Frame-Based Semantic Representation and Similarity Analysis in Audio Description Scripts

Maucha Andrade Gamonal[1], **Adriana Silvina Pagano**[2], **Tiago Timponi Torrent**[1], **Ely Edison Matos**[1]

[1] Departamento de Letras – Universidade Federal de Juiz de Fora (UFJF)

[2] Departamento de Tradução – Universidade Federal de Minas Gerais (UFMG).

{maucha.andrade@visistante.ufjf.br, apagano@ufmg.br, tiago.torrent@ufjf.br,ely.matos@ufjf.br}

Abstract. *We present a semantic similarity analysis between two versions of audio description scripts for a Brazilian short film, based on Frame Semantics and the FrameNet Brasil annotation model. Our study applies manual frame-based semantic annotation, identifying lexical units, frames, and frame elements. A similarity metric combining spread activation and cosine similarity is employed to measure semantic overlap, showing that variations in frame evocation reflect different narrative and descriptive choices across scripts. This work demonstrates how frame-based models can capture semantic overlap and divergence in multimodal translation tasks such as audio description.*

1. Introduction

Semantic similarity is central to many Natural Language Processing (NLP) tasks, such as machine translation, textual inference, and semantic search. While most approaches rely on distributional semantics [Lenci et al. 2022], neural embeddings [Turton et al. 2021), or lexical ontologies [Saedi et al 2018], frame-based approaches offer a cognitively motivated alternative, that models meaning as structured conceptual knowledge.

This study investigates the use of frame-based semantic annotation to measure semantic similarity between two versions of audio description (AD) scripts produced for the Brazilian short film *Eu Não Quero Voltar Sozinho* [Lacuna Filmes, 2010]. Audio description is a form of intersemiotic translation designed to make audiovisual content accessible to blind and visually impaired audiences [Pagano et al. 2021]. Different AD versions for the same film often reflect distinct linguistic choices and interpretive strategies, raising questions about their semantic equivalence.

We apply the FrameNet Brasil (FN-Br) annotation model [Torrent et al, 2022], based on Frame Semantics [Fillmore 1982], to manually annotate the lexical units, frames, and frame elements in both AD scripts. Semantic similarity is then measured using a metric that combines spread activation over the frame network with cosine similarity [Gouws et al. 2010; Viridiano et al., 2022]. Our contributions are: (i) demonstrating the applicability of frame-based semantic similarity in a multimodal translation task, and (ii) expanding the FN-Br data with manually annotated AD material.

2. Frame Semantics and Its Application to Audio Description

Frame Semantics is a cognitive linguistic theory that models meaning as structured knowledge organized into schematic representations of situations called frames. Each frame includes a conceptual structure involving participants, entities, and other contextual

elements known as Frame Elements (FEs). Words evoke frames, and their meanings are understood through these structured networks.

FrameNet Brasil extends the original FrameNet [Baker et al., 1998] to Brazilian Portuguese, providing a lexical database annotated with frames, lexical units (LUs), and FEs. The model supports applications in semantic parsing [Das et al., 2014], information extraction [Dutra, 2023], and knowledge representation [Torrent et. al., 2022].

In multimodal contexts, the same theoretical model can represent meaning beyond verbal language, aligning textual elements with visual and other perceptual elements. The ReINVenta[1] project (Research and Innovation Network for Vision and Text Analysis) adopts this perspective, assuming that different modalities, such as text, images, gestures, and video, can evoke frames, enabling cross-modal semantic alignment.

Audio Description (AD) is an intersemiotic translation that makes visual content accessible to blind and visually impaired audiences [Naves et al., 2016]. It narrates actions, settings, and other visual information essential for following a narrative. While AD has been widely studied in translation and accessibility research [Fryer, 2016; Pagano et al., 2016], applying frame-based semantic annotation to AD enables systematic modeling of how linguistic choices represent perceptual and narrative content. This also provides a foundation for quantifying semantic similarity between different AD versions, capturing both preserved and shifted meanings.

3. Related Work

Recent studies have applied frame-based semantic representation to multimodal and multilingual similarity tasks. Viridiano et al. (2022) applied FrameNet-based annotations on the *Multi30k* and *Flickr30k Entities* datasets, measuring similarity across captions in different languages and between captions and images. They showed that cosine-based metrics, combined with frame relations, effectively capture cross-modal meaning alignment. Samagaio et al. (2024) used frame-based similarity to compare audio transcriptions and subtitles, highlighting how subtitling constraints influence frame evocation and semantic overlap.

Similarly, Souza et al. (2024) combined spread activation and soft cosine similarity in Portuguese audio and English subtitles, relating these results to corresponding images. Their findings show that translation choices significantly affect semantic alignment in the short film.

Finally, Dornelas et al. (2022) provided multimodal annotations of *Eu Não quero Voltar Sozinho*, aligning AD transcripts with visual scenes. Although not focused on similarity, their data form the basis for our analysis.

4. Method and Material

4.1. Frame-based Semantic Annotation

Semantic annotation followed the FN-Br multimodal framework, assigning **frames**, **lexical units** (LUs), and **frame elements** (FEs) to text segments. Each LU evokes a frame that structures the semantic interpretation of the scene, through its participants, entities, properties, and other contextual elements. Frames cover events, states, entities, attributes and spatial relations.

[1] https://www2.ufjf.br/framenetbr/reinventa/, accessed on August 9, 2025.

For example, the Delivery[2] frame models a transfer in which a DELIVERER transfers a THEME to a RECIPIENT, or, indirectly, to a GOAL. In *O menino cego **ENTREGA** suas chaves para a colega, que abre o portão*. (*The blind boy HANDS OVER his keys to his classmate, who opens the gate), entregar.v* evokes Delivery, with "*o menino cego"* as DELIVERER, "*suas chaves"* as THEME, and "*a colega"* as RECIPIENT. Other FEs, such as GOAL or SOURCE, are not expressed in the sentence.

4.1.1 Frame-to-frame relations and Top-level frames

In FrameNet Brasil, frame-to-frame relations are formally defined to ensure the coherence of the semantic network (frame+net), allowing to represent conceptual hierarchies, compositional structures, and inferential links between situations. These relations include:

- **Inheritance**: a more specific frame inherits core elements from a more general one (e.g.: Product_delivery inherits from Delivery);
- **Subframe**: a complex event decomposes into their constituent sub-event;
- **Using**: situations that conceptually depend on others (e.g.: Delivery uses Sending to describe the physical transfer);
- **Causative_of:** causal link between two frames (e.g.: Cause_motion is causative of Motion)
- **Inchoative_of:** transition of state (e.g.: Becoming_dry is inchoative of Being_dry
- **Perspective_on**, which reflects different viewpoints on the same situation (e.g., Giving vs. Transfer).

By formally encoding these relations, the network supports structured semantic representations essential for both linguistic analysis and computational applications, including semantic similarity measurement.

Beyond frame-to-frame relations, frame types also constitute important representations that add to the set of semantic information derived from linguistic annotations. FrameNet's linguistic annotation model is based on the semantic and syntactic valence of the target LU. There is a systematic relationship between frame types and the valence patterns, since many frames are organized into abstract conceptual categories known as top-level frames. These top-level frames structure more specific ones in the network and include categories such as Event, State, Attribute, and Entity.

However, not all frames in the FrameNet Brasil network are currently assigned to top-level semantic categories. This lack of full categorization may limit broader generalization processes and the ability to group frames into higher-order conceptual domains. While this does not affect frame-to-frame local similarity measurements, it constrains analyses that rely on hierarchical relationships, such as semantic clustering or category-based inferences.

For instance, although the frame Delivery clearly represents an event, it is not formally linked to the top-level frame Event in the network. This gap may be attributed to several factors: the constant expansion of the network, the complexity of manual classification, and the high specificity of certain frames, which makes immediate categorization challenging.

4.2. Corpus and statistics

[2] Following established conventions, frame names are set in Courier, and FEs in VERSALETE.

The corpus used in this study consists of two AD scripts of the Brazilian short film *Eu Não Quero Voltar Sozinho*. One is the official AD produced by the TRAMAD group, and the other is an alternative version created by Vieira (2015) as part of an academic project on accessible audiovisual translation. The semantic annotation of the official script was previously presented in Dornelas et al. (2024), whereas the annotation of the alternative AD script was conducted specifically for this study.

The official audio description (AD) script contains 787 tokens, 283 types, and 100 sentences, while the alternative AD has 785 tokens, 282 types, and 88 sentences. Both scripts consist of a single document each. The corpus excludes the closing credit of the ADs, which contain information about the description team and production details. Due to the limited scope of this study, sentence alignment was carried out manually, based on the narrated visual content, resulting in 78 aligned sentences.

A total of 489 frames were annotated in the official AD and 476 in the alternative version. The most frequently instantiated frames[3] in both scripts include `Body_parts`, `Motion`, `Perception_active`, `Architectural_part`, `Manipulation`, and `People_by_age`. Notably, `Body_parts` and `People_by_age` frames appear prominently across both versions, reflecting core elements of the visual and narrative content.

Regarding nominal lexical units, the most common in both versions is *adolescente.n* (teenager), linked to the `People_by_age` frame. Other frequent nominal units in the official AD include *portão.n* (gate), *cabeça.n* (head), *livro.n* (book), and *mão.n* (hand), while the alternative AD highlights *adolescente.n*, *menino.n* (boy), *quarto.n* (room), *livro.n* (book), and *rua.n* (street). This lexical overlap indicates similar focal points, though with some variation in object and character emphasis.

In terms of frame types, events are the most frequent, with 116 in the official AD and 117 in the alternative, followed by entities (98 and 78), states (71 and 67), relations (12 and 17), and attributes (23 and 15). These distributions provide an empirical foundation for comparing the semantic strategies between the two scripts. These distributions provide an empirical foundation for comparing semantic strategies in both AD versions.

4.3. Semantic Similarity Metric

To evaluate the semantic overlap between the two audio description versions, we applied a frame-based similarity metric that combines relational knowledge from FrameNet with vector-based measurement. This metric follows three-step, adapted from Gouws et al. (2010) and Viridiano et al. (2022).

i. **Associate Table Construction:** using FrameNet's network graph, we derive frame-to-frame associative strengths by considering semantic relations such as inheritance, subframe, perspective_on, and uses. Each relation receives a weight based on semantic proximity, resulting in an associative table that captures both direct frame matches and indirect semantic connections.
ii. **Spread Activation Algorithm:** activation starts from the directly evoked frames and propagates through the FrameNet graph to related frames. Activation strength decays with graph distance, assigning higher weights to closely related frames and

[3] All frames are accessible in both the English FrameNet database (https://framenet.icsi.berkeley.edu/) and the Brazilian Portuguese FrameNet Brasil (https://webtool.frame.net.br/).

lower weights to more distant ones. This process models how semantic similarity extends beyond exact matches, accounting for conceptual proximity.

iii. **Cosine Similarity Calculation:** the spread activation process generates a vector for each annotation. Each position in this vector corresponds to the frame and its aggregated activation value. Semantic similarity is then calculated as the cosine similarity between the two activation vectors, producing a normalized score between 0 (no similarity) and 1 (identical frames).

This hybrid algorithm captures both exact semantic matches and related concepts within the frame network - enhancing the metric's ability to detect paraphrases and conceptually related descriptions (e.g., `Delivery` and `Sending`). Such sensitivity is crucial for evaluating interpretative differences in audio description. Each version of the audio description was semantically annotated with frames, lexical units, and frame elements. The metric quantifies the degree of semantic overlap between their frame-based representations, capturing variations in linguistic encoding that preserve or shift meaning in translation.

5. Discussion and Results

5.1. Semantic similarity and Conceptual Structures in Audio Description Scripts

Mapping the prevalence of frame types - such as Event, State, Attribute, or Entity – reveals how dynamic visual content is anchored in the verbal modality and which aspects of the filmic experience are prioritized. A predominance of event frames, for example, may indicate a focus on narrating actions and dynamic sequences, while a higher frequency of entity frames suggest emphasis on descriptive details or static elements.

This frame-based analysis helps explain how semantic choices shape the experiential affordances of AD, mediating between visual and linguistic representation [Bateman et al. 2017]. It also underscores the role of semantic architecture in promoting narrative coherence, spatial orientation, and character interaction for visually impaired audiences, ultimately shaping their access to the audiovisual experience.

5.2. Semantic Similarity Metric and Analysis

Figure 1 presents the distribution of semantic similarity scores between the two AD scripts. The mean similarity is 0.5 (SD=0.2), indicating variability in how semantic content aligns across sentence pairs.

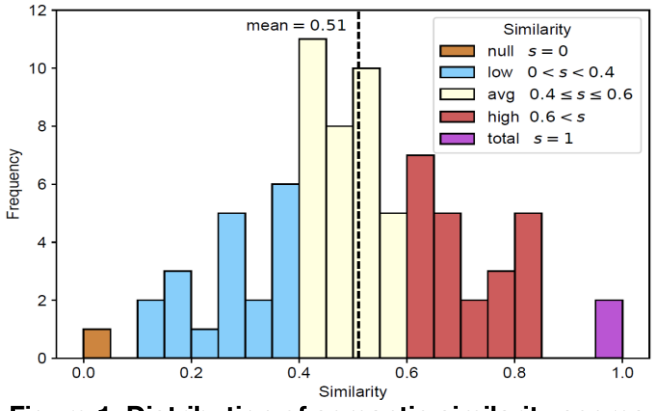

Figure 1. Distribution of semantic similarity scores

Total similarity (s = 1) occurs when all frames evoked in both versions match exactly. It appears in 2.56 % of pairs. For example:

(1) Leo **FECHA**[CLOSURE] o **LIVRO**[TEXT]. (Official AD) – *Leo closes the book*.
(2) Eles **FECHAM**[CLOSURE] o **LIVRO**[TEXT]. (Alternative AD) – *They close the book.*

High similarity (0.6 < s < 1) represents 28.21% of the material. Even when the number of frames differs, shared frames raise the score:

(3) **BEBE**[INGESTION] um squeeze (Official AD) – *(Leo) drinks (from) a squeeze (bottle).*
(4) Leo **BEBE**[INGESTION] **ÁGUA**[FOOD] **NA**[INTERIOR_PROFILE_RELATION] **GARRAFA**[RECIPIENT]. (Alternative AD) – *Leo drinks water from the bottle.*

Average similarity (0.4 ≤ s ≤ 0.6) accounts for 43.59% of the data:

(5) **RAPIDAMENTE**[TAKING_TIME] Gabriel **PEGA**[TAKING] sua **MOCHILA**[ARTIFACT], o **MOLETOM**[CLOTHING] e os **TRÊS**[CARDINAL_NUMBERS] **SAEM**[MOTION]. (Official AD) – *Gabriel quickly grabs his backpack, (his) sweatshirt and the three (of them) leave.*
(6) Gabriel **PEGA**[TAKING] o seu **MOLETOM**[CLOTHING] e **ACOMPANHA**[ACCOMPANIMENT] o **CASAL**[PERSONAL_RELATIONSHIP]. (Alternative AD) – *Gabriel grabs his sweatshirt and follows the couple.*

Low similarity (0 < s < 0.4) appears in 24.36% of pairs:

(7) Seu **ROSTO**[BODY_PARTS] **SORRIDENTE**[FACIAL_EXPRESSION] se **ILUMINA**[LOCATION_OF_LIGHT]. (Official AD) – *Your smiling face lights up.*
(8) Ele **DEMONSTRA**[CAUSE_TO_PERCEIVE] **ALEGRIA**[EMOTION_DIRECTED] **COM**[HAVE_ASSOCIATED] um **SORRISO**[FACIAL_EXPRESSION]. – *He expresses happiness with a smile.*

Null similarity (s = 0) is rare 1.28% and occurs when no frames match:

(9) Gabriel **ABAIXA**[BODY_MOVEMENT] a **CABEÇA**[BODY_PARTS] e dá um **SORRISINHO**[FACIAL_EXPRESSION]. (Official AD) – *Gabriel lowers his head and gives a little smile.*
(10) Gabriel ESTÁ **ENVERGONHADO**[EMOTION_DIRECTED.] (A. AD) – *Gabriel is embarrassed.*

6. Conclusions and Future work

This study applied a frame-based semantic similarity analysis to two audio description scripts of a Brazilian short film, using the FrameNet Brasil model for manual annotation of lexical units, frames, and frame elements. Combining frame relations with spread activation and cosine similarity, the method quantified semantic overlap between aligned sentences.

Results showed that semantic similarity often extends beyond exact frame matches, as related frames also contribute to shared meaning. The analysis revealed how different linguistic choices in AD impact the representation of events, entities, and relationships, thus shaping audience access to the filmic experience.

Future work will explore automated approaches for frame identification and similarity computation, enabling scalability for larger datasets. We also plan to compare frame-based similarity with other semantic models to assess complementarity and improve applications in accessibility-oriented NLP tools.

7. Acknowledgment

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). 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.

8. References

Baker, C. F., Fillmore, C. Jay, and Lowe, J. B. (1998) “The Berkeley FrameNet Project”. In: Proceedings of the 17th International Conference on Computational Linguistics (COLING) (Vol. 1, pp. 86–90). Montreal, Canada.

Bateman, J., Wildfeuer, J., and Hiippala, T. (2017) “Multimodality: Foundations, Research and Analysis - A Problem-oriented Introduction”. Walter de Gruyter GmbH & Co KG. DOI https://doi.org/10.1515/9783110479898.

Das, Dipanja, Chen, Desai, Martins, André; Schneider, Nathan and Smith, Noah A, (2014) “Frame-semantic parsing,” Computational linguistics, vol. 40, no. 1, pp. 9–56.

Dornelas, L. D., Gamonal, M. A., and Pagano, A. S. (2024) “Análise semântica de audiodescrição em curta metragem: uma abordagem multimodal a partir da Semântica de Frames”. Domínios de Linguagem, Uberlândia, v. 1866, p. 2-30. DOI: https://doi.org/10.34019/1808-9461.2022.v23.38564.

Dutra, L. V. (2024) “Evaluating the contribution of FrameNet to gender-based violence identification: How semantic annotation can be used as a resource for identifying patterns of violence”. Master’s Thesis (Master in Language Technology) – Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Göteborg. Available in: hdl.handle.net/2077/81763.

EU NÃO QUERO VOLTAR SOZINHO. Curta Metragem. Direção: Daniel Ribeiro. Produção: Lacuna Filmes. Brasil: [s. n.], 2010. Disponível em: https://www.youtube.com/watch?v=FkNoXubidmk.

Fillmore, C. J. (1982). “Frame semantics”. In: Linguistic society of Korea (Ed.), Linguistics in the morning calm (pp. 111–137). Hanshin Publishing Co.

Gouws, S., van Rooyen, G.-J., and Engelbrecht, H. A. (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, August. Coling 2010 Organizing Committee.

Lenci, A., Sahlgren, M., Jeuniaux, P. et al. (2022) “A comparative evaluation and analysis of three generations of Distributional Semantic Models”. Lang Resources & Evaluation 56, 1269–1313. https://doi.org/10.1007/s10579-021-09575-z

Pagano, A. S., Teixeira, A. L. R. and Mayer, F. A. (2021) “Accessible Audiovisual Translation”, In: Meng Ji, and Sara Laviosa (eds), The Oxford Handbook of Translation and Social Practices, England. https://doi.org/10.1093/oxfordhb/9780190067205.013.4.

Saedi, C, Branco, A, Rodrigues, J. R., and Silva, J. (2018). “WordNet Embeddings”. In Proceedings of the Third Workshop on Representation Learning for NLP, pages 122–131, Melbourne, Australia. Association for Computational Linguistics.

Samagaio, M., Torrent T. T., Matos, E., and Lorenzi A. (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, p. 168–176, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Souza, M. M. S., Gamonal, M. A, and Pagano, A. S. (2025) “Permanência semântica

entre áudio original e legenda: um estudo sobre anotação semântica multimodal em obra audiovisual". Caligrama: Revista De Estudos Românicos, 30(1), 52-73. https://doi.org/10.35699/2317-2096.2025.57566

Viridiano, M., Torrent T. T., Czulo, O., Lorenzi A., Matos, E. and Belcavello, F. (2022) "The case for perspective in multimodal datasets". In: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, pp. 108–116, Marseille, France. European Language Resources Association.

Torrent, T. T, Matos, E., Belcavello, F. Viridiano, M., Gamonal, M. A., DINIZ, A. and Coutinho, M. M. (2022) "Representing context in framenet: A multidimensional, multimodal approach". In: Frontiers in Psychology, v. 13. https://doi.org/10.3389/fpsyg.2022.838441

Turton J., Smith R. E.;,Vinson D. (2021). Deriving Contextualised Semantic Features from BERT (and Other Transformer Model) Embeddings. In Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), pages 248–262, Online. Association for Computational Linguistics.