir	1
295	Empregar	1
326	Desejabilidade	1
338	Desejar	1
354	Sofrer_ferimento_corporal	1
382	Intérpretes_e_papéis	1
384	Danificar	1
392	Cor	1
454	Reunir-se	1
462	Fazer_cara	1
466	Causar_fazer_barulho	1

483	Ser_localizado	1
486	Atributos_mensuráveis	1
490	Peça_arquitetônica	1
515	Padrão_temporal	1
757	Criar_representação	1
828	Abertura	1
846	Local_animado	1
858	Locais_por_evento	1
957	Representação	1
1063	Revolução	1
1140	Nível_de_luz	1
1197	Colocação_espacial	1
1441	Instrumentos_musicais	1