Frame^2: a FrameNet-based multimodal dataset for tackling text-image interactions in video

Frederico Belcavello[1], Tiago Timponi Torrent[1,2], Ely Matos[1], Adriana Pagano[2,3], Maucha Gamonal[1,3], Natalia Sathler Sigiliano[1], Lívia Vicente Dutra[1,4], Helen de Andrade Abreu[1], Mairon Samagaio[1], Mariane Carvalho[1], Franciany Campos[1], Gabrielly Azalim[1], Bruna Mazzei[1], Mateus Fonseca de Oliveira[1], Ana Carolina Luz[1], Lívia Pádua Ruiz[1], Júlia Bellei[1], Amanda Pestana[1], Josiane Costa[1], Iasmin Rabelo[3], Anna Beatriz Silva[3], Raquel Roza[3], Mariana Souza[3] and Igor Oliveira[3]

[1] FrameNet Brasil, Federal University of Juiz de Fora
[2] Brazilian National Council for Scientific and Technological Development – CNPq
[3] LETra, Federal University of Minas Gerais
[4] Masters in Language Technology, Gothenburg University
{fred.belcavello, tiago.torrent@}ufjf.br

Abstract

This paper presents the Frame^2 dataset, a multimodal dataset built from a corpus of a Brazilian travel TV show annotated for FrameNet categories for both the text and image communicative modes. Frame^2 comprises 230 minutes of video, which are correlated with 2,915 sentences either transcribing the audio spoken during the episodes or the subtitling segments of the show where the host conducts interviews in English. For this first release of the dataset, a total of 11,796 annotation sets for the sentences and 6,841 for the video are included. Each of the former includes a target lexical unit evoking a frame or one or more frame elements. For each video annotation, a bounding box in the image is correlated with a frame, a frame element and lexical unit evoking a frame in FrameNet.

Keywords: multimodal dataset, FrameNet, audiovisual semantic annotation

1. Introduction

Since the late 1990's, FrameNet (Fillmore et al., 2003) has been developed as a language resource. It correlates lexical items with background scenes: the frames. Frames include participants and props, whose conceptualization is deemed necessary for accessing the meaning of lexical items. FrameNet is composed of frames and their associated roles in a network of typed frame-to-frame relations (Ruppenhoffer et al., 2016). Over the years FrameNet was expanded to several languages other than English (Subirats-Rüggeberg; Petruck, 2003; You; Liu, 2005; Ahlberg et al., 2014; Boas; Ziem, 2018; Gruzitis et al., 2018; Ohara et al., 2018; Torrent et al., 2018a; Hahm et al., 2020). More recently, the FrameNet model has also been applied to the annotation of other communicative modes, i.e., images (Belcavello et al., 2020; Torrent et al., 2022; Viridiano et al., 2022; Belcavello et al., 2022; Luz et al., 2023).

The main claim behind the application of the FrameNet model to other communicative modes is that, in a way equivalent to linguistic material, images – and the elements in them – may also evoke frames or work together with verbal text in the process of meaning making (Belcavello et al., 2020).

In this paper we present the first release of the Frame^2 dataset[1] as an expansion of FrameNet into the multimodal domain. The goal is to offer a new gold standard semantically enriched fine-grained resource for multimodal NLP tasks.

2. Designing the Dataset

Frame^2 is a dataset composed by multimodal objects. These objects are a result of an annotation task carried out for a specific audiovisual corpus using FrameNet categories. The annotated data accounts for both the verbal language and the video image of a Brazilian TV Travel Series, named "*Pedro pelo Mundo*[2]". The verbal language mode is, in turn, composed of two types of text: (i) the audio spoken during the episodes of the TV show–which was transcribed for annotation purposes, and (ii) the subtitles present in those segments of the show in which the host conducts interviews in English.

The data also includes the relations between the annotated data as mediated by the semantic structure modeled in the FrameNet Brasil database (Torrent et al., 2022). That means that information about the frames, their frame elements, their relations with other frames, and relations between lexical units is included in the Frame^2 dataset.

In as much as Frame^2 was built to serve as a gold standard dataset for multimodal NLP tasks, the notion of gold standard dataset should be regarded here with caution. The annotations in the Frame^2 dataset are meant to represent possible perspectives on the meaning construction processes. These perspectives may be triggered by the combination of different communicative modes. Therefore, more than one set

[1] https://github.com/FrameNetBrasil/frame-squared

[2] Pedro around the World.

LREC-COLING 2024, pages 7429–7437
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

of annotations is possible and even different annotation methodologies can be proposed to account for the multiperspectivized nature of meaning construction. Such an approach to data annotation follows the Perspectivized NLP approach, as defined by The Perspectivist Data Manifesto[3] and by Basile et al. (2021).

Frame[2] comprises data that accounts for the frame-based semantic representation of verbal language and its interaction with a frame-based interpretation of video sequences–i.e., sequences of visual frames related with audio (especially when it contains spoken material), forming a video. Therefore, the dataset reflects audio and video combination possibilities in terms of frames, – in the way they were defined by Fillmore (1982), as structured representations of interrelated concepts.

2.1 The Corpus

The corpus to which FrameNet annotations were added comprises the ten episodes of the first season of the Brazilian TV Travel Series "*Pedro pelo Mundo*". The show premiered in 2016 on GNT, a cable channel dedicated to entertainment and lifestyle productions.[4] Four seasons of "Pedro pelo Mundo" were aired until 2019. The first season has 10 episodes of 23 minutes each. The second, third and fourth are also composed by 10 episodes each, but these are 48 minutes long. For the purposes of this dataset, the corpus was limited to the 10 episodes of the first season, which means a total of 230 minutes of video.

The plot of each episode focusses on getting in contact and exploring social, economic, and cultural aspects of a location which has experienced some kind of recent transformation. Thus, what the viewer sees is Pedro Andrade, the host, trying to connect with locals, instead of merely proposing a touristic view of popular places of interest. The format of the show combines standups, voice-over sequences, short interviews, and video clip sequences. It thus offers rich material as an exemplar of complex audiovisual composition for meaning making.

The 10 episodes in the first season were pre-processed for annotation following the pipeline proposed in Belcavello et al. (2022). In this methodology, two separate files are extracted from the videos for annotation: (i) a text file containing all the time-stamped audio transcriptions and subtitles in the episode, and (ii) a set of image files extracted at a 25 image frames per second rate. The resulting corpus is composed by 2,195 sentences, transcribed from the 230 minutes of video.

We now turn to the description of the annotation task carried out in the corpus.

2.2 The annotation task design

The annotation task was devised as divided into two parts: text annotation and image annotation. Annotation was carried out by undergraduate students trained in the task. Training strategies employed varied according to the different kinds of annotation teams – permanent or temporary – assembled for the task. The permanent annotation team was composed by 12 students hired to perform several annotation tasks, including the one described here. They received monthly stipends of R$ 700.00 for 20 hours of work a week. The per hour value paid to the students is circa 15% higher than the minimum wage in Brazil. Stipend values are defined by Brazilian research funding agencies.

The temporary annotation teams were assembled among the students enrolled in undergraduate division hands-on annotation workshops. Each workshop is composed of 45 hours of academic work, comprising tutoring and annotation practice. Two classes of the workshops contributed to the annotation of the dataset. A total of 32 undergraduate students were part of the temporary annotation team.

The teams conducted the annotation tasks using tools specifically designed for both full-text and video annotation. For the full-text annotation, annotators used the same web-based annotation tool used for the Global FrameNet Shared Annotation Task (Torrent et al., 2018). For the image annotation, a semi-automatic, human-in-the-loop tool for annotating static and dynamic images for semantic frames was used (Belcavello et al., 2022). This tool was developed to annotate visual objects, correlate them with textual data and label frames and frame elements evoked by them. The tool is compatible with the full-text annotation tool and is composed of two modules: a static mode, for annotating picture-text pairings, and a dynamic mode – which was the one used in the task reported on in this paper – for annotating images in video.

The task followed two major annotation guidelines:

(i) when annotating text for audiovisual corpora, annotators should always watch the video and see the sentences in their multimodal context.

(ii) in the same way, when annotating images, annotators should always listen to the spoken audio and should also read its transcribed sentences made available in the video annotation workspace.

In the methodology used, there were two possibilities to carry out the full-text annotation of the corpus: (i) annotators would annotated all sentences of an episode first, and then start annotating images; or (ii) annotators would complete the annotation of the sentences that correspond to a sequence[5], then

[3] http://pdai.info/

[4] Authors have been granted written permission from the show copyright owners to use the first season for research purposes and distribute them together with the dataset.

[5] We define sequence for these purposes as a set of scenes which presents a distinctive unit in terms of the topic presented as a subtopic of the episode's theme.

annotate image in the respective sequence, and go back to the sentences of the following sequence.

2.2.1 Text Annotation

When building the Frame[2] dataset the annotation process started by following FrameNet's guidelines for full-text annotation (Ruppenhofer et al., 2016). Going sentence by sentence in the corpus, annotators created Annotation Sets (AS) for each word for which there is a Lexical Unit (LU) in FrameNet. For instance, a word such as the adjective creative – *creative.a* – has its meaning defined based on a scene in which a Protagonist acts with a particular Behavior. This scene is the `Mental_property` frame (Figure 1).

Sentence (1) is in the Frame[2] dataset and shows an occurrence of the LU *creative.a* annotated for the `Mental_property` frame:

(1) People started being more creative again.

Annotation plays a key role in FrameNet, to the extent that it provides evidence supporting the analysis in the model. Figure 2 shows five ASs created for (1). Note that, for each of them, there are three layers of annotation: (i) Frame Element (FE), which indicates the role other words or phrases have in relation with the target LU – for instance, 'People' is the Protagonist in the `Mental_property` frame evoked by the LU *creative.a*; (ii) Grammatical Function (GF) and (iii) Phrase Type (PT). The NI column is used for indicating that core FEs are not instantiated in the sentence but can be inferred, or that they are incorporated by the stem of the LU.

Mental_property

[@Abstract_attribute] [@Psychology] [@Lexical] [#24]

Definition

The adjectives and nouns in this frame are all based on the idea that mental properties may be attributed to a person (Protagonist) by a (usually implicit) Judge on the basis of that person s Behavior, as broadly understood. Though on a conceptual level these words always attribute mental properties to people, they may be applied to Protagonist s Behaviors as well, with the understanding that the Behavior is revealing a (usually temporary) property of the Protagonist responsible for it. For example, while we may speak of a stupid person, we may also speak of a stupid thing to say/do/think. In addition, we may mention both the Protagonist and the Behavior, as in It was stupid of me to do that. Some of the words in this frame also have slightly different uses, in which there is a constituent expressing the Practice with respect to which the mental property holds of the Protagonist, as in: She is astute

Figure 1: Definition of the `Mental_property` frame

Figure 2: Text annotation example

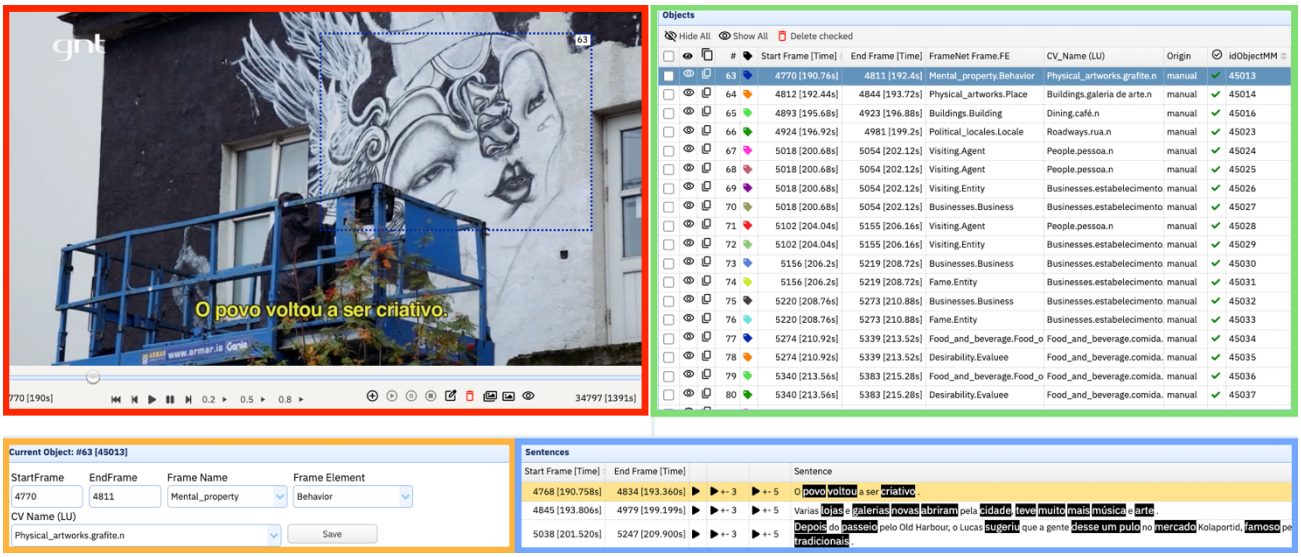


Figure 3: Example of the image annotation screen.

2.2.2 Video Annotation

Figure 3 shows the dynamic image annotation workspace. Highlighted in red is the 'video panel' in which annotators control video playback, draw and edit bounding boxes. Bellow it, highlighted in orange, is the 'annotation panel' in which annotators see the indication of start and end point of the object being annotated in the video. In this panel, they associate a frame with the object, select the frame element and indicate a Computer Vision (CV) name for the object.

The CV Name categorization was created for matching the object with pre-trained computer vision categories. In principle, the CV Name field is to be filled in from the automatic labeling of the visual objects using classes from the Open Images Dataset v6 dataset, which were all associated to LUs evoking frames in the database.[6]

Moving right, highlighted in green, is the 'objects panel' which presents the list of objects created – both manually and automatically – and the metadata associated with it during annotation. In the bottom right corner, highlighted in blue, is the 'sentences panel'. It shows the sentences annotated in the full-text annotation stage, associated with its timestamps and playback controls for the 'video panel'. Those controls allow annotators to visualize the sentences in action in the episode segment.

The image annotation proposed here refers to the selection of part of the screen by using a bounding box. Such a selection is understood as a correspondent visual demonstration of a frame element in a frame. In this sense, a visual object is defined as a set of bounding boxes in a time interval that is associated with a frame element. For instance, in Figure 3 looking at the video panel and at the objects panel, object 63 stores the information that:

(i) that portion of the image refers to the Behavior FE in the `Mental_property` frame.

(ii) the bounding box list starts at the video frame 4770 – which is also correspondent to second 190.76 – and ends at the video frame 4811 – second 192.4.

(iii) it is also associated with the LU *grafite.n* (*graffiti.n*) in the `Physical_artworks` frame for the CV Name categorization.

The multimodal text-oriented approach for this annotation can be explained as follows. When looking for correspondences between text and image, object 63 (Figure 3) was annotated as the visual manifestation of the Behavior in the `Mental_property` frame. On the other hand, as what is visually recognizable is a graffiti, the CV Name chosen for the object was *grafite.n* (*graffiti.n*) in the `Physical_artworks` frame. What is interesting here is the fact that this annotation makes it possible to associate a concrete art manifestation with the intangible idea of a mental property. And that was probably what motivated the video editor when this shot was chosen to illustrate the sentence in (1). Therefore, this example shows how the Frame^2 dataset covers the addition of meaning layers and granularity to the FrameNet semantic representation by having annotated visual data in correspondence with textual data in a corpus.

On top of the two general annotation guidelines stated in section 2.2, the following annotation methodology guidelines were also proposed:

(i) The locality of each bounding box is a shot. No bounding box should last more than one shot. If one object is present on screen throughout multiple sequential shots one different bounding box should be drawn for each shot.

6 https://github.com/DmitryRyumin/OIDv6/blob/master/oidv6/classes.txt

(ii) The beginning of a bound box coincides with the beginning of a shot or the first appearance of the object in the shot, even if it occurs before the beginning of the sentence or the pronunciation of the target LU in the sentence.
(iii) One visual object can be duplicated as many times as necessary if it instantiates different FEs – either in one same frame or in different frames.
(iv) The limit of asynchrony for considering a relation between a bounding box and a target LU in a sentence is the video sequence. Bounding boxes can be created and annotated as referring to lexical units that are 'n' seconds prior to or ahead of the presence of the LU in the audio, if they are both located within the same video sequence and/or if there is not a better connection with a closer LU.
(v) The bounding box size and position should be adjusted from frame to frame – if not automatically adjusted – to match changes in object size and position.
(vi) CV Names should always be chosen taking the most empirical and concrete LU possible to designate what is seen on screen.

3. Annotation outcomes

The 2,195 sentences in the corpus generated 11,796 full-text annotation sets, while the images have been annotated for 6,841 visual objects (VOs). To the best of our knowledge, this is the first dataset that combines a multimodal approach and Frame Semantics for video annotation of visual objects. In the remainder of this section, we present the annotation outcomes in terms of the total amount of metadata associated with the corpus. We also also present a qualitative analysis of the kinds of correlations made possible by the multimodal annotation method used for building Frame^2.

3.1 Semantic annotation totals

The number of ASs, sentences, AS per sentence and VOs per episode is shown in Table 1. The numbers are very close to the ones estimated in the pilot study introducing the extension of the FrameNet model to the multimodal domain by Belcavello et al. (2020), with a variation of 1.7% less annotation sets and 9.75% more sentences. For the visual objects, however, the result of 6,841 VOs represents an increase of 36.82% on the number estimated.

Episode	AS	Sent.	AS/Sent.	VO
01	1164	226	5.1504	593
02	890	205	4.3415	805
03	1029	208	4.9471	638
04	1011	199	5.0804	562
05	1385	248	5.5847	657
06	1191	226	5.2699	503
07	1087	218	4.9862	698
08	1373	227	6.0485	545
09	1403	215	6.5256	779
10	1263	223	5.6637	1061
TOTAL/AVG	**11,796**	**2,195**	**5.3598**	**6,841**

Table 1: Corpus annotation totals and averages

The average of annotations per sentence ranged from 4.34 – the lowest value – to 6.53 – the highest value. The corpus average of AS per sentence was 5.3598. This result is below the full-text annotation average of 6.1 AS per sentence found in the FrameNet Brasil database (Belcavello et al., 2020). We can empirically associate this reduction with the perception of a great presence of short sentences in the corpus. Moreover, this can be explained by the oral and very colloquial origin of the sentences in the corpus, which include a relevant number of greetings and other more pragmatic level operators that are not yet covered by FrameNet frames.

Concerning the variability of the corpus, Table 2 shows how many discrete frames and LUs – in the case of the CV Name – were used in each episode and in the corpus.

Episode	Frames in Text	Frames in Image	Frames in CV Name	LUs in CV Name
01	279	163	42	91
02	256	91	29	88
03	243	110	55	93
04	257	89	31	52
05	284	103	33	73
06	278	110	30	55
07	265	123	24	53
08	298	141	39	82
09	291	106	28	49
10	292	136	51	81
CORPUS	**611**	**393**	**129**	**478**

Table 2: Numbers of discrete frames and LUs used in the annotation of text and image

The numbers in Table 3 show that frame variability in textual annotation is much higher than in the annotation of visual objects. It is true that the number of annotations sets per episode is always higher than the number of visual objects – the ratio of VOs per ASs is 0.57. However, the ratio of VO discrete frame per AS discrete frame is 0.64 – higher than the VOs/ASs value. On the other hand, the ratio of CV Name discrete frame per AS discrete frame is 0.21 – much lower than the VOs/ASs value. We have empirical elements to believe that this difference may be related to the predominance of entities annotated for CV Name and to the high rate of repetition of some frames during annotation, especially those evoked by many LUs, such as `Food_and_beverage`, for example.

The average number of discrete LUs used as CV Name per episode is 71.7. The total number of 478 discrete LUs used as CV Name in the corpus can be taken as the number of different categories of objects annotated as a way of comparing Frame^2 with other datasets. The number is considerably higher than the 80 categories of the MS COCO (Lin et al., 2014). It is also close to the 600 boxable classes of the Open Images Dataset v7 (Kusnetzova et al., 2020), but with the difference that the classes in Frame^2 classes are not merely hierarchized, but organized in a more complex network of concepts that is FrameNet, using 129 different frames, as presented in Table 2.

Finally, Table 3 presents another aspect that supports the improved granularity of the Frame^2 dataset: the

matching ratio of only 1.61 between the frames used for the VO annotation and the ones used for the CV Name. This means that 98.39% of the VOs have been associated with two different frames at the annotation level, which indicates that they are semantically enriched objects from the start, even before the establishment of the other relations that form the network of frames and LUs in FrameNet.

Episode	VO to CV Name frame matching ratio
01	3.54
02	2.88
03	6.76
04	4.92
05	4.42
06	4.14
07	2.9
08	5.32
09	2.58
10	3.28
AVG	**1.61**

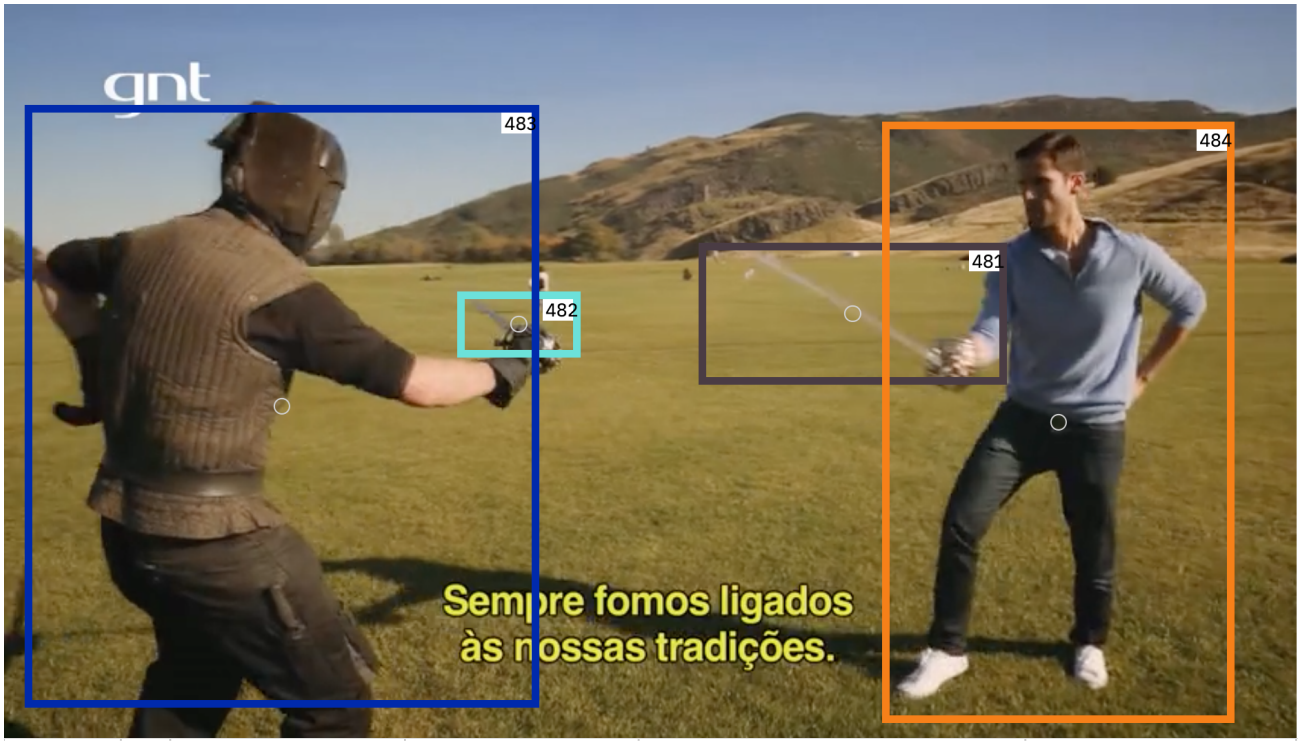
Table 3: Matching ratio of frames used for image annotations

3.2 Relations in the annotated data

The example presented in section 2.2.2 demonstrated that visual elements in video shots may also evoke frames and organize their elements on the screen or work complementarily with the frame evocation patterns of the sentences narrated simultaneously to their appearance on screen, providing different profiling and perspective options for meaning construction. In the case where the `Mental_property` and the `Physical_artworks` frames were connected, there was a blending of an entity from a visual object to instantiate a FE in the text.

FrameNet and all its sister projects in other languages are composed of frames and their associated roles in a network of typed relations such as inheritance, perspective, and use (Ruppenhoffer et al., 2016). These are frame-to-frame relations traditionally used in most – if not every – FrameNet. FrameNet Brasil has also developed other types of relations aimed at enriching the database structure (Torrent et al., 2022). One of these relations links FEs to the frames licensing the lexical items that typically instantiate those elements. Another relation connects core FEs to non-core FEs in the same frame when the latter can act as metonymic substitutes for the first (see Gamonal, 2017).

An additional set of relations emerges from the intricate connections among LUs, drawing inspiration from the concept of qualia roles established by Pustejovsky (1995). Derived from Pustejovsky's four fundamental qualia categories, namely agentive, constitutive, formal, and telic, FrameNet Brasil has devised a framework of frame-mediated ternary relationships (Torrent et al., 2022). In this framework,



481		30017 [1200.64s]	30034 [1201.32s]	Custom.Descriptor	Weapon.espada.n
482		30017 [1200.64s]	30034 [1201.32s]	Custom.Descriptor	Weapon.espada.n
483		30017 [1200.64s]	30034 [1201.32s]	Custom.Protagonist	People.pessoa.n
484		30017 [1200.64s]	30034 [1201.32s]	Custom.Protagonist	People.pessoa.n

Figure 4 : VO annotation for the fencing sequence.

a particular LU is intricately associated with another LU through a subcategory of quale, which is specified by a frame.

In Figure 4, we show an example in which a VO annotated for one frame instantiates another frame evoked by the LU in the text the image specifies the text, the connection between them being made by qualia relations. The VO annotation shown in Figure 4 comes from a fencing sequence in episode 6 – Edinburgh. It shows the host exploring the highlanders' way of fencing as a tradition kept by Scots. He meets Paul McDonald, presented as one of the great Scotland's authorities in the history of medieval battle and fencing instructor. They talk about Scottish traditions and McDonald offers the host a practical fencing lesson. During this sequence, the subtitles in Portuguese – see (2) – translate the original English spoken audio in (3).

(2) Sempre[Frequency] fomos ligados[Social_connection] às nossas tradições[Custom].
(3) We have always[Frequency] been connected[Social_connection] to our traditions[Custom].

In sentence (2), *tradições.n* (*traditions.n*) is annotated for the `Custom` frame. The FE Behavior is incorporated in the LU, while its Protagonist is annotated in the video – objects 483 and 484. Objects 481 and 481are annotated for the FE Weapon in the `Weapon` frame and designated as *espada.n* (*sword.n*) also in the `Weapon` frame for the CV Name (Figure 4). The arising issue would be how to represent the connection between the art of fencing, previously mentioned in (2) and triggered in the shot by the sword – objects 481 and 482 – combined with the `Custom` frame, evoked by *tradições.n*. The kinds of ternary qualia relations present in the FrameNet Brasil database make this combination possible – see Figure 5.

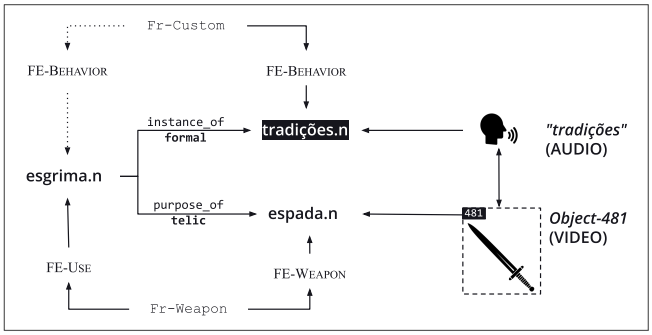


Figure 5: Ternary qualia relations in the multimodal annotation of sentence (2)

Note that the existence of a ternary qualia relation mediated by the `Exemplar` frame connects *esgrima.n* (*fencing.n*) to *tradição.n* (*tradition.n*), while another relation, mediated by the `Tool_purpose` frame connects *esgrima.n* to *espada.n* (*sword.n*). Those two relations allow for the inference that, in the multimodal setting, the behavior is that of practicing fencing.

Other different interactions between VOs and LUs can be found in the Frame^2 dataset (Belcavello, 2023). All of them demonstrate how connecting text and image is a way of enriching a language resource with visual material.

4. Conclusions and future work

The Frame^2 dataset exploits the complexity of an enriched FrameNet model to create meaningful connections between semiotic modes. The dataset offers the means for FrameNet to diversify its ways of representing meaning, once it incorporates image as a token for establishing relations and then for meaning-making. The multimodal approach to the dataset keeps the linguistic anchorage to the way the elements in it may be analyzed, explored, and used. However, the research conducted to culminate in this dataset shows that the path to approach image in meaning-making processes is broad and offers other possibilities worth exploring.

The next stage of the Frame^2 dataset development includes its usage to train a model for improving and start to automatically identify and tag Visual Objects for LUs and frames in the CV Name label. In parallel, we anticipate other ways of annotating other elements of visual composition. One of them is currently being designed in a pilot study: the annotation of the events perceived in image. This approach relies on the hypothesis that it is possible to identify one or more frames that account for the event(s) shown on screen in a shot within the locality of a scene or a video sequence.

5. Acknowledgments

Authors acknowledge the support of the Graduate Program in Linguistics at the Federal University of Juiz de Fora. Research presented in this paper is developed by ReINVenTA - Research and Innovation Network for Vision and Text Analysis of Multimodal Objects. ReINVenTA is funded by FAPEMIG grant RED 00106/21, and CNPq grants 408269/2021-9 and 420945/2022-9. Belcavello's research was funded by CAPES PDSE PhD exchange grant 88881.362052/2019-01. Torrent is an awardee of the CNPq Research Productivity Grant number 315749/2021-0. Pagano is an awardee of the CNPq Research Productivity Grant number 313103/2021-6. Gamonal's research was funded by CAPES/PRINT grant 88887.936139/2024-00.

6. Bibliographical References

Ahlberg, M., Borin, L., Dannélls, D., Forsberg, M., Toporowska Gronostaj, M., Friberg Heppin, K., Johansson, R., Kokkinakis, D., Olsson, L. J., Uppström, J. (2014). Swedish framenet++ the beginning of the end and the end of the beginning. In: *Proceedings of the Fifth Swedish Language Technology Conference*. Uppsala, 13-14 November 2014

Belcavello, F. ; Viridiano, M. ; Diniz Da Costa, A. ; Matos, E. E. ; Torrent, T. T. (2020). Frame-Based Annotation of Multimodal Corpora : Tracking (A)Synchronies in Meaning Construction. In : *Proceedings of the LREC International FrameNet Workshop 2020* : Towards a Global, Multilingual FrameNet. 1Marseille, France : ELRA, p. 23-30.

Belcavello, F. ; Viridiano, M. ; Matos, E. ; Torrent, T. T. (2022). Charon : A FrameNet Annotation Tool for Multimodal Corpora. In : *Proceedings of The 16th Lingusitic Annotation Workshop (LAW-XVI) within LREC2022*. Marseille, France : ELRA, p. 91-96.

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

Boas, H. C. and Ziem, A. (2018). Constructing a constructicon for german. In: Lyngfelt, B., Borin, L., Ohara, K., Torrent, T. (Eds), *Constructicography: Constructicon development across languages.* John Benjamins Publish-ing Company, Amsterdam, pp. 183–228.

Fillmore, C. J. (1982). Frame semantics. IN: Linguistic society of Korea (org*). Linguistics in the morning calm*. (1982) pp. 111-137. Hanshin Publishing Co., Seoul.

Fillmore, C. J., Petruck, M. R., Ruppenhofer, J., & Wright, A. (2003). FrameNet in action: The case of attaching. *International journal of lexicography, 16*(3), 297-332.

Fillmore, C. J. and Baker, C. (2009). A frames approach to semantic analysis. In Bernd Heine et al., editors, *The Oxford Handbook of Linguistic Analysis*, pages 791–816. Oxford University Press, Oxford, UK, December.

Gamonal, M. A. (2017). *Modelagem Linguístico Computacional de Metonímias na Base de Conhecimento Multilíngue (m.knob) da FrameNet Brasil*. Ph.D. Dissertation in Linguistics. Universidade Federal de Juiz de Fora. Juiz de Fora.

Gruzitis, N., Nespore-Berzkalne, G., Saulite, B. (2018). Creation of latvian framenet based on universal dependencies. In: Torrent, T. T., Borin, L., Baker, C. F. (Eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* European Language Resources Association (ELRA), pp 23–27.

Hahm, Y., Noh, Y., Han, J. Y., Oh, T. H., Choe, H., Kim, H., Choi, K. S. (2020). Crowdsourcing in the development of a multilingual framenet: A case study of korean framenet. In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. European Language Resources Association (ELRA). pp 236–244.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., ... & Ferrari, V. (2020). The open images dataset v4 : Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, *128*(7), 1956-1981.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco : Common objects in context. In *Computer Vision–ECCV 2014 : 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740-755). Springer International Publishing.

Luz, A. C. L. ; Braz, G. ; Ruiz, L. P. ; Pinto, M. C. ; Belcavello, F. ; Sigiliano, N. S. ; Torrent, T. T. (2023). Anotação do Dataset Multimodal da ReINVenTA. In : *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*. Porto Alegre : SBC, p.360 - 364.

Ohara, K., Kawahara, D., Sekine, S., Inui, K. (2018). Linking japanese framenet with kyo-to university case frames using crowdsourc-ing. In: Torrent, T. T., Borin, L., Baker, C. F. (Eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* European Language Resources Association (ELRA), pp. 56–61.

Pustejovsky, J. & Ježek, E. (2016). Qualia Structure. In : Pustejovsky, J. & Ježek, E. Integrating Generative Lexicon and Lexical Semantic Resources (pp. 11–55). *Tutorial at The Language Resources and Evaluation Conference* (LREC 2016), Protoroz, Slovenia.

Ruppenhofer, J., Ellsworth, M., Schwarzer-Petruck, M., Johnson, C. R., Baker, C. F., and Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. Berkeley, CA: International Computer Science Institute.

Subirats-Rüggeberg, C. and Petruck, M. R. (2003). Surprise: Spanish FrameNet! In: Hajivcová, E., Kotévsovcová, A., Mirovsky, J. (Eds), *Proceedings of the Workshop on Frame Semantics*, XVII International Congress of Linguists (CIL). Matfyzpress, Mat-fyzpress, Prague.

Torrent, T. T., da Silva Matos, E. E., Lage, L. M., Laviola, A., da Silva Tavares, T., de Almeida, V. G., Sigiliano, N. S. (2018a). Towards continuity between the lexicon and the constructicon in FrameNet Brasil. In: Lyngfelt, B., Borin, L., Ohara, K., Torrent, T. T. (Eds), *Constructicography: Constructicon development across languages*. John Benjamins, Amsterdam, pp 107–140.

Torrent, T. T. ; Ellsworth, M. ; Baker, C. F. ; Matos, E. E. (2018b). The Multilingual FrameNet Shared Annotation Task : a Preliminary Report. In : *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : ELRA, p. 62-68.

Torrent, T. T. ; Matos, E. E. S. ; Belcavello, F. ; Viridiano, M. ; Gamonal, M. A. ; Costa, A. D. ; Marim, M. C. (2022). Representing Context in FrameNet : A Multi-Dimensional, Multimodal Approach. *Frontiers in Psychology*, v. 13, article 838441.

Torrent, Tiago Timponi; Matos, Ely Edison da Silva; Costa, Alexandre Diniz da; Gamonal, Maúcha Andrade; Peron-Corrêa, Simone; Paiva, Vanessa Maria Ramos Lopes. (forthcoming) A Flexible Tool for a Qualia-Enriched FrameNet: The FrameNet Brasil WebTool. *Language Resources and Evaluation*.

Viridiano, M.; Torrent, T. T.; Czulo, O.; Lorenzi, A.; Matos, E.; Belcavello, F. (2022). The Case for Perspective in Multimodal Datasets. In: *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. Marseille, France: ELRA, p. 108-116.

You, L. and Liu, K. (2005). Building chinese framenet database. In: *2005 international conference on natural language processing and knowledge engineering*. IEEE, pp 301–306.