Audition: A Frame-Annotated Multimodal Dataset for Accessible Audiovisual Content

Maucha Andrade Gamonal[1], **Tiago Timponi Torrent**[1,2], **Ely Edison Matos**[1],
Adriana S. Pagano[2,3], Frederico Belcavello[1], Flávia Affonso Mayer[3,4],
Arthur Lorenzi[1], Natália S. Sigiliano[1], Helen de Andrade Abreu[1],
Lívia Vicente Dutra[1,5], Marcelo Viridiano[1] , André Coneglian[3],
Victor A. S. Herbst[1] , Franciany O. Campos[1],
Kenneth Brown[1], Lívia Pádua Ruiz[1], Lisandra Carvalho Bonoto[1],
Luiz Fernando Pereira[1] and Yulla Liquer Navarro[1]
[1]*FrameNet Brasil, Federal University of Juiz de Fora*
[2]*Brazilian National Council for Scientific and Technological Development – CNPq*
[3]*Observatory for Language and Inclusion, Federal University of Minas Gerais*
[4]*Federal University of Paraíba*
[5]*Department of Multilingualism, Gothenburg University*
maucha.andrade@visitante.ufjf.br; tiago.torrent@ufjf.br

Abstract

This paper presents a multimodal semantic analysis of accessible Brazilian short films using a frame-based annotation approach. We introduce a subset of the *Audition* dataset, comprising six short films from the animation and documentary genres. We analysed three communicative modes: original audio, audio description, and visual content. Trained annotators semantically annotated each mode following the FrameNet Brazil multimodal methodology. To compare meaning across modalities, we used cosine similarity over frame-semantic representations. Results show that audio description aligns more closely with video content than original audio, reflecting its role in translating visual meaning into language. Our findings demonstrate the effectiveness of frame semantics in modelling meaning across modalities and provide quantitative evidence of audio description as a bridge between visual and verbal communication. The dataset and annotation strategies are a valuable resource for research on multimodal representation, semantic similarity, and accessible media.

1 Introduction

Large Language Models (LLMs) have significantly propelled research in Natural Language Processing (NLP). However, these models are still predominantly trained on textual data, while human communication is inherently multimodal, construing meaning through spoken language, sound, gestures, and visual content (Li et al., 2022; Cánovas et al., 2020; Radford et al., 2021). Modelling multimodality is essential not only for advancing NLP in general, but also for developing accessible technologies.

A critical domain for multimodal understanding is production of accessible audiovisual content (Ma et al., 2024; Ye et al., 2024; Lee et al., 2024; Hendricks et al., 2017; Han et al., 2023). Creating inclusive media requires systems capable of interpreting and generating meaning across modalities — particularly between audio and visual content — to support users with visual impairment.

This paper reports an experiment on cross-modal semantic similarity using *Audition*, a frame-annotated multimodal dataset of accessible Brazilian Portuguese short films. We compute similarity across three communicative modes — original audio, audio description and video content — using a hybrid metric that combines frame-based spread activation and cosine similarity.

Our work aims to advance semantically grounded methods for multimodal alignment, contributing both to assistive technologies and to multimodal NLP.

2 Background

2.1 Theoretical foundation: Frame Semantics and FrameNet Brazil

Frame Semantics is a theory of meaning representation developed by (Fillmore, 1982), which posits that understanding linguistic expressions requires access to schematic representations of situations, known as *frames*. Each frame models a conceptual scene that involves participants, entities and relevant objects, referred to as frame elements (FEs).

Words evoke frames and their meanings are interpreted in relation to these structured conceptualizations. A central tenet is that meaning emerges from

the relation between lexical items and the frames they evoke, as well as from contextual factors such as world knowledge and the communicative situation (Fillmore, 1985). For example, in the sentence "*He put the keys in the drawer*," the verb *put* evokes the `Placing`[1] frame, which presupposes the presence of frame elements such as AGENT (*He*),THEME (*the keys*), and GOAL (*the drawer*).

Frame Semantics has been computationally implemented through FrameNet, a large lexical database of English based on the annotation of frames, lexical units and their syntactic and semantic valences (Baker et al., 1998). Lexical Units (LU) are words or expressions that evoke frames. The annotation process begins with a LU linked to a frame annotated in real linguistic contexts and captures semantic information through frame elements, and their syntactic realizations and grammatical functions.

FrameNet includes a set of frame-to-frame relations that ensure conceptual interconnection — such as *Inheritance*, *Subframe*, and *Perspective_on*. These relations demonstrate how lexical meaning is structured within an articulated conceptual system. For instance, the *Cause_motion* frame inherits from *Transitive_action*, decomposes into `Placing` and `Removing` as subevents, and serves as background for frames such as `Bringing`, `Excreting`, `Gathering_up`, and `Ingestion`.

FrameNet Brazil (henceforth FN-BR) builds on this structure, adapting and expanding it for Brazilian Portuguese (Torrent and Ellsworth, 2013). FN-BR introduces additional dimensions to the FrameNet architecture: (i) *frame elements-to-frame* relations, which connect frame elements to other frames and enable recursive modelling of complex scenes; (ii) a mechanism for *metonymy modelling* to represent semantic shifts in which a frame elements evokes a related entity; and (iii) *ternary qualia relations* inspired by the Generative Lexicon (Pustejovsky, 1995), which formalize inferential links among concepts. These extensions significantly expand the model's semantic representational capacity (Torrent et al., 2022).

Additionally, FN-BR extends frame-based modelling to include **multimodal semantic representation**. Drawing on the premise that not only verbal expressions, but also images, gestures, and visual scenes may evoke frames, frame activation operates across multiple communicative modalities, with their elements instantiated by linguistic expressions, visual cues, or bodily signals, thereby enhancing the interpretative scope of the semantic network. This is central to the *ReINVenTA* network described in the following section.

2.2 ReINVenTA and related datasets

Building on the theoretical and architectural advances of FN-BR, ReINVenTA — Research and Innovation Network for Vision and Text Analysis — integrates NLP and accessibility research to advance computational semantic modelling of multimodal meaning. Grounded in Frame Semantics, the network develops gold-standard annotated[2] datasets and methods for multimodal meaning representation.

Two major datasets have already been developed: Frame^2 (Belcavello et al., 2024) and Framed Multi30k (Viridiano et al., 2024).

Frame^2 is a frame-annotated multimodal dataset based on 230 minutes of a Brazilian travel television program. It includes 11,796 semantic annotations for transcribed speech and subtitles, and 6,841 semantic annotations for video segments. Through bounding boxes, each video annotation is associated with a frame and one or more frame elements, linked to a lexical unit.

A recent study based on Frame^2 (Samagaio et al., 2024) investigates the notion of semantic permanence in interlingual subtitling. By comparing frame-semantic annotations in original audio transcriptions and their translated subtitles, it demonstrates how cosine similarity can detect semantic shifts due to translation strategies, as well as to the temporal and spatial constraints in the subtitling process. Ongoing work with Frame^2 includes the development of a multimodal model for turn organization in conversation in audiovisual discourse (Abreu and Matos, 2025).

Framed Multi30k expands the widely used Multi30k dataset (Elliott et al., 2016) with 158,915 image captions in Brazilian Portuguese — both originally created in Portuguese and translated into this language — and more than 4.5 million frame and frame element annotations for English and Portuguese. It also enriches the Flickr30k Entities dataset (Plummer et al., 2015) by aligning phrase-

[1] Following established conventions, frame names are set in `Courier`, and FEs in SMALL CAPS.

[2] For details on the semantic annotation process see sections 3.2 and 4.

to-region correlations with semantic frames.

Building on Framed Multi30k, a new multimodal corpus is currently being constructed for the journalistic domain. Based on image-text pairs extracted from online news portals, it is designed to be automatically processed by NLP and computer vision systems to identify visual entities, events, and semantic relations in real-world discourse-situated contexts.

2.3 Motivation for Audition

Accessible Audiovisual Translation encompasses a set of practices, such as audio description, subtitling, and closed captioning, designed to make audiovisual content accessible to audiences with hearing and visual impairment. This is an inherently multimodal task, requiring the translation of meaning across spoken and visual modalities in an accurate and accessible manner, which poses a significant challenge to current NLP systems, primarily trained on textual data with little regard for accessibility.

The Audition dataset was created to address this gap, by providing a semantically annotated multimodal resource covering original audio, audio description, and video content from accessible Brazilian Portuguese short films. *Audition*'s frame-based structure supports the development and evaluation of NLP models for assistive technologies grounded in multimodal semantic understanding.

3 Design of the Audition dataset

3.1 Corpus composition and communicative modes covered

Audition[3] is a multimodal dataset comprising Brazilian short films with accessibility features spanning a variety of cinematic genres[4], including animation, fiction, autobiography, performance and documentary. Its full version, currently under finalization, comprises more than 240 minutes (over four hours) of audiovisual material.

The dataset includes semantic annotation across verbal and non-verbal communicative modes: original audio (dialogue and narration), audio description, subtitles, closed captions, overlaid on-screen text, and video content. These modes are defined as follows:

- **Original audio (OA)**: spoken language of the film's original soundtrack, including dialogue and narration.
- **Audio description (AD)**: scripted additional narration verbalizing visual information such as actions, gestures, body language, emotional states, settings, and costumes.
- **Subtitles**: written translations of the original dialogue into another language, usually synchronized with spoken content and displayed on screen.
- **Closed captions (CC)**: on-screen textual representations of spoken language and relevant sound cues (e.g., music, sound effects).
- **Overlaid on-screen text (text-overlays)**: written language integrated into the visual composition of the film for artistic or informational purposes, such as titles, narrative inserts, or stylized captions.
- **Video content (VC)**: moving image stream segmented into meaningful visual scenes or shots conveying narrative, spatial, and emotional information.

This work focuses on a subset of the dataset, comprising 71 minutes of semantically annotated audiovisual content extracted from six Brazilian short films belonging to the animation and documentary genres.

Our analysis centres on three communicative modes: **Original Audio**, **Audio Description** and **Video Content**, selected because of their strong semantic interplay: each provides a distinct but interdependent representation of a given narrative sequence. While original audio and audio description operate in the verbal channel, dynamic visuals convey non-verbal semantic content. Their alignment enables a robust investigation of cross-modal semantic similarity and frame-level coherence.

3.2 Methodology and annotation tool

Semantic annotation in Audition follows FrameNet's full text methodology, complemented by the multimodal guidelines established by FN-BR. This framework ensures systematic labelling of frames, lexical units, frame elements, and visual entities in all communicative modalities of the dataset.

[3]The first version of the dataset is available at: https://huggingface.co/datasets/FrameNetBrasil/Audition.

[4]we classify animation as a cinematographic genre distinct from fiction based on(Gordeef, 2023)

We used Webtool [5], a web-based multimodal platform developed within FN-BR to support the integrated annotation of text, audio, images, gestures, and video. This tool allows annotators to work simultaneously with verbal and visual data, grounding their decisions in the conceptual structure defined by FN-BR.

Its interface supports: (i) synchronized visualization of aligned modalities (e.g., audio, audio description, and video segments); (ii) selection of lexical units or visual entities that evoke frames; (iii) assignment of frame elements to verbal spans or visual regions.

Annotation was performed by seven undergraduate junior researchers[6], who were trained on both the annotation tool and FN-BR multimodal guidelines. Expert linguists supervised the process, reviewed annotations when necessary, and solved ambiguous cases collaboratively. Tasks were assigned and completed within comparable timeframes, annotation time per instance not being measured.

We adopted a perspectivist approach to semantic annotation (Basile et al., 2021), which acknowledges that frame selection may vary depending on annotators' perspectives, background knowledge, and contextual interpretation.

4 Semantic Annotation across Communicative Modes

The semantic annotation process covered the three communicative modes under analysis: audio, audio description, and video. The dataset was segmented into aligned analytical units: sentences for verbal modalities and bounding boxes representing events, states and other coherent visual segments for the video. This alignment enables a cross-modal comparison of the semantic structures instantiated in each communicative mode.

Annotation comprised identifying the frames evoked in each modality and labelling their associated frame elements. For verbal content, both in the original audio and in the audio description, frames were assigned based on lexical units identified in the spoken language. For the visual modality, frames were annotated through delimitation of bounding boxes that mark and categorize the salient visual elements, including participants, actions, spatial configurations, and perceptually relevant events.

Whereas in verbal text annotation, frames are directly evoked by lexical units, in video annotation frames are instantiated based on their frame elements within the scene. In addition, visual entities are also annotated and linked to semantic frames, refining object recognition through alignment with conceptual structures.

Semantic annotation was performed at three levels: (i) identification of the evoked frame; (ii) labelling of the associated frame elements; (iii) linking to a lexical unit (in verbal modalities) or to a visual entity (in the video) responsible for frame evocation.

This structure was applied to segments from audio transcriptions and audio descriptions, as well as to video segments involving characters, actions, and objects. Although each modality was independently annotated, using a shared conceptual structure based on FN-BR enabled semantic alignment between communicative modes. Alignment was foundational for the cross-modal similarity analyses conducted in our experiment.

For video annotation, we used bounding boxes that identify scene entities associated with frame evocation. YOLOv3 was initially tested to assist this process, but its generic object detection categories were not adequate for the situational entities required in frame-based annotation. Bounding boxes had to be predominantly created manually by trained annotators, ensuring consistency with the semantic framework.

First, we collected and preprocessed the audiovisual materials, transcribing the audio and aligning transcriptions with the corresponding video segments. Next, we conducted semantic annotation, assigning frames, frame elements, lexical units, and visual entities to both textual and video content.

4.1 Audio annotation

Verbal content derived from audio transcriptions is annotated in two communicative modes - original audio (OA) and audio description (AD) - using the same method, *full text annotation* (Ruppenhofer et al., 2016), in which all semantic frames evoked within the scope of the sentence are identified and labelled.

Figure 1 illustrates audio semantic annotation

[5] Both the Webtool annotation software and the annotated data for FN-BR are available at: https://webtool.frame.net.br/. The FN-BR repository on GitHub can be accessed at: https://github.com/FrameNetBrasil.

[6] All student annotators were funded by Brazilian research agencies.

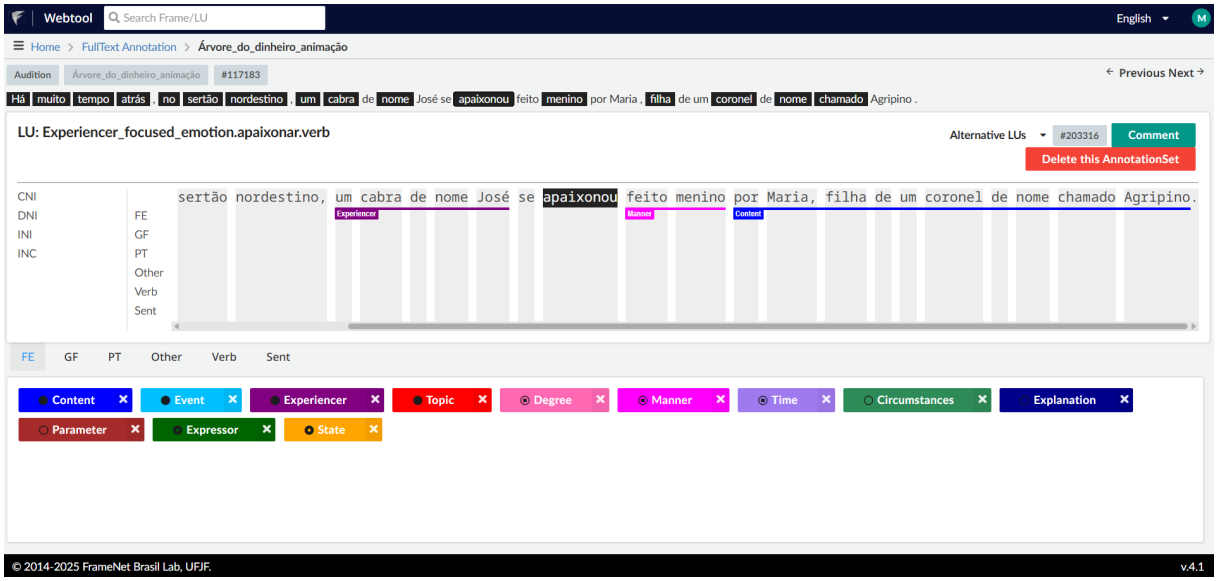


Figure 1: Multimodal annotation of audio: full text annotation

using the sentence[7]:

Há muito tempo atrás, no sertão nordestino, um cabra de nome José se apaixonou feito menino por Maria, filha de um coronel de nome chamado Agripino.

(*A long time ago, in the backlands of the Northeast, a guy named José fell in love like a boy with Maria, the daughter of a colonel named Agripino.*)

Several Lexical Units (LUs) were annotated in this sentence. Figure1̃ highlights the LU *apaixonar-se.v* (*fall in love.v*), which evokes the frame `Experiencer_focused_emotion`. The Frame Elements annotated in the sentence are: EXPERIENCER (*José*), MANNER (*feito menino*), and CONTENT (*por Maria, filha de um coronel de nome chamado Agripino*).

4.2 Video annotation

Video content (VC) is annotated through the creation of bounding boxes across the entire duration of the audiovisual material. These boxes are linked to the Frame Elements of semantically relevant frames. The procedure follows a *text-oriented* approach (Belcavello et al., 2024), semantic annotation being guided by transcribed audio. Our study annotated the visual representation of states, events, processes, and relations that may contribute to the audience cinematic experience.

Since the annotation task is text-oriented, annotators have access to the previous completed textual annotation as well as to the full video. Figure 2 shows an example of this annotation process.

The video sequence corresponds to the narration presented in Figure 1. The annotation consists in marking the visual entity that refers to the EXPERIENCER in the frame `Experiencer_focused_emotion`. In addition to this annotation, the visual entity classified as a Framed Entity is also marked and linked to the appropriate semantic frame, in this case, *homem.n* the frame `Person`.

Bounding boxes semantically anchor visual elements that contribute to narrative construction in both original audio and audio description. Depending on the situations perceived throughout the integration of audio, audio description, and video segments, the annotator can choose to duplicate the same bounding box and associate it with more than one frame by assigning different frame elements.

This is the case in Figure 3. The bounding box delimiting the person is annotated twice: first with the FE SUPPLIER in the frame `Service_client_supplier_interaction`, and second with the FE AGENT in the frame `Manipulation`, since the worker is holding a razor, the instrument used to perform the service. The audio description sequence[8] is as follows:

[7]Transcribed audio retrieved from the original audio of the short film *Árvore do Dinheiro* (Genre: Animation) Available at: https://cinematecapernambucana.com.br/filme/?id=2551. Access on: August 13, 2025.

[8]This transcribed audio is derived from audio description of the short film *Cinema Glória* (Genre: Documentary). Available at: https://cinematecapernambucana.com.br/filme/?id=3267. Acessed on: August 13, 2025.

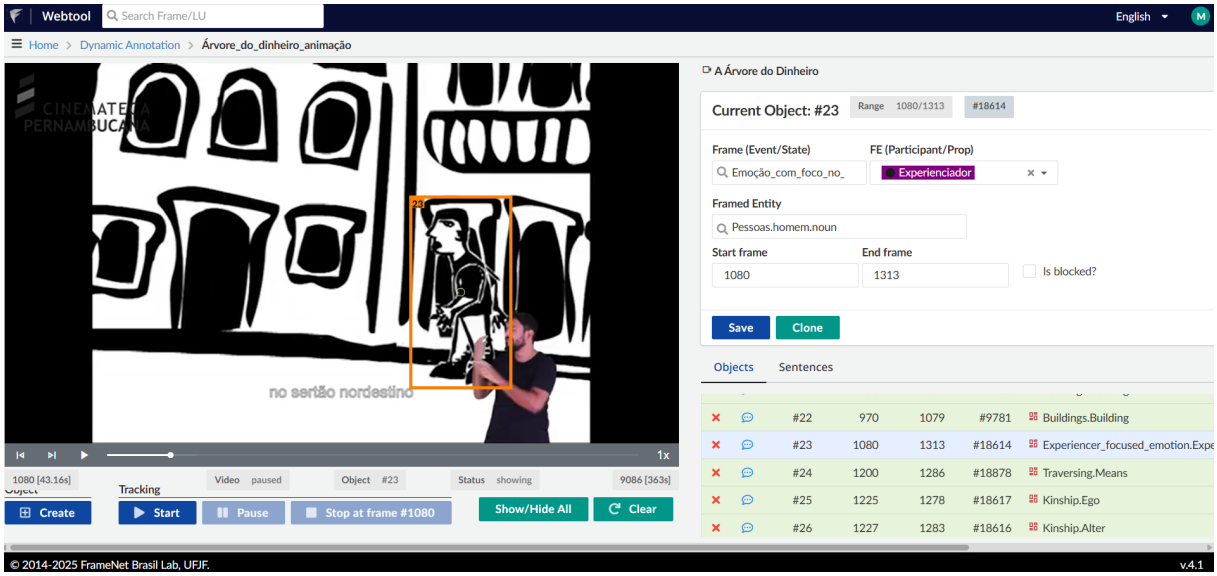


Figure 2: Multimodal annotation of video: bounding boxes anchored to frame element

Embaixo da sombra da árvore, um barbeiro atende um cliente./ O homem está com a cabeça recostada na cadeira de barbeiro./ Um babador branco sobre a camisa e o rosto com espuma./O barbeiro usa uma navalha.

(Under the shade of the tree, a barber attends to a client. / The man has his head on the barber's chair. / A white cape over his shirt and foam on his face. / The barber uses a razor.")

5 Dataset Metrics

5.1 Annotation Totals

Table 1 shows the total number of sentences, annotation sets, frame elements and semantic boxes used in this experiment, based on the selected subset of the Audition dataset, split by film genre. The dataset comprises 894 sentences in the verbal modality (464 from OA and 430 from AD), with 3,979 corresponding semantic annotation sets. Each **annotation set** includes a Lexical Unit (LU), an evoked frame, and the corresponding Frame Elements (FEs) identified in the sentence, totalling 8,189 FEs across both OA and AD. The documentary genre accounts for the majority of the data across all modalities, contributing over 69 percent of the total annotations[9]

The video modality includes 1,103 semantic boxes, representing visual entities anchored to the frame elements of the annotated frames. Each visual entity is associated both with a frame that aligns with the auditory narrative and with a generic entity type and its corresponding frame, supporting automatic visual entity recognition with semantic refinement.

	anim.	doc.	total
AO			
Sentences	142	322	464
Annotation sets	605	1597	2202
FEs	1172	3120	4292
AD			
Sentences	139	291	430
Annotation sets	501	1276	1777
FEs	1604	2293	3897
VC			
Semantic boxes	304	799	1103

Table 1: Data overview of animation, documentary, and total counts for AO, AD, and VC.

5.2 Semantic Similarity Across Modalities

The frame-based semantic similarity metric used in Viridiano et al. (2024) is particularly suitable for this study, as it quantifies semantic alignment across heterogeneous and multimodal data while being grounded in the cognitive principles of Frame Semantics. In this work, FrameNet categories are used to semantically annotate the communicative modes present in the dataset, namely, verbal language and visual representations. The metric evaluates how original audio, audio description and

[9]Dataset balancing will be addressed in future work, after all semantic annotations are completed. Issues related to genre diversity and modality distribution will be explored in subsequent studies, including the validation of the results reported here.

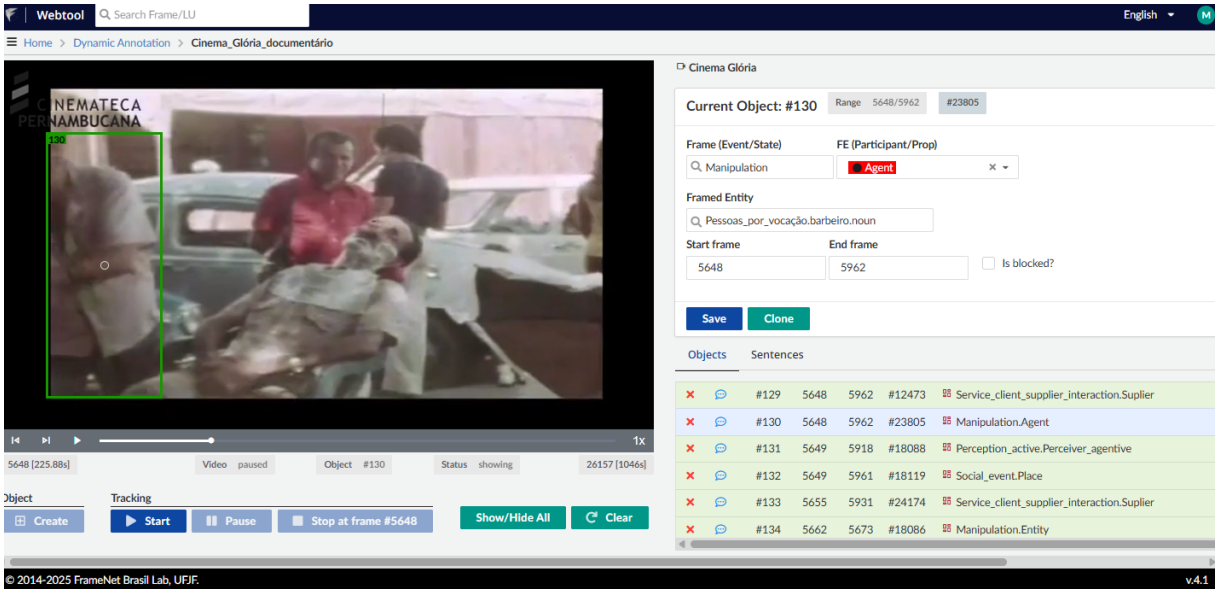


Figure 3: Multimodal annotation of video: duplicated bounding box anchored in more than one semantic frame

video content converge or diverge in terms of frame activation.

This method is crucial for multimodal analysis, where meaning is distributed between semiotic modes with distinct representational properties. The spread activation mechanism enables graded comparisons of semantic similarity, accounting not only for direct overlap but also for conceptual proximity mediated by FrameNet's relational structure. Specifically, the metric quantifies the degree of semantic adherence between video content and original audio and audio description. In this way, it provides an objective measure of how these modalities jointly contribute to filmic meaning construction.

The metric operates in three steps. First, we construct an **association matrix** that encodes frame-to-frame relations defined in FrameNet, including Inheritance, Subframe, Perspective_on, Causative_of, Inchoative_of, and Using. Each relation is weighted according to its semantic proximity, with stronger semantic links receiving higher weights (Gouws et al., 2010). Second, a **spread activation algorithm** propagates activation from the frames directly evoked in the annotations throughout the FrameNet graph.

Activation strength decays exponentially with graph distance, giving higher weights to semantically closer frames and lower weights to distant frames. Third, the spread activation process generates a **vector representation** for each annotated instance, where each vector dimension corresponds to a frame and its cumulative activation value. The semantic similarity between two instances is then computed as **cosine similarity** between their respective activation vectors, producing a normalized score between 0 (no similarity) and 1 (identical semantic content).

As described in 4, the verbal modes selected for comparison are annotated for frames and frame elements derived from the lexical units identified in the transcriptions. The visual mode, in turn, is annotated through bounding boxes, where each box is linked to one or more frame element, depending on the frame evoked. In addition, a framed entity is labelled to designate the entity delimited by the bounding box. This labelling adds semantic refinement to the model by associating a frame from the entity category with its corresponding lexical unit, as illustrated in Figure 2.

Therefore, besides comparing modes with one another, possible variation in the annotations can be compared as seen in Table 2. The comparison can target exclusively the lexical units in the annotations (AD LU / OA LU), which would indicate the extent to which the same entities are present in the verbal modes and in the video. It can also target the frames and FEs used in the annotations (AD Frame and FE / OA Frame and FE), which focuses on measuring the extent to which similar events, states, processes and relations are mentioned across modes. Finally, the comparison can target all annotations simultaneously (AD Full / OA Full).

Cosine similarities obtained for each comparison type are shown in Table 2, which indicates

number of pairs of semantic representations compared (pairs), average normalized cosine similarity (avg), variance (var) and standard deviation (stdev) obtained[10].

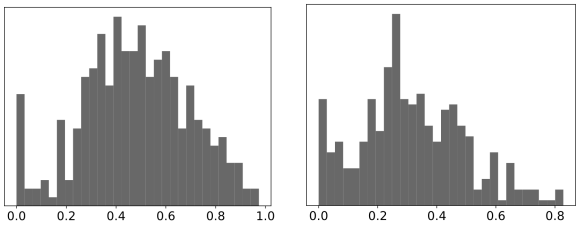
	Video Content			
	pairs	avg	var	stdev
AD LU	323	0.42	0.05	0.22
AD Frame and FE	323	0.38	0.05	0.22
AD Full	323	0.49	0.05	0.21
OA LU	357	0.23	0.03	0.17
OA Frame and FE	357	0.29	0.04	0.19
OA Full	357	0.32	0.03	0.18

Table 2: Cosine similarity between frame-based semantic representations of verbal language modes and video content.

From Table 2, we observe that comparisons including both events, states, processes, relations, and entities mentioned in verbal language with those shown in the video segments yield the highest average cosine similarities. This result is statistically significant across all comparisons, except for the comparison between OA Frame and Frame Element and OA Full (with test statistic $t(317828) = -1.82$, $p < 0.0067$).

These results indicate that audio description is more similar to visual content than to original audio. This result was expected, given that the purpose of audio descriptions is precisely that of providing some sort of access to the content shown in video through audio. By contrast, original audio comes from the film script, which, for most cases, assumes that the video content will be accessed by the person experiencing the film.

The distributions of the cosine similarity values shown in Figure 4 corroborate this claim.



(a) AD Full×VC (b) OA Full×VC

Figure 4: Distribution of similarity values between Audio Description Full, Original Audio Full and Visual Content.

6 Conclusion and Future Work

This paper introduced the first release of the *Audition* dataset, a multimodal corpus of accessible Brazilian short films annotated within a frame-based semantic approach. This initial subset established a methodological basis for cross-modal comparisons and provided quantitative evidence of how accessibility resources such as AD mediate between non-verbal and verbal communication.

We applied FN-BR's multimodal semantic annotation methodology, labelling Lexical Units, frames, frame elements, and visual entities. Cosine similarity analyses show that audio description aligns more closely with video content than with original audio, confirming its role as a mediating modality. These results point to the effectiveness of Frame Semantics for modelling meaning across communicative modes.

A central contribution of our work is the frame-based similarity metric for multimodal data. This methodological innovation has direct implications for accessibility studies, audio description evaluation, and semantic modelling. Its architecture draws on cognitive models of semantic activation spreading, maintaining theoretical alignment with Frame Semantics, which reinforces the interpretability and the conceptual consistency of our analyses.

Future developments will extend the analysis to the full *Audition* dataset, including additional film genres, as well as subtitles, closed captions, and overlaid on-screen text to investigate their roles in multimodal meaning construction. From an NLP perspective, we plan to train and evaluate models for the automatic identification of frames and frame elements across modalities, and explore applications such as the automatic generation and evaluation of audio descriptions grounded in visual content.

7 Acknowledgments

The development of the Audition dataset is one of the initiatives of ReINVenTA—Research and Innovation Network for Vision and Text Analy-

[10]The reference point for cosine similarity is the audio time span as determined by the annotated transcriptions. Thus, all visual entities annotated within the time span of the sentences are considered in the similarity measurement. The alignment between sentences and video annotations is not strictly one-to-one.

sis of Multimodal Objects—, funded by the Minas Gerais State Agency for Research and Development (FAPEMIG – grant RED-00106-21) and the Brazilian National Council for Scientific and Technological Development (CNPq – grant 420945/2022-9). M. A. Gamonal is a postdoctoral fellow supported by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES – grant 88887.015648/2024-00). The construction of the dataset was also funded through grants 88887.936139/2024-00 (CAPES) and 151361/2023-1 (CNPq). T. T. Torrent has a grant from CNPq (311241/2025-5). A. S. Pagano has grants from CNPq (404722/2024-5; 313103/2021-6) and FAPEMIG's program for internationalization of scientific, technological and innovation institutions of Minas Gerais. F. Belcavello was supported by CNPq (200270/2023-0). The authors acknowledge the reviewers of the *Beyond Language: Multimodal Semantic Representations II* Workshop for their valuable feedback and suggestions.

References

Helen de Andrade Abreu and Ely Edison da Silva Matos. 2025. A framenet brasil approach to annotation of pragmatic frames evoked by turn organization gestures. *Caligrama: Revista de Estudos Românicos*, 30(1):94–109.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Frederico Belcavello, Tiago Timponi Torrent, Ely E. Matos, Adriana S. Pagano, Maucha Gamonal, Natalia Sigiliano, Livia Vicente Dutra, Helen de Andrade Abreu, Mairon Samagaio, Mariane Carvalho, Franciany Campos, Gabrielly Azalim, Bruna Mazzei, Mateus Fonseca de Oliveira, Ana Carolina Loçasso Luz, Lívia Pádua Ruiz, Júlia Bellei, Amanda Pestana, Josiane Costa, Iasmin Rabelo, Anna Beatriz Silva, Raquel Roza, Mariana Souza, and Igor Oliveira. 2024. Frame2: A FrameNet-based multimodal dataset for tackling text-image interactions in video. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7429–7437, Torino, Italia. ELRA and ICCL.

Cristóbal Pagán Cánovas, Javier Valenzuela, Daniel Alcaraz Carrión, Inés Olza, and Michael Ramscar. 2020. Quantifying the speech-gesture relation with massive multimodal datasets: Informativity in time expressions. *PLoS ONE*, 15.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Charles Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di semantica*, 6:222–254.

Charles J. Fillmore. 1982. Frame Semantics. In Linguistics Society of Korea, editor, *Linguitics in the morning calm*. Hanshin Publishing Co., Seoul, South Korea. Pages:111–138.

Eliane M. Gordeef. 2023. Avaliação sobre animação e cinema de vida real: semelhanças e diferenças. *Diálogo com a Economia Criativa*, 8(24):50–63.

Stephan Gouws, G-J van Rooyen, and Herman A. Engelbrecht. 2010. Measuring conceptual similarity by spreading activation over Wikipedia's hyperlink structure. In *Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 46–54, Beijing, China. Coling 2010 Organizing Committee.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023. Autoad ii: The sequel – who, when, and what in movie audio description.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language.

Seon-Ho Lee, Jue Wang, David Fan, Zhikang Zhang, Linda Liu, Xiang Hao, Vimal Bhat, and Xinyu Li. 2024. Nowyousee me: Context-aware automatic audio description.

Zhenhao Li, Marek Rei, and Lucia Specia. 2022. Multimodal conversation modelling for topic derailment detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5115–5127, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jian Ma, Wenguan Wang, Yi Yang, and Feng Zheng. 2024. MS2SL: Multimodal spoken data-driven continuous sign language production. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7241–7254, Bangkok, Thailand. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Mairon Samagaio, Tiago Torrent, Ely Matos, and Arthur Almeida. 2024. Semantic permanence in audiovisual translation: a FrameNet approach to subtitling. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 168–176, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Tiago Timponi Torrent and Michael Ellsworth. 2013. Behind the labels: Criteria for defining analytical categories in framenet brasil. *Veredas-Revista de Estudos Linguisticos*, 17(1):44–66.

Tiago Timponi Torrent, Ely Edison da Silva Matos, Frederico Belcavello, Marcelo Viridiano, Maucha Andrade Gamonal, Alexandre Diniz da Costa, and Mateus Coutinho Marim. 2022. Representing context in framenet: A multidimensional, multimodal approach. *Frontiers in Psychology*, Volume 13.

Marcelo Viridiano, Arthur Lorenzi, Tiago Timponi Torrent, Ely E. Matos, Adriana S. Pagano, Natália Sathler Sigiliano, Maucha Gamonal, Helen de Andrade Abreu, Lívia Vicente Dutra, Mairon Samagaio, Mariane Carvalho, Franciany Campos, Gabrielly Azalim, Bruna Mazzei, Mateus Fonseca de Oliveira, Ana Carolina Luz, Livia Padua Ruiz, Júlia Bellei, Amanda Pestana, Josiane Costa, Iasmin Rabelo, Anna Beatriz Silva, Raquel Roza, Mariana Souza Mota, Igor Oliveira, and Márcio Henrique Pelegrino de Freitas. 2024. Framed Multi30K: A frame-based multimodal-multilingual dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7438–7449, Torino, Italia. ELRA and ICCL.

Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. 2024. MMAD:multimodal movie audio description. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11415–11428, Torino, Italia. ELRA and ICCL.