A Frame and Canvas-Based Perspective-Encoding Methodology for Multimodal Semantic Annotation of Classroom Settings

Cláudia Ferraz[1,2], Ely Matos[1], Frederico Belcavello[1], Júlia Gasparetto[1], Juliana de Oliveira[1], Janina Wildfeuer[2], Tiago Timponi Torrent[1,3]

[1] Federal University of Juiz de Fora | FrameNet Brasil, [2] University of Groningen,
[3] Brazilian National Council for Scientific and Technological Development – CNPq
{claudia.ferraz, julia.gasparetto, juliana.oliveira}@estudante.ufjf.br, {ely.matos, fred.belcavello, tiago.torrent}@.ufjf.br, j.wildfeuer@rug.nl

Abstract

We propose a methodology for the multimodal semantic annotation of classroom interactions that takes interactional canvases and semantic frames as its core analytical categories. The approach enables the systematic recording of semantic correlations among interactants, communicative modes, and material supports involved in situated meaning-making processes. The methodology encodes participant perspective by relying on the temporal alignment of multiple video recordings of the same instructional event captured from different viewpoints, allowing for the representation of how meaning construction unfolds across perceptual and interactional positions. To operationalize the proposal, we introduce an annotation tool that implements the scheme and supports the integration of multimodal data streams within a unified semantic representation framework. We conclude by discussing the limitations of the current proposal and the possibilities for extending it to other interactional settings.

Keywords: Multimodal Semantic Annotation, FrameNet, Canvas, Perspective

1. Introduction

In recent years, there has been a growing demand for more comprehensive approaches to multimodal data annotation, capable of offering systematic, consistent, and interconnectable annotation schemes that capture the rich diversity of communicative situations and all media and modes involved (Bateman et al., 2017a; Pflaeging et al., 2021). Given the increasing complexity of multimodal data analyzed in different domains, this demand is driven by academic and social motivations.

As a particular challenge, the immense diversity and complexity of multimodal communicative situations makes it necessary to develop tools and schemes that are tailored to the different types of data at question: page-based data, such as documents, diagrams, or comics, are analyzed with regard to their spatial arrangement, while linear data, such as films or face-to-face interactions, request a temporal and dynamic annotation. At the same time, there is a shortage of annotation methodologies that explicitly consider the real contexts of data production and interpretation, particularly those related to education. In such environments, where multiple participants engage in joint meaning-making processes, the perspective of each participant towards the communicative event is key.

Existing FrameNet-based annotation schemes allow for a myriad of multimodal analyses (Belcavello et al., 2024; Viridiano et al., 2024; Abreu and Matos, 2025; Sigiliano, 2025). They are based on an adaptation of the FN-Br WebTool (Torrent et al., 2024),

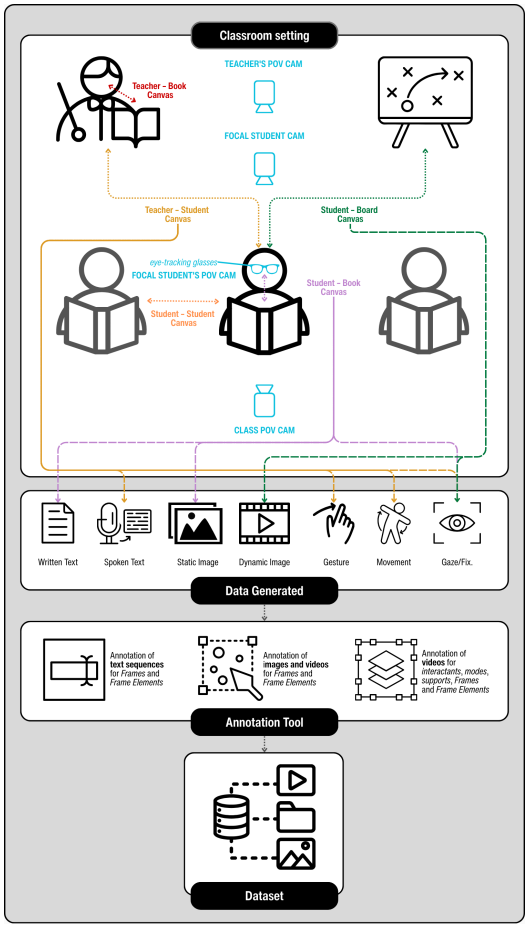

Figure 1: Annotation scheme and resulting dataset

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 22–32
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

which enables fine-grained semantic annotation of multimodal corpora, particularly combinations of text and image (Belcavello et al., 2022). The tool allows annotators to choose frames and locate Frame Elements (FEs) in both text and images. When annotating videos, it allows tracking the time interval in which these elements are active in the image and in the text transcript of spoken audio (Belcavello et al., 2020).

Such annotation schemes were aimed at analyzing multimodal corpora built from mostly two-dimensional media, but this foundation allows for the development of new annotation frameworks that focus on even more complex communicative situations, such as those in educational settings. In-classroom language teaching, for example, involves multiple modes of expression, such as speech, gesture, image, and writing, while at the same time using material (books, blackboard) and digital (tablets or computers) supports, which again feature a diversity of semiotic modes. This multidimensional environment combining both static and dynamic materialities that make space for direct, indirect, or subordinate modes (Bateman et al., 2017a) has brought three new challenges:

1. How to annotate the simultaneous and dynamically unfolding use of multiple semiotic modes (such as speech, gesture, image, etc.) and their interrelationships between interactants and objects in order to capture and represent information from frame evocation processes?

2. How to delimit the analysis considering the co-occurrence and interdependence of these modes, enabling semantic analysis of granularity resulting from the combination of different modalities?

3. How to encode each participant's perspective towards the multimodal meaning-making processes?

So far, only a few empirical studies have systematically approached how different interactants (teachers, students, artifacts) relate and coordinate in authentic teaching-learning situations (Thomas, 2018; Fillietaz et al., 2022, e.g.). Furthermore, although there are relevant methodological proposals in the field of multimodality research, there is still no widely standardized approach capable of dealing with the overall complexity of these interactions in real educational contexts. Given this scenario, this paper presents a proposal for a multimodal annotation scheme for real classroom interactions (Figure 1), based on the multimodal approach proposed by Bateman et al. (2017a) in combination with frame semantics (Fillmore, 1982), as systematized within the scope of FrameNet Brasil (Torrent et al., 2022). With a particular focus on the materiality and canvases of the communicative situation and by implementing the conceptual structure of Frame Semantics used in the FN-Br WebTool (Torrent et al., 2024), the proposal aims to capture multimodal interactions in an authentic educational environment in an integrated manner, promoting reproducibility, analytical consistency, and interoperability between annotation schemes. Moreover, by systematically annotating the same communicative events from different points of view, the scheme incorporates data perspectivism in its strong variety (Cabitza et al., 2023), to the extent that multiple annotation labels are assigned to one same piece of data depending on whose perspective — the teacher's, the student's, the group's — is being encoded.

2. Frame Semantics and FrameNet

The proposed annotation methodology adopts Frame Semantics as its main theoretical foundation, which includes context as a central aspect of the theory. According to this approach, frames are defined as the background scenes against which the meaning of a given event, situation, or linguistic utterance must be constructed. Such scenes include the participants and props that make up the situation, each of which is conceived as a Frame Element (FE). In this sense, FEs are related in such a way that the presence of one triggers the participation of the others (Fillmore, 1982, p.111).

Frames can be instantiated in two ways: by evoking Lexical Units (LUs) that trigger certain frames or by invocation performed by comprehenders based on the combination of contextual clues present in the text. Thus, the comprehension of a statement depends not only on linguistic structure but also on knowledge of the world, the immediate context, and the communicative conditions in which the message occurs (Torrent et al., 2022).

FrameNet was established as a computational resource for semantic annotation based on English language corpora (Baker et al., 1998), providing a formalized inventory of frames, FEs, and annotated LUs. Thus, the corpus analyzes performed in FrameNet follow the fundamental principle that *"meaning is relativized to scenes"* (Fillmore, 1977). In FrameNet, each frame is composed of a name, an in-prose definition, and a set of FEs and their definitions. Because it is a FrameNET, not a frame list, frames are linked to one another via frame-to-frame relations. Consider, as an example, the `Manipulate_into_doing` frame in (1).

(1) **`Manipulate_into_doing`**
Definition: A Manipulator gets a Patient to perform a Resulting_action.

Core Frame Elements:
Manipulator: The person who gets the Patient to act.
Patient: The person who the Manipulator gets to do something they would not have done had without some request or pressure.
Resulting_action: What the Manipulator intends to make the Patient do and the Patient actually does
Frame-to-Frame Relations:
inherits from: `Intentionally_affect`
is inherited by: `Talking_into`
uses: `Influencing_potential`
Lexical Units: *lure.v; manipulate.v...*

The Inheritance relation is a Parent-Child hierarchy in which the daughter frame represents an equally or more specific version of the scene portrayed by the mother frame, inheriting all its roles and constraints. Every FE in the mother frame must correspond to an FE in the daughter frame, although these elements may be assigned different names or more detailed definitions in the daughter. In this example, the `Manipulate_into_doing` frame inherits from the more general `Intentionally_affect` frame, and it is itself inherited by the even more specialized `Talking_into` frame. The MANIPULATOR FE in the `Manipulate_into_doing` frame corresponds to the AGENT in the `Intentionally_affect` frame and to the SPEAKER in the `Talking_into` frame. Similar relations hold for the other two FEs.

The Using relation, in turn, is a frame-to-frame connection indicating that a daughter frame presupposes the background knowledge of a mother frame as a necessary context for its own meaning. Unlike the Inheritance relation, where the child is a more specific version of the parent, the Using relation implies that only part of the scene in the daughter frame refers to the conceptual structure of the mother frame. In (1), the `Manipulate_into_doing` frame uses the `Influencing_potential` frame because the specific act of manipulation presupposes the broader concept of influence. Other types of relations in FrameNet include Perspective_on, Subframe, Preceedes, Inchoative_of and Causative_of (Ruppenhofer et al., 2016).

The FrameNet model has been implemented elsewhere in the world for several other languages. Among them, FrameNet Brasil was established for Brazilian Portuguese, considering the specific lexical, grammatical, and semantic properties of the language, ensuring interlinguistic comparability and theoretical consistency (Torrent and Ellsworth, 2013). More recently, the scope of semantic annotation within FrameNet Brasil has expanded beyond exclusively verbal data, giving rise to multimodal datasets that integrate linguistic and visual modes (Belcavello et al., 2024; Viridiano et al., 2024; Gamonal et al., 2025). To operationalize this integration, multimodal annotation tools have been developed that allow the alignment of frame-based semantic annotations with temporal and spatial representations of dynamic multimodal data, enabling the analysis of how frames are evoked through speech, gestures, gaze, visual objects, and other visual resources. This establishes conditions for the development of empirical research covering the study of event frames (Pinto, 2025), pragmatic frames (Abreu and Matos, 2025), and deictic center frames (Sigiliano, 2025) in audiovisual corpora. Such empirical studies have shown that, similar to the way words in a sentence evoke frames and organize their elements in syntactic locality, other elements in various communicative modes can also do so or work complementarily with frame evocation patterns across modalities (Belcavello et al., 2020).

3. Empirical Multimodality Research

In multimodality studies, there is currently a focus on empirical data analysis and the comprehensive examination of larger data sets (Pflaeging et al., 2021; Bateman et al., 2026). This trend builds on a productive context of theoretical, methodological, and analytical developments over three decades with an initial interest in the combination of written language with images and a rapid extension to all expressive forms and the question of how different types of meaning-making, previously studied separately in diverse disciplines, combine into an integrated, multimodal whole.

In order to capture the complex nature of multimodal communicative situations, many current works aim for a comprehensive analysis of larger corpora and therefore develop multi-level standoff annotation schemes that combine descriptions of the actual material with interpretations of the communicative structure of the analytical object(s). This method has been used for several kinds of artifacts, ranging from multi-dimensional page-based documents (Bateman, 2008), packaging (Thomas, 2009) and tourist brochures (Hiippala, 2015) to comics (Bateman et al., 2017b), TV series (Drummond and Wildfeuer, 2020), commercials (Wildfeuer and Coffie, 2022), news videos (Bateman and Tseng, 2023), and short form video content (Grzenkowicz and Wildfeuer, 2025).

As a starting point for the development of such annotation schemes, the concept of the material canvas (Bateman et al., 2017a) provides a way to capture the initial perceptual unit of analysis. A canvas functions as the material basis carrying and

realizing semiotic modes. Each canvas is structured and analyzed according to a classification scheme involving the following dimensions (Bateman et al., 2017a, chap. 3): (i) **temporality**, which defines whether the canvas is static or dynamic; (ii) **spatial dimensionality**, which determines whether the canvas is two-dimensional (a screen or paper) or three-dimensional (e.g. architectural spaces or face-to-face interactions); (iii) **transience**, concerning the permanence of information and distinguishing transient canvases, such as the dynamic body, hands, and face of a person speaking in a conversation, from permanent ones, such as a printed book page; (iv) **roles of participant/observer**, which considers how the message recipients interact with the canvas or act simply as observers; and (v) **ergodicity**[1], which addresses the level of effort or contribution of the reader/user in constructing the meanings of the text.

In relation to the analytical objects in question in this paper, Bateman et al. (2017a, chap. 7) discuss the communicative situation of classroom interaction as one of the most complex multimodal situations that can be "sliced" into several canvases and subcanvases. For example, a teacher may, at a given moment, explain content from a textbook while simultaneously pointing to the book and looking at a particular student. These are different activities that occur simultaneously on different canvases with different materialities that make space for different semiotic modes. The teacher's spoken language, for example, operates on a dynamic, three-dimensional, and transient canvas of interaction with the students where both the teacher and the students are participants that can interact and interrupt each other. In contrast, the book the teacher uses as a medium is a static, two-dimensional, permanent canvas for which teacher and students are observers and which features several other semiotic modes such as images, written language, page layout, etc.

In the annotation methodology proposed in this paper, we integrate the canvas analysis with FrameNet annotation in order to provide a detailed understanding of how meaning is constructed in classroom interactions. The main canvas, the overall classroom situation, includes several subcanvases that provide multiple dynamic interactional spaces — making visible the characteristics of the multimodal environment that shape social interaction. On the teacher-student interaction canvas, for example, speech and gesture operate as semiotic modes in a face-to-face interaction, which is mutable and transient. For a teacher, interacting with a whiteboard, in contrast, it is relevant to identify modes supported by the dynamic canvas of the whiteboard (e.g. moving images in a video or written language from a text displayed on the whiteboard screen). Frames establish semantic units that allow meanings to be mapped within this particular context of interaction. In order to analyze these frames as nodes in a network of relations, it is necessary to determine the different interactants on the canvas as well as the modes and processes that enable meanings to be dynamically constructed from the different connections between frames and FEs. The annotation scheme devised to support such an analysis is presented next.

[1] Ergodic canvases are those that require a non-trivial effort to be traversed. The reader of a complex comic page, for example, does not simply follow a predetermined linear sequence but actively participates in choosing the path for generating the content. This also happens in video games, for example, where the game also reacts to how players interact with the interface. Bateman et al. (2017a) therefore distinguish between different layers of ergodicity. Non-ergodic canvases are those in which the reading path is fixed, and the reader has a more passive role in determining the structure of the text, as in a traditional novel, for example.

4. A Multilevel Annotation Scheme for Classroom Interactions

The theoretical basis for the construction of the annotation scheme proposed in this paper is based on Systemic-Functional Linguistics (SFL), which conceives language as a resource for meaning-making, in which "*meaning resides in systemic patterns of choice*" (Halliday and Matthiessen, 2004, p.23). In this context, system networks present the structured options available to the language user, allowing for the selection of combinations to convey specific meanings. The application of this approach to multimodal artifacts was already proposed in early multimodality research and has served as the foundation for the development of several multilevel annotation schemes for the analysis of larger corpora (see Section 3).

Building on the concept of canvases introduced in Section 3 and on the advances of multimodal FrameNet analysis summarized in Section 2, the annotation methodology proposed here focuses on classroom interactions as the primary object of analysis. These interactions unfold across multiple layers of semiotic activity, including spoken language, gestures, visual resources, technological artifacts, and so on, each of which can take part in a subcanvas within the broader classroom interaction canvas. In addition, each interaction may be regarded from at least two different perspectives — see Figure 1.

Annotating and analyzing these perspectivized interactions presents several challenges. First, classroom communication is typically asymmetrical, with the teacher leading the progression of activities. However, multiple agents (the teacher

and the students) are simultaneously present in and, therefore, imposing their perspectives on the interaction, and varied teaching strategies create dynamic layouts that both reorganize and facilitate interaction in different ways. This allows for the creation of multiple layers of annotation. In addition, the main classroom canvas is mutable and fully ergodic, evolving over time as participants engage in different communicative activities (Bateman et al., 2017a). This significantly shapes the ways in which participants "*establish and maintain embodied copresence and mutual orientation to each other, relative to the unfolding activity and the multimodal and material environment*" (Bateman et al., 2026, p.404). By developing an annotation scheme that directly aligns with this hierarchy of canvases, our goal is to systematically and structurally capture the multimodal complexity of in-classroom interactions.

Addressing such complexities required the development of a multilevel annotation scheme capable of representing semiotic modes at different levels of granularity. To test the feasibility of implementing the proposed annotation scheme, we built a dataset of in-classroom interactions, developed a tagset, and implemented an annotation task by adapting the FN-Br WebTool (Torrent et al., 2024). We present each of these stages next.

4.1. Dataset

To test the feasibility of the implementation of the proposed methodology, we built a dataset comprising video recordings and eye-tracking data[2] from 10 Portuguese language lessons in elementary school, each lasting approximately 50 minutes.

In line with a strong perspectivist approach to dataset construction (Cabitza et al., 2023), five cameras were strategically positioned to capture the perspectives of the several participants involved in the classroom canvas: (i) one in the back of the room facing the front, capturing the point of view of ALL_STUDENTS; (ii) one in the front facing ALL_STUDENTS and capturing the point of view of the TEACHER; (iii) one mounted on a pair of eye-tracking glasses, which is worn by a randomly selected student in each class (the FOCAL_STUDENT); (iv) one focused on the FOCAL_STUDENT or on a GROUP_OF_STUDENTS, depending on the type of pedagogical activity being recorded and (v) one security camera installed on the ceiling, which was used only as a reference. The lessons were designed to provide a variety of teaching strategies, generating several forms of interaction — subcanvases — and allowing analyzes in terms of the interactants, modes and supports involved from different points of view.

The video files from all cameras were aligned for time intervals and the audio was automatically transcribed using an AI tool.[3] Audio transcriptions were then manually checked by two annotators with access to the original video files. Once validated, the transcriptions were uploaded to the FN-Br WebTool text annotation module (Torrent et al., 2024) to be annotated for frames and FEs, following FrameNet's fulltext annotation methodology (Ruppenhofer et al., 2016). The instructional materials used during the classes were scanned and uploaded to the same tool for multimodal annotation according to the methodology proposed in (Torrent et al., 2022). The annotation of the other communicative modes involved in classroom interactions, in turn, required both the development of a new tagset and adaptations to the WebTool, which are the core of the methodology presented here.

4.2. Units of analysis and the resulting annotation layers

The perspectivized nature of the dataset allows for the annotations to focus on the interactions occurring in the classroom from the viewpoint of different participants. The main canvas therefore corresponds to the complete physical situation in which the interaction takes place, constituted not only by the setting, objects, and environment, but also by verbal utterances, facial expressions, bodily postures, and distances maintained during interactions (Bateman et al., 2017a, p. 96).

The definition of the analytical focus articulates higher-level actions and levels of materiality through the concept of canvases. This directs the analysis toward the affordances available for meaning-making in the observed situation and allows for a clear delimitation of what is included or excluded from the analysis. In line with Bateman et al. (2017a) and Grzenkowicz and Wildfeuer (2025), the formulation of the annotation scheme begins with the identification of specific analytical units within these canvases. The scheme is then "*applied to such units, producing a set of classificatory features and, where necessary, additional segmentations, in order to ensure a satisfactory description of the investigated phenomenon*" (Bateman et al., 2017b, p. 13).

In this sense, the most general canvases distinguished in our annotation scheme are those of the ongoing **Interactions** in the classroom, which appear to be the most universal units of analysis within

[2] In the context of FrameNet multimodal annotation, eye-tracking data are used for the psycholinguistic validation of the image annotations. Research by Belcavello (2023) demonstrates that bounding boxes annotated for frames and FEs in a dataset composed of episodes of a TV Travel Series present higher fixation durations than other regions of the image less semantically-relevant.

[3] https://www.notta.ai

the communicative situation under study and are equally applicable to other face-to-face interaction contexts. The notion of **Interaction** also facilitates articulation with Frame Semantics, where it can be evoked through event frames, allowing a direct mapping to eventualities (Grzenkowicz and Wildfeuer, 2025) in the multimodal context.

At the level of **Interactions**, the subcanvases (layers) that proved most relevant for the analysis are the following:

Interactants The participants involved in the interactional situation, including vocally active participants and co-presence in the communicative situation, focusing specifically on interaction through speech, gaze, gestures, and other relevant semiotic modes supported by the canvas and its subcanvases. For this layer, we define the following labels:

- TEACHER: the participant fulfilling the role of mediator in the pedagogical situation, often guiding the interaction and regulating participation.
- FOCAL_STUDENT: the student specifically selected as the central observation point for analysis. This category is useful for examining, in detail, co-presence, mutual orientation, multimodal behaviors, and appropriation of semiotic modes by the focal student. Given the eye-tracking data, it also allows tracking of their actions, verbal and non-verbal responses, and interaction with other participants and objects in the classroom.
- SECONDARY_STUDENT: another student who is not the main focus of analysis but whose presence and participation influence or are influenced by the interactional dynamic.
- GROUP_OF_STUDENTS: a set of students (including the FOCAL_STUDENT) who interact around a specific activity, allowing analysis of group-level interaction, coordination of actions, role distribution, and shared use of multimodal and material resources (e.g., notebooks, laptops, educational games) in joint meaning-making.
- ALL_STUDENTS: the entire class.
- NONE: indicates the absence of an active participant in the analytical situation or moments when no meaningful action can be attributed to specific participants.

Mode The ways in which different semiotic modes are activated by interactants as socially and culturally organized systems of resources for meaning-making, endowed with specific materiality and their own expressive potentials. Within the scope of multimodal analysis, each mode constitutes a relatively stable set of conventions, forms, and possibilities of articulation that allow participants in interaction to perform communicative actions and construct meaning. In this sense, within our multilevel scheme, this category makes it possible to identify, segment, and describe which semiotic modes are in use at a given moment of the interaction and how they are articulated — whether mediated or not by an object — in the realization of pedagogical communicative events. Its attributes correspond to the main embodied and verbal resources identifiable in the interaction. This layer can be annotated for the following labels and sublabels:

- GAZE_DIRECTION: refers to the orientation of the interactant's gaze during the interaction. The direction of gaze functions as a semiotic marker of focus, engagement, and distribution of authority, allowing identification of how the participant positions themselves and others within the interactional space. The sublabels under this attribute are as follows: (i) LOOKING AT FOCAL_STUDENT; (ii) LOOKING AT SECONDARY_STUDENT; (iii) LOOKING AT GROUP_OF_STUDENTS; (iv) LOOKING AT ALL_STUDENTS; (v) LOOKING AT SUPPORT; (vi) LOOKING AWAY; (vii) NONE.
- HEAD_MOVEMENT: refers to the movements of the interactant's head during interaction, acting as a semiotic bodily mode that expresses attitudes and regulates interactional participation. The sublabels under this attribute are: (i) AGREEMENT; (ii) DISAGREEMENT; (iii) ATTENTION/ENGAGEMENT; (iv) DISBELIEF (v) SELF-REGULATION; (vi) NONE.
- BODY_MOVEMENT: refers to observable bodily movements of the interactant that assumes semiotic relevance in the interaction. This attribute encompasses broader displacements and postural changes involving the body as a whole or larger body segments such as the trunk (in positional changes), spatial displacement (e.g., walking around the classroom), and global movements. The values attributed are: (i) LEANING FORWARD; (ii) TURNING TO THE RIGHT; (iii) TURNING TO THE LEFT; (iv) MOVING IN SPACE; (v) FACING PARTICIPANT; (vi) NONE.
- GESTURES: specific observable movements performed with hands and arms that carry semiotic relevance in interaction, contributing to meaning construction in communication. The values of this attribute are: (i) DEICTIC; (ii) ICONIC; (iii) METAPHORIC; (iv) CONVENTIONAL; (v) SELF-REGULATORY; (vi) NONE.

- SPOKEN_LANGUAGE: everything that is said during the interactions. This layer is linked to the sentences transcribed from the audio recordings and is annotated for frames and FEs in the existing fulltext annotation module of the WebTool.

- WRITTEN_LANGUAGE: everything that is written in any of the support materials used during the class. Sentences in this layer are also annotated for frames and FEs.

- STATIC_IMAGE: pictures, drawings and other images in the support materials. This layer is annotated in the existing multimodal annotation module for frames and FEs.

- DYNAMIC_IMAGE: videos used during the classes. This layer is also annotated for frames and FEs.

Support Support refers to materials or technological resources visibly present and potentially mobilized in the interaction, functioning as support for participants. The values attributed are: PRINTED PEDAGOGICAL MATERIAL (teacher's book, student book, handout, poster, notebook); DIDACTIC GAME MATERIAL (board game, game cards, game manual, handheld choice boards); TECHNOLOGICAL SUPPORT (laptop, projector, projection screen); BOARD; or NONE.

Viewpoint This layer refers to the perspective from which the interaction is filmed, indicating the point of view represented in the framing (e.g. the focal student's perspective, teacher's perspective, whole-class perspective, or overhead view).

4.3. Annotation interface

In this section, we present how the proposed annotation scheme is used in the FN-Br WebTool (Torrent et al., 2024). As indicated in Section 4.2, different modules of the WebTool are used to annotate different communicative modes. The transcribed SPOKEN_LANGUAGE and WRITTEN_LANGUAGE modes are annotated for frames and FEs using the fulltext annotation module. For instance, a sentence such as (2), which is uttered by the teacher in the moment of one of the classes shown in Figure 3, can be annotated for the `Manipulate_into_doing` frame in (1).

(2) [Eu$_{\text{MANIPULATOR}}$] gostaria$^{\texttt{Manipulate_into_doing}}$ [que [vocês$_{\text{PATIENT}}$] abrissem o livro na página 16$_{\text{RESULTING_ACTION}}$].
[I$_{\text{MANIPULATOR}}$] would like$^{\texttt{Manipulate_into_doing}}$ [that [you$_{\text{PATIENT}}$] open the book on page 16$_{\text{RESULTING_ACTION}}$].

Visual aids, in turn, are annotated by means of bounding boxes which are labeled for frames, FEs and an additional label indicating the nature of the object inside the bounding box. Figure 2 shows one example of this type of annotation.

For the other layers described in Section 4.2 to be annotated, a new module had to be created in the WebTool. This module — Figure 3 — is organized in three panels:

- The top left panel has a video player from which annotators can watch the recordings from each of the cameras.

- The top right panel reproduces the video timeline and has four types annotation layers — interactant, modes, support and viewpoint. This panel records that, given the **viewpoint** of one of the participants — in the case of the example, the FOCAL_STUDENT —, one or more **interactants** — e.g. the TEACHER, the FOCAL_STUDENT and others — mobilize communicative **modes** — Body_movement, FACIAL_EXPRESSION, GAZE_DIRECTION, GESTURE, HEAD_MOVEMENT and SPOKEN_LANGUAGE — and a **support** — the TEXT_BOOK — in favor of meaning construction.

- The bottom left panel opens every time a new label is added to the panel on the right. In this panel, annotators can specify sublabels, associate frames and FEs to them and also indicate which visual element in the video triggers the annotation. In the example, the teacher's Deictic gesture of pointing to the book while speaking the sentence in (2) is annotated for the `Manipulate_into_doing` frame, with the TEACHER label for the **Interactant** layer being labeled as the MANIPULATOR FE. In addition, the correlation between the pointing gesture and the **support** is also noted, and the fact that the element of the image triggering the gesture annotation is the teacher's finger is noted in the CV Name field.[4]

The interface organized in multilevel layers allows for the simultaneous capture of different semiotic modes mobilized by the participants. Also, layers can be connected to one another, so that the mode layer can be associated with both the interactant and the support layers. Each interactive event is also associated with a frame and its corresponding FEs, which are evoked by these events. Additionally, a CV name is assigned to the image delimited by a bounding box, incorporated into a

[4]The CV Name field identifies the entity in the image as it would be labeled by a standard computer vision (CV) algorithm.

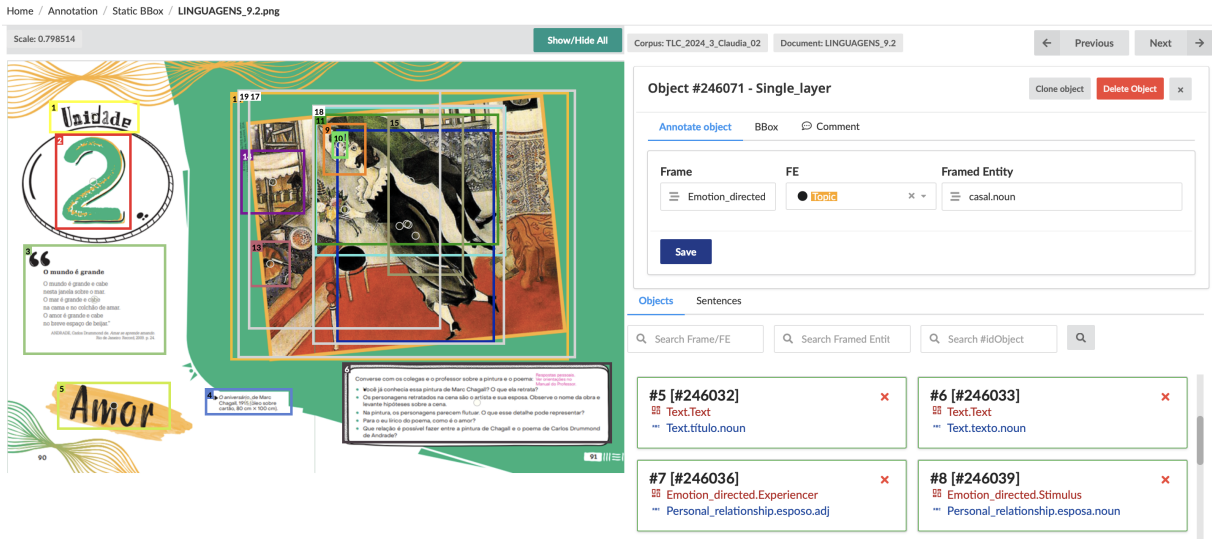

Figure 2: The visual annotation module

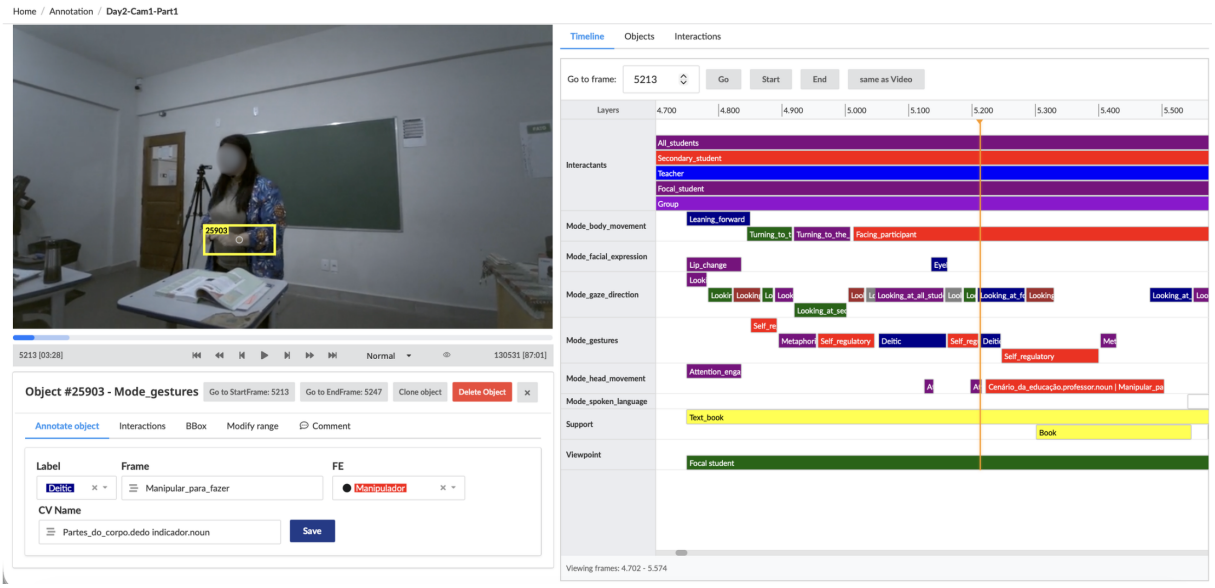

Figure 3: The canvas annotation module

frame-evoking LU. This bounding box spans the timeline until the event — for example, a gesture — ends.

Note also that the fact that frame, FE and CV Name labels are nested within the labels in the multiple layers allows the annotator to capture the complexities of multimodal communication in an organized fashion. In the example, the annotation scheme implemented in the WebTool records:

- In the BODY_MOVEMENT layer, that the TEACHER, annotated as the Manipulator in the `Manipulate_into_doing` frame is in the FACING PARTICIPANT position;
- In the GESTURE layer, that she is using her finger in a DEICTIC gesture to point to the TEXT_BOOK while using the SPOKEN_LANGUAGE mode in the form of the sentence in (2);
- In the GAZE_DIRECTION layer, that, some seconds later, the TEACHER will reinforce the instruction given in (2) by LOOKING_AT_FOCAL_STUDENT first, then by LOOKING_AT_SECONDARY_STUDENT and finally by LOOKING_AT_SUPPORT, that is, at the TEXT_BOOK.

Hence, the annotation methodology presented integrates two multimodal analysis methodologies, allowing for both the articulation of multilayers and the annotation of frames. As Fillmore (1985) argues, frames structure the understanding of events,

defining the roles and relationships between participants and objects. In our methodology, each interactive event evokes frames that organize the perception of meaning and articulate the mobilized semiotic modes, ensuring that the analysis captures not only the semiotic modes in isolation but also their potential for meaning-making within the communicative situation.

5. Conclusion

In this paper, we present an annotation scheme and a multimodal dataset of classroom interactions. Our work advanced in the construction of an annotation tool with multilevel layers, integration between the mode, interactant, and support layers, and association of events with frames and their elements, enabling the detailed capture of movements, gestures, speech, and support resources from a temporal and interactional perspective. These records will allow not only the qualitative analysis of interactions, but also an analysis of the frames evoked during the lessons. The annotation scheme also incorporates perspective, since one same interactional event can be annotated from different points of view.

Although the methodology has been devised for classroom settings, it can be extended to other types of interaction scenarios. As next steps, we plan to conduct inter-annotator reliability tests, evaluate the robustness of the scheme, and refine the tool to allow for more precise and replicable annotation. In the long term, our goal is to consolidate the methodology and datasets deriving from it as a gold-standard resource, which can be used in research on multimodality, classroom learning, and Machine Learning applications, such as the automatic analysis of semantic roles.

6. Ethical considerations and limitations

Research presented in this paper was approved by the Ethics in Research Committee of the Federal University of Juiz de Fora, process number 87876425.9.0000.5147. The research protocol includes strategies for participant anonymization. All students and the teacher taking part in the recorded classes, as well as the students' parents, signed a term of consent for volunteer participation in the experiment.

All annotation used in the experiments, including the revising of the audio transcription, was carried out by trained annotators who were paid a monthly stipend, which is, at least, equivalent to the minimum wage according to local regulations. All annotators involved in the annotation of the corpus used in the evaluation experiment reported here are co-authors of this paper.

Among the limitations of the methodology described, it is worth noting that the annotation categories were defined based on the configuration used for video-recording the classes. Different camera configurations may require different annotation categories.

7. Acknowledgements

Research reported in this paper was developed under the ReINVenTA—Research and Innovation Network for Vision and Text Analysis of Multimodal Objects—initiative, 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). The resulting dataset will be part of the data collection of the National Science and Technology Institute for Responsible Artificial Intelligence, Computational Linguistics and Information Treatment and Dissemination (INCT-TILDIAR, CNPq grant 408490/2024-1). Ferraz was supported by CAPES/PDSE (grant 88881.127655/2025-01). Torrent is a CNPq research productivity grantee (grant 311241/2025-5).

8. Bibliographical 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 *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

John Bateman, Janina Wildfeuer, and Tuomo Hiippala. 2017a. *Multimodality: Foundations, research and analysis–A problem-oriented introduction*. Walter de Gruyter GmbH & Co KG, Berlin.

John A. Bateman. 2008. *The GeM Model: Treating the Multimodal Page as a Multilayered Semiotic Artefact*, pages 107–142. Palgrave Macmillan UK, London.

John A. Bateman and Chiao-I Tseng. 2023. Multimodal discourse analysis as a method for revealing narrative strategies in news videos. *Multimodal Communication*, 12(3):261–285.

John A Bateman, Francisco OD Veloso, Janina Wildfeuer, Felix HiuLaam Cheung, and Nancy Songdan Guo. 2017b. An open multilevel classification scheme for the visual layout of comics and graphic novels: Motivation and design. *Digital Scholarship in the Humanities*, 32(3):476–510.

John A. Bateman, Janina Wildfeuer, and Tuomo Hiippala. 2026. *Multimodality. A Hands-On Guide*. de Gruyter.

Frederico Belcavello. 2023. *FrameNet Annotation for Multimodal Corpora: devising a methodology for the semantic representation of text-image interactions in audiovisual productions*. Ph.D. Thesis in Linguistics, Universidade Federal de Juiz de Fora, Juiz de Fora.

Frederico Belcavello, Tiago Timponi Torrent, Ely E. Matos, Adriana S. Pagano, Maucha Gamonal, Natalia Sigiliano, Lívia 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.

Frederico Belcavello, Marcelo Viridiano, Alexandre Diniz da Costa, Ely Edison da Silva Matos, and Tiago Timponi Torrent. 2020. 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*, pages 23–30, Marseille, France. European Language Resources Association.

Frederico Belcavello, Marcelo Viridiano, Ely Matos, and Tiago Timponi Torrent. 2022. Charon: A FrameNet annotation tool for multimodal corpora. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 91–96, Marseille, France. European Language Resources Association.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Tamara Drummond and Janina Wildfeuer. 2020. *The Multimodal Annotation of Gender Differences in Contemporary TV Series*, pages 35–58. De Gruyter, Berlin, Boston.

Laurent Filliettaz, Stéphanie Garcia, and Marianne Zogmal. 2022. Video-based interaction analysis: A research and training method to understand workplace learning and professional development. In Michael Goller, Eva Kyndt, Susanna Paloniemi, and Crina Damşa, editors, *Methods for researching professional learning and development: Challenges, applications and empirical illustrations*, pages 419–440. Springer.

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

Charles J. Fillmore. 1977. *The Case for Case Reopened*, pages 59 – 81. Brill, Leiden, The Netherlands.

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

Maucha Andrade Gamonal, Tiago Timponi Torrent, Ely Edison Matos, Adriana S. Pagano, Frederico Belcavello, Fl avia Affonso Mayer, Arthur Lorenzi, Natália S. Sigiliano, Helen de Andrade Abreu, Lívia Vicente Dutra, Marcelo Viridiano, André Coneglian, Victor A. S. Herbst, Franciany O. Campos, Kenneth Brown, Lívia Pádua Ruiz, Lisandra Carvalho Bonoto, Luiz Fernando Pereira, and Yulla Liquer Navarro. 2025. Audition: A Frame-Annotated Multimodal Dataset for Accessible Audiovisual Content. In *Proceedings of the 21st Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-21) including of a contribution to the Second Workshop on Multimodal Semantic Representation (MMSR II)*, page 95—104.

Maciej Grzenkowicz and Janina Wildfeuer. 2025. Addressing tiktok's multimodal complexity: a multi-level annotation scheme for the audiovisual design of short video content. *Digital Scholarship in the Humanities*, 40(4):1143–1166.

Michael A. K. Halliday and Christian M. I. M. Matthiessen. 2004. *An Introduction to Functional Grammar*, 3rd edition. Edward Arnold, London.

Tuomo Hiippala. 2015. *The Structure of Multimodal Documents: An Empirical Approach*, volume 13 of *Routledge Studies in Multimodality*. Routledge, London.

Jana Pflaeging, Janina Wildfeuer, and John A Bateman. 2021. *Empirical multimodality research:*

Methods, evaluations, implications. Walter de Gruyter GmbH & Co KG, Berlin.

Mariane de Carvalho Pinto. 2025. Anotação Multimodal para Copilotos de Produção de Tecnologias Assistivas: uma proposta para a audiodescrição. M.A. Thesis in Linguistics, Universidade Federal de Juiz de Fora, Juiz de Fora.

Josef Ruppenhofer, Michael Ellsworth, Miriam Petruck, Christopher Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute (ICSI).

Natália Sathler Sigiliano. 2025. Multimodal Frame Semantics: Expanding the Analytical Categories of FrameNet Brasil Multimodal Datasets. *Caligrama: Revista de Estudos Românicos*, 30(1):110–138.

Chinchu Thomas. 2018. Multimodal teaching and learning analytics for classroom and online educational settings. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 542–545.

Martin Thomas. 2009. Developing multimodal texture. In Eija Ventola and Arsenio Jesús Moya Guijarro, editors, *The world told and the world shown: multisemiotic issues*. Palgrave Macmillan, Basingstoke.

Tiago Timponi Torrent and Michael Ellsworth. 2013. Behind the Labels: Criteria for Defining Analytical Categories in FrameNet Brasil. *Revista Veredas*, 17(1).

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*, 13.

Tiago Timponi Torrent, Ely Edison da Silva Matos, Alexandre Diniz da Costa, Maucha Andrade Gamonal, Simone Peron-Corrêa, and Vanessa Maria Ramos Lopes Paiva. 2024. A flexible tool for a qualia-enriched FrameNet: the FrameNet Brasil WebTool. *Language Resources and Evaluation*, pages 1–29.

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.

Janina Wildfeuer and Joseph Adika Coffie. 2022. # socialiseresponsibly. analyzing the rhetorical structure of heineken tv commercials during the pandemic. *Frontiers in Communication*, 7:887706.