Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet, pages 23–30
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Frame-Based Annotation of Multimodal Corpora: Tracking (A)Synchronies in Meaning Construction

Frederico Belcavello, Marcelo Viridiano, Alexandre Diniz da Costa, Ely Edison da Silva Matos, Tiago Timponi Torrent
FrameNet Brasil – Federal University of Juiz de Fora
Juiz de Fora, Brazil
{frederico.belcavello, alexandre.costa, ely.matos, tiago.torrent}@ufjf.edu.br
marcelo.viridiano@gmail.com

Abstract

Multimodal aspects of human communication are key in several applications of Natural Language Processing, such as Machine Translation and Natural Language Generation. Despite recent advances in integrating multimodality into Computational Linguistics, the merge between NLP and Computer Vision techniques is still timid, especially when it comes to providing fine-grained accounts for meaning construction. This paper reports on research aiming to determine appropriate methodology and develop a computational tool to annotate multimodal corpora according to a principled structured semantic representation of events, relations and entities: FrameNet. Taking a Brazilian television travel show as corpus, a pilot study was conducted to annotate the frames that are evoked by the audio and the ones that are evoked by visual elements. We also implemented a Multimodal Annotation tool which allows annotators to choose frames and locate frame elements both in the text and in the images, while keeping track of the time span in which those elements are active in each modality. Results suggest that adding a multimodal domain to the linguistic layer of annotation and analysis contributes both to enrich the kind of information that can be tagged in a corpus, and to enhance FrameNet as a model of linguistic cognition.

Keywords: Frame Semantics, Multimodal Annotation, FrameNet

1. Introduction

The FrameNet Brasil Lab has been engaged in developing resources and applications for Tourism (Torrent el al., 2014; Diniz da Costa et al., 2018) using Frames – in the way they were defined by Fillmore (1982) – as structured representations of interrelated concepts. Frames are, then, the pivot structures for Frame Semantics, in which words are understood relative to the broader conceptual scenes they evoke (Fillmore, 1977). As the computational implementation of Frame Semantics, FrameNet has been developed as a lexicographic database that describes the words in a language against a computational representation of linguistic cognition based on frames, their frame elements (FEs) and the relations between them. The analysis is attested by the annotation of sentences representing how lexical units (LUs) instantiate the frames they evoke. FrameNet projects have been started producing databases in many languages, such as Brazilian Portuguese.[1]

In order to make FrameNet Brasil able to conduct multimodal analysis, we outlined the hypothesis that similarly to the way in which words in a sentence evoke frames and organize their elements in the syntactic locality accompanying them, visual elements in video may, then, (i) evoke frames and organize their elements on the screen or (ii) 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.

To test the hypothesis, we designed a pilot experiment for which we selected a Brazilian television travel show critically acclaimed as an excellent example of good practices in audiovisual composition. The TV format chosen also configures a novel experimental setting for research on integrated image and text comprehension, since, in this corpus, text is not a direct description of the image sequence, but correlates to it indirectly in a myriad of ways.

The methodology defined was to:

1. annotate the audio transcript using the FrameNet Brasil Annotation WebTool (Matos and Torrent, 2018), that allows for the creation of frames and relations between them, as well as for the annotation of sentences and full texts;
2. considering audio as the controlling modality in this corpus, annotate the frames evoked by visual objects or entities that are grounded on or related to the auditory guidance;
3. analyze synchronies and asynchronies between the annotations.

To accomplish the steps (ii) and (iii) we developed a Multimodal Annotation Module for the FrameNet Brasil Webtool.

The results achieved so far suggest that, at least for this TV format but maybe also for others, a fine grained semantic annotation tackling the (a)synchronous correlations that take place in a multimodal setting may provide data that is key to the development of research in Computational Linguistics and Machine Learning whose focus lies on the integration of computer vision and natural language processing and generation. Moreover, multimodal annotation may also enrich the development of FrameNets, to the extent that correlations found between modalities can attest the modeling choices made by those building frame-based resources.

[1] See https://www.globalframenet.org/partners.

2. Computational Processing of Multimodal Communication

Multimodal analyses have been growing in importance within several approaches to both Cognitive Linguistics and Natural Language Understanding, changing the scenario depicted by McKevitt (2003), according to whom little progress had been made in integrating the areas of Natural Language Processing (NLP) and Vision Processing (VP), although there had been much success in developing theories, models and systems in each of these areas separately.

Aksoy et al. (2017) present a review of the state of art on linking natural language and vision, highlighting that the related literature mostly focuses on generating descriptions of static scenes or object concepts. They, then, offer an unsupervised framework which is able to link continuous visual features to textual descriptions of videos of long manipulation activities. The results show interesting capacity of semantic scene understanding, although the linguistic material is limited to automatically generated text descriptions.

Sun et al. (2019), on the other hand, report the development of a joint model for video and language representation learning, VideoBERT, in which the text processed is captured from the original audio of the videos that integrate the corpus. Therefore, this model is capable of learn bidirectional joint distributions over sequences of visual and linguistic inputs. Although it is shown that the model learns high-level semantic features, it should be pointed out that the genre of videos selected – cooking instructions or recipe demonstrations – offers a very straightforward correlation between visual and auditory content, when compared with many other TV, audiovisual or cinematography genres.

Turner (2018) explains that multimodality is traditionally expressed in three different forms of communication and meaning construction: auditory, visual and text. Steen et al. (2018) highlight that multimodal corpora have been annotated for correlations involving mainly gesture communication and text data, and that computational infrastructure for dealing with large multimodal corpora has been under development. Both Turner and Steen lead an effort on this direction through the collaborative works of The International Distributed Little Red Hen Lab™[2], in terms of establishing tools and methodology for analyzing large multimodal corpora, mostly exploring correlations between spoken and gesture communication.

FrameNet Brasil, then, aims to establish an approach complementary to these works, since it is based on the establishment of fine-grained frame-based relations between the auditory and visual modalities, which is not restricted to human gestures. Moreover, it builds on Cohn’s (2016) systematization of the semantic investigation in multimodal data, according to the grammaticality of the modalities involved. It was used as a reference to evaluate the relation expressed by audio and video in the selected corpus. This aspect will be discussed next.

3. Multimodal Grammars

Based on Jackendoff’s (2002) parallel architecture of language, Cohn (2016) focuses on how grammar and meaning coalesce in multimodal interactions, extending beyond the semantic taxonomies typically discussed within the domain of text–image relations. He thus classifies the relations between text and image in visual narratives, evaluating the presence or absence of grammar structuring each of the modalities and also the presence or absence of semantic dominance by one of the modalities.

The first step of this method for analyzing multimodal interactions would be to determine if one of the modalities controls the other in terms of meaning, that is, if there is a semantic dominance according to which one of the modalities plays a preponderant role in determining the meaning expressed by the media. If the answer is yes, there will be a relation of assertiveness or dominance. If the answer is no, the relation will be of co-assertiveness or co-dominance.

Cohn’s model considers that there is assertiveness (or co-assertiveness) when both modalities have grammar - in the case of text modality, the grammar is expressed in terms of syntax; in the case of image, what counts as grammar is the narrative. The dominance (or co-dominance) will occur when one of the modalities has grammar and the other doesn't.

In our study we consider that, throughout the TV show, audio plays a controlling role in establishing meaning, although there are significant visual sequences in the form of video clips that express a linear narrative.

Although Cohn's (2016) model offers a coherent framework to approach multimodal data, the author does not incorporate any sort of fine-grained semantics into his model. Nonetheless, he recognizes the importance of using one for adequately tackling the interrelations and interactions between modalities and its components.

Given the lack of research incorporating fine-grained models of semantic cognition into multimodal analyses, the research presented in this paper aims to tackle the issue of meaning construction in multimodal settings, specifically on what concerns the interaction between audio (verbal expression transcribed into text) and video (not necessarily gesture communication), based on a principled structured model of human semantic cognition: FrameNet. Such a model is presented next.

4. FrameNet and Frame-Based Semantic Representation

Frames have a long history in both AI (Minsky, 1975) and linguistics (Fillmore, 1982) as structured representations of interrelated concepts. In Frame Semantics, words are

[2] See http://www.redhenlab.org

understood relative to the broader conceptual scenes they evoke (Fillmore, 1977). Hence, the expression *child-safe beach*, for example, is understood only in the context of a scene in which an Asset (the child) is exposed to some potentially Harmful_event (a strong sea current, for example).

This theoretical insight is the basis for lexicographic resources such as Berkeley FrameNet and its sister projects in other languages. Currently, there are FrameNet projects for several languages besides English, including Chinese, French, German, Italian, Japanese, Korean, Spanish, Swedish and Brazilian Portuguese. These frame-based resources have been applied to different Natural Language Understanding problems, such as conversational Artificial Inteligence (Vanzo et al. 2019) and paraphrase generation (Callison-Burch and Van Durme 2018).

4.1 Frame-to-Frame and Frame Element-to-Frame Relations

All framenets are composed of frames and their associated roles in a network of typed relations such as inheritance, perspective and subframe. The `Risk_scenario` frame alluded to above, for example, is an umbrella frame for several more specific perspectivized frames such as `Being_at_risk` (in which the Asset is exposed to a risky situation) and `Run_risk` (in which a Protagonist puts an Asset at risk voluntarily). Each perspective may be evoked by different words or by one same lexeme with different syntactic instantiation patterns.

`Being_at_risk`, for example, is evoked by adjectives such as unsafe.a and nouns such as risk.n in constructions like X is at risk. On the other hand, `Run_risk` is evoked by verbs such as risk.v and also by risk.n, but in a different construction: Y has put X at risk (Fillmore and Atkins 1992). The database structure also features annotated sentences, which attest the use of a given word in the target frame.

On top of the frame-to-frame relations traditionally used in most – if not every – FrameNet, FrameNet Brasil also developed other types of relations aimed at enriching the database structure. One of these relations links FEs to the frames licensing the lexical items that typically instantiate those elements. Hence, the FE Tourist, in the `Touring` frame, for instance, is linked via and FE-to-frame relation to the `People_by_leisure_activity` frame. 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).

Another group of relations developed by FrameNet Brasil holds between LUs and is inspired by qualia roles, based on Pustejovsky's (1995) categorization. From the four original qualia types – agentive, constitutive, formal and telic – FrameNet Brasil has developed frame-mediated ternary relations in which a given LU is linked to another LU via a subtype of quale elaborated on by a frame. Those relations will be discussed next.

4.2 Frame Mediated Ternary Qualia Relations

Although frame-to-frame and frame element-to-frame relations already provide a fine-grained semantic representation, they are unable to capture differences in the semantics of a group of lexical units within one same frame. Such differences are relevant for the semantic representation of (multimodal) texts, as the pilot analysis in this paper will demonstrate.[3]

The Generative Lexicon Theory (GLT) (Pustejovsky, 1995) arises as an approach to lexical semantics focusing on the combinatorial and denotational properties of words, as well as on peculiar aspects of the lexicon such as polysemy and type coercion. The advance of the theory is due to a dissatisfaction of many theoretical and computational linguists with the characterization of the lexicon as a closed and static set of syntactic, morphological and semantic traits.

Qualia roles emerged as characteristics or different possible context predication modes of a lexical item. Pustejovsky and Jezek (2016) argue that qualia roles "indicate a single aspect of a word's meaning, defined on the basis of the relation between the concept expressed by the word and another concept that the word evokes". There are four main qualia roles:

1. The Formal quale is the relation that distinguishes an entity within a larger domain. Like a taxonomic categorization, it includes characteristics like orientation, shape, dimensions, color, position, size etc.
2. The Constitutive quale is established between an object and its constituents and the material involved in its production.
3. The Telic quale is associated with the purpose or function of the entity. We can expand this role to a persistent and prototypical property (function, purpose or action) of the entity (object, place or person).
4. The Agentive quale refers to the factors that are involved in the origin or "coming into existence" of an entity. Characteristics included in this relation are the creator, the artifact, the natural type and a causal chain.

Figure 1 exemplifies these qualia roles for the word pizza.n.

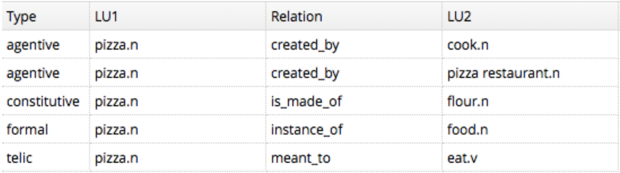
$$\begin{bmatrix} pizza.n & \\ QUALIA & \begin{bmatrix} F & = \boldsymbol{food.n} \\ T & = \boldsymbol{eat.v} \\ C & = \boldsymbol{flour.n} \\ A & = \boldsymbol{cook.n, pizza\ restaurant.n} \end{bmatrix} \end{bmatrix}$$

Fig. 1. Qualia roles for pizza.n

[3] One could argue that the creation of even more fine-grained frames would solve the problem mentioned here. Nonetheless, such a solution would be more time consuming and, at the same time, split unnecessarily into different frames, plus sharing the same background semantics and the same valence properties.

In Figure 1, we see that food.n is represented as formal_of pizza.n, being a more general category to which pizza belongs. The word eat.v is telic_of pizza.n since the latter is made to be eaten. Because it is an ingredient used in it, flour.n is constitutive_of pizza.n, while cook.n and pizza restaurant.n are agentive_of pizza.n, because they represent the person who causes the pizza to come into existence, and the place that prototypically sells it, respectively. Through qualia roles, a semantic relation is established between two words, providing a specific word with semantic features.

One recurrent problem of working with qualia is that the four relations just presented above are too generic. This has led to the proposal of long lists of subtypes for each relation (Lenci et al. 2000). However, instead of incorporating another list of relations to the FN-Br database, we use frames in this same database as mediators of ternary qualia relations to address both the lack of direct links between LUs in the framenet model and the poor specificity of qualia relations. In this innovative type of ternary relation, two LUs, 1 and 2, are linked to each other via a given quale using the background structure of frames as a way to make the quale role denser in terms of semantic information. For each quale, a set of frames was chosen from the FN-Br database based on the aspects of such quale they specify. LU1 would be related to an FE of the background frame, whereas LU2 would be related to another FE of the same frame. The frame would specify the semantics of the relation. The relations are represented in a directional fashion, that is, they are to be interpreted as unidirectional, although it is possible to create inverse relations.



Type	LU1	Relation	LU2
agentive	pizza.n	created_by	cook.n
agentive	pizza.n	created_by	pizza restaurant.n
constitutive	pizza.n	is_made_of	flour.n
formal	pizza.n	instance_of	food.n
telic	pizza.n	meant_to	eat.v

Fig. 2. Ternary qualia relations for pizza.n in the FrameNet Brasil database

Figure 2 provides an example of this implementation. In the FrameNet Brasil database the LU pizza.n has relation with five other LUs via qualia. The LU pizza.n has an Agentive relation (created_by) with pizza restaurant.n and cook.n. This relation is mediated by the Cooking_creation frame, which relates pizza.n to the FE Produced_food and pizza restaurant.n and cook.n to the FE Cook.[4] The LU pizza.n has also a Constitutive relation (is_made_of) with the LU flour.n, which is mediated by the Ingredients frame, pizza.n being related to the FE Product and flour.n to the FE Material. The Formal relation (instance_of) is established via the Exemplar frame, pizza.n being related to the FE Instance and food.n to the FE Type. Finally, the Telic relation (meant_to) establishes that pizza.n is related to the FE Tool, i.e. the object or process that has been designed specifically to achieve a purpose, in the Tool_purpose frame. As for eat.v, it is related to the FE Purpose in the same frame.

Figure 3 presents a diagram which details the ternary relations described for pizza.n .

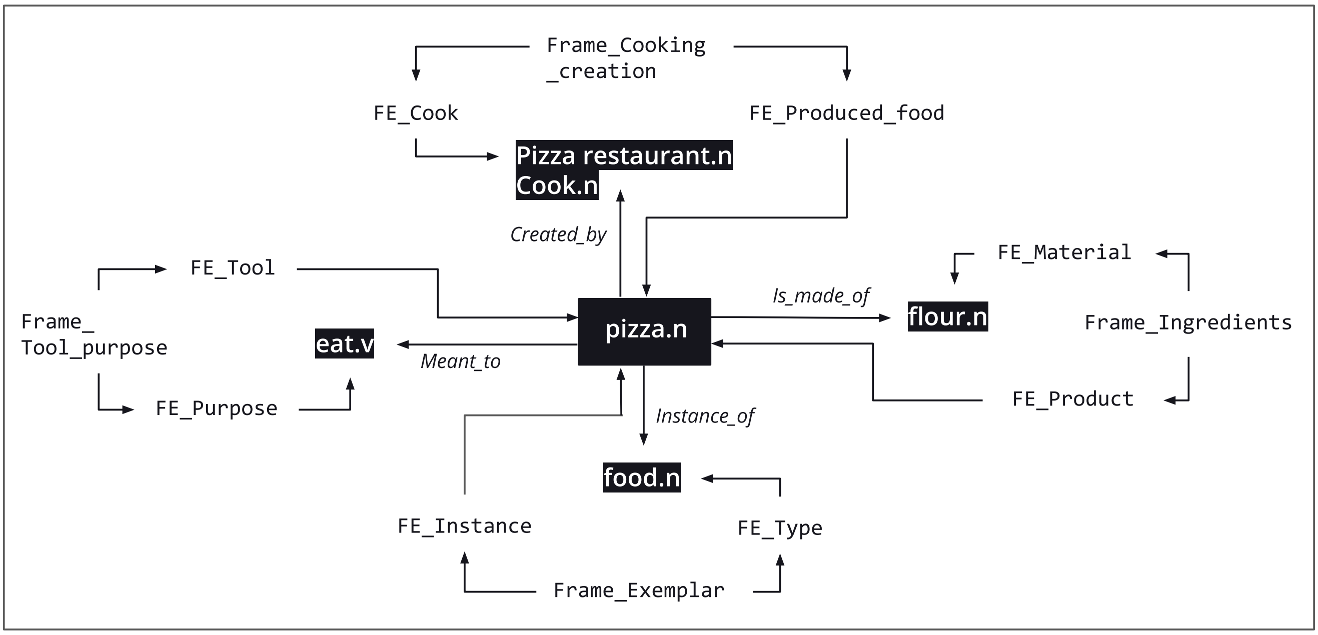


Fig. 3. Diagram of the ternary qualia relations for *pizza.n*

[4] Because we also implement metonymy relations between FEs, the peripheral FE Place can stand for the core FE Cook in the Cooking_creation frame.

As a general policy, only core – and core unexpressed – FEs can be recruited as ternary qualia mediators. The reason behind this policy relates to the very distinction between core and non-core FEs in FrameNet methodology: only core FEs are absolutely frame-specific, hence, they are the only ones that actually differentiate one frame from another.

The other policy refers to the degree of generality of frames recruited as mediators for the ternary qualia relations. Frames should be as general as possible, provided that they do not conflict with or overgeneralize the quale. For example, there are two more general frames in the inheritance chain leading to the `Tool_purpose` frame in the FrameNet Brasil database: `Inherent_purpose` and `Relation`. The `Relation` frame overgeneralizes the Telic quale, since it states that two Entities are related via a Relation_type. Because no constraints are posited for the Relation_type, it could actually refer to any type of qualia.

On the other hand, `Inherent_purpose` and `Tool_purpose` differ in terms of the nature of the LU1. In the former, it is a natural entity or phenomenon, while, in the latter, it is created by a living being. Such a difference relates to Pustejovsky's (2001) discussion on the difference between natural and functional types, and, therefore, the `Tool_purpose` frame should be used as the mediator for the Telic relation between some manmade item and its intended purpose, while the `Inherent_purpose` frame should be used for the Telic relation between a natural entity and the purpose that may be imposed to it in some context.

Given the possibilities enabled by the language model just described, as pointed out before, the hypothesis being investigated in this work is that, similarly to the way in which words in a sentence evoke frames and organize their elements in the syntactic locality accompanying them, video scenes may also either (i) evoke frames and organize their elements on the screen, or (ii) complement the frame evocation patterns of the sentences they are attached to, providing different profiling and perspective options for meaning construction, while also exploring alternative connections between concepts in the FrameNet Brasil model. To test the validity of this hypothesis and, therefore, the potential relevance of the project, an exploratory corpus study was conducted and is described in the next section.

5. Exploratory corpus study and annotation tool

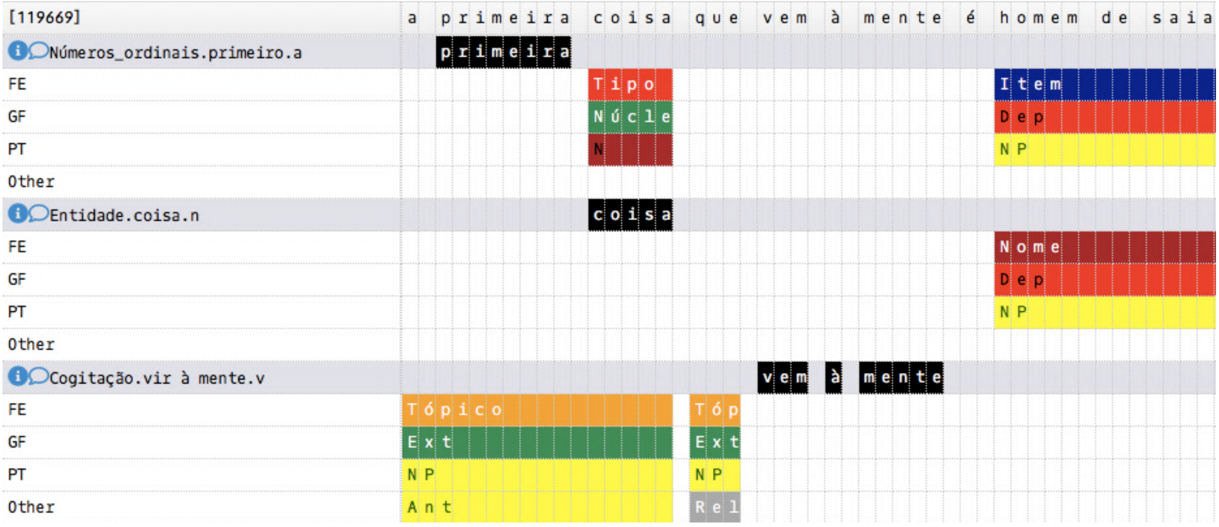
FrameNet Brasil has been building a fine-grained semantic infrastructure and developing resources and applications for the Tourism domain (Torrent et al., 2014; Diniz da Costa et al. 2018). Therefore, this exploratory study reported in this paper refers to the same such domain.

5.1 The Corpus

The corpus is composed by the first season of the Brazilian television travel show "Pedro pelo Mundo" (Pedro around the world). There are 10 episodes, of 23 minutes each. In each episode we see the host exploring a city, region or country, highlighting its cultural and socioeconomic aspects. The TV format combines voice-over sequences, short interviews and video clip sequences in a well-integrated script that offers rich composition of audio and video. For each episode, the audio transcription generates approximately 200 sentences, which means 2000 sentences for the entire season. Following the FrameNet Brasil full-text annotation average of 6.1 annotation sets per sentence, the annotation of the whole textual part of the corpus should yield, when complete, about 12,200 lexical annotation sets.

5.2 Annotation Method

In the first step for the analysis conducted in the study, one annotator manually annotated the audio transcript of one random episode of the first season, using the FrameNet Brasil Web Annotation Tool (Matos and Torrent, 2018) – an open source database management and annotation tool that allows for the creation of frames and relations between them – and following FrameNet's guidelines for full-text annotation. An example of the sort of annotation carried out in this project is shown in Figure 4.



[119669]	a	primeira	coisa	que	vem	à	mente	é	homem	de	saia
Números_ordinais.primeiro.a		primeira									
FE			Tipo						Item		
GF			Núcle						Dep		
PT			N						NP		
Other											
Entidade.coisa.n			coisa								
FE									Nome		
GF									Dep		
PT									NP		
Other											
Cogitação.vir à mente.v					vem	à	mente				
FE	Tópico			Tóp							
GF	Ext			Ext							
PT	NP			NP							
Other	Ant			Rel							

Fig. 4. Example of a sentence annotated for frames in the FN-Br WebTool

LEXICAL UNIT	AUDIO FRAME	AUDIO TIME	VIDEO FRAME	VIDEO TIME	SYNC
quando *(when)*	`Temp_collocation`	32.03 to 32.08	-	-	async
pensa *(think)*	`Cogitation`	32.18 to 32.29	-	-	async
primeira *(first)*	`Ordinal_numbers`	33.20 to 34.02	-	-	async
coisa *(thing)*	`Entity`	34.03 to 34.11	-	-	async
vem à mente *(come to mind)*	`Cogitation`	34.14 to 35.02	-	-	async
homem (*man)*	`People`	35.03 to 35.14	`People_by_origin`	36.12 to 37.12	async
saia *(skirt)*	`Clothing`	35.17 to 35.29	`Clothing`	36.12 to 37.12	async
whisky	`Food`	36.00 to 36.10	`Food`	35.02 to 36.12	sync
escocês *(Scottish)*	`Origin`	36.11 to 36.23	–	-	async
gaita de fole *(bagpipe)*	`Noise makers`	36.24 to 37.23	`Noise makers`	36.12 to 37.12	sync

Table 1. Audio (text) and video annotation comparison.

After the annotation of the audio transcript has been carried out, the same annotator annotated the video superimposed in the episodes for the same categories. Next, we contrasted the annotations, searching for matching frames while also considering the synchronicity or asynchronicity of the frames instantiated in both. The time stamps associated to the audio transcripts and the video were taken as the correlational unit between the two modalities.

5.3 Sample Annotation Discussion

In the remainder of this section, we present and discuss the data obtained from the multimodal annotation of one sentence in the corpus, transcribed in (1).

(1) Quando a gente pensa na Escócia, a primeira coisa que vem à mente é homem de saia, whisky escocês e gaita de fole.
'When we think of Scotland, the first thing that comes to mind is man in skirt, Scottish whisky and bagpipe'.

The full annotation of (1) yielded ten lexical annotation sets, while the annotation of the video it is superimposed to generated four visual annotation sets. Table 1 presents these data and how they synchronize – or not.

The six lines in white present frames found only in the audio. Because the annotation is audio-oriented, we did not annotate the frames that were present only in the video for this pilot study, although we plan to include them in the near future. The four lines highlighted in grey show the matches between frames annotated for both text/audio and video, although there is asynchrony in two of them and an indirect match in one of those two. The asynchrony is due to the fact that although evoked by both text/audio and video, the occurrences do not coincide in terms of time. In both cases the text/audio evocation occurs before the elements appear visually on screen.

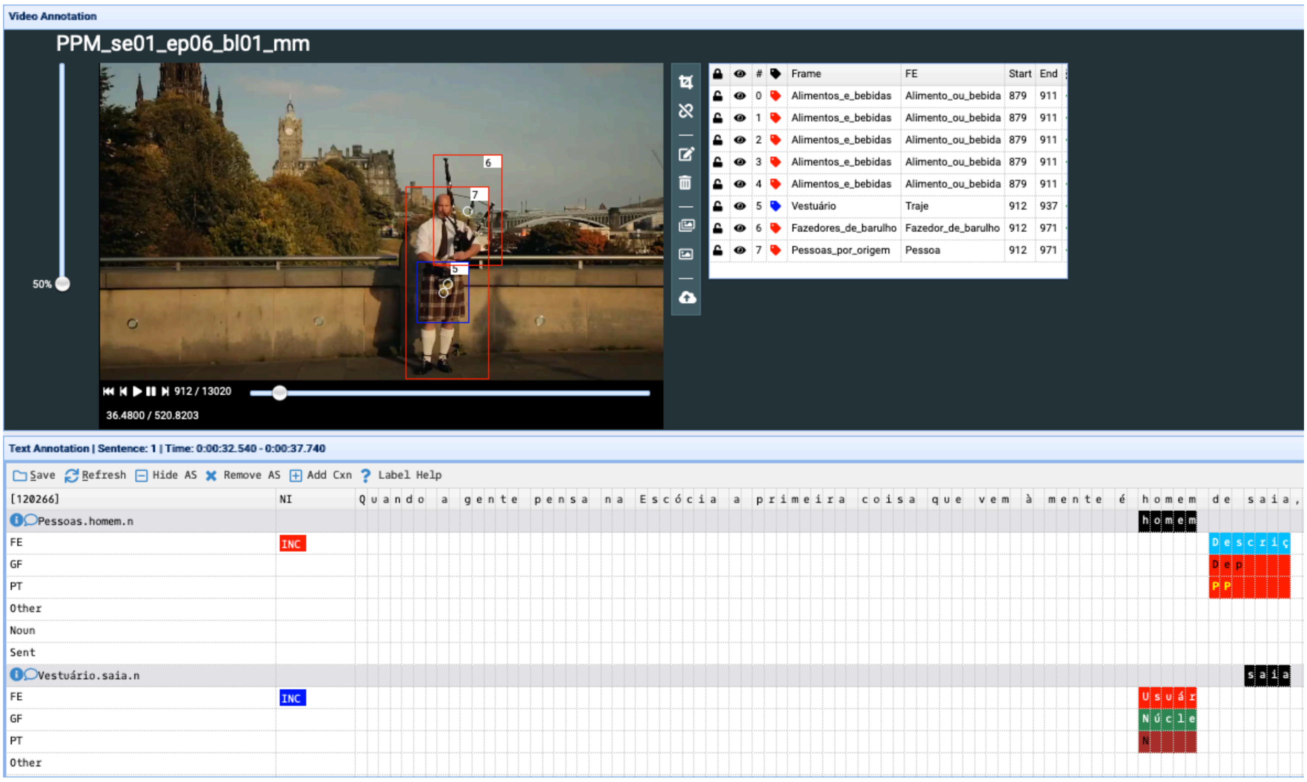
The indirect correspondence between the frames `People`, annotated for text/audio, and `People_by_origin`, annotated for the video is more interesting though. Although the latter inherits the first, this seems to be only one of the correspondences between them.

The LU evoking the `People_by_origin` frame is homem.n *'man'*. This LU does not bring any information on the origin of the person, therefore, the frame evoked is the most general of the People family of frames in FrameNet Brasil. Nonetheless, in the video annotation, the annotator chose the `People_by_origin` frame, which is evoked by the Object 7, as shown in Figure 5. The reason behind this choice is the fact that the man depicted in the video right after the audio mentions homem de saia *'man in* skirt' is wearing a kilt and playing a bagpipe, which are a typical clothing and musical instrument of Scotland, respectively. This combination of factors makes it very likely to infer that what we see is a Scottish person. Therefore, it makes possible to the annotator to choose the `People_by_origin` frame instead of the `People` frame.

The first question that arises from this sample annotation is how such a reasoning could be captured by some non-human tagger. Moreover, one could wonder whether this kind of annotation is supported by the FrameNet Brasil language model. Ternary qualia relations provide the answer to both of them (see Figure 6).

First, a subtype of the formal quale, mediated by the `Type` frame connects the LUs kilt.n and saia.n *'skirt'* in FrameNet Brasil. Second, a subtype of the constitutive quale mediated by the `Idiosyncrasy` frame connects the LU kilt.n, instantiating the FE Idiosyncrasy to the LU escocês.n *'Scot'*, instantiating the FE Entity in this frame. Finally, the LU escocês.n evokes the `People_by_origin` frame, which is precisely the one evoked by the Object 7 in Figure 5.

Figure 6 presents a summary of the connections between the multimodal elements annotated for (1), which can be found in FrameNet Brasil enriched language model.



Video Annotation

PPM_se01_ep06_bl01_mm

50%

912 / 13020

36.4800 / 520.8203

🔒	👁	#	🏷	Frame	FE	Start	End
🔓	👁	0	🏷	Alimentos_e_bebidas	Alimento_ou_bebida	879	911
🔓	👁	1	🏷	Alimentos_e_bebidas	Alimento_ou_bebida	879	911
🔓	👁	2	🏷	Alimentos_e_bebidas	Alimento_ou_bebida	879	911
🔓	👁	3	🏷	Alimentos_e_bebidas	Alimento_ou_bebida	879	911
🔓	👁	4	🏷	Alimentos_e_bebidas	Alimento_ou_bebida	879	911
🔓	👁	5	🏷	Vestuário	Traje	912	937
🔓	👁	6	🏷	Fazedores_de_barulho	Fazedor_de_barulho	912	971
🔓	👁	7	🏷	Pessoas_por_origem	Pessoa	912	971

Text Annotation | Sentence: 1 | Time: 0:00:32.540 - 0:00:37.740

Save Refresh Hide AS Remove AS Add Cxn ? Label Help

[120266]	NI	Quando a gente pensa na Escócia a primeira coisa que vem à mente é homem de saia,
Pessoas.homem.n		homem
FE	INC	Descriç
GF		Dep
PT		PP
Other		
Noun		
Sent		
Vestuário.saia.n		saia
FE	INC	Usuár
GF		Núcle
PT		N
Other		

Fig. 5. Screenshot of the FN-Br Webtool Multimodal Annotation Module

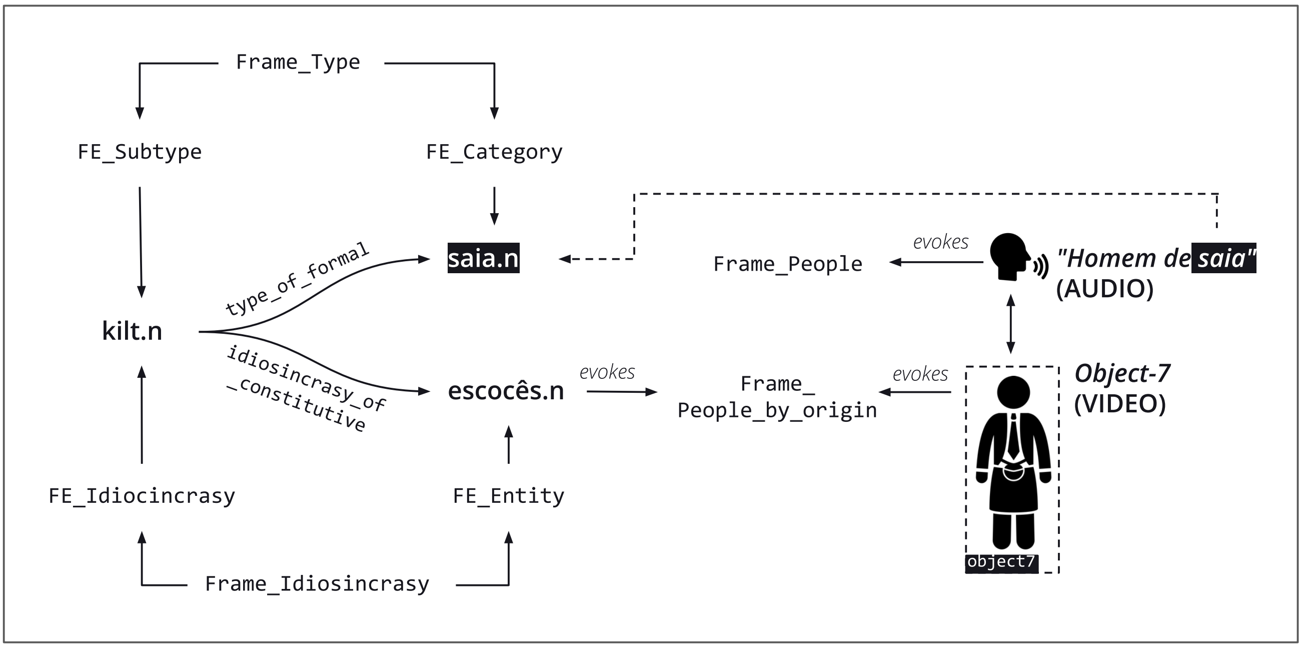


Fig. 6. Summary of connections

6. Conclusions

In this paper, we presented a tool and annotation scheme for fine-grained annotation of multimodal corpora. Such a tool controls for the synchronicity between different media types and allows for cross-modality annotation, yielding, as an annotation product, material that can shed light on the role of multimodality in language comprehension. This new annotation module was projected to run combined with the original FN-Br WebTool, which could annotate only text. The combination of both modules is crucial to multimodal annotation, since timing has demonstrated to be a key issue in measuring frame correlations across different media. Thus, the Multimodal Module allows annotators to choose frames and locate frame elements both in the text and in the images, while keeping track of the time span in which those elements are active in the video and in the audio.

There are several text annotation tools and several video and/or image annotation tools. However, they do not control for the synchronicity between different media types nor allow for cross-modality annotation. Also, none of them are frame-based and, therefore, none of them yield, as an annotation product, material that can shed light on the role of multimodality in language comprehension.

Future work includes the creation of a gold standard multimodal annotated corpus that may be used in Machine Learning applications such as Automatic Visual Semantic Role Labeling and video indexing.

7. Acknowledgements

The FrameNet Brasil Computational Linguistics Lab is funded by the CAPES PROBRAL (grant # 88887.144043/2017-00) and the CAPES STINT (grant # 99999.009910/2014-00) programs. Mr. Belcavello's PhD research is funded by the CAPES PDSE program (grant # 88881.362052/2019-01). Mr. Diniz da Costa's PhD research was funded by the CAPES PROBRAL program (grant # 88887.185051/2018-00).

8. Bibliographical References

Aksoy, E. E., Ovchinnikova, E., Orhan, A., Yang, Y., & Asfour, T. (2017). Unsupervised linking of visual features to textual descriptions in long manipulation activities. *IEEE Robotics and Automation Letters*, *2*(3), 1397-1404.

Callison-Burch, C., & Van Durme, B. (2018). *Large-Scale Paraphrasing for Natural Language Understanding*. Johns Hopkins University Baltimore United States.

Cohn, N. (2016). A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition*, *146*, 304-323.

Diniz da Costa, A., Gamonal, M. A., Paiva, V. M. R. L., Marção, N. D., Peron-Corrêa, S. R., de Almeida, V. G., ... & Torrent, T. T. (2018). FrameNet-Based Modeling of the Domains of Tourism and Sports for the Development of a Personal Travel Assistant Application. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (pp. 6-12).

Fillmore, C. J. (1977). The case for case reopened. *Syntax and semantics*, *8*(1977), 59-82.

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., & Atkins, B. T. (1992). Toward a frame-based lexicon: The semantics of RISK and its neighbors. *Frames, fields, and contrasts: New essays in semantic and lexical organization*, *103*, 75-102.

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.

Jackendoff, R., & Jackendoff, R. S. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press, USA.

Lenci, A., Busa, F., Ruimy, N., Gola, E., Monachini, M., Calzolari, N. and Zampolli, A. (2000). *Simple Linguistic Specifications, Deliverable D2, 1.* Istituto di Linguistica Computazionale Antonio Zampolli, Pisa, Italy.

Matos, E. E. S., Torrent, T. T. (2018) *FN-Br WebTool: FrameNet Brasil Web Annotation Tool.* INPI Registration Number BR512018051603-3

McKevitt, P. (2003, January). MultiModal semantic representation. In *First Working Meeting of the SIGSEM Working Group on the Representation of MultiModal Semantic Information* (pp. 1-16).

Minsky, M. (1975). A Framework for Representing Knowledge: In Winston, PH (eds.), The Psychology of Computer Vision, 211–277.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, USA, MIT Press.

Pustejovsky, J. (2001). Type construction and the logic of concepts. In: P. Bouillon and F. Busa. *The language of word meaning* (pp. 91–123). Cambridge University Press, New York, USA.

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

Steen, F. F., Hougaard, A., Joo, J., Olza, I., Cánovas, C. P., Pleshakova, A., ... & Turner, M. (2018). Toward an infrastructure for data-driven multimodal communication research. *Linguistics Vanguard*, *4*(1).

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). Videobert: A joint model for video and language representation learning. arXiv preprint arXiv:1904.01766.

Torrent, T., Salomão, M. M., Campos, F., Braga, R., Matos, E., Gamonal, M., ... & Peron, S. (2014, August). Copa 2014 FrameNet Brasil: a frame-based trilingual electronic dictionary for the Football World Cup. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations* (pp. 10-14).

Turner, M. (2018). The Role of Creativity in Multimodal Construction Grammar. *Zeitschrift für Anglistik und Amerikanistik*, *66*(3), 357-370.

Vanzo, A., Bastianelli, E., & Lemon, O. (2019). Hierarchical multi-task natural language understanding for cross-domain conversational ai: HERMIT NLU. *arXiv preprint arXiv:1910.00912.*