v. 30, n. 1, p. 110–138, 2025
ISSN: 2238-3824 · DOI: 10.35699/2238-3824.2025.57570
Submetido em: 14/02/2025 · Aprovado em: 31/03/2025

caligrama
revista de estudos românicos

Multimodal Frame Semantics: Expanding the Analytical Categories of FrameNet Brasil Multimodal Datasets

Semântica de Frames Multimodal: expandindo as categorias de análise dos datasets multimodais da FrameNet Brasil

Natália Sathler Sigiliano
Universidade Federal de Juiz de Fora (UFJF) | Juiz de Fora | MG | BR
natalia.sigiliano@ufjf.br
https://orcid.org/0000-0002-8460-5546

Abstract: In this paper, I propose an extension of the joint attention scenes models to include mediated multimodal communicative settings. From this extension, I discuss how the FrameNet analytical model can be used for the annotation of multimodal genres so as to include categories capable of accounting for invited shifts in joint attention in visual narratives. I revisit the literature on deixis and claim that, despite the fact that much of the foundational research on this topic proposes correlations between deixis in verbal language and other semiotic modes, the focus of research lies on the linguistic material in which the expression of deixis is grounded. By bringing together contributions from human communication studies, from the analysis of deixis in narratives and from Frame Semantics, I claim that semiotic devices of different types can be mobilized in filmic narratives to invite viewers to promote shifts in the center of the joint attention scene. To illustrate the proposed typology, I provide example analyses of deictic center shifts in films and describe the annotation methodology incorporated to the FrameNet model to account for them.

Keywords: deixis; joint attention; multimodality; Frame Semantics; filmic narrative.

Resumo: Neste artigo, proponho uma extensão dos modelos de cena de atenção conjunta para incluir cenários comunicativos multimodais mediados. A partir dessa extensão, discuto como o modelo analítico da FrameNet pode ser usado para a anotação de gêneros multimodais, de modo a incluir categorias capazes de explicar as mudanças convidadas na atenção conjunta em narrativas visuais. Revisito a literatura sobre dêixis e afirmo que, apesar de grande parte da pesquisa fundamental sobre esse tópico propor correlações entre dêixis na linguagem verbal e outros modos semióticos, o foco da pesquisa está no material linguístico no qual a expressão de dêixis está fundamentada. Ao reunir contribuições dos estudos de comunicação humana, da análise de dêixis em narrativas e da Semântica de Frames, proponho que dispositivos semióticos de diferentes tipos podem ser mobilizados em narrativas fílmicas para convidar os espectadores a promover mudanças no centro da cena de atenção conjunta. Para ilustrar a tipologia proposta, apresento exemplos de análises de deslocamentos do centro dêitico em filmes e descrevo a metodologia de anotação incorporada ao modelo FrameNet para explicar esses deslocamentos.

Palavras-chave: dêixis; atenção conjunta; multimodalidade; Semântica de Frames; narrativa fílmica.

1 Introduction

More than one century ago, Ferdinand de Saussure (1916) defined the object of Linguistics as a system composed of linguistic signs – langue – separating the field from other areas of study, such as Psychology and Semiotics. More recently, however, as the discussion on the relation between language use and multimodality advances, research has been emphasizing the need for a multidisciplinary approach to semantics, given that the analysis of verbal text as the sole system for meaning construction is not enough. This is not to say that phenomena beyond verbal text analysis have been ignored by linguists, rather, taking into consideration the field to which they are circumscribed, there has been a natural emphasis on the proposition of analytical categories based on verbal text. Nonetheless, the multimodal turn in Linguistics has been highlighting the need for rethinking the emphasis on verbal text. Such a communicative mode should be taken as one of the aspects of language, and the multiple semioses mobilized for using language for meaning construction should be taken into consideration.

The multimodal turn alluded to above is present in several subfields of Linguistics. The New London Group has pioneered in highlighting the notion of multiliteracy. This concept broadens the idea of literacy by looking at elements other than verbal language, while introducing metalanguage to describe and interpret the design elements of different modes for meaning construction (The New London Group, 1996). Analytical apparatus for multimodal phenomena built on contributions from diverse fields of research, such as Semiotics, Communication Studies and Film Analysis, has been proposed by the Bremen-Groningen group led by Bateman – see Bateman, Wildfeuer and Hiippala (2017) for an overview. Also, building on the parallel architecture model proposed by Jackendoff (2002) for the analysis of verbal language, Cohn proposes a "grammar" for image comprehension, applying it first for comics (Cohn, 2013) and later extending it to film analysis (Cohn, 2016b).

In Cognitive Linguistics, research on co-speech gesture – see, among others, Sweetser (2007), Steen et al. (2018) and Cienki (2022) – has expanded the analytical coverage of models such as Mental Spaces and Blending Theory and Construction Grammar. Such an expansion is a natural path in Cognitive Linguistics, since, as pointed out by Turner (2018, p. 357):

> Construction Grammar [in particular and Cognitive Linguistics in general] accepts responsibility to account for forms of creativity otherwise almost entirely ignored in linguistics. This commitment is wise, given that creativity is the engine that develops systems of communication.

As for Frame Semantics, since the original propositions of the theory, Fillmore (1976, 1982, 1985) has recruited the notion of visual scenes to explain the role of frames in organizing knowledge. Moreover, the proposed theory has highlighted the fact that frames play a role in understanding that goes beyond lexical semantics. At some point in the theory, distinctions between (lexical) semantic frames, interactional frames and cognitive frames (or scenes) have been proposed. However, those were later abandoned by Fillmore (2008) and merged into one single notion of frame. Nonetheless, such a diversity of possibilities for the application of frames to the analysis of meaning construction has been restricted – for the sake of feasibility – to lexical semantics, as the theory was implemented in FrameNet (Fillmore *et al.*, 2003).

Such an implementation restriction found a welcoming home in Computational Linguistics, which, back in the early 2000s, was exclusively centered on the analysis of strings of characters (see Dannélls et al., 2022). Therefore, the original FrameNet methodology had as its first step transforming text used in the analysis into those types of strings, stripping out of it all kinds of information that could not be encoded as UTF-8 characters. Such a method was – and still is to a large extent – standard in Computational Linguistics. Even considering that a multimodal turn has been also taking place in Computational Linguistics, the approach to multimodality is still very reductionist, to the extent that it mostly focuses on the compositional analysis of the elements depicted in the image. Such an analysis takes images as some sort of ground truth in reference to which the text can be analyzed (see Viridiano et al., 2022 for discussion).

More recent research has been extending the FrameNet methodology to the analysis of both textual genres (Dutra; Sigiliano, 2021) and static and dynamic images (Belcavello et al., 2020, 2022; Torrent et al., 2022; Viridiano et al., 2022; Belcavello, 2023). Those analyses have been uncovering correlations which are fundamental for the development of a Multimodal Frame Semantics. Nonetheless, they still take verbal language as a starting point, since they seek to associate entities depicted in the images to verbal language superimposed to them.

Considering the varied resources that the different communicative modes recruit for meaning construction, I claim that two additional extensions of the analytical framework of Frame Semantics – and FrameNet – are crucial: first, the very nature of the phenomena usually analyzed should be broadened; second, it is important to build multidisciplinary teams whose role will be that of defining analytical categories for multimodal texts.

In this context, I propose that Frame Semantics should be repositioned to account for key aspects of multimodal texts. Given the importance of deixis for meaning construction, in this paper, I will be focusing on the proposition of improvements to the Frame Semantics model, more specifically to its implementation as FrameNet, so that it can account for key deictic phenomena in multimodal texts.

2 Language and Deixis

In this section, I discuss the centrality of the notions of shared intentionality, shared attention and cooperative motivation to the model of language assumed in this paper. I later point out the connections between this model and the study of deixis, emphasizing that such a connection is not limited to linguistic conventions, but is also present in other communicative semioses.

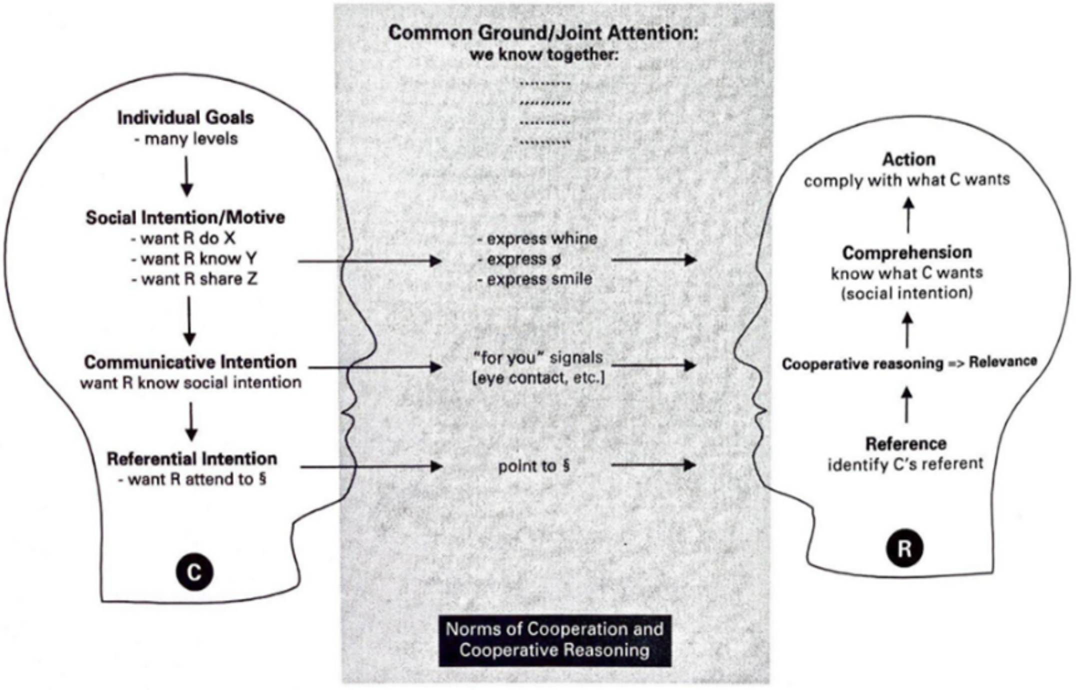
2.1 The Cooperative Foundations of Human Communication

Every linguistic use involves intentionality (Grice, 1957), and shared intentionality "presupposes a background sense of the other as a candidate for cooperative agency" (Searle, 1990, p. 414-415). As pointed out by Wittgenstein (1953), language is based on a non-linguistic structure comprising intentional knowledge and shared concepts. In this context, Tomasello (2008) proposes that the analysis of human communication must depart from the observation of non-conventionalized, non-coded communication, as well as from forms of connection other than language, such as natural gestures – e.g. pantomiming and nodding –, for example. Despite being simple and natural, such gestures reveal great communicative power.

Therefore, communication in general – be it linguistic or not – requires shared intentionality and a context of common knowledge. Tomasello (2018) proposes a model for human communication that involves different language manifestations. Although he frames one of the participants as the "recipient", Tomasello argues for the existence of cooperation norms and cooperative reasoning throughout human communication. Moreover, he claims that social, communicative and referential intentionalities are the individual goals of communication, as seen in Figure 1.

From Figure 1, communicative acts involving different semiotic modes can be present in any regular communicative setting. Smiles, gesture and eye gaze, for example, can be used for reinforcing the demonstration of intentionalities, according to the model proposed.

Figure 1- Summary of cooperative model of human communication



Source: Tomasello (2018, p. 98).
Caption: C = communicator; R = recipient.

Another important aspect of communication, according to Tomasello, is the necessity of a joint attention scene for linguistic learning to take place. Tomasello (2003) defines language as a joint action whose fulfillment is tied to shared attention. When discussing the stages of child development, Tomasello (2003) situates joint attention in the period when children start to perceive themselves, the other and the world surrounding them. At this point, children also start to recognize their interlocutors as intentional agents. Both the child and their interlocutor may intentionally define external referents and activities that compose the joint attention frames. Also, children understand the role fulfilled by themselves, their caregiver and the external referent, as well as the interchangeability of such roles, allowing them to adopt an external perspective to build concepts about themselves or to understand the intentions of caregivers when using linguistic symbols to manipulate their attention (Tomasello, 2003). Together with the ability to build shared attention frames, children develop communicative behaviors using triadic deictics, manipulating the attention of their interlocutor towards external referents, changing them (imperative gesture) or indicating them (descriptive gesture). Pointing gestures can be regarded, then, as the foundational manifestation of deixis in human communication. I will turn to them next.

2.2 Shared Intentions, Shared Attention and Deixis

The use of deictic gestures constitutes the first pillar of reference processes underlying the development of intentional communication (Goldin-Meadow, 2007; Liszkowski; Brown;

Callaghan; Takada; de Vos, 2012; Tomasello; Carpenter; Liszkowski, 2007). Tomasello, Carpenter and Liszkowski (2007) claim that:

> [...] these early pointing gestures are used not merely to direct attention either to the self or to an object, but to influence the mental states of others. [...] Pointing may thus represent a key transition, both phylogenetically and ontogenetically, from non-linguistic to linguistic forms of human communication (Tomasello; Carpenter; Liszkowski, 2007, p. 720).

The basic function of directing someone's attention to something through gesture is present in every known human society (Kita, 2003). Such an ability can still be used in support of language, in some contexts, after the latter is acquired (Tomasello, 2008). However, pointing gestures alone are not capable of expressing the whole range of intentionality. Tomasello (2018, p. 91) states that "[...] a pool of common ground between the pointing infant and her recipient, including shared assumptions about why she wants to establish joint attention in the first place" is required.

Rodríguez, Moreno-Núnez, Basilio and Sosa (2015) advocate in favor of the existence of ostensive gestures preceding and providing the foundation for the acquisition of deictic gestures. They argue that the gestures for showing – i.e. holding an object so that one can see it – and giving occur before that of pointing and are intentional. Wilkins (2003), when analyzing the gesture of pointing with the index finger as not being universal in sociocultural and semiotic terms, concludes that "[...] pointing (i.e., the use of some part of the body to make deictic gestural reference) appears to be universal. However, the use of the index finger for pointing does not appear to be universal" (Wilkins, 2003, p. 212). Therefore, regardless (a) of being used to refer to internal or external objects that are given to the interlocutor, (b) of the way those objects are presented to the interlocutor (with open hands or while pointing), or (c) of the motion associated with pointing, it is important to look at the pointing gesture as a signal of intentionality in the joint attention scene and as a foundation for human communication, coming before language acquisition.

The study of pointing translates into Linguistics as deixis. Levinson (2004) describes the issues revolving around how deixis is accounted for in Linguistics as follows:

> For those who want to treat language as a generative system for objectively describing the world, deixis is one hell of a big black fly in the ointment. Deixis introduces subjective, attentional, intentional and of course context-dependent properties into natural languages. Further, it is a much more pervasive feature of languages than normally recognized, and is theoretically puzzling in many regards. All this makes difficult a tidy treatment within formal theories of semantics and pragmatics (Levinson, 2004, p. 97).

Levinson (2004) differentiates deixis and indexicality, explaining that, while the first comes from the Linguistic tradition, the latter belongs to Philosophy. He uses indexicality to label broader phenomena of contextual dependency, and deixis to refer to narrower linguistic aspects of indexicality, namely, deictic or indexical expressions in language such as *you, here, now. He stresses the fact that indexicality cannot be reduced to such a study, since any referential* *expression can be used indexically. When discussing deixis, Levinson includes examples where this phenomenon includes elements beyond the limits of verbal language (Levinson, 2004, p. 106).*

Levinson's (2004) analyses indicate that a pragmatic resolution for deictics is crucial, because "there is a close relation between exophora and anaphora—in both cases we have contextual resolution of semantically general expressions, in the one case in the physical space-time context of the speech event, and in the other in the ongoing discourse" (Levinson 2000, p. 268 apud Levinson, 2004, p. 107). Moreover, he frames indexicality as both an intentional and attentional phenomenon, revolving around the spatio-temporal center of verbal interaction (cf. Bühler, 1934).

While discussing uses of linguistic expressions that are semantically insufficient for successfully producing reference without contextual support, Levinson (2004) explains that such a support is provided by interlocutors' mutual attention and by their ability to reconstruct the referential intentions indicated by the speaker through clues in context. Such clues include gesture or eye gaze, which may take part in the indexical sign. Nonetheless, such an analysis would not be enough to account for Bühler's Deixis am Phantasma ('deixis in the imagination'). In those cases, the deictic origo is shifted in a series of transpositions triggered by the process of imagining oneself at a different place. Deixis am Phantasma demonstrates how easily human cognition can change joint attention scenes (cognitively speaking) to, through the use of deictics, achieve intended goals.

Levinson (2004) lists yet other challenges for the analysis of deictics, which are related to the fact that deixis is much broader than what a description of purely verbal language aspects could capture, even when context is taken into consideration. This is because deixis is a cognitive tool for shifting attentional focus, which can be performed through language, gesture, gaze, body movement, among others, including even the simulation of spatial, temporal and eventive motion. Even analyses focusing on linguistic deixis recognize that other communicative modes and strategies can be recruited. When discussing the use of demonstratives in language, Levinson (2004) highlights that deictic expressions arise from a necessity to locate the intended reference, directing the attention of the interlocutor.

Similar discursive strategies are found in other authors discussing deixis in Linguistics. Bühler (1934) describes deixis as a mental imagery system, where an imaginary referential line along which elements are positioned is built in the mind of the interlocutor. Similarly, as per Hanks (2008), the deictic system of the language is composed of a field of spatial-temporal coordinates, which can be phenomenologically expanded to include the perceptual coordinates from constructive fantasy, namely from ontological memories of the past.

When differentiating deixis and non-deixis, Fillmore (1975) proposes an analogy where a sculpture is seen as non-deictic, since it can be observed in its entirety when standing on a square, for example. On the other hand, a photograph would be deictic, since a point of view was adopted when it was taken. The analogy proposed by Fillmore makes clear the possibility of extending the concept of deixis to other semioses. Moreover, while analyzing the linguistic expression of deixis, similarly to Levinson (2004) and the other authors discussed so far, Fillmore claims that certain lexical and grammatical items can only be interpreted when the sentences where they occur are understood against some social context, defined so as to identify the participants in the communicative act and their location in space and time for the duration of said act.

Those contributions demonstrate that the study of deixis should not be restricted to the analysis of verbal language. Kita (2003, p. 325) corroborates this claim by proposing that "the coordination of gaze, torso movement, gesture, and speech is motivated by communicative factors but also by the interlinkage among various cognitive processes". She proceeds by stating that "Both interactional and cognitive sides of the story need to be taken into account in order to reach a full understanding of various body movements and speech as a unified system of communication" (Kita, 2003, p. 327). Therefore, looking at deixis as a phenomenon involving multiple physical, linguistic and cognitive tasks provides clues for understanding its role beyond strictly linguistic processing. Deixis plays a key role in our ability to read the world and understand diverse interaction situations, during which we use deictics to indicate our intentions.

Theoretical constructs central to the cognitive models of human communication – such as intentionality and the ways to express it, joint attention and cooperation – and their relation to deixis are therefore also relevant for the analysis of multimodal text. This is so because in a text the choices reflected on its organization – be they linguistic, typographic, musical, or imagetical, among others – are intentional and rely on the cooperation between author and reader via text to be properly understood. In other words, deixis plays a key role in manipulating the reader's attention so that they can understand the intentions motivating the text and reconstrue the joint attention scene needed for cooperative communication to take place.

Therefore, concerning multimodal text and, more specifically in this paper, filmic narratives, I claim that expanding our understanding of deixis can provide means to analyze how deictic operators trigger cognitive rearrangements that allow comprehenders to construe meaning from sequences of events. To investigate this claim, I will propose an analytical model built upon the existing work developed by FrameNet Brasil for the annotation of multimodal text for Frame Semantics. Multimodal texts abound with elements functioning as clues inviting the comprehender to shift the deictic center of the joint attention scene. I will turn to the relation between deixis and filmic narrative next.

3 Multimodal Filmic Narrative and Deixis

It is a fact that multimodal text has always existed and that we are immersed in multimodality in our everyday lives. It is also a fact that multimodal features in text have been enriched and made more available with the advancement of technology (Bateman, 2008). Multimodality, according to Bateman, Wildfeuer and Hiipala:

> [...] is a way of characterizing communicative situations (considered very broadly) which rely upon combinations of different 'forms' of communication to be effective—the TV programme uses spoken language, pictures and texts; the book uses written language, pictures, diagrams, page composition and so on; talking in the cafeteria brings together spoken language with a host of bodily capabilities and postures; and the computer game might show representations of any of thesethings and include movement and actions as well (Bateman; Wildfeuer; Hiipala, 2017, p. 7).

Such a concept makes the complexity of multimodal phenomena explicit, especially if we consider the diversity of resources, textual genres and communicative settings involved in the analyses of the phenomena. Multimodal analysis is not about putting communicative modes side by side in a sum. Rather, it is a multiplication operation (Lemke, 1998). Analyzing such multiplying factors in context – aiming at describing them at work – is crucial for understanding multimodal phenomena, despite the associated challenge, which is "to find ways of characterising the nature of such interdependencies and to develop methodologies for investigating them empirically" (Bateman; Wildfeuer; Hiipala, 2017, p. 17).

Besides the multiplication of meaning construction possibilities emerging from the combination of communicative modes, research points to the sharing of meaningful properties across modes – see Jackendoff and Lerdahl (2006) for language and music, Steedman (2002) for language and body action, and Cohn (2016a) for language and visual sequences in comic strips. In this context, investigating, as proposed by Belcavello *et al. (2020), how different modes interact during human comprehension of multimodal genres is key.*

I subscribe to their claim that Frame Semantics is an adequate model for representing such interactions. In this theory, "meaning is relativized to scenes" (Fillmore, 1977, p. 59), and, although this maxim has been largely explored only in relation to language, it does not mean that it cannot be extended to other communicative modes. Sound, image and language, representative of distinct modes, bear similarities in their organization and in how they are perceived cognitively. If we consider sound, for example, it allows humans to gather information on space, distance, directions, emotional and physical states, among other features. Our cognitive capacity for the perception of sound sequences also allows humans to remember musically arranged sequences and predict harmonic progression (Bateman; Wildfeuer; Hiipala, 2017, p. 28). The knowledge humans acquire by experiencing textual genres throughout life also allows them to make predictions about language use patterns (Bakhtin, 1984). Therefore, it is only natural that the confluence of communicative modes in a multimodal genre allows humans to combine different types of formal clues in the process of re-construing intended meaning.

Frames work precisely in this process. They provide schematic structures that can be accessed and brought into play by comprehenders for making sense of varied formal clues found in (multimodal) text. When we isolate communicative modes, we are surely still able to analyze how each type of formal pattern – linguistic, musical, visual – can serve as a clue for activating the frames needed for interpretation. However, important aspects of the meaning construction process will be lost. Hence, approaching multimodal genres requires trying to look at all the diversity of communicative modes involved, and such an effort, although challenging, is crucial for the adequacy of the analyses.

Therefore, even phenomena that have been traditionally associated with their linguistic manifestations, such as deixis, must be considered under the perspective of the combination of communicative modes for meaning construction. When it comes to studying filmic narratives, this proposition can be pervasively illustrated. Fillmore (1975) already recognized deictic centering to be both more general and more abstract than the linguistic material pointing to shifts in deixis, such as here, now and you. While buy and sell and loan and borrow can demonstrate the perspectival nature of linguistic symbols (Fillmore, 1975; Tomasello, 1999), go and come can do so for the abstractness of the deictic center in language (Fillmore, 1975). In all those cases, verb alternations encode different vantage points adopted while describing the relations between the participants in the scene, or whether motion happens toward or away from the speaker. Similarly, shot composition, camera positions and editing choices in a filmic narrative can serve as clues for meaning construction.

It is thus necessary to recognize that deixis goes beyond the linguistic expression of categories such as person, space and time, for example. It is a foundational phenomenon in human perception and in the organization of semioses contributing to our ability to make sense of the world from constantly rebuilding (joint) attention scenes. This is so because we engage in meaning construction by individually observing the world and sharing our perspective on it through joint attention scenes, which involve not only individual, but shared intentionalities. Through pointing – which can manifest via language, gesture, gaze, body motion – we can intentionally guide our interlocutor to new joint attention scenes. Therefore, motion and change are also foundational to approaches to meaning construction involving multiple semioses.

How does pointing occur in filmic narratives? In this case, would it be limited to the analysis of gesture and language? How are intended attention foci defined and how is the need for viewers to reframe their attention focus signaled? Those are the questions I aim to investigate.

3.1 Deixis in Narrative

Segal (1995), when approaching the relation between narrative and deixis, proposes that:

> when one reads a narrative as it is meant to be read, he or she is often required to take a cognitive stance within the world of the narrative. A location within the world of the narrative serves as the center from which the sentences are to be interpreted (Segal, 1995, p. 15).

In this scenario, deixis is then seen, as per Zubin and Hewitt (1995), as the structuring framework around which narrative emerges. These authors proceed with the following claim:

> stories are made possible because readers can import knowledge of the everyday world and of other possible worlds into the current story world; this provides the listener/reader with the illusion of mentally inhabiting a fully specified and coherent world (Zubin; Hewitt, 1995, p. 130).

Zubin and Hewitt (1995) situate deixis in narrative from the perspective of the Deictic Center (DC) Theory, which attempts to model the consequences of deictic shifts beyond the here and now of real-world interaction and into fictional text. Such a theory is founded on Deictic Shift Theory (DST), which, in turn, claims that the DC frequently changes from the here and now to a locus in the mind model representing the discourse world (Segal, 1995). According to this later theory, in fictional narrative, readers and authors change their deictic centers and cognitively project themselves to a place inside the world of the story, taking over other deictic centers. According to Segal's (1995) view, the DC is a structure providing coherence to a text, even when it is not directly represented in the lexicon or syntax, and which changes according to the progression of the story.

Zubin and Hewitt (1995) illustrate deictic change in narrative through the observation of typical oral storytelling styles, stating that:

> The story is not addressed to the audience in the way conversation or a lecture is; rather, it opens a conceptual window through which the story world can be glimpsed. The story is self-enclosed. Its deictic structure presupposes its own story world, and not the current interactional context of the teller and audience. In fact, the listener's deictic perspective becomes the one chosen for him or her by the teller. In a successful story, we have the illusion of experiencing the fictional world directly, because we unconsciously adopt the deixis of the DC as our own (Banfield, 1982; Wiebe, this volume) (Zubin; Hewitt, 1995, p. 131).

The illustration provided by Zubin and Hewitt aligns with Turner's (2017) idea of Blended Classic Joint Attention (BCJA) Scenes. According to Turner, the everyday experience of watching the news on TV requires a complex blend:

> The news anchor is not actually in a scene of classic joint attention with the viewer; [...] *here* for the participants in the news interaction is not actually a single shared space ("It's good to have you here," says the news anchor, but where is "here"?); *now* for the participants in the news interaction need not be a particular moment ("Now we have a special announcement coming up for you here," says the news announcer, but perhaps it was recorded, perhaps it is meant to be viewed at many different times, perhaps the announcer did not even know what the special announcement would be; in addition, who is "we," and again, where is "here"?). But we can blend all these elements into a scene of *blended classic joint attention*, which is tractable and familiar because it draws on our understanding of classic joint attention. Most of the language that is available for running a scene of classic joint attention can be projected, adapted, and used for BCJA. BCJA is a generic integration template—that is, a well-known general pattern of blending that can guide the mental construction of indefinitely many specific networks (Turner, 2017, p. 3).

Human ability to take over different DCs in narrative helps narrative progression, and the DC window provides the listener-reader with two shifting foci, defined by Zubin and Hewitt (1995) as "focalizing perspective" – origin – and "focalized perspective" – content. The first refers to the reader's point of view, to what is being shown to them. The latter to what can be seen inside the deictic window in terms of motion in space, time and people inside the story world, considering the purpose of the focalizing perspective.

To analyze how DCs are organized in written narratives, Zubin and Hewitt (1995) propose four basic concepts: WHO, WHERE, WHEN and WHAT, which are subject to several operations allowing the narrative to be introduced, maintained, changed or suspended during text. The authors map DC-devices in written text and define them as instruments for signaling DC stability or change, while maintaining textual cohesion. DC-devices manifest as either morphemic or syntactic structures building the DC and guiding the reader in reconstruing it (Zubin; Hewitt, 1995, p. 141). They also propose that deictic operations in the DC are carried out by the reader-listener during the process of interpreting a narrative text, including: (a) introducing people, objects, places and time; (b) maintaining the stability of the DC; (c) changing the WHO, WHAT, WHEN or WHERE to another one; (d) voiding one or more of the components of the DC, which become no longer relevant at that point of the narrative. Finally, Zubin and Hewitt (1995) highlight that listeners-readers build DCs not only from devices present in the text, but also based on shared world knowledge and on cooperative principles.

Analyses, propositions and descriptions such as the ones summarized in the previous paragraphs are crucial for understanding narrative structure. Nonetheless, their scope is limited to language. In a context where multimodal filmic narratives are in focus, there is still a lack of studies and models aimed at accounting for the correlations between linguistic and non-linguistic aspects of deixis. The next section presents a proposal for that area.

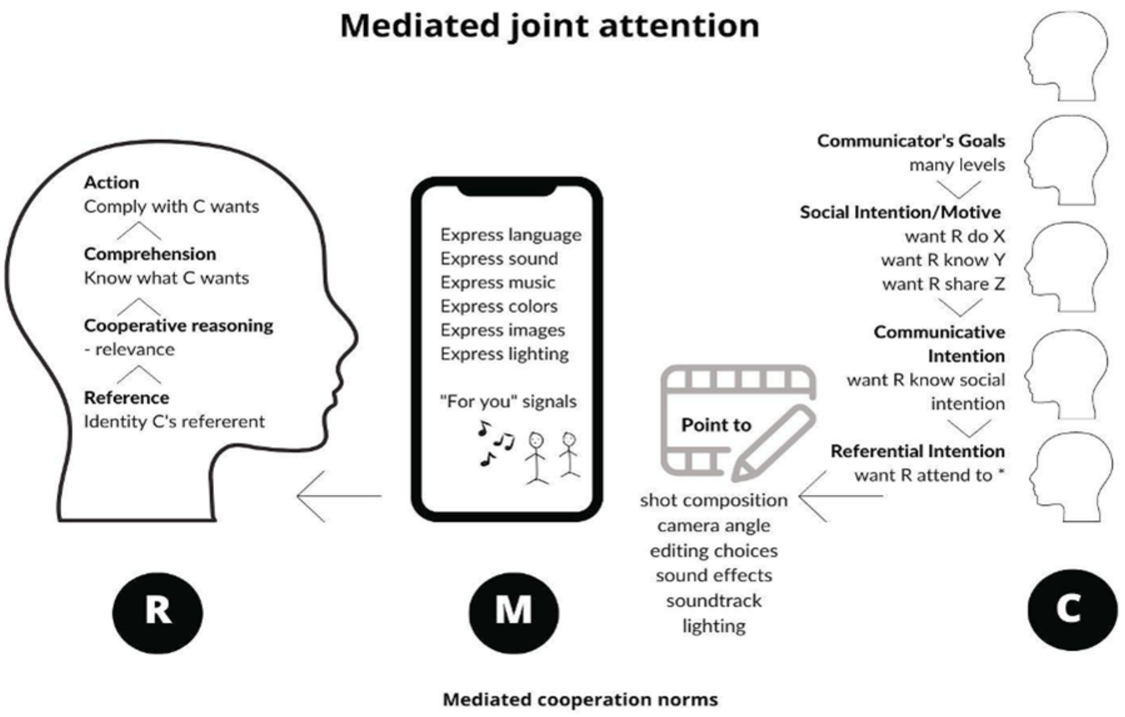
3.2 A Model for Analyzing Multimodal Deixis

Based on the cooperative human communication model by Tomasello (2018) – Figure 1 – and drawing on Turner's (2017) notion of BCJA, I propose a new model – Figure 2 – where multiple communicators (authors) choose a screen – the medium – (M) to occupy the position of the communicator (C). Together with the comprehender (R), and sharing intentionality with them, the joint attention scene is built and the comprehender of the multimodal text is expected to interact with it, deploying their knowledge about the world. The comprehender is supposed to construe their comprehension of the text by following the clues provided in the text by the multiple communicators using clues from different semioses, which are broadcast via the screen.[1]

The cooperation principle still holds between the communicator and the comprehender. However, it may be the case – and it usually is – that the comprehender is not aware of the communicators involved. There is a prominence of the text – broadcast via the screen – as an indicator of the communicative intentions. The situation is similar to that of reading a written novel, where the reader interacts directly with the narrative, where intentionalities have been indicated by linguistic material. In multimodal text, however, many other elements, besides the linguistic ones, are at play and operate on the DC, especially in filmic narratives. If, on the one hand, in written narratives, one has to pay attention to morphological, lexical and syntactic aspects to analyze deixis in narrative progression, in multimodal narratives, visual and sound aspects must be also analyzed.

[1] It may be the case that R's meaning construction ignores C's intentions completely. However, it does not mean that C has no communicative goals that they want R to recognize.

Figure 2 - Summary of model for mediated communication focusing on the role of DC shifts in meaning construction



Source: The author. Based on Tomasello (2018).

In this context, the comprehender engages in a mediated cognitive interaction with the communicators in the narrative world. In a multimodal joint attention scene mediated by devices, joint attention depends on the intentionality of the comprehender to be established. They must turn on the device and pay attention to it. Next, it is expected that the comprehender aligns their own deictic perspective to the one that was intentionally chosen by the communicators, who, throughout the narrative, establish joint attention micro scenes guiding textual understanding.[2]

While building these micro scenes, communicators use deictic devices which may or may not comprise language, and go beyond it, becoming multiplying agents in meaning construction. Such devices collaborate to the maintenance or alteration of the joint attention micro scenes. The alteration of the micro scenes prompts alterations in the frames invoked in the mind of the comprehender and promotes narrative progression.

As previously indicated, pointing – or deixis – is an important element in everyday joint attention scenes. We use pointing by means of several communicative modes related to

[2] Note that, in the case of filmic narratives, it is usually the case that the narrator is not present in the medium used to broadcast the story, which makes the process of building the blended joint attention scene more complex, to the extent that the projection of the dialogue between C and R requires the construction of an abstract or idealized C.

cognitive processes, to refer to an element external to the interaction, directing the attention of our interlocutor and marking intentionality. The coordination for a new attention focus of the interaction arises from deictic pointing actions. One of the participant's attention in the interaction is redirected to some specific point in space and in time. In filmic narratives, similarly, new joint attention scenes are previously established by communicators every time shifts in the DC are intended. Such shifts can be motivated by the WHO, WHAT, WHERE and WHEN features (cf. Zubin; Hewitt, 1995), and the forms they manifest linguistically have been described. Nonetheless, their manifestation in other semioses has not been the focus of descriptions and discussions. Considering the prominent role of deictics in the various semioses, in the means used for calling attention and indicating intentions, analyzing their manifestation in multimodal narratives is certainly relevant.

In the following section, I will further explain the model summarized in Figure 2, providing elicited and didactic examples of multimodal deictic devices in filmic narratives, especially in what concerns the establishment of joint attention micro scenes. I will also explain why the incorporation of such a model into FrameNet brings both new challenges and new possibilities.

4 Incorporating Deixis into Multimodal FrameNet Analysis: Challenges and Possibilities

Linguistics was defined as a science focusing on the analysis of verbal language. To do so, it considers the surrounding context of the element being analyzed to provide descriptions and explanations on several aspects of such elements: how they developed in time, how they are acquired by infants, how they interact with other elements in language, how they contribute to meaning construction, how they function in textual genres and communicative settings, and so on.

Genres take communicative intentionality as a foundational aspect of their organization, of what one can expect from them. Moreover, genres occur in communicative situations where they are relevant. Therefore, the genre a communicator chooses to express their intentionality is not fortuitous. In this context, genres can be regarded as large deictic elements signaling expected intentionalities to the interlocutor. In what specifically concerns filmic narratives, it is expected that one or more stories will be told, forming a plot, and that elements such as character, time, objects and space will interact to build the narrative. Any movement or change involving any of those elements may result in a shift in the narrative's deictic center. Such shifts are expected in a narrative, since the plot comprises constant reframing allowing for textual progression.

In Frame Semantics, as already pointed out in this text and elsewhere (Dannélls *et al.*, 2022), the analysis is mostly circumscribed to the lexicon and to grammatical constructions. Only recently has the model been extended to include analyses going beyond sentence level, allowing for the annotation of genre-relevant structure and multimodal text. Such an extension has brought new challenges to the model. This is because signaling meaning correlations between what is expressed in the text and what is expressed by the image is only part of the analysis of texts whose organization presupposes multimodality.

If, in a text purely composed by means of verbal language, the use of some specific lexical item may lead to the activation of a frame, in a multimodal text, such as a filmic narrative, the mere act of a character closing their eyes can (a) evoke a frame, (b) serve as a clue for the comprehender to invoke a frame, and (c) invite the comprehender to reframe their interpretation of the text, requiring them to cognitively move towards a new joint attention micro scene. Those three functions may also be associated with a lexical item, but in a much more conventionalized fashion. One of the parade examples of frame invocation by Fillmore (1985, p. 232) demonstrates (a) and (b). I reproduce the example in (1).

(1) We never open our presents until morning.

If this sentence were to be annotated in FrameNet, it would result in the evocation of the following frames: `Negation`, by *never.adv*; `Closure`, by *open.v*; `Giving`, by *present.n*; `Time_vector`, by *until.prep*; and `Calendric_unit`, by *morning.n*.[3] Nonetheless, anyone familiarized enough with the western culture would say that this is a sentence about Christmas, even though there is no LU evoking a frame for modeling the meaning of this festivity. This phenomenon is what Fillmore called frame invocation. In other words, while the LUs listed above are conventionally associated with the frames mentioned, there is an additional frame that plays a key role in the interpretation of (1) which is reconstrued by the comprehender without being grounded in any specific LU. I will claim that frame invocation is even more common in multimodal text. This is because the act of closing one's eyes, in a multimodal context, may evoke, for example, the `Body_movement` frame, but may also be a clue for the invocation of the Death or the `Fall_asleep` frames. Moreover, it can function as a device signaling the starting point for a new segment of the narrative, where the deictic center is moved towards another setting. However, two key aspects of semiotic modes other than language are at play in this elicited example.

First, it is important to consider that the eye closing per se is not as conventionally associated with the `Death` frame as the verb *to die* is. Language is a more conventionalized mode than the other ones involved in a filmic narrative. The character in question could be going to sleep, or passing out, or, if the shot is in slow motion, blinking. For the invocation of the `Death` frame to take place in this case, other clues are needed in a filmic narrative, such as where the character is located, what the comprehender knows about the character's previous health conditions, the soundtrack and lighting composition choices for the scene, among others. Therefore, the first important question to ask when it comes to the association of visual clues to frames is: which elements in the scene are responsible for that frame to be invoked in the comprehender's mind?

Second, while annotating multimodal text, it is also important to investigate the DC-devices that point to either the constitution of a new joint attention micro scene or to their maintenance in different textual genres. Such devices may be linguistic or not and must

[3] All frames mentioned in `Courier` font in this paper can be found at FrameNet Brasil Lab (c2014-2025). The use of such a font is a convention adopted in most FrameNet-related publications.

be accounted for in terms of their central contribution to the progression of the meaning construction process beyond a sentence or a sequence of shots in a film.

Therefore, while annotating multimodal text, mapping the correlations between the frames evoked in the image and those evoked in the text only covers part of the meaning construction process. In a film, the act of closing the eyes may invoke the death of a character or just a kid going to sleep after listening to a bedtime story. Regardless of which of the two is invoked, what comes next in the sequence will rely on the activation of new joint attention micro scenes to be interpreted so that the narrative can advance. In this context, several pragmatic elements and shared knowledge between the comprehender and the communicators would be brought into play to guide expectations regarding a new joint attention micro scene.

The act of closing the eyes, if performed by a character in the narrative, could have its invoked meaning multiplied by other elements in the multimodal composition, such as the slow rhythm in the background soundtrack, a blueish or noir color palette in the image, a teardrop on the face of another character, a dog lowering its ears. In that case, the act of closing the eyes would trigger the invocation of the Death frame and the sadness atmosphere would both be explained by such invocation and reinforce it.

It is also important to point out that, as much as the act of closing one's eyes may be associated with the invocation of different frames, diverse image sequences may also lead to the invocation of the `Death` frame. Suppose the filmic narrative, at some point, shows one of the characters taking care of a garden in what, at that point, appears to be a shot in an opening sequence that is not key for the narrative. At a later point of the narrative, a shot of an abandoned garden may serve as a deictic device situating the comprehender at a new attention focus, guiding them towards a new joint attention micro scene. Note that no linguistic or imagetic element would need to be used to explicitly evoke the Death frame in this case. The invocation of such a frame would derive from a reframing of the role of that initial gardening shot in contrast with the new aspect of the garden. Therefore, non-verbal deictic devices would have been used to signal to the comprehender the need to seek an explanation for the change in the state of the garden, identifying the possible intentionality behind the communicators' choice. If, from the opening sequence, the comprehender has learnt that the character takes good care of the garden, and now the garden is no longer in good shape, some element in the frame referring to the gardening process must be no longer present. In this case, the death of the gardener could be one of the hypotheses for the current state of the garden and the progression of the narrative would confirm or not the possibility of invoking the Death frame.

Of course, there is an individual dimension to the interaction between the comprehender and the text, meaning that the effects intended by the communicators may not be successfully achieved, while other ones which were not predicted may occur. However, this possibility is not exclusive to multimodal text and the analytical focus should rely on the intentionality behind the elements chosen to enable the progression of the narrative. Focusing on that will of course bring new challenges to FrameNet annotation, namely that of looking beyond the immediate co-text of the element evoking a frame and looking more closely into frame invocation, an aspect of Frame Semantics not properly explored in its computational implementation (Torrent *et al.*, 2022). Nonetheless, it could open a series of new possibilities in using FrameNet annotation for building semantic representations of multimodal genres which capture key aspects of textual progression.

In the final section of this paper, I will provide a sample analysis of real data based on the model sketched so far.

5 Sample Analysis

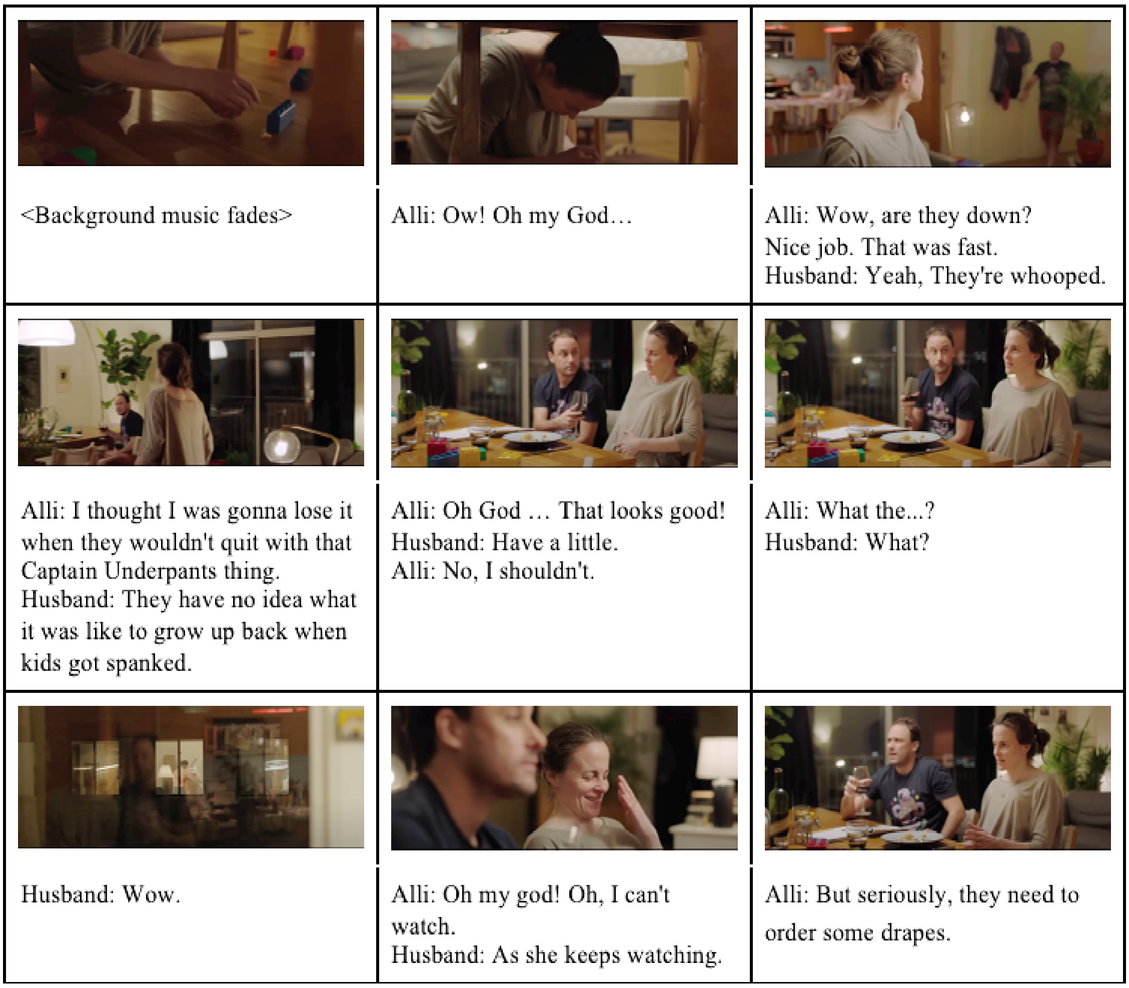
The incorporation of the mediated joint attention scene model and the consequent multimodal approach to deictic center shifts to FrameNet multimodal annotation requires new analytical categories to be included in the model. To demonstrate the feasibility of this proposal, in this section I analyze *The Neighbors Window*, an Academy winning short film by Marshall Curry.[4] I start by showing the aspects of multimodal composition currently accounted for by the FrameNet Brasil multimodal annotation model (Belcavello *et al.*, 2020, 2022; Belcavello, 2023), and then, present the additional analytical layers made possible by the proposal devised in this paper.

5.1 FrameNet Multimodal Annotation of Dynamic Images

For exemplifying the methodology proposed by Belcavello *et al.* (2022), I will consider the sequence of the short film in Figure 3. In the sequence, which is the first of the film, we see the main character Alli cleaning up the floor amidst kids toys. She bangs her head on the table when standing up. Next, her husband enters the room, after successfully putting the kids to sleep. The couple sits to have dinner, and Alli's attention turns to something in front of them. We realize next that it is a scene of her neighbors having sex.

[4] Available at The Neighbors (2019).

Figure 3: Sequence of stills for The Neighbors Window

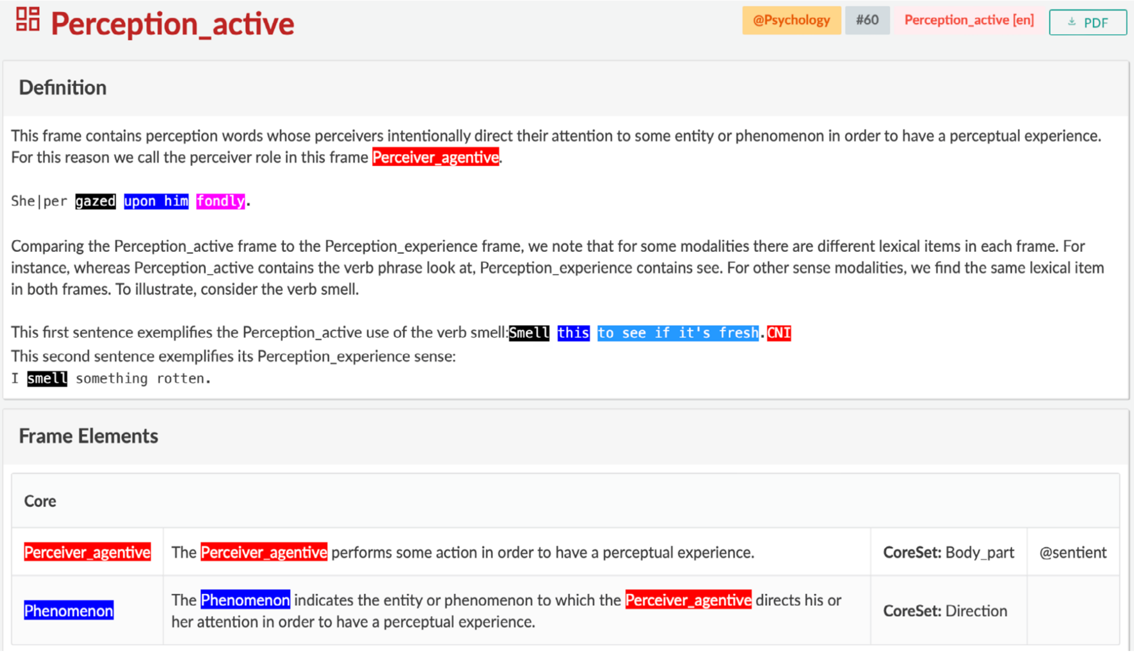


<Background music fades>	Alli: Ow! Oh my God…	Alli: Wow, are they down? Nice job. That was fast. Husband: Yeah, They're whooped.
Alli: I thought I was gonna lose it when they wouldn't quit with that Captain Underpants thing. Husband: They have no idea what it was like to grow up back when kids got spanked.	Alli: Oh God … That looks good! Husband: Have a little. Alli: No, I shouldn't.	Alli: What the...? Husband: What?
Husband: Wow.	Alli: Oh my god! Oh, I can't watch. Husband: As she keeps watching.	Alli: But seriously, they need to order some drapes.

Source: The Neighbors (2019).

Following Belcavello's (2023) methodology, the first step in the analysis is that of annotating the audio transcription for frames and frame elements (FEs). For instance, the sentence spoken by Alli when the eighth video frame in Figure 3 is shown could be annotated for the `Perception_active` frame, shown in Figure 4. This frame is evoked by *watch.v*. The resulting annotation is seen in Figure 5. Note that the Perceiver_agentive FE is assigned to I, while the Phenomenon is noted as a definite null instantiation, meaning that it can be inferred or found in the context of the annotation, but not in the syntactic locality of the lexical unit that is the target of the annotation. In this specific case, because this is a multimodal annotation, and annotators are instructed to analyze the text while watching the video, they know that both the Perceiver_agentive and the Phenomenon FEs will be shown in the video.

Figure 4: The `Perception_active` frame in FrameNet Brasil



Perception_active

Definition

This frame contains perception words whose perceivers intentionally direct their attention to some entity or phenomenon in order to have a perceptual experience. For this reason we call the perceiver role in this frame Perceiver_agentive.

`She|per gazed upon him fondly.`

Comparing the Perception_active frame to the Perception_experience frame, we note that for some modalities there are different lexical items in each frame. For instance, whereas Perception_active contains the verb phrase look at, Perception_experience contains see. For other sense modalities, we find the same lexical item in both frames. To illustrate, consider the verb smell.

This first sentence exemplifies the Perception_active use of the verb smell:`Smell this to see if it's fresh.CNI`
This second sentence exemplifies its Perception_experience sense:
`I smell something rotten.`

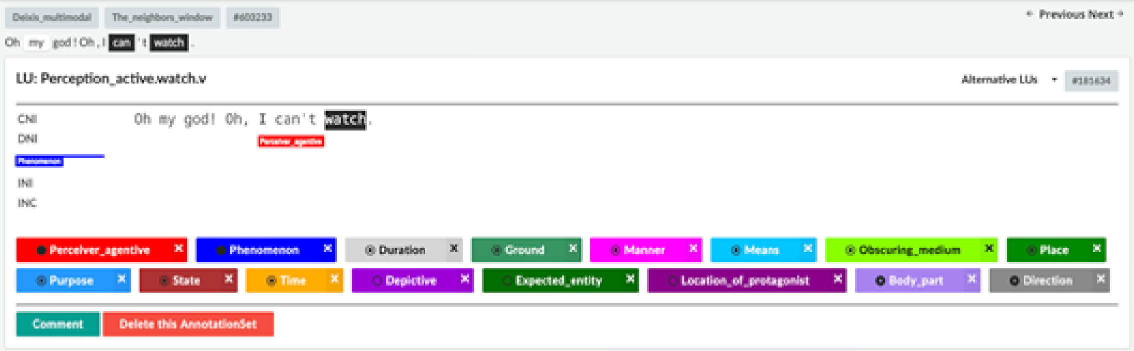
Frame Elements

Core

Perceiver_agentive	The Perceiver_agentive performs some action in order to have a perceptual experience.	**CoreSet:** Body_part	@sentient
Phenomenon	The Phenomenon indicates the entity or phenomenon to which the Perceiver_agentive directs his or her attention in order to have a perceptual experience.	**CoreSet:** Direction	

Source: FrameNet Brasil Lab (c2014-2025).

Figure 5: Sentence annotated for the `Perception_active frame`



Source: FrameNet Brasil Lab (c2014-2025).

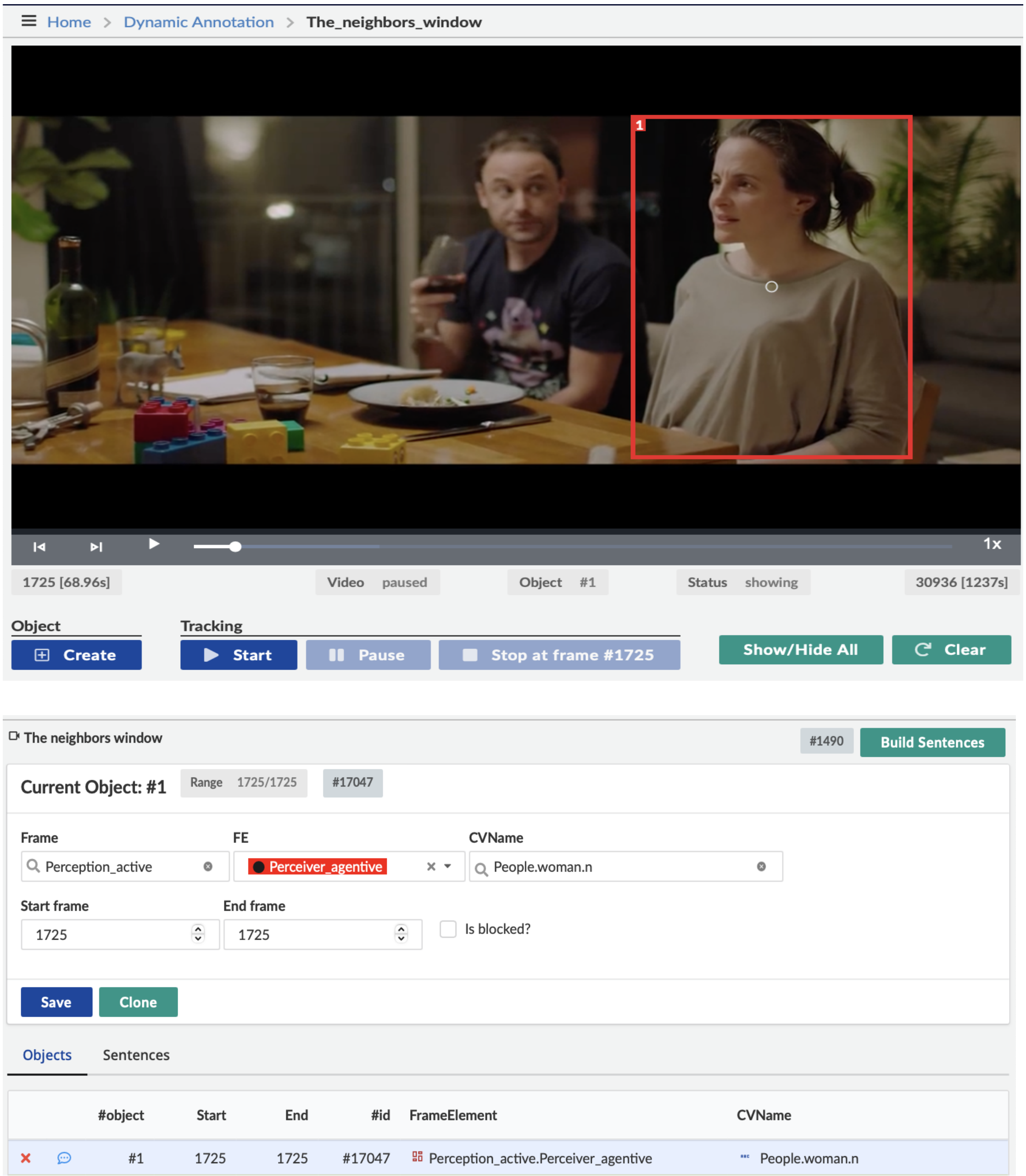
Once the transcribed audio is annotated for frames and FEs, the methodology states, as the next step, that the video sequences are annotated. The video annotation is shown in Figures 6 and 7. In Figure 6, the annotation represents the moment where Alli notes the neighbors having sex in the other building and focuses her attention on that scene. Note that, on top of recording that this portion of the image represents a FE in the `Perception_active` frame, the annotation also indicates that the object being annotated is a *woman.n*, a lexical unit of the People frame.

Next, in Figure 7, Objects 2 and 3 are annotated at the same portion of the image. Object 2 indicates the instantiation of the Sex frame and marks the Participants FE for this frame, which is labeled using the *couple.n* lexical unit evoking the `Personal_rela-`

tionship frame. Object 3 is a clone of Object 2 in terms of spatial and temporal properties in the video, but registers the fact that the sexual act happening in the other apartment is the Phenomenon FE of the `Perception_active` frame which was a definite null instantiation (DNI) in the text annotation.

As the example shows, the methodology proposed by Belcavello (2023) already captures several complexities of the meaning-making process involved in the interpretation of the filmic narrative presented to the comprehender. Next, I propose an augmentation of the analytical categories FrameNet Brasil uses for the annotation of multimodal genres to capture the role of deictic center shifts in the process of interpreting filmic narratives.

Figure 6: Annotation of the Perceiver_agentive FE as Object #1



1725 [68.96s] | Video paused | Object #1 | Status showing | 30936 [1237s]

The neighbors window #1490

Current Object: #1 Range 1725/1725 #17047

Frame: Perception_active

FE: Perceiver_agentive

CVName: People.woman.n

Start frame: 1725

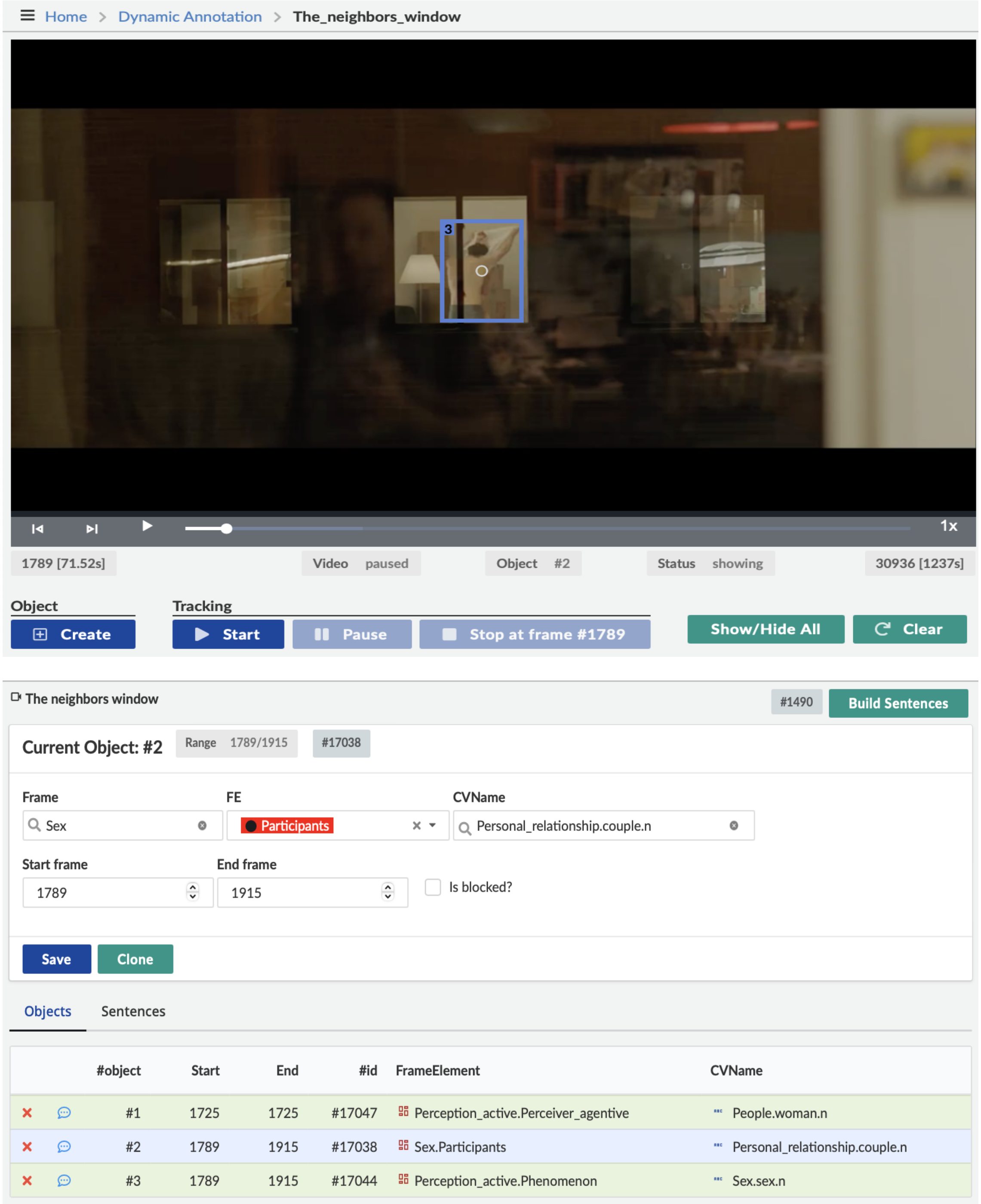
End frame: 1725

☐ Is blocked?

		#object	Start	End	#id	FrameElement	CVName
×		#1	1725	1725	#17047	Perception_active.Perceiver_agentive	People.woman.n

Source: FrameNet Brasil Lab (c2014-2025).

Figure 7: Annotation of the Participants FE in the `Sex` frame and of the Phenomenon FE in the `Perception_active` frame as Objects #2 and #3



Frame: Sex

FE: Participants

CVName: Personal_relationship.couple.n

Start frame: 1789

End frame: 1915

□ Is blocked?

		#object	Start	End	#id	FrameElement	CVName
×	💬	#1	1725	1725	#17047	Perception_active.Perceiver_agentive	People.woman.n
×	💬	#2	1789	1915	#17038	Sex.Participants	Personal_relationship.couple.n
×	💬	#3	1789	1915	#17044	Perception_active.Phenomenon	Sex.sex.n

Source: FrameNet Brasil Lab (c2014-2025).

5.2 Annotating for Deictic Center Shifts in FrameNet

The methodology for the annotation of deictic center shifts I present in this paper was based on the categories devised by Zubin and Hewitt (1995). Such categories, in turn, were based on studies focusing on written narratives and the role of deictic centers in their progression. They served as the basis for the definition of the types of deixis to be annotated. Next, the idea was to find ways through which changes in the deictic center could be annotated for different communicative modes. The notion of communicative modes presented by Kalantzis, Cope and Pinheiro (2020) proved to be a good foundation for the establishment of categories for annotation.

Nonetheless, the methodology still lacked granularity. For the visual mode, for instance, video editing choices and resources needed better definition. I used the work by Bordwell and Thompson (2013) as the basis for the definition of filmic categories for the visual mode, such as framing, cut, among others. The categories proposed are presented in Table 1.

Table 1: Categories for the annotation of deictic center shifts in filmic narratives

Macro category	Types	Trigger/ Object	Specification	How to annotate
Deictic center	Reader's DC Character's DC Reader's and character's DC			Checkbox
Deictic operations	DC Introduction DC Maintenance DC Change DC Annulment			Checkbox
Types of deixis	Place Person Time			Checkbox

Meaning -making mode	Gestural			Bounding box
	Tactile			Bounding box
	Spatial	Scene		Bounding box
		Region		
	Visual	Character		Bounding box
		Artifact		
		Camera angle	Horizontal	Checkbox
			High	
			Low	
		Shot	Wide	Checkbox
			Full	
			Cowboy	
			Medium	
			Medium close-up	
			Close-up	
			Extreme close-up	
			Insert	
		Camera motion	Horizontal pan	Checkbox
			Vertical pan	
			Crane	
			Traveling	
			Zoom in	
			Zoom out	
			Dolly in	
			Dolly out	
		Cut	Cut	Checkbox
			Fade in	
			Fade out	
			Fusion	
			Jump cut	
			Cut in	
			Final cut	

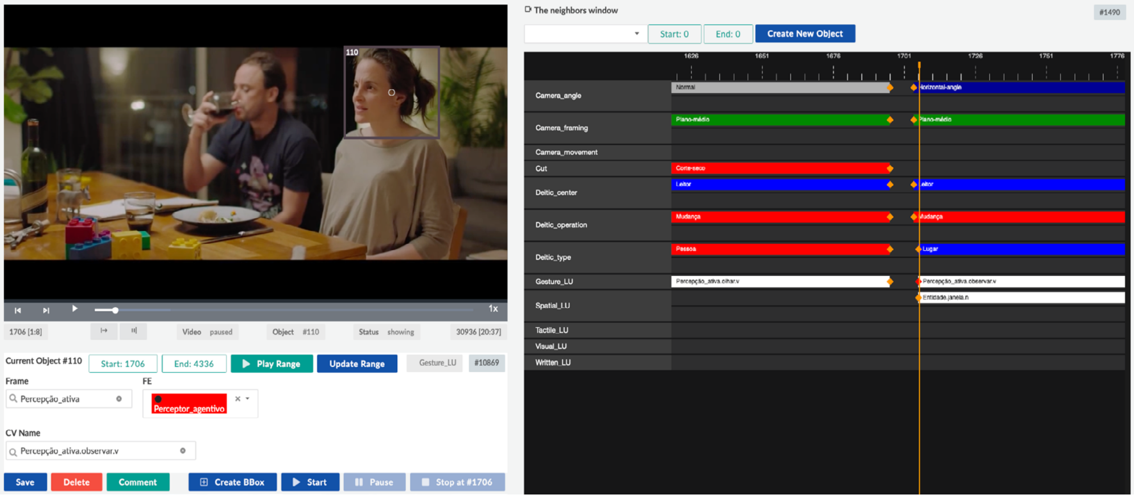
	Written	In scene	Bounding box
		Overlay	Bounding box
	Oral		Text annotation
	Audio	Sound effect	Checkbox
		Soundtrack	

Source: elaborated by the author.

Caption: blank cells indicate the lack of intermediate categories for a given level of analysis.

Once categories were defined, a first round of annotation of the short film *The Neighbors Window* was carried out using a spreadsheet. This experiment showed that the timespan was the main axis along which annotation needed to be performed, since, in the demarcation of changes to the deictic center, one needs to not only indicate when the DC changes, but also how long it takes until another change happens. The need for indicating specific elements in the video which could trigger the changes in the DC was also clear. The resulting interface is shown in Figure 8.

Figure 8: Annotation of the deictic center shift marked by the change in Alli's focus of attention towards the neighbors' window



1706 [1:8]
Video paused
Object #110
Status showing
30936 [20:37]
Current Object #110
Start: 1706
End: 4336
Gesture_LU
#10869
Frame: Percepção_ativa
FE: Perceptor_agentivo
CV Name: Percepção_ativa.observar.v

The neighbors window
#1490
Start: 0
End: 0

Camera_angle	Normal; Horizontal-angle
Camera_framing	Plano-médio; Plano-médio
Camera_movement	
Cut	Corte-seco
Deitic_center	Leitor; Leitor
Deitic_operation	Mudança; Mudança
Deitic_type	Pessoa; Lugar
Gesture_LU	Percepção_ativa.olhar.v; Percepção_ativa.observar.v
Spatial_LU	Entidade.janela.n
Tactile_LU	
Visual_LU	
Written_LU	

Source: FrameNet Brasil Lab (c2014-2025).

Figure 8 shows the annotation for the moment the DC changes from the conversation between the couple over dinner to the event at the neighbor's apartment. Note that the change in this case refers to the reader's DC, that is, the intention of the shot composition is to invite the reader to pay attention to the fact that something in the dynamics of the story will change. Alli's focus is no longer on her husband, or on the dinner, but on something else. The horizontal camera angle in a medium shot of the living room helps build the setting for the change. The meaning-making mode used as the main trigger for the DC shift is gestural, meaning that it is the change in Alli's gaze towards the window that provides the main clue for the reader to notice the DC shift. Note that the annotation system allows for the specification of the gesture in terms of frame and FE, since it can be annotated as a bounding box.

The annotation exemplified in Figure 8 adds to the meaning representation of multimodal objects currently used in FrameNet Brasil multimodal annotation, to the extent to which it allows for frame semantic information to be correlated with shifts in the narrative progression. In the near future, the main goal is that of augmenting and refining the categories related to meaning-making modes such as audio, for example.

6 Conclusions and Outlook

In a scenario where text analysis must go beyond the analysis of sequences of characters, linguists agree on the need to broaden the scope of the science they practice. More specifically, in areas such as Computational Linguistics, where the demand for multimodal datasets is growing, research initiatives have been undertaken as a means of grounding language analyses and what machines can learn from them. Even if the kind of multimodal analysis currently implemented by FrameNet Brasil represents an advance in the area, limiting the annotation to what is said and directly seen in multimodal genres leaves out foundational aspects of textual progression.

In this context, I have argued in favor of repositioning Frame Semantics so that its computational implementation – FrameNet – can embrace the analysis of deictic devices and their role in textual progression. Considering the foundational role of such devices in human communication (cf. Tomasello, 2003), analyzing their behavior in multimodal texts is key in the process of representing how such texts create meaning.

To illustrate this claim, I proposed exemplar analyses of filmic narratives in which I showed how shifts in the deictic center triggered by different communicative modes in interaction allow for the progression of the narrative. These analyses were performed within a model of Mediated Joint Attention, based on Tomasello's (2018) model of in presence communication and drawing from Turner's (2018) notion of blended joint attention scenes. Within this new model, in a filmic narrative, for example, the joint attention scene is established through some medium in which the intentionalities of one or more communicators are projected. Those communicators intentionally act to produce a multimodal text aimed at potential comprehenders. In this process, communicators make use of several communicative modes appealing to human sensory perception, mostly vision and hearing. All those modes collaborate for meaning making in multimodal texts and can be used for promoting shifts in the deictic centers, creating new joint attention micro scenes. Looking at only one of these semioses would be ignoring important aspects of these texts and their meaning.

The analytical model proposed reinforces the need for FrameNet multimodal analyses to consider (a) the inclusion of other communicative modes in the annotation methodology, (b) the need to build multidisciplinary teams including experts in Music, Film Analysis, Semiotics and Linguistics, at least, and (c) the need to consider other annotation levels that make room for categories lying closer to the pragmatic end of the Semantics-Pragmatics continuum.

I hope the proposed model can contribute to broadening the Frame Semantics perspective on the field of Multimodal Analysis, with replications for both Computational

Linguistics and Language Pedagogy. I believe that more comprehensive representations of how texts invite readers to construe meaning and of how such meaning-making process progresses as they read may help students in thinking critically about multimodal texts and further analyzing them.

Acknowledgements

The analytical proposal devised in this paper was developed during the author's sabbatical period as a visiting researcher at the Department of Swedish, Multilingualism and Language Technology of the University of Gothenburg. The author acknowledges the contributions of the undergraduate student annotators involved in the pilot analyses aimed at testing the analytical model: Rian Costa and Taís Reis, whose work was supported by scholarships provided by the Federal University of Juiz de Fora and FAPEMIG, respectively. The author is also thankful to Dr. Ely Matos for all the support in developing the module of the FrameNet Brasil WebTool used for the deixis annotation.

References

BAKHTIN, M. M. *Esthétique de la création verbale*. Paris: Gallimard, 1984.

BATEMAN, J. *Multimodality and Genre*: A Foundation for the Systematic Analysis of Multimodal Documents. Berlin: Springer, 2008.

BATEMAN, J.; WILDFEUER, J.; HIIPALA, T. *Multimodality*: Foundations, Research and Analysis – A Problem-Oriented Introduction. Berlin: Walter de Gruyter GmbH & Co, 2017.

BELCAVELLO, F. *FrameNet Annotation for Multimodal Corpora*: Devising a Methodology for the Semantic Representation of Text-Image Interactions in Audiovisual Productions. 2023. 134 f. Thesis (Doctorate in Linguistics) – Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora, 2023.

BELCAVELLO, F.; VIRIDIANO, M.; COSTA, A. D. da; MATOS, E. E. da S.; TORRENT, T. T. Frame-Based Annotation of Multimodal Corpora: Tracking (A) Synchronies in Meaning Construction. *In:* INTERNATIONAL FRAMENET WORKSHOP: TOWARDS A GLOBAL, MULTILINGUAL FRAMENET, 2020, Marseille. *Proceedings* [...]. Marseille: European Language Resources Association (ELRA), 2020. p. 23-30.

BELCAVELLO, F.; VIRIDIANO, M.; MATOS, E.; TORRENT, T. T. Charon: A FrameNet Annotation Tool for Multimodal Corpora. *In:* LINGUISTIC ANNOTATION WORKSHOP (LAW-XVI), 16. Marseille. *Proceedings* [...]. Workshop presented at LREC2022. Marseille: ELRA, 2022. p. 91-96.

BORDWELL, D.; THOMPSON, K. *A arte do cinema*: uma introdução. Tradução: Roberta Gregoli. Campinas, SP: Editora da Unicamp; São Paulo: Editora da USP, 2013.

BÜHLER, K. *Sprachtheorie.* Jena: Fischer, 1934.

CIENKI, A. The Study of Gesture in Cognitive Linguistics: How It Could Inform and Inspire Other Research in Cognitive Science. *Cognitive Science*, v. 13, n. 6, e1623, 2022.

COHN, N. A Multimodal Parallel Architecture: A Cognitive Framework for Multimodal Interactions. *Cognition*, v. 146, p. 304-323, 2016a.

COHN, N. From Visual Narrative Grammar to Filmic Narrative Grammar: The Narrative Structure of Static and Moving Images. *In:* WILDFEUER, J.; BATEMAN, J. (ed.). *Film Text Analysis*: New Perspectives on the Analysis of Filmic Meaning. London: Routledge, 2016b. p. 94-117.

COHN, N. Visual Narrative Structure. *Cognitive Science*, v. 37, n. 3, p. 413-452, 2013.

DANNÉLLS, D.; TORRENT, T. T.; SIGILIANO, N. S.; DOBNIK, S. Beyond Strings of Characters: Resources Meet NLP – Again. *In*: VOLODINA, E.; DANNÉLLS, D.; BERDICEVSKIS, A.; FORSBERG, M.; VIRK, S. (ed.). *Live and Learn*: Festschrift in Honor of Lars Borin. Gothenburg: Institutionen för Svenska, Flerspråkighet och Språkteknologi – Göteborgs Universitet, 2022. p. 29-36.

DUTRA, L. V.; SIGILIANO, N. S. Ferramenta linguístico-computacional como facilitadora para o ensino de gramática na escola. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 2021, [*S. l.*]. *Anais [...]. [S. l.]:* SBC, 2021. p. 432-436.

FILLMORE, C. J. Frame Semantics and the Nature of Languages. *In*: ANNALS OF THE NEW YORK ACADEMY OF SCIENCES. *Conference on the Origin and Development of Language and Speech*, v. 280, p. 20-32, 1976. Work presented at the Conference on the Origin and Development of Language and Speech, 1976, New York.

FILLMORE, C. J. Frame Semantics. *In:* THE LINGUISTIC SOCIETY OF KOREA (ed.). *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Company, 1982. p. 111-137.

FILLMORE, C. J. Frames and the Semantics of Understanding. Q*uaderni di Semantica*, v. 6, n. 2, p. 222-254, 1985.

FILLMORE, C. J. *Santa Cruz Lectures on Deixis, 1971*. Bloomington, IN: Indiana University Linguistics Club, 1975.

FILLMORE, C. J. The Case for Case Reopened. *In*: COLE, P.; SADOCK, J. M. (ed.). *Syntax and Semantics*: Grammatical Relations. New York: Academic Press, 1977. v. 8, p. 59-81.

FILLMORE, C. J. The Merging of Frames. *In*: ROSSINI FAVRETTI, R. (ed.). *Frames, Corpora, and Knowledge Representation*. Bologna: Bononia University Press, 2008. p. 1-12.

FILLMORE, C. J.; PETRUCK, M. R.; RUPPENHOFER, J.; WRIGHT, A. FrameNet in Action: *The Case of Attaching*. International Journal of Lexicography, v. 16, n. 3, p. 297-332, 2003.

FRAMENET BRASIL LAB. *WebTool*. c2014-2025. Available at: http://webtool.frame.net.br/. Accessed on: 31 Jan. 2025.

GOLDIN-MEADOW, S. Pointing Sets the Stage for Learning Language – and Creating Language. *Child Development*, v. 78, n. 3, p. 741-745, 2007.

GRICE, H. P. Meaning. *The Philosophical Review*, v. 66, n. 3, p. 377-388, 1957.

HANKS, W. F. *Língua como prática social*: das relações entre língua, cultura e sociedade a partir de Bourdieu e Bakhtin. Textos selecionados por Hanks, traduzidos por Anna Christina Bentes, Marco Antônio Rosa Machado, Marcos Rogério Cintra e Renato C. Rezende. São Paulo: Cortez, 2008.

JACKENDOFF, R. *Foundations of Language*. Oxford: Oxford University Press, 2002.

JACKENDOFF, R.; LERDAHL, F. The Capacity for Music: What is It, and What's Special About It? *Cognition*, v. 100, p. 33-72, 2006.

KALANTZIS, M.; COPE, B.; PINHEIRO, P. *Letramentos*. Campinas: Unicamp, 2020.

KITA, S. *Pointing*: Where Language, Culture and Cognition Meet. Mahwah, NJ: Lawrence Erlbaum, 2003.

LEMKE, J. Multiplying Meaning. *In*: MARTIN, J. R.; VEEL, R. (ed.). *Reading Science*: Critical and Functional Perspectives on Discourses of Science. London: Routledge, 1998. p. 87-114.

LEVINSON, S. C. Deixis. *In*: HORN, L. R.; WARD, G. (ed.). *The Handbook of Pragmatics*. Malden, MA: Blackwell Publishing, 2004. p. 97-121.

LISZKOWSKI, U.; BROWN, P.; CALLAGHAN, T.; TAKADA, A.; DE VOS, C. A. Prelinguistic Gestural Universal of Human Communication. *Cognitive Science*, v. 36, p. 698-713, 2012.

RODRÍGUEZ, C.; MORENO-NÚNEZ, A.; BASILIO, M.; SOSA, N. Ostensive Gestures Come First: Their Role in the Beginning of Shared Reference. *Cognitive Development*, v. 36, p. 142-149, 2015.

SAUSSURE, F de. *Cours de linguistique générale*. Paris: Payot, 1916.

SEARLE, J. R. Collective Intentions and Actions. *In:* COHEN, O.; MORGAN, J.; POLLACK, M. (ed.). *Intentions in Communication*. Cambridge, MA: MIT Press, 1990.

SEGAL, E. M. Narrative Comprehension and the Role of Deictic Shift Theory. *In:* DUCHAN, J.; BRUDER, G. A.; HEWITT, L. E. (ed.). *Deixis in Narrative*: A Cognitive Science Perspective. London: Psychology Press, 1995. p. 3-18.

STEEDMAN, M. J. Plans, Affordances, and Combinatory Grammar. *Linguistics and Philosophy*, v. 25, n. 5-6, p. 723-753, 2002.

STEEN, F.; HOUGAARD, A.; JOO, J.; OLZA, I.; CÁNOVAS, C.; PLESHAKOVA, A.; RAY, S.; UHRIG, P.; VALENZUELA, J.; WOŹNY, J.; TURNER, M. Toward an Infrastructure for Data-Driven Multimodal Communication Research. *Linguistics Vanguard*, v. 4, n. 1, e20170041, 2018.

SWEETSER, E. Looking at Space to Study Mental Spaces: Co-Speech Gesture as a Crucial Data Source in Cognitive Linguistics. *In:* GONZALEZ-MARQUEZ, M.; SPIVEY, M. J.; COULSON, S.; MITTELBERG, I. (ed.). *Methods in Cognitive Linguistics*. Amsterdam: John Benjamins, 2007. p. 201-224.

THE NEIGHBORS Window. Written and directed by Marshall Curry. New York: Marshall Curry Productions, 2019. Available at: http://www.theneighborswindow.com/. Accessed on: 31 Jan. 2025.

THE NEW LONDON GROUP. A Pedagogy of Multiliteracies: Designing Social Futures. *Harvard Educational Review*, v. 66, n. 1, p. 60-93, 1996.

TOMASELLO, M. *Becoming Human*: A Theory of Ontogeny. Cambridge, MA: The Belknap Press of Harvard University Press, 2018.

TOMASELLO, M. *Constructing a Language*: A Usage-Based Theory of Language Acquisition. Cambridge, MA: Harvard University Press, 2003.

TOMASELLO, M. *Origins of Human Communication*. Cambridge, MA: MIT Press, 2008.

TOMASELLO, M. *The Cultural Origins of Human Cognition*. Cambridge, MA: MIT Press, 1999.

TOMASELLO, M.; CARPENTER, M.; LISZKOWSKI, U. A New Look at Infant Pointing. *Child Development*, v. 78, n. 3, p. 705-722, 2007.

TORRENT, T. T.; MATOS, E. E. S.; BELCAVELLO, F.; VIRIDIANO, M.; GAMONAL, M. A.; COSTA, A. D. da; MARIM, M. C. Representing Context in FrameNet: A Multidimensional, Multimodal Approach. *Frontiers in Psychology*, v. 13, e838441, 2022.

TURNER, M. Multimodal Form-Meaning Pairs for Blended Classic Joint Attention. *Linguistics Vanguard*, v. 3, n. s1, e20160043, 2017.

TURNER, M. The Role of Creativity in Multimodal Construction Grammar. *Zeitschrift für Anglistik und Amerikanistik*, v. 66, n. 3, p. 357-370, 2018.

VIRIDIANO, M.; TORRENT, T. T.; CZULO, O.; LORENZI, A.; MATOS, E.; BELCAVELLO, F. The Case for Perspective in Multimodal Datasets. *In*: WORKSHOP ON PERSPECTIVIST APPROACHES TO NLP, 1., 2022, Marseille. *Proceedings […]*. Workshop presented at LREC2022. Marseille: European Language Resources Association, 2022. p. 108-116.

WILKINS, D. Why Pointing with the Index Finger is not a Universal (in Sociocultural and Semiotic Terms). *In*: KITA, S. (ed.). *Pointing*: Where Language, Culture and Cognition Meet. Mahwah, NJ: Lawrence Erlbaum, 2003. p. 171-216.

WITTGENSTEIN, L. *Philosophische Untersuchungen*. London: Kegan Paul, 1953.

ZUBIN, D. A.; HEWITT, L. E. The Deictic Center: A Theory of Deixis in Narrative. *In*: DUCHAN J. F.; BRUDER G. A.; HEWITT, L. E. (ed.). *Deixis in Narrative*: A Cognitive Science Perspective. New York: Routledge: Taylor & Francis Group, 1995. p. 129-155.