

**FEDERAL UNIVERSITY OF JUIZ DE FORA**  
**FACULTY OF LETTERS**  
**GRADUATE PROGRAM IN LINGUISTICS**

**Frederico Belcavello**

**FrameNet Annotation for Multimodal Corpora:** devising a methodology for the semantic representation of text-image interactions in audiovisual productions

Juiz de Fora

2023

**Frederico Belcavello**

**FrameNet Annotation for Multimodal Corpora:** devising a methodology for the semantic representation of text-image interactions in audiovisual productions

Dissertation Project presented to the Graduate Program in Linguistics at the Federal University of Juiz de Fora as a partial requisite for being awarded the title of Doctor in Linguistics.  
Focus area: Linguistics.

Advisor: Dr. Tiago Timponi Torrent  
Co-advisor: Dr. Mark Turner

Juiz de Fora  
2023

Ficha catalográfica elaborada através do programa de geração automática da Biblioteca Universitária da UFJF, com os dados fornecidos pelo(a) autor(a)

Belcavello, Frederico.

FrameNet annotation for multimodal corpora : devising a methodology for the semantic representation of text-image interactions in audiovisual productions / Frederico Belcavello. -- 2023. 134 p. : il.

Orientador: Tiago Timponi Torrent

Coorientador: Mark Turner

Tese (doutorado) - Universidade Federal de Juiz de Fora, Faculdade de Letras. Programa de Pós-Graduação em Linguística, 2023.

1. FrameNet. 2. Frame Semantics. 3. Multimodality. I. Torrent, Tiago Timponi, orient. II. Turner, Mark, coorient. III. Título.



FEDERAL UNIVERSITY OF JUIZ DE FORA  
RESEARCH AND GRADUATE PROGRAMS OFFICE



Frederico Belcavello Guedes

**FrameNet Annotation for Multimodal Corpora:** devising a methodology for the semantic representation of text-image interactions in audiovisual productions

Thesis submitted to the  
Graduate Program in Linguistics  
of the Federal University  
of Juiz de Fora as a partial  
requirement for obtaining  
a Doctorate degree in Linguistics.  
Concentration area: Linguistics

Approved on 27 of June of 2023.

EXAMINING BOARD

Prof(a)Dr(a) Tiago Timponi Torrent - Orientador  
Universidade Federal de Juiz de Fora

Prof(a)Dr(a) Mark Turner  
Case Western Reserve University

Prof(a)Dr(a) Aline Alves Fonseca  
Universidade Federal de Juiz de Fora

Prof(a)Dr(a) Janina Wildfeuer  
University of Groningen

Prof(a)Dr(a) André Vinícius Lopes Coneglian  
Universidade Federal de Minas Gerais

Prof(a)Dr(a) Ely Edison da Silva Matos  
Universidade Federal de Juiz de Fora



Documento assinado eletronicamente por **Tiago Timponi Torrent, Coordenador(a)**, em 27/06/2023, às 16:51, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Aline Alves Fonseca, Professor(a)**, em 27/06/2023, às 16:55, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Ely Edison da Silva Matos, Técnico Administrativo em Educação**, em 27/06/2023, às 16:55, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Andre Vincius Lopes Coneglian, Usuário Externo**, em 27/06/2023, às 16:58, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Janina Wildfeuer, Usuário Externo**, em 27/06/2023, às 16:59, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Mark Bernard Turner, Usuário Externo**, em 27/06/2023, às 17:09, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf ([www2.ufjf.br/SEI](http://www2.ufjf.br/SEI)) através do ícone Conferência de Documentos, informando o código verificador **1319039** e o código CRC **7F962631**.

## ACKNOWLEDGEMENTS

This work, formally authored by a single person, is in fact the result of the commitment of dozens of collaborations. First of all, to be a member of the FrameNet Brasil Research Laboratory means to produce science collectively, and I am very grateful for that. Thus, there were many people who participated directly, very directly or extremely directly in the process of construction and development of the research that here takes the form of a dissertation. Thanks to all of you!

I also thank the Graduate Program in Linguistics at UFJF – faculty, secretariat and coordination – which offered all the academic support for each one of the important steps of study and research.

Thanks also to the UFJF Communications Office which allowed me, at the first moment, to start the doctorate course, and to the Faculty of Letters, which welcomed me during the process and offered me the conditions to conciliate institutional and academic tasks.

My gratitude also goes to CAPES for funding my visiting student researcher period at Case Western Reserve University (grant number 88881.362052/2019-01). And many thanks to the Department of Cognitive Science at CWRU for receiving me with affection and attention and offering me unique and valuable opportunities to develop my research.

To the UFJF's Research Ethics Committee (CEP) and the Center for Studies in Language Acquisition and Psycholinguistics at (NEALP) I thank for the support in the design, processing, and execution of the eye-tracking experiment.

Special thanks to Dr. Patrícia Nora and Dr. Peter Uhrig for their invaluable contributions during the qualifying committee meeting for this dissertation; to Dr. Janina Wildfeuer for the extra contributions at the time of qualification and for accepting to participate in the final evaluation committee of this dissertation; to Dr. André Coneglian for being part of the evaluation committee; to Dr. Aline Fonseca for her guidance in the design and execution of the eye tracking experiment and for participating in the evaluation committee; to Dr. Ely Matos for his tireless support in the development and improvement of the entire technical infrastructure for this project and for accepting to participate in the committee; to Dr. Mark Turner for so

promptly receiving me at CWRU CogSci, sharing his knowledge, and for honoring me with his participation as a co-advisor.

Thank you, Dr. Tiago Torrent, for trusting me to conduct this research and for guiding it with such dedication, commitment, and care. Thank you for all the opportunities you have offered, encouraged, and supported me.

Thank you to my boys.

Thank you to my brother.

Thank you, Mom and Dad.

Thank you Mônica.

## ABSTRACT

Multimodal analyses have been growing in importance within several approaches to Cognitive Linguistics and applied fields such as Natural Language Understanding. Nonetheless fine-grained semantic representations of multimodal objects are still lacking, especially in terms of integrating areas such as Natural Language Processing and Computer Vision, which are key for the implementation of multimodality in Computational Linguistics. In this dissertation, we propose a methodology for extending FrameNet annotation to the multimodal domain, since FrameNet can provide fine-grained semantic representations, particularly with a database enriched by Qualia and other interframal and intraframal relations, as it is the case of FrameNet Brasil. 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 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. The corpus annotated for testing the hypothesis is composed of episodes of a Brazilian TV Travel Series critically acclaimed as an exemplar of good practices in audiovisual composition. The TV genre 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 with it indirectly in a myriad of ways. The dissertation also reports on an eye-tracker experiment conducted to validate the approach proposed to a text-oriented annotation. The experiment demonstrated that it is not possible to determine that text impacts gaze directly and was taken as a reinforcement to the approach of valorizing modes combination. Last, we present the Frame<sup>2</sup> dataset, the product of the annotation task carried out for the corpus following both the methodology and guidelines proposed. The results achieved demonstrate that, at least for this TV genre but possibly also for others, a fine-grained semantic annotation tackling the diverse correlations that take place in a multimodal setting provides new perspective in multimodal comprehension modeling. Moreover, multimodal annotation also enriches the development of FrameNets, to the extent that correlations found between modalities can attest the modeling choices made by those building frame-based resources.

Keywords: FrameNet. Frame Semantics. Multimodality.

## RESUMO

Análises multimodais vêm crescendo em importância em várias abordagens da Linguística Cognitiva e em diversas áreas de aplicação, como o da Compreensão de Linguagem Natural. No entanto, há significativa carência de representações semânticas refinadas de objetos multimodais, especialmente em termos de integração de áreas como Processamento de Linguagem Natural e Visão Computacional, que são fundamentais para a implementação de multimodalidade no campo da Linguística Computacional. Nesta tese, propomos uma metodologia para estender o método de anotação da FrameNet ao domínio multimodal, uma vez que a FrameNet pode fornecer representações semânticas refinadas, particularmente com um banco de dados enriquecido por Qualia e outras relações interframe e intraframe, como é o caso do FrameNet Brasil. Para tornar a FrameNet Brasil capaz de realizar análises multimodais, delineamos a hipótese de que, assim como as palavras em uma frase evocam frames e organizam seus elementos na localidade sintática que os acompanha, os elementos visuais nos planos de vídeo também podem evocar frames e organizar seus elementos na tela ou trabalhar de forma complementar aos padrões de evocação de frames das sentenças narradas simultaneamente ao seu aparecimento na tela, proporcionando diferentes perfis e opções de perspectiva para a construção de sentido. O corpus anotado para testar a hipótese é composto por episódios de um programa televisivo de viagens brasileiro aclamado pela crítica como um exemplo de boas práticas em composição audiovisual. O gênero televisivo escolhido também configura um novo conjunto experimental para a pesquisa em imagem integrada e compreensão textual, uma vez que, neste corpus, o texto não é uma descrição direta da sequência de imagens, mas se correlaciona com ela indiretamente em uma miríade de formas diversa. A Tese também relata um experimento de rastreamento ocular realizado para validar a abordagem proposta para uma anotação orientada por texto. O experimento demonstrou que não é possível determinar que o texto impacta diretamente o direcionamento do olhar e foi tomado como um reforço para a abordagem de valorização da combinação de modos. Por fim, apresentamos o conjunto de dados Frame2, produto da tarefa de anotação realizada para o corpus seguindo a metodologia e as diretrizes propostas. Os resultados obtidos demonstram que, pelo menos para esse gênero de TV, mas possivelmente também para outros, uma anotação semântica refinada que aborde as diversas correlações que ocorrem em um ambiente multimodal oferece uma nova perspectiva na modelagem da compreensão multimodal. Além disso, a anotação multimodal também enriquece o desenvolvimento de FrameNets, na medida em que as correlações encontradas entre



as modalidades podem atestar as escolhas de modelagem feitas por aqueles que criam recursos baseados em frames.

Palavras-chave: FrameNet. Semântica de Frames. Multimodalidade.

## LIST OF FIGURES

Figure 1 – Barthes' classification of text image relations represented graphically .....	15
Figure 2 – Modality quotes – what might a ‘mode’ be?.....	18
Figure 3 – An organic compound: image or text? .....	20
Figure 4 – Visual poem in Portuguese.....	21
Figure 5 – A photograph of a street with verbal and non-verbal signs .....	38
Figure 6 – The parallel architecture expanded to allow for multimodal interactions.....	40
Figure 7 – The canonical constituency phase of visual narrative in Arc.....	41
Figure 8 – Narrative grammar applied to a sequence from “ <i>Pedro pelo Mundo</i> ”.....	43
Figure 9 – Gross differences in dimensions between prototypical cases of drawn and filmed	44
Figure 10 – Model of the Scene Perception & Event Theory (SPECT) theoretical framework .....	45
Figure 11 – Step-by-step method for the analysis of multimodal interactions.....	47
Figure 12 – Audio guided sequence from “ <i>Pedro pelo Mundo</i> ” .....	48
Figure 13 – Child safe and tornado safe signs.....	50
Figure 14 – Risk_scenario frame in Berkeley FrameNet.....	51
Figure 15 – Frame to frame relations for the Risk_scenario.....	53
Figure 16 – Frame-to-frame relations legend.....	54
Figure 17 – Lexical Annotations in Berkeley FrameNet.....	55
Figure 18 – Full-text annotation in the FrameNet Brasil database.....	56
Figure 19 – Example of annotation layers in the FrameNet Brasil WebTool .....	57
Figure 20 – Qualia structure representation .....	59
Figure 21 – Qualia roles for pizza.n .....	59
Figure 22 – Frame-mediated ternary qualia relations for pizza.n in FN-Br .....	60
Figure 23 – Multimodal Corpus Import Pipeline .....	65
Figure 24 – Example of instructional sentence on screen .....	68
Figure 25 – Heat map comparison for ‘homem de saia’ .....	72
Figure 26 – Heat map comparison for ‘gaita de fole’ .....	72
Figure 27 – Gaze plot for ‘gaita de fole’ .....	73
Figure 28 – Gaze plot for the haggis arrival .....	74
Figure 29 – Heat map for the dish arrival shot.....	74
Figure 30 – Gaze plot for Gordon Nicolson revelation.....	75
Figure 31 – Heat map for Nicolson and his kilt .....	75
Figure 32 – Initial framing and heat map for the back view of the kilt.....	76
Figure 33 – In point, out point and heat map for 'Pedro de saia' .....	77
Figure 34 – Pedro in a skirt heat map 2.....	78
Figure 35 – Heat map for the beer in the pub.....	79
Figure 36 – Heat map and gaze plot for the pub .....	80
Figure 37 – Heat maps for England similarities .....	81
Figure 38 – Full-text annotation example 2.....	84
Figure 39 – Charon’s dynamic mode annotation workspace .....	85
Figure 40 – Video panel detail for creating new objects.....	86
Figure 41 – Annotation panel detail for labeling new objects.....	87
Figure 42 – Comparative shot board of image annotation and gaze fixations .....	90
Figure 43 – Image annotation for People_by_origin.....	97
Figure 44 – Frame-mediated ternary qualia relations for <i>homem de saia.n</i> and object 24 .....	98
Figure 45 – Sankey diagram for qualia relation between <i>saia.n</i> and <i>kilt.n</i> at FN-Br WebTool Report Module.....	99

Figure 46 – Sankey diagram for potential qualia relations derive from LUs in a time interval .....	100
Figure 47 – Report example of the 'Time interval' menu in the FN-Br WebTool Report Module.....	101
Figure 48 – Report example of the 'Sentence' menu in the FN-Br WebTool Report Module	102
Figure 49 – Report example of the 'Synchronicity menu' in the FN-Br WebTool Report Module.....	103
Figure 50 – Report example of the 'Frame-frame menu' in the FN-Br WebTool Report Module .....	104
Figure 51 – Visual objects annotation of (14).....	106
Figure 52 – Ternary qualia relations in the multimodal annotation of sentence (14) .....	107
Figure 53 – Multiple coincident visual objects .....	108
Figure 54 – The Cathedral annotated as ATTRACTION in Attracting_tourists.....	110
Figure 55 – Image annotation for Attraction_tourism.....	110
Figure 56 – Sankey diagram report on frame-to-frame relation between Attraction_tourism and Attracting_tourists frames.....	111
Figure 57 – Image annotation for the Ingestion frame .....	112
Figure 58 – Example of multiple objects for a single lexical unit.....	113
Figure 59 – Multiple objects instantiating FOOD_OR_BEVERAGE FE.....	114
Figure 60 – Visual object embodies a FE.....	115

## LIST OF TABLES

Table 1 – Corpus annotation totals .....	93
Table 2 – Corpus annotation averages.....	94
Table 3 – Numbers of discrete frames annotated .....	94
Table 4 – Discrete LUs as CV Name per episode .....	95
Table 5 – Matching ratio of frames annotated for images.....	96

## TABLE OF CONTENTS

<b>1 INTRODUCTION</b> .....	<b>9</b>
<b>2 MULTIMODAL ANALYSYS FOR LANGUAGE TECHNOLOGY</b> .....	<b>13</b>
2.1 FUNDAMENTALS OF MULTIMODAL ANALYSIS .....	13
2.2 LITERACY AND GENRE STUDIES .....	23
2.3 TRAVEL SHOW ON TV AS A MULTIMODAL GENRE.....	30
2.4 RELEVANCE OF MULTIMODAL OBJECTS FOR COMPUTATIONAL APPLICATIONS.....	33
<b>3 THE GRAMMAR AND THE SEMANTICS OF MULTIMODAL COMMUNICATION</b> .....	<b>38</b>
3.1 VISUAL AND FILMIC NARRATIVE GRAMMARS .....	39
3.2 FRAMENET BRASIL: A FRAME-BASED MEANING REPRESENTATION ENRICHED WITH QUALIA STRUCTURE.....	49
<b>3.2.1 FrameNet Basics</b> .....	<b>49</b>
<b>3.2.2 FrameNet Brasil</b> .....	<b>57</b>
<b>4 CORPUS AND METHODS</b> .....	<b>63</b>
4.1 CORPUS.....	63
4.2 THE CORPUS IMPORT PIPELINE .....	64
4.3 SPOKEN AUDIO DOMINANCE INVESTIGATION EXPERIMENT .....	66
4.4 CORPUS ANNOTATION .....	69
<b>5 REFINING THE ANNOTATION METHODOLOGY</b> .....	<b>71</b>
5.1 RESULTS OF THE AUDIO-DOMINANCE VERIFICATION EXPERIMENT .....	71
5.2 FRAME-BASED ANNOTATION METHODOLOGY OF AUDIOVISUAL MULTIMODAL CORPORA.....	83
5.3 CONTRASTING ANNOTATIONS AND FIXATIONS.....	89
<b>6 THE FRAME<sup>2</sup> DATASET</b> .....	<b>93</b>

6.1 IMAGE SPECIFIES TEXT .....	96
6.2 VISUAL OBJECT ANNOTATED FOR ONE FRAME INSTANTIATES ANOTHER FRAME EVOKED BY THE LEXICAL UNIT IN THE TEXT .....	105
6.3 DUPLICATED VISUAL OBJECT FOR MATCHING DIFFERENT INSTANCES OF DIFFERENT LEXICAL UNITS .....	107
6.4 PERSPECTIVE AND TEXT-ORIENTATION SHIFTS VISUAL OBJECT ANNOTATION .....	109
6.5 VISUAL OBJECT INSTANTIATES WHAT IS A NULL INSTANTIATION IN TEXT .....	111
6.6 MULTIPLE VISUAL OBJECTS MATCHING A SINGLE LEXICAL UNIT .....	113
6.7 BLENDING ENTITY FROM VISUAL OBJECT TO INSTANTIATE FRAME ELEMENT IN TEXT .....	114
6.8 REMARKS ON THE FRAME <sup>2</sup> DATASET .....	116
<b>7 CONCLUSION .....</b>	<b>117</b>
<b>REFERENCES .....</b>	<b>120</b>

## 1 INTRODUCTION

The concept of *frame* has some decades of history in Linguistics, since Charles J. Fillmore formalized a definition for *semantic frame* as “any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits” (FILLMORE, 1982, p. 11). The idea of conceptual units evoked by words, however, was already present in Fillmore’s work throughout the 1970’s (FILLMORE 1975, 1976, 1977a, 1977b, 1977c) and also in the field of Artificial Intelligence with Marvin Minsky, to whom “A *frame* is a data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child’s birthday party” (MINSKY, 1974, p. 1).

In the 1990’s, Frame Semantics gained a computational implementation with the launching of the Berkeley FrameNet project, 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 for many languages, such as Brazilian Portuguese.

The FrameNet Brasil Lab has started developing its own database in 2009. 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 evoked by 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 (GAMONAL, 2017). A third group of relations developed by FrameNet Brasil holds between LUs and is inspired by *Qualia* roles, based on Pustejovsky’s (1995) proposal. From the original qualia types, FrameNet Brasil has developed *frame-mediated ternary qualia* relations in which a given LU is linked to another LU via a subtype of quale elaborated on by a frame (TORRENT et al., 2022).

Since the beginning, the focus of FrameNet has always been on textual data. FrameNet Brasil, however, has envisioned the possibility of analyzing multimodal data as a new way of enriching its database and starting to take part in this important field of research. Multimodal aspects of human communication have been empirically perceived for a long time in history, cited as parallel aspects of language, but not given the proportional attention in research. Recently, however, in the wake of the Social Semiotics approach (HODGE; KRESS, 1998), Multimodality has been not only on the horizon, but at the center of many efforts to account for the combination of modes in the process of meaning construction. That is the case, for example,

in several applications of Natural Language Processing, such as Machine Translation (ELLIOT et al. 2015; SPECIA et al. 2016; CALIXTO et al., 2016; WEISS et al., 2017; BÉRARD et al., 2018; ZHENG et al., 2018; SANABRIA et al.; WANG et al. 2019) and Natural Language Generation (DEVLIN et al., 2015; FANG et al., 2015; BATRA, HE and VOGIATZIS, 2016; AKSOY et al., 2017; NIKOLAUS et al., 2019; SUN et al., 2019; YANG and OKAZAKI, 2020). This context has, then, boosted the interest of the FrameNet Brasil Lab in adding multimodal phenomena, corpora, data and analyses to its scope, which is the case of this dissertation. This interest was officially embraced when the FrameNet Brasil Lab became part of ReINVenTA, the Research and Innovation Network for Vision and Text Analysis. ReINVenTA is a research network in computational semantic processing of multimodal objects in the state of Minas Gerais, Brazil. As such, it brings together research projects dedicated to building and evaluating a computational model for representing objects such as TV programs and pairings of static images and text. To this end, it mobilizes laboratories and research groups from UFJF, UFMG, UFU and PUC-MG with expertise in Model Development for Natural Language Understanding, Artificial Intelligence, Knowledge Discovery and Assistive Technologies. With this confluence of expertise and projects, the ReINVenTA network hopes to achieve: (i) the expansion of the coverage of the FrameNet model to Brazilian Portuguese; (ii) the constitution of a gold standard dataset of semantically annotated and psycholinguistically validated multimodal objects; (iii) the development of artificial intelligence algorithms for automatic labeling and knowledge discovery in multimodal objects; and (iv) the proposition of best practices for video audiodescription.

Considering the principles of frame semantics and the FrameNet architecture, the driving research question we address in this dissertation is: how can we build a frame-based model to address the coalescence of image and audio in the process of meaning construction prompted by audiovisual media stimuli? The hypothesis we investigate as the possible answer to this question is that visual elements in a video sequence may (i) evoke frames, similarly to the way in which words in a sentence do, and (ii) combine with the words in sentences of the simultaneous spoken audio to offer a complementing role in the frame evocation patterns – which provide different profiling and perspective options for meaning construction, while also exploring alternative connections between concepts in the FrameNet Brasil model.

Considering the hypothesis, we developed a frame-based annotation methodology for audiovisual corpora. To achieve this, we built a corpus composed by the first season (ten



episodes) of the Brazilian TV Travel Series “*Pedro pelo Mundo*”<sup>1</sup>, which offers 230 minutes of rich combination of spoken audio and image sequences. Moreover, we developed a pipeline to process the videos in the corpus, separating spoken audio from images, making it possible to annotate both of them, abiding to the specific requisites each may demand. The pipeline also allows for the choice of whether any of the modes should be annotated first and/or whether any of them should guide the annotation process, exercising a role of dominance over the other in the process of meaning construction.

To investigate whether one of the modes annotated for in the corpus exercised a dominant role over the other, we designed an eye-tracking experiment. Two different groups of participants watched two different versions of corpus’ extracts – one complete, one modified by the removal of spoken audio –, as a way of evaluating whether the spoken audio guides the viewers gaze to visual elements on screen. Upon the conclusion of the experiment and the analysis of the resulting data, we determined the guidelines for instructing the annotation task which comprised the totality of the corpus for both spoken audio and visual modes.

The resulting annotated dataset is composed by 2,195 sentences, transcribed from 230 minutes of video. The sentences generated 11,796 annotation sets, while the images have been annotated for 6,841 visual objects. To the best of our knowledge, this is the first dataset that combines multimodal approach and Frame Semantics for video annotation of visual objects.

This dissertation reports on the whole path of this research and is organized in seven chapters, the first being this introduction. In chapter 2, we present the foundations of multimodal analysis and its application possibilities – from literacy to computational analysis – with a special attention to multimodality and genre studies. In this sense, we present the characteristics of the TV Travel Series as a multimodal genre to be explored as the corpus.

Chapter 3 discusses the grammar and the semantics of multimodal communication, taking Neil Cohn’s categorization of Visual Narrative Grammar (COHN, 2013, 2016a, 2019) and Filmic Narrative Grammar (COHN, 2016b) as a means to identify the semantic elements and patterns for the combination of spoken audio and image as elaborated in the main hypothesis. The chapter also presents FrameNet as a model for fine-grained semantic representation from its original Berkeley version to the enriched database of FrameNet Brasil.

In chapter 4, we present the “*Pedro pelo Mundo*” corpus built for achieving the goals of this dissertation, the tool and the pipeline developed to construct it. Chapter 4 also reports

---

<sup>1</sup> Pedro around the World.

on the spoken audio dominance investigation experiment conducted as a step for defining the methodology to be used in the annotation task.

Chapter 5 presents the results of the experiment, discusses how they impact on the on the annotation and, then, presents the refined methodology proposed for annotating both text and image in the “*Pedro pelo Mundo*” corpus. In this chapter we also present an evaluation of the annotation based on the experiment results.

In chapter 6 we present the Frame<sup>2</sup> dataset, the product of the annotation task carried out in this dissertation, and demonstrate how the data provides a fine-grained semantic representation of the complex combination of spoken-audio and images in terms of meaning-making in the experience of watching the TV Travel Series chosen.

The results indicate that a FrameNet enriched by other relations such as Frame-based Ternary Qualia allows for modeling rich and complex audiovisual integration, as offered by the selected corpus. This means that applying a fine-grained semantic annotation tackling the synchronous and asynchronous correlations that take place in a multimodal setting provide data that is key to the development of research in multimodal approaches for Computational Linguistics. The development of a frame-based multimodal annotation methodology also enriches the development of FrameNets, expanding the possibilities of their use into meaning construction research and frame-based computational solutions of various types.

## 2 MULTIMODAL ANALYSYS FOR LANGUAGE TECHNOLOGY

The ability of human beings to communicate with each other using sequences of linguistic elements is almost (if not) always associated with other and diverse modalities of expression. This perception, though, has not been taken into consideration as a key factor by many linguistic theories and/or language models, which kept their approaches aligned with the fundamentals postulated by Ferdinand du Saussure when establishing the object of Linguistics as a science: the study of purely linguistic semiosis (SAUSSURE, 1959 [1916]). Steen et al. (2018) point out that it is not difficult to understand why the systematic study of human communication is historically connected to the analysis of written representation of language, once writing represents a highly structured way of expression.

This does not mean that the presence of multimodal stimuli or inputs in human daily life can be considered a novelty. Steen and Turner (2013, p. 1) highlight that “multimodal communication predates and contextualizes language, and extends into a series of social, artistic, and technological innovations, from dance to cave painting, from theater to cinema, from town criers to television news”. Even Saussure understood that the external elements of language had an important role in communication, but they should be the object of other sciences, such as Ethnography, Sociology or Psychology, but also a new field should be built in order to take care of rites, culture and signs: Semiology. However, there is an empirical perception that the recently protagonism of digital devices, the internet and multimodal Artificial Intelligence emphasizes the combinations of diverse modes of expression as the norm in media nowadays. Therefore, there is a broadly recognized growth of multimodal analysis, that is, the combined analysis of at least two of the following aspects of human communication: verbal/textual, gestural, auditory, and visual. The remainder of this chapter will approach multimodal analyses from the perspective of their applicability to language technology.

### 2.1 FUNDAMENTALS OF MULTIMODAL ANALYSIS

Bateman, Wildfeuer and Hiippala (2017) define multimodality as “a way of characterizing communicative situations (considered very broadly) which rely upon combinations of different ‘forms’ of communication to be effective”. Adami and Kress (2014) state that “multimodality is an approach is not a theory” in the sense that investigating multimodal aspects of communication can be a goal of a broad range of theories. It is noteworthy that the concepts of multimodality and multimodal communication are taken for

granted. The historical contextualization of the origins of the concepts is not usually a focal point in publications on the subject. Likewise, this is not the key point of this dissertation, but we consider it important to dig briefly in the direction of the foundations of the concept.

The existence of multimodality as an empirical phenomenon can be invoked from the classical Greek philosophers and their attempts to both understand truth, the world or even the language, and summarize the appropriate and successful ways of presenting themselves in public life, through rhetoric and poetics, as written by Aristotle, for example. These approaches have often emphasized the role of voice, gesture and expressions in public speeches and in how they would interfere in meaning construction and suasion.

In modern science, within linguistics and semiotics, we can claim the early presence of the notion of multiple modes both in Saussure's and in Peirce's contributions to the foundation of these fields of knowledge, pursuing the systematization of how signs are used to convey meaning. Despite the distinctions of their respective theories, both Saussure (1959 [1916]) and Peirce (1931; 1977 [1953]) considered that the manifestation of semioses does not occur in isolation and, in this sense, they paved the way to a multimodal approach, as summarized by Roswell and Collier:

Saussure developed a formalized approach to semiotics that described how signs have meaning relationships to each other. Peirce, on the other hand, believed that people use semiotic resources at hand to communicate. One of Peirce's well-known phrases is "we think only in signs." Along with Saussure, he discussed the signifier as the form and the signified as the concept one derives from the form. In multimodal parlance, the signifier is the material and mode and the signified is how meaning is made. Both theories are complex, and this summary does not do them justice; nonetheless, on the whole, what both semioticians foreground in their work is an opening up of what text is or can be, and the germs of their theories grew into multimodality. (ROSWELL; COLLIER, 2017, pp. 313-314).

We can also find foundations for the contemporary multimodal approach in Roland Barthes' (1977[1964], p.38) proposed taxonomy to analyze "the functions of the linguistic message with regard to the (twofold) iconic message": anchorage and relay. Anchorage would be the denominative function in which a visual object has its meaning denoted recursively to a nomenclature – as in the case of a descriptive caption of a photograph, for example. In all different ways of anchorage, the text has a "repressive value" in determining the limits of interpretation of an image. On the other hand, the function of relay would refer to the situations in which text and image share, in some way, a complementary relationship: "the words, in the same way as the images, are fragments of a more general syntagm and the unity of the message

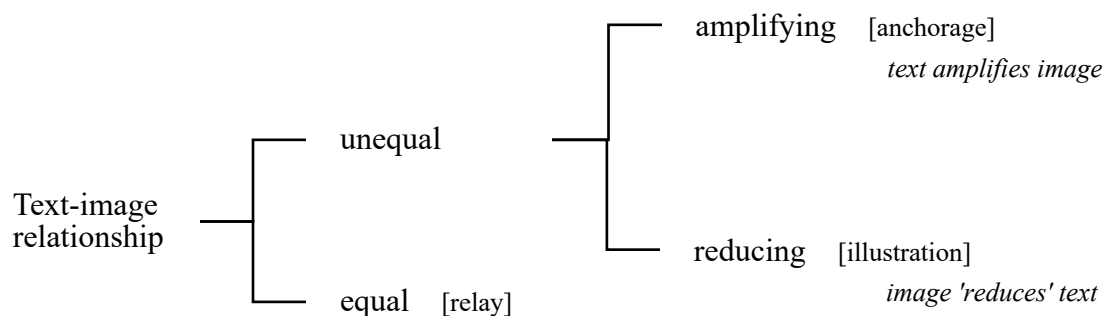
is realized at a higher level, that of the story, the anecdote, the diegesis” (BARTHES, 1977[1964], p. 41). Barthes also considers that the co-existence of both functions can also occur, but generally there would be some kind of dominance.

Prior to the proposition of the twofold anchorage/relay taxonomy, Barthes (1977 [1961]) had written about illustration, when examining specifically “the Photographic Message”. He took as his main example press photographs accompanied by text. In this case, Barthes points out to a change in the way image has been historically related to text in terms of illustration in mid XX Century:

[...] the image no longer illustrates the words; it is now the words which, structurally, are parasitic on the image. The reversal is at a cost: in the traditional modes of illustration the image functioned as an episodic return to denotation from a principal message (the text) which was experienced as connoted since, precisely, it needed an illustration; in the relationship that now holds, it is not the image which comes to elucidate or 'realize' the text, but the latter which comes to sublimate, patheticize or rationalize the image. [...] Formerly, the image illustrated the text (made it clearer); today, the text loads the image, burdening it with a culture, a moral, an imagination. Formerly, there was reduction from text to image; today, there is amplification from the one to the other. (BARTHES, 1977[1961], p. 25-26).

Bateman (2014) presents a summary of text-image relations in Barthes. He, then, combines illustration with anchorage and relay, to offer the graphically systemic classification in Figure 1:

Figure 1 – Barthes' classification of text image relations represented graphically



Source: BATEMAN, 2014, p. 35.

In the 1970s, we can find an important contribution to the foundations of multimodal approaches in Michael Alexander Kirkwood Halliday's work. Halliday plays a leadership role

in developing social semiotics, theorizing about the social negotiation and construction of language. He, then, claims that there is a “dynamic process of sign making” (HALLIDAY, 1985) in which meaning arises from social interaction. In this sense, based on the situation and the audience, individuals make choices from different modes of representation and expression. “Halliday’s language of description provided more granular ways of describing meaning-making and showed how pivotal social mediation and subjective choices are in sign-making.” (ROWSELL; COLLIER, 2017, p. 314).

The perception of different modes or modalities in communication processes is a key factor for the development of multimodal approaches. The investigation of the combination of different modes or modalities is the next challenge. Jay Lemke (1998) analyzes mode combination in terms of meaning multiplication through visual and verbal integration in scientific text.

In multimedia genres, meanings made with each functional resource in each semiotic modality can modulate meanings of each kind in each other semiotic modality, thus multiplying the set of possible meanings that can be made (and so also the specificity of any particular meaning made against the background of this larger set of possibilities). (LEMKE, 1998, p. 92).

Bateman (2014) interprets Lemke’s idea of meaning multiplication as a metaphor that states that, in some cases, the combination can be more valuable than the information that can be obtained from the modes when used alone. “In other words, text ‘multiplied by’ images is more than text simply occurring with or alongside images” (BATEMAN, 2014, p. 6). When it comes to the combination of audio and video in audiovisual productions, such as films or TV shows – as it is the case with the dataset presented in this dissertation – the superior potential of combinatorial meaning-making can be empirically perceived in contrast to the experience of solely listening to the audio or merely seeing the images.

Gunther Kress, influenced by Halliday’s work, is usually associated with the foundation of multimodality. Forceville (2010, p. 3624) states that “he can be considered a founding father of the discipline that is nowadays often referred to as visual and multimodality studies”. Rowsell and Collier point out that “Kress often stands as a harbinger of multimodality”, this consideration being given to him because of his longtime work on theorizing multimodality and design. A conceptualization of modality in visual or semiotic systems is offered by Hodge and Kress (1988) as derived from the modality systems of language and adapted to semiotic phenomena. In this sense, modality would work as modal auxiliaries or modal markers,

determining the weight attached to an utterance, the truth value or credibility of statements about the world.

Later, when developing a “grammar of visual design”, Kress and Van Leeuwen (2006 [1996]) return to this concept of modality in visual communication, emphasizing that it is essential to navigate through the many layers of reality, truth or naturalism. But they also focus on the concept of modes: the semiotic modes other than language or the multiple modes that are combined and act in multimodal communication, especially in visual design.

We can summarize this discussion in the form a set of hypotheses: (a) human societies use a variety of modes of representation; (b) each mode has, inherently, different representational potentials, different potentials for meaning-making; (c) each mode has specific social valuation in particular social contexts; (d) different potentials for meaning-making may imply different potentials for the formation of subjectivities; (e) individuals use a range of representational modes, and therefore have available a range of means of meaning-making, each affecting the formation of their subjectivity; (f) the different modes of representation are not held discretely, separately, as strongly bounded autonomous domains in the brain, or as autonomous communicational resources in culture, nor are they deployed discretely, either in representation or in communication; (g) affective aspects of human beings and practices are not discrete from other cognitive activity, and therefore never separate or absent from representational and communicative behaviour; (h) each mode of representation has a continuously evolving history, in which its semantic reach can contract or expand or move into different areas of social use as a result of the uses to which it is put. (KRESS; VAN LEEUWEN, 2006 [1996], p.41).

In this sense, Kress and Van Leeuwen examine the different modes in a very broad way, taking the writing mode and the visual mode as the most important for the proposed discussion. But, in many examples, what we see are subtypes of these broader modes being pointed out: language as speech, visual mode of drawing, visual/spatial mode, etc. This emphasizes that there is some flexibility in the definition of the concept, what is also claimed by many authors as a key point in the challenge of organizing multimodal communication as a theory. Nevertheless, they consider a mode to be “a systematically organized resource” and offer the following definition: “A mode is a means for making representations, through elements (sounds, syllables, morphemes, words, clauses) and the possibilities of their arrangement as texts/messages.” (KRESS; VAN LEEUWEN, 2006 [1996], p.226).

The concepts of mode and multimodality, although still in discussion about their precise definition, have been gradually spread since the 1990s. Bateman, Wildfeuer and Hiippala (2017), when evaluating the challenges for multimodality studies, address that the identification of the modes are, usually, still superficial, a simple list of examples being made, such as written

text, spoken language, gesture, facial expressions, pictures, drawings, diagrams, music, moving images, comics, dance, typography, page layout, intonation, voice quality, etc.

Attempts to systematise this complex area of what semiotic modes there are and what they do paint a less than clear picture, particularly when we move from general statements of what a mode may be and proceed instead to concrete practical analysis of actual cases (BATEMAN; WILDFEUER; HIIPPALA; 2017, p. 17-18).

Bateman, Wildfeuer and Hiippala (2017) demonstrate the wide range of definitions offered for the notion of ‘mode’ with a list of eight different quotes, which is reproduced in Figure 2.

Figure 2 – Modality quotes – what might a ‘mode’ be?

1.	“Mode is used to refer to a regularized organized set of resources for meaning-making, including image, gesture, movement, music, speech and sound-effect. Modes are broadly understood to be the effect of the work of culture in shaping material into resources for representation.” (JEWITT; KRESS, 2003, p.1-2)
2.	“the use of two or more of the five senses for the exchange of information” (GRANSTRÖM ET AL. 2002, p. 1).
3.	“[Communicative mode is a] heuristic unit that can be defined in various ways. We can say that layout is a mode, which would include furniture, pictures on a wall, walls, rooms, houses, streets, and so on. But we can also say that furniture is a mode. The precise definition of mode should be useful to the analysis. A mode has no clear boundaries.” (NORRIS, 2004a, p. 11)
4.	“[Mode is] a socially shaped and culturally given resource for making meaning. Image, writing, layout, gesture, speech, moving image, soundtrack are examples of modes used in representation and communication.” (KRESS, 2010, p. 79)
5.	“we can identify three main modes apart from the coded verbal language. Probably the most important, given the attention it gets in scholarly circles, is the visual mode made up of still and moving images. Another set of meanings reach us through our ears: music, diegetic and extradiegetic sound, paralinguistic features of voice. The third is made up of the very structure of the ad, which subsumes or informs all other levels, denotes and connotes meaning, that is, lecture-type ads, montage, mini-dramas.” (PENNOCK-SPECK; DEL SAZ-RUBIO, 2013, p. 13-14)
6.	“image, writing, gesture, gaze, speech, posture” (JEWITT, 2014b, p. 1).
7.	“There is, put simply, much variation in the meanings ascribed to mode and (semiotic) resource. Gesture and gaze, image and writing seem plausible candidates, but what about colour or layout? And is photography a separate mode? You will find different answers to these questions not only between different research publications but also within.” (JEWITT ET AL., 2016, p. 12)
8.	“[i]n short, it is at this stage impossible to give either a satisfactory definition of ‘mode’, or compile an exhaustive list of modes.” (FORCEVILLE, 2006, p. 382)

Source: adapted from Bateman; Wildfeuer; Hiippala (2017).



The authors provide this list with the aim to emphasize the challenge involved in transferring multimodality theorization into analytical practice. They consider none of these definitions capable of offering a sustainable model to be applied to the analysis of a wide range of phenomena.

On the other hand, we can summarize important attributes presented in each definition to characterize the approach we are developing for multimodal analysis under the scope of Frame Semantics. In this sense, we should start with the notion expressed in definition 3 which claims that the mode considered should be useful to the analysis. Therefore, the modes we are dealing with are audio and video, but we have subcategorized them, making an option to consider only (i) the verbal communication segments of the audio – the spoken audio, and (ii) the video objects that compose each of the frames shot, what excludes lettering, for example.

These choices also corroborate the regularized and organized aspects of the modes, expressed in definition 1, once we have Portuguese Language and video sequences broadly recognized as resources of representations in meaning-making. They are also, clearly, empirical recognized parts of the contemporary social culture.

Audiovisual material also fills the requirements of definition 2, which states the use of at least two of the five human senses, once hearing and vision are required to a complete experience of consuming a TV show – although this definition might be highly contested if we consider the occurrence of different modes in visual media, when text and photos are combined in advertising, for example.

From definition 4, we highlight the socially shaped and culturally given aspects of the mode which justify our choice of TV shows as the basis for building a corpus, once they represent a broadly recognized genre of communication. This is also a fundamental reason why we think it is relevant to propose a multimodal semantic representation based on these exemplars of contemporary communication.

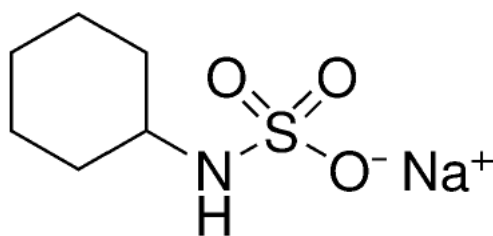
The categorization of four types of modes described in definition 5 can be taken as a recommendation on which kind of modes should be considered in our research: the coded verbal language mode and the visual mode – expressed by moving images, since those are the categories, we think can generate structured semantic representation.

The definitions 6 and 7 bring examples and reinforce the possibility of variation on what should be considered a mode and/or when. In this sense we can see that gesture or color, for instance, although plausible in many multimodal analyses, are not part of the goals of this research, and will not be considered.

The definition number 8 consolidates that we are far from a final and conclusive definition, what can, so far, make essential to consider the utility for the analysis as a key parameter.

Bateman (2014) debates the difficult task of understanding how different ways of producing meaning can be harmoniously combined and easily understood, even when they are systemically very different. For instance, this is usually the case of text-image relation. Although we can frequently take the combination for granted, sometimes, if we go deep in analyzing it, some issues arise. As an example, we reproduce a mixture of text and visual material in the representation of an organic compound in Figure 3.

Figure 3 – An organic compound: image or text?



Source: BATEMAN, 2014, p. 10.

This example highlights how challenging it is to determine the boundaries between modes or modalities sometimes, once we have letters combined with lines in a particular way, blurring what we can categorize as text and what we can categorize as visual material. Moreover, Bateman (2104) points out that it is impossible to understand it properly without some background of the social conventions that determine the way this representation is built.

Superficially we might take it as a depiction of a molecule because the letters look familiar from chemistry classes. Unless we have also learnt about organic chemistry and its conventions, however, we will not know what the particular lines connecting these labels, their shapes and whether they are single or double, mean. These are also no longer pictorial representations at all in the naïve sense of ‘resemblance’ – the lines and their connections have very specific conventional meanings that would be impossible to decode without more information about how such molecules are currently assumed to ‘work’. (BATEMAN, 2014, p. 11).

We can find similar challenges in analyzing examples of visual poetry. Figure 4 shows a visual poem in Portuguese. In red we read *cale-se*, which means ‘shut up’, written four times. In black we read *sim senhor não senhor*, which means “yes sir no sir”, although it is not possible

to pinpoint whether the sentence starts with *sim*(yes) or *não*(no). In this example there is clearly a text, recognizable words and meaning. But the way in which the words are written, or actually drawn, gives us a representation of a quiet sign, with the index finger positioned in front of the lips.

Figure 4 – Visual poem in Portuguese



Source: <https://twitter.com/TorrentTiago/status/1350189431567880198/photo/1>

It is then clear that we have text and image, but what are the boundaries between them? In this case, image only exists because of text. The same graphic units depict what we understand as textual and what we understand as visual. However, in terms of meaning making, is it possible to separate them?

Therefore, any definition of mode or modality encompasses somewhat artificial boundaries that are proposed for analytical purposes, leading inexorably to some simplification of the actual phenomena. With that in mind, the definition employed in this dissertation can be stated as follows:

Mode or Modality is an experientially recognized (set of) resource(s) for meaning-making, which is shaped by a society/culture, and captured and systematized so as to be useful to the analysis of a phenomenon.

This definition could lead us to consider any kind of combination of two or more modes of expression as multimodal objects. We can think of a broadcasted orchestra concert as a multimodal phenomenon since it combines the sound of the instruments played – one mode – with sequenced images – another mode –, which, in turn, are composed by a set of resources, such as framing, angle, camera movement, objects photographed, etc. The same could be empirically perceived from a fashion show, in which we usually have music setting up the atmosphere for a variety of visual stimuli from models walking on a runway. Both can, then, be examples of multimodal objects that combine audio and video and are suitable for semiotic analysis.

However, we are interested in characterizing linguistic multimodal objects and, so, we have to limit our scope to objects in which at least one of the modes is language. If we imagine the experience of watching a football match live inside a stadium we will be definitely exposed to auditory and visual stimuli combined, in what we could classify as a multimodal phenomenon. Now, if we transpose this same event to a mediated broadcasted version that shows the images from the field, captures the sound from the field and from the stands and adds the voice of a narrator, we then have an example of a multimodal object suitable to a linguistic analysis. For instance, the experience of watching the match with the combined guidance offered by the camera framing and selection plus the traditional descriptive narration can be labeled as a case of anchorage, in Barthes' (1977[1964]) classification.

These are, then, instances of multimodal objects suitable for linguistic analysis. Thus, a linguistic multimodal object can be defined as a communicative phenomenon in which there

are two or more modes actively combined for meaning-making, and at least one of the modes is language based. In this dissertation, we examine a specific kind of linguistic multimodal object, namely a TV travel series, in the context of building semantic representations in a computational language model, as discussed in the last section of this chapter. However, let us first turn our attention to how subfields of Linguistics address the issue of multimodality so as to extract contributions to define the object of our study.

## 2.2 LITERACY AND GENRE STUDIES

Since the 1990s, the field of Literacy Studies has been receiving an important push into the multimodal approach direction. Rowsell and Colier (2017) highlight the formation of the New London Group and the publication of the pedagogy of multiliteracies as a key factor for building an agenda for transformation of literacy practice. The intention was to “(1) shift what is counted as literacy and (2) acknowledge the multimodality of literacy practices. [...] The New London Group (1996) pushed for use of their pedagogical manifesto to reframe and expand literacy – and the importance of multimodality as a primary idea – in both research and educational contexts.” (ROWSELL; COLIER, 2017, p. 315-316)

As a result, it emerged a way of looking into text considering screens, its design aspects, and the combination of modes. Those aspects were considered new in the 1990s, and they are gaining more and more relevance in the 2020s.

The New London Group (1996) argue that the notions of design, available designs and redesign are fundamental to how we make meaning with modern texts. Designing on-screen has not only transformed how we make meaning, but also, transformed ways of reconstructing and renegotiating our identities. Multimodality comes first in that it informs how we make meaning and multiliteracies, as a possible pedagogy, gives us tools for doing so. Multiliteracies scholars claim that the screen governs our understanding of the world and curricula needs to reflect this dramatic shift in our ideological and interpretative frame. Situating teaching based on student needs and competencies, teaching students overtly based on the skills that they have when they enter our classrooms, and most importantly and what students do not necessarily possess, are ways of critically framing their learning to think about multiple modes, issues of power, ruling passions, communities of practices, home and community literacy, the role of their race, culture, religion, and social class in their literacy learning. Multiliteracies as a pedagogy simultaneously accounts for linguistic diversity and the use of multimodalities in communication. (ROWSELL; WALSH, 2011, p. 56).

In the context of New Literacies, the usage of multimodal texts as methodology for language learning is widely spread. Rojo and Moura (2012) collect several experiences of multimodal practices by Brazilian teachers in the classroom, reported by papers on, for instance, the creation of a blog for rewriting and illustrate classic stories (LORENZI; DE PÁDUA, 2012); video analysis for the creation of video parodies of fairy tales (TEIXEIRA; MOURA, 2012); reading and writing multimodal flash fiction (DIAS et al., 2012); reading hyperfiction to produce podcasts (DIAS, 2012), among others. Experiences on second language learning using multimodality are discussed by Gilakjani, Ismail and Ahmadi (2011) and Zheng, New Garden and Young (2012) who report on the experience of English learning through videogame playing. Lirola (2013) reports on the case in which students of an English language course in Spain analyze immigrant minority representations in local newspaper multimodal texts; and Lo Bianco (2000) discusses multimodal communication in the context of multilingualism.

The study of genres (which is frequently associated with language learning) is also an example of field in which multimodal analyses are prominent. Hiippala (2014, p.111) states that we can see the concept of genre frequently invoked in multimodal analysis to describe multimodal phenomena and their properties, which means an effort to circumscribe the analyzed phenomenon to a somehow stable concept. The multimodal approach expands the possible elements taken into consideration to characterize a genre. In this sense, Bateman, Wildfeuer and Hiippala (2017, p. 129) define genre as follows: “Essentially ‘genre’ is a way of characterising patterns of conventions that some society or culture develops to get particular kinds of ‘communicative work’ done.”. The challenge, however, is to surpass the general debate on how to categorize genres, so that multimodal genres can be singularized and used analytically.

The idea of genres can be traced back to ancient Greece. The differentiation of literary categories can be found in Aristotle’s classification of comedy, tragedy, epic and ballad, for example. Following this ancient heritage, Glen Creeber (2015, p.1) argues that genre is not only a way of classifying literary or artistic expression, but it has to do with how meaning is created for (and by) an audience in literature, cinema, or television. In all these different fields, the discussion about genre frequently encounters Mikhail Bakhtin’s genre conceptualization. This is mostly likely since Bakhtin was the author who first pushed the boundaries of genre from a literary to a universal category to analyze uncountable communication phenomena. The proposed distinction between primary and secondary genres and also the prominence of the notion of sphere of communication (BAKHTIN, 1986 [1952-1953]) are indicators of such a

widening of the field promoted by Bakhtin in comparison to his Russian Formalists contemporaries<sup>2</sup>.

Bakhtin's main idea of genre can be captured as expressed in this quote:

Language is realized in the form of individual concrete utterances (oral and written) by participants in the various areas of human activity. These utterances reflect the specific conditions and goals of each such area not only through their content (thematic) and linguistic style, that is, the selection of the lexical, phraseological, and grammatical resources of the language, but above all through their compositional structure. All three of these aspects—thematic content, style, and compositional structure—are inseparably linked to the **whole** of the utterance and are equally determined by the specific nature of the particular sphere of communication. Each separate utterance is individual, of course, but each sphere in which language is used develops its own **relatively stable types** of these utterances. These we may call **speech genres**. (BAKHTIN, 1986 [1953], p. 60, author's emphasis).

Bernard Schneuwly (2004 [1994]), building on that argues:

- i. that Bakhtin's definition emphasizes that genres are instruments which are chosen in face of a specific discursive action within a sphere of communication;
- ii. that the choice is made within a particular sphere which offers a possible set of genres; and
- iii. that, although flexible, genres are to some extent stable.

This flexibility can lead to another legacy of Bakhtin's approach, which is the perspective of genre transmutation (1997 [1929]): the modification of any genre into a new one throughout the years, decades or centuries, and also the possibility of one genre incorporating another one and, then, generating a new one. These cases are particularly stimulated in periods of technological development or creation of new media. Novels, for example, when "transposed" from books to radio in the first half of the Twentieth Century got new perspectives and inaugurated the Soap Operas. Decades later, this genre would be adapted to TV, and it is still extremely popular and aired in prime time on TV Channels throughout Latin America. In this sense, Luiz Antônio Marchuschi (2002) argues that genres are highly flexible and dynamic textual events, although they play a primary role in stabilizing daily life communicative activities.

Genres are frequently grouped into text types. Although these groupings can vary widely, since this notion is used in very different linguistic traditions, the general idea is that

---

<sup>2</sup> According to Thomson (1984) Russian Formalists such as Tynyanov and Tomashevsky advocated for a genre theory based on the proposition of elaborate typologies or generic categories, in an attempt to build an organized and coherent taxonomy system for texts.

each text type is characterized by the predominance of certain language marks. In this sense, Marchuschi (2002) argues that a text type is a theoretical construction defined by the intrinsic linguistic nature of its composition: the lexical choices, the syntactic aspects, the use of verb tenses, the logical relations established. Text types would be expressed in five categories: narration, argumentation, exposition, description and injunction. In a similar way, Schneuwly and Dolz (2004 [1996], p. 60-61) offer a way of grouping genres that take into consideration not only text types, but also the social communication domains and the global language capacities. The five types, according to these authors are: narrate, report, argue, expose, describe actions.

It is important to highlight that there is significant consensus among linguists that texts are not uniformly adherent to one text type. Typological variation across sentences, paragraphs or sequences is very common in texts. Then it is possible to determine prototypical matches for some text types, taking the predominance of the type in a text. On the other hand, it is also possible to look for text subparts instantiating typological sequences (ADAM, 2011).

In the field of Multimodal Communication, this same sort of effort in the direction of outlining genres by taking into consideration stable patterns incorporates and emphasizes the way in which different modes are combined, especially visual and textual elements. A list of multimodal genres based on a broad sense of genre can quickly be produced if we think of websites, newspaper front pages, social media posts or comics. A more refined approach for genre in multimodal analysis should consider the “range of possibilities open to documents” and also the “materiality of multimodal artefacts” (BATEMAN, 2008). The first aspect would make it possible to group some similar, but also slightly different, documents under the same genre. The second aspect highlights the fact that the manifestation of a document (either physical or electronic) would impact the way people interact with it, as stated by Bateman (2008, p. 11): “Much more than in the case with verbal texts, therefore, the actual artefactual nature of a document will impact on its association with a multimodal genre”. The Genre and Multimodality framework – GeM – (BATEMAN, 2008; HIIPPALA, 2017) proposes a corpus-based method for annotation, in which multiple analytical layers support empirical research on page-based documents. One of the challenges is to determine how genres (or one specific genre) manifest through the mode “layers”.

Moreover, when it comes to television, the study of genres faces yet another challenge. The concepts of TV show and broadcast programming were built, in general, on the foundations of radio and cinema, making it extremely difficult to propose the existence of genuine original TV genres. On the other hand, the genres that have been observed, labeled and studied



throughout the years were described based on genre theories from literature and cinema. Bakhtin's genre theory is usually the foundation for any taxonomic effort to categorize TV shows (MACHADO, 2005 [2000]; FECHINE, 2001; ARONCHI DE SOUZA, 2004; DE MELO, ASSIS, 2016). The problem, then, is that there is very little agreement on and consistency in the categories proposed. Telenovelas or Soap operas, for instance, are sometimes taken as a TV genre, but also as a subtype (or subgenre) of the fictional series genre, which would include other subcategories such as sitcoms and miniseries.

Part of this inconsistency also came from the fact that there is notable difference between the theoretical approach over TV categories and the way that Television, as an industry, refers to its own programming, trying to use genres as empirical labels to identify programs in a schedule. Fachine (2001) argues that the self-institutional presentation of TV genres brings an idealistic approach of pure categorization, hierarchization and labeling that are far from the real conceptual richness of TV language. This strengthens the option for applying Bakhtin's approach to the analysis, as Machado (2005 [2000]) defines genre as:

[...] an agglutinating and stabilizing force within a given language, a certain way of organizing ideas, means and expressive resources, sufficiently stratified in a culture, in order to guarantee the communicability of the products and the continuity of that form in future communities. In a certain sense, it is the genre that guides all the use of language within a given medium, as it is there that the most stable and organized expressive tendencies in the evolution of a medium are manifested, accumulated over several generations of enunciators. (MACHADO, 2005 [2000], p. 68)<sup>3,4</sup>.

There is another element that makes the categorization of TV content extremely challenging: the concept of format. TV shows are not only labeled by the genre they are affiliated to, but they can also be circumscribed to a TV format. Serafina Fusco and Marta Perrotta (2008) offer this definition of TV format:

A format can be defined as an original explanatory structure of any type of show, accomplished in a detailed and exhaustive articulation of its sequential and thematic phases, suitable for transposition into one or more products intended for public use, also through adaptation, elaboration, transformation or translation. Expressive tendencies in the evolution of a medium are

---

<sup>3</sup> “[...] uma força aglutinadora e estabilizadora dentro de uma determinada linguagem, um certo modo de organizar ideias, meios e recursos expressivos, suficientemente estratificado numa cultura, de modo a garantir a comunicabilidade dos produtos e a continuidade dessa forma junto às comunidades futuras. Num certo sentido, é o gênero que orienta todo o uso da linguagem no âmbito de um determinado meio, pois é nele que se manifestam as tendências expressivas mais estáveis e organizadas na evolução de um meio, acumuladas ao longo de várias gerações de enunciadores.”

<sup>4</sup> All translations from Portuguese into English in this dissertation are the responsibility of the author.

manifested, accumulated over several generations of enunciators. (FUSCO; PERROTTA, 2008, p. 91).

We can point out that the key factor of this definition is “explanatory structure”. It means that the combination of elements and the sequential way in which they are expressed defines the form of the program. Albert Moran (2004) calls attention to the fact that the term “format” was originally used to designate the form of a page on which content was published in the printing industry. The author, then, suggests that the concept of TV format came from radio and relates to the notion of producing content in series to the TV industry.

A format can be used as the basis of a new program, the program manifesting itself as a series of episodes, the episodes being sufficiently similar to seem like instalments of the same program and sufficiently distinct to seem like different episodes [...] television format is that set of invariable elements in a program out of which the variable elements of an individual episode are produced. (MORAN, 2004, p. 258).

Moran also uses a pie metaphor to describe the concept of TV format: the format would be the crust, a stable, regular form that evolves the filling. The filling of the pie can vary from episode to episode, from an old version to a new version of the program or even for different adaptations of the program in different countries.

However, when we have genre and format as categories in action, labeling does not look so clear. For instance, the concept of a talk show can be globally recognizable. It refers to a kind of TV program in which a TV host interviews a guest. This general idea would be enough to characterize the talk show as genre. But the empirical perception is that a talk show is a program that matches the format conceived for late evening American television, or “The late-night entertainment talk show” (TIMBERG, 2002). In this format there is a stage at a theater or studio, with a live performing band, and where a TV host performs a stand-up comedy sequence, then presents some other joke-based segment, then interviews one or two guests and the program is finalized with a musical guest performing live. This description is commonly taken into TV industry as a manifestation of a format and can be found in authors who consider format a particular way of organizing a program or a subgenre, like Aronchi de Souza (2004) and Gomes (2002; 2011).

There is also a view of format as an asset in global television trade market (CHALABY, 2011; 2012; MORAN, 2004; 2004b; MORAN, MALBON, 2006). In this sense, not only the organization and sequence of the elements are taken into consideration, but also its brand and trade value. This is the approach that makes us capable of taking different TV shows that are

very similar, like Got Talent, Idols, The X Factor and The Voice, and consider each of them a specific format, instead of taking all of them as instances of one format of music contest into the journey to success and stardom. FRAPA, The Format Recognition and Protection Association, emphasizes the perspective of a format as a specific program, defining format as follows:

A specific type of intellectual property that allows for and guides the replication of the original idea in subsequent iterations across media, platforms and territories. In television (or any audio and/or video medium), a clear and repeatable set of elements that, when combined, enable the production of a programme. Elements may include, but are not limited to, narrative structure, character descriptions, set and lighting plans, graphic and audio designs, music and sound effects, rules, production procedures and anything else that permits subsequent users to reproduce the original concept. (FRAPA, 2020).

Considering all this, the conclusion here is that the category format is relevant to the areas of television producing and marketing, as well as intellectual property. Therefore, because we are interested in the way television organizes its contents into meaningful utterances, we will analyze the multimodal aspects of TV using genres as a category. Moreover, because television programs combine audio and video with the purpose of communicating something, it is clear that television programs are instances of multimodal genres. When it comes to identifying multimodal genres, we could take two different approaches:

- i. consider television programs in general as a multimodal genre, focusing on the distinguishing elements that make a TV program different from other audiovisual multimodal genres such as film, videoclip, videoart, video ad.
- ii. consider each kind of television program as a multimodal genre, focusing on the distinguishing elements – which, in general, are multimodal or relative to a mode – that make each one singular.

In this dissertation we take the second approach, once we pay attention to specific possibilities of meaning construction through multimodal semantic representations within a selected TV genre. This will emphasize certain specific communicative purposes and will make the analysis capable of identify the discourse semantics of the modes provided by the corpus. In accordance with Bateman, Wildfeuer and Hiippala (2017, p. 131), we take genres as:

[...] bundles of strategies for achieving particular communicative aims in particular ways, including selecting particular media. They thus necessarily bind notions of social action, particularly communicative action, and styles of presentation on the one hand, and semiotic modes as ways of achieving those actions, on the other. (BATEMAN; WILDFEUER; HIIPPALA, 2017, p. 131).

The corpus we use in this dissertation is composed by episodes of a TV show characterized as an exemplar of TV Travel Show, Travel Series or Travelogue genre. Next section presents the elements that characterize this genre.

### 2.3 TRAVEL SHOW ON TV AS A MULTIMODAL GENRE

As a proof of concept for the multimodal annotation system with fined-grained semantics devised in this dissertation, we built a corpus based on the episodes of “*Pedro pelo Mundo*”, a Brazilian TV show broadcasted by GNT, a Brazilian cable channel, from 2016 to 2019. In each of the 40 episodes, Pedro Andrade, the host, guides the viewer in a tour through a country, city or even a specific city area, like Brooklyn in New York City. The itinerary is not focused on popular tourist attractions, landmarks or sightseeing. The purpose of the show is to explore cultural, social and economic aspects of the destination, usually paying attention to some recent change process that may have occurred. This makes its producers call it a “travel/current affairs show”<sup>5</sup>. We consider it, then, an exemplar of a travel show.

Travel Shows, Travel Programmes, Travel Series or Travelogues are different terms that commonly refer to a same genre – some nuances will be pointed out next. This genre is characterized by an audiovisual production in which, prototypically, a host guides the viewer to a specific destination (or a specific tour through some destinations) highlighting aspects related to the experience of being in that particular place. These aspects may include touristic attractions, cultural manifestations, social contexts etc.

Anne Marit Waade (2009, p. 46) defines Travel Series, highlighting the role of the host and the aesthetics: “Travel series are characterized by being a series format in which the host typically guides the viewer to new destinations every week, and his/her capacity to create a good mood and the audiovisual pleasure given are important concepts.”. The author also points out that Travel Series in the broad sense is a hybrid genre, once it may highly combine elements from other genres such as documentary film and lifestyle series among others.

Maja Sonne Damkjaer and Anne Marit Waade (2014) list ten subgenres of Travel Series, considering how much they get close to three other genres: Journalistic Documentaries, Factual Entertainment or Consumer Information. The subgenres are:

---

<sup>5</sup> See <https://www.producingpartners.com/about> and <https://www.producingpartners.com/pedro-pelo-mundo> .

- i. Travelogue
- ii. Popular Science
- iii. Convivial
- iv. Group Travel
- v. Tourist Guide
- vi. Backpacker
- vii. Sports and Adventure
- viii. Culinary
- ix. Expat
- x. Meta

For the purpose of this dissertation, it is not the case of detailing the tentative typology proposed by the authors or the specificities of each subgenre. On the contrary, we are interested in the key common elements of the broader genre. However, one of them stands out, since it is the inaugural term that refers to audiovisual productions about travel, inspired the name of a literary genre and, sometimes, is taken as the authentic broad genre for TV shows based on travels: Travelogue. This would be the most common travel series type, usually identified with journalistic documentaries. Damkjaer and Waade (2014) describe Travelogues as a TV Series in which a host takes a roundtrip with a particular purpose in mind and, once in sight, he or she “acts as a tour guide, gives lectures on location and engages with nature, the local culture and population” (DAMKJAER; WAADE, 2014, p. 49).

Creeber (2015, p. 156) also defines the Travelogue in association with the documentary genre: “The TV travelogue is a documentary which involves presenters travelling to distant and often exotic places around the world, pointing out the sights, meeting some of the local people and sampling the native customs and cuisine”. The author, however, when listing examples and describing the historic evolution of the genre argues that Travelogues also appeared in different hybrid forms, which match, in general, the subgenres described by Damkjaer and Waade (2014). This makes clear that Creeber (2015) sees the Travelogue as a genre in the same way as Damkjaer and Waade (2014) see Travel Series.

Travelogue is a term coined by Elias Burton Holmes in 1904 to name his amused illustrated travel lectures series in Chicago’s and New York City’s theatres (Barber, 1993). The new term at that time was taken to emphasize the step forward Holmes was doing when introducing film clips to the lectures – until then the very popular travel lectures only presented lantern slides. Travelogue is a combination of Travel and Monologue that results in a blend.

Ozola (2014) points out that Travelogue is an Americanism that appeared in dictionaries in the beginning of 20<sup>th</sup> Century and referred to the film making industry, but later gained popularity as a literary genre. “The emergence of the travelogue as a genre is associated with the desire to describe the journey undertaken by the human, to remember everything that happened on the way, and to record it”, argues Kislova (2019, p. 128).

The original Burton Holmes Travelogues lectures as described by Barber (1993) give us the foundational elements of the genre that we can see adapted to main Travel TV productions:

[...] Holmes made a striking appearance on the stage. He was elegantly tailored and had a sophisticated manner. Early in his career he sported a beard, but later he became known for his goatee. He had a well-modulated voice, and his delivery was frequently described as 'crisp'. Indeed, Holmes considered himself a performer. The lecture platform, in his view, was a stage to which he brought the theatre of life as he had experienced it around the world. Furthermore, he felt that his travel exhibitions were a natural extension of the magic shows he had given as a young man. Through his lectures, he tried to present the illusion of actual travel, and he downplayed the fact that an optical contrivance produced the views that the audience saw. [...] Holmes considered slides to be element in his shows. In constructing his presentations, he first selected interesting views from his collection and then built his talks around these. [...] Holmes spoke in a more informal fashion meant to keep the audience's attention focused on the visual imagery. His lectures were not tightly organized and moved freely from topic to topic. [...] it was Holmes's use of motion pictures, first introduced in his shows in the fall of 1897, that helped him achieve wide fame. (BARBER, 1993, p. 80-81).

We then can summarize:

- i. The role of the host as the guide.
- ii. The importance of visual aesthetics.
- iii. The synchronicity in talking about what is seen.

As described by many authors, these core characteristics present at the original travelogue lecture events were transposed both to travelogues in literature and in television. This summary also emphasizes the perception of the genre as multimodal. The explicit relevance of both visual and speech elements and also the importance of their combination are reasons why we took this genre to work with. Moreover, the synchronicity between what is heard and what is seen is a key analytical category in this dissertation for which we have been seeking a computational way of representing it and then, performing multimodal analysis in computational linguistics.

## 2.4 RELEVANCE OF MULTIMODAL OBJECTS FOR COMPUTATIONAL APPLICATIONS

There are many ways of seeing multimodal objects in relation with computational applications. Bateman, Wildfeuer and Hiippala (2017) give a broad vision of how this encounter occurs:

The idea of a computational approach to multimodality research refers to how the data is processed by a computer, that is, by performing calculations. The prerequisite for performing any kind of calculation at all, of course, is that the data under analysis can be represented numerically. [...] computational methods are highly effective in converting multimodal data such as photographs or entire documents into numerical representations. Even more importantly, algorithms that manipulate these numerical representations are getting better at forming abstractions about them in a manner similar to humans, learning to recognize objects, their shape, colour and texture. (BATEMAN; WILDFEUER; HIIPPALA, 2017, p. 163)

In this sense, we point out that multimodal analyses have been growing in importance within several computational approaches to both Cognitive Linguistics and Natural Language Understanding. There has been much success in developing theories, models, and systems to both Natural Language Processing (NLP) and Computer Vision (CV), fields that are related to the research reported in this dissertation. However, there is still a long path in the challenge of integrating NLP and CV to establish comprehensive multimodal semantic representations.

Traditional work on caption generation has focused on improving the capacity of detection and description of still images. For instance, Fang et al. (2015) present an approach for automatically generating image descriptions with a system which trains on images and corresponding captions, and learns to extract nouns, verbs, and adjectives from regions in the image. Devlin et al. (2015) present a discussion of the methodologies of image-conditioned models, using convolutional neural network (CNN), maximum entropy (ME) and a recurrent neural network (RNN) in different pipelines that combine them to achieve better results on captioning generation. Nikolaus et al. (2019) propose a model for compositional generalization which is capable of combining unseen concepts.

News-image captioning is an example of recent focus in caption generation that integrates NLP and CV. In this case, considering the expectation of a caption that is more interpretative than descriptive, the inputs are usually news articles and their accompanying images. Yang and Okazaki (2020) present a Transformer model that integrates text and image

to generate captions. *Batra, He and Vogiatzis (2016)* propose a methodology for automatically generating captions for newspaper articles consisting of a text paragraph and an image, processed through several deep neural network architectures built upon RNNs.

There are also many works focusing on caption generation for video. *Aksoy et al. (2017)* 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 learning 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.

There is also relevant work on the other way, which means text-to-image multimodal tasks, or natural language text to image generation. In general, this type of tasks aims to generate photo-realistic images as a visual output to given text description inputs. Text-to-image generation models have been recurrently developed by applying generative adversarial networks (GAN) as proposed by *Goodfellow et al. (2014)*. That is the case of *Xu et al. (2018)* – AttnGAN model –, *Zhang et al. (2017)* – StackGAN++ – and *Zhu et al. (2019)* – DM-GAN – which focus on generating high resolution images. *Han et al. (2020)* propose VICTR as a model attachable to these previous to add rich visual semantic information of objects from the text input. More recently, models adopting machine learning techniques relying on correspondences between raw text and images (*RADFORD et al., 2021*) and stable diffusion models such as DALL.E and DALL.E 2 (*RAMESH et al., 2021; 2022*) have inaugurated a new chapter of text-to-image generation applications of multimodal datasets, allowing downstream users to use elaborate textual descriptions to generate complex scenes produced according to specific graphic styles.

Another prominent field of computational application of multimodality is multimodal machine translation (MMT). The general principle that sustains MMT is using information from more than one modality as a way of offering alternative views of the input data, processing data or disambiguating results. *Sulubacak et al. (2019)* highlight three prominent tasks of MMT: (i) spoken language translation (SLT) refers to the process of translating speech in a source



language to text in a target language (e.g. WEISS et al., 2017; BÉRARD et al., 2018); (ii) image-guided translation (IGT) cover the cases of contextual grounding tasks, in which text data are combined with still images that offer semantic correspondence in a way to avoid or resolve ambiguities or to reinforce context in the process (e.g. ELLIOT et al. 2015; SPECIA et al. 2016; CALIXTO et al., 2016; ZHENG et al., 2018); (iii) video-guided translation (VGT) is similar to IGT, but it tackles video clips instead of still images associated with textual input, which, very often, is the transcript of the audio from the video (e.g. SANABRIA et al.; WANG et al. 2019).

Gesture recognition is another category of computational task involving multimodal objects. The interest in gesture can vary in a wild range, but it primarily refers to hands and arms and research on co-speech gestures. Katsamanis et al. (2017) summarize the task in three main modules: (a) tracking of human movements and recognition of characteristic patterns; (b) detection of synchronic speech; and (c) combination of other audio-visual information. The challenges in this are enormous and defy both qualitative research, in terms of analyzing one pattern, and quantitative research, in terms of corroborating a pattern – identifying three dimensional gestures in two dimensional data, recognizing patterns, developing annotation tools, tracking movements, stablishing connections with speech etc. Work on developing models for multimodal gesture recognition can be find (among many others) in Escalera et al. (2013), Wu et al. (2016), Katsamanis et al. (2017), Pitsikalis et al. (2017), Turchyn et al. (2018).

Francis Steen and Mark Turner (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. Steen and Turner lead an effort on this direction through the collaborative works of The International Distributed Little Red Hen Lab (Red Hen), in terms of establishing tools and methodology for analyzing large multimodal corpora, mostly exploring correlations between spoken and gesture communication.

Red Hen is an initiative designed as a global laboratory and consortium of researchers and institutions developing multiple research efforts on multimodal communication based on audiovisual data throughout the world, using multidisciplinary teams who work on ecologically-valid datasets. One of the key principles of the initiative is to collect data on multimodal communication on a large scale, facilitating, then, research based on large-scale corpora. Steen et al. (2018) highlights that the lab provides computational and storage tools to manage data and also promote the integration of the results and feedback of all different research projects, enabling iterative improvement.

Red Hen’s main dataset is the The NewsScape Archive of International Television News ([newsscape.library.ucla.edu](http://newsscape.library.ucla.edu)). It is an international TV news archive built from live recording<sup>6</sup> of real time news streams. By mid 2020, it included broadcasts from 51 networks, totaling over 400,000 hours of recordings in nineteen languages.

The system automatically ingests and processes roughly 150 hours of television news each day from miniature capture stations. To make the data accessible, we extract closed captions, tag them for various linguistic features (grammar, parts of speech, lemmas, para-linguistic elements, conceptual frames, etc.), and then parse the news stream data into different topics by clustering news stories using multimodal cues. We automatically detect and recognize faces, objects, and other entities in images and texts, semantically represented (e.g., who, when, where) and organized into hierarchical topic structures. These operations generate a great variety of visual and textual metadata for use in subsequent studies. (JOO; STEEN; TURNER, 2017, p. 358-359).

Steen and Turner (2013) state that the Red Hen project aims to integrate this multimodal data collection with different types of data enhancement, making it possible for researchers to conduct transdisciplinary investigation in multimodal communication, linguistics, computational linguistics, cognitive science, neuroscience, education, statistics, media effects studies, political communication, and library science. Research is usually related to main questions such as “how humans rely on both verbal and non-verbal cues in order to communicate and interact with other humans?”; or “what meanings are carried by such multimodal messages”; and “how we decode and perceive these meanings from multimodal cues”.

Red Hen’s annotation process relies on a multi-level feedback process between linguists and computer scientists, aimed at training computers to perform tasks that generate annotations according to the linguist’s specifications. [...] A series of pipelines process these data, using customized open-source software (STEEN et al., 2018, p. 4)

FrameNet Brasil has been engaged with some of Red Hen’s projects for the past couple of years, mentoring the development of pipelines which annotate for frames. The collaboration with the Red Hen Lab has also influenced in the move of inaugurating multimodal research in FrameNet Brasil and, so, in the effort to include multimodal data into the FrameNet Brasil database and develop an annotation methodology of annotation for it. The foundations for the

---

<sup>6</sup> Steen et al. (2018) explain that this procedure is assured by section 108 of the U.S. Copyright Act which authorizes libraries and archives to record and store any broadcast of any audiovisual news program and to loan those data, within some limits of due diligence for the purpose of research.

development of this methodology point to the perception of grammar in multimodal communication, the topic of the next chapter.

### 3 THE GRAMMAR AND THE SEMANTICS OF MULTIMODAL COMMUNICATION

The investigation of multimodal data has its first step based on the perception of the different modes that operate in the process of meaning-making. This can be a challenging task for many multimodal objects, but when it comes to objects composed by text-image or verbal-visual binary there is a more explicit call for mode detection. This does not mean a straightforward task, once there are many nuances on what elements should be considered in both text and image, and also some communicative situations in which a mode can be composed by both visual and verbal elements. In the Figure 5, as an example, we have a photograph in which there is a lot of written text, but they may be analyzed in multimodal research as instances of sign mode, which include some other pictorial signs.

Figure 5 – A photograph of a street with verbal and non-verbal signs



Source: Low Ze Yann on Unsplash<sup>7</sup>

We have already highlighted the role of Kress as a founder of the field of multimodality. Kress (2010, p. 6-7) argues that instead of perceiving the concept of grammar as “a stable system of rules”, it should be shifted in the sense of “relative regularity of a semiotic mode”. In

<sup>7</sup> <https://unsplash.com/photos/bKJqs72nJgI>

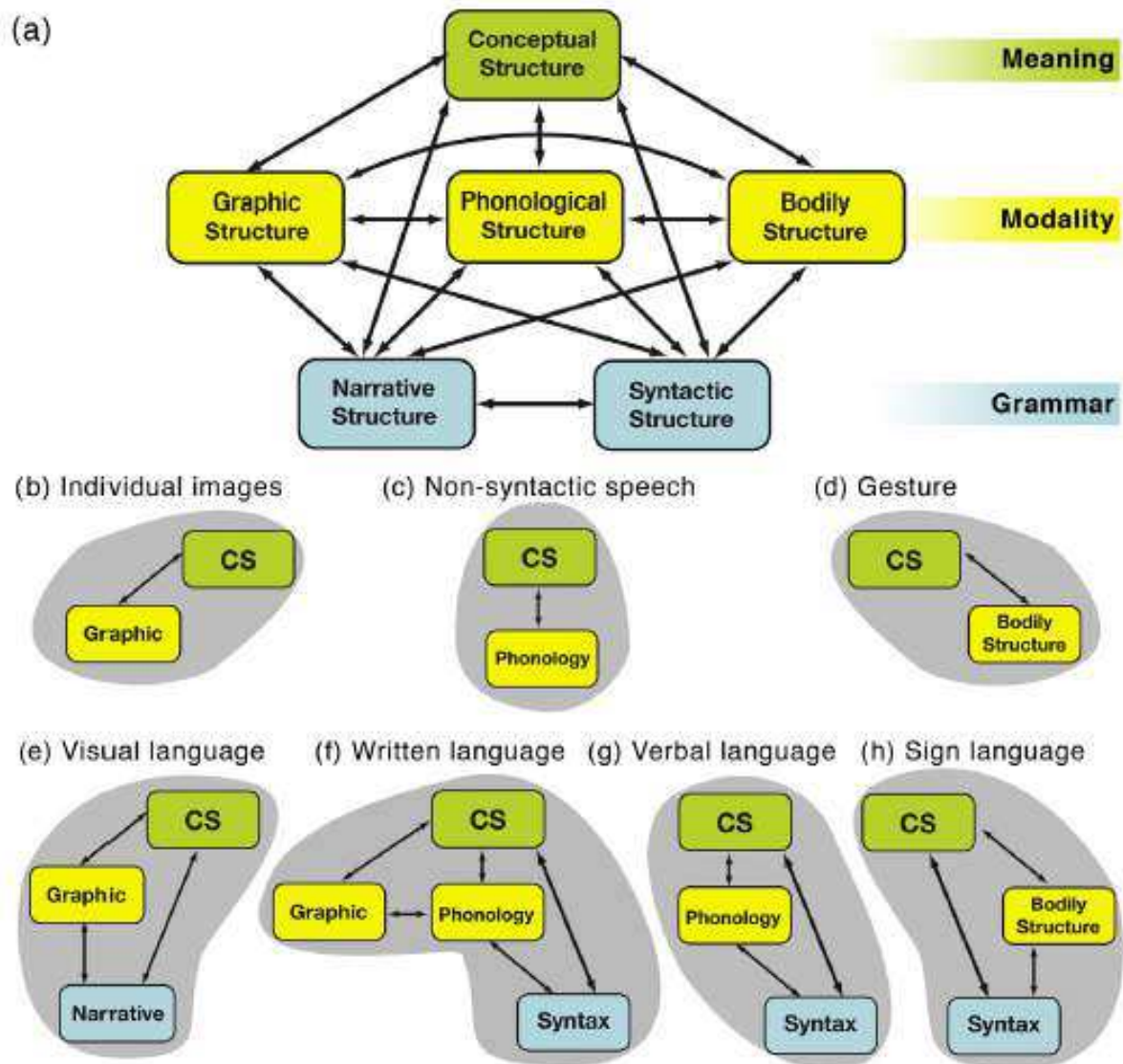
terms of the existence of a grammar in visual communication, Kress (2006, p. 1) states that it is related to “the way in which depicted elements – people, places, and things – combine in visual ‘statements’ of greater or lesser complexity and extension”. This means to look for the way in which these elements are combined into meaningful wholes. Following the social semiotics theoretical framework, the author highlights that the visual grammar is not universal but circumscribed to cultural dynamics. In this sense, for instance, Kress (2006) defines his analyses of the language of contemporary visual design in Western cultures or, even more precisely, Western European cultures, although trying to encompass a wide range of manifestations such as oil painting, magazine layout, comic strip, and scientific diagram. The author, then, focuses more on interpreting the grammar present in some types of visual expression, than proposing a framework that could be applied by researchers to other analyses.

In this dissertation we present multimodal research that establishes fine-grained frame-based relations between the auditory and visual modalities through the annotation in a TV Travel Series corpus. To determine the relation of auditory and visual elements within a Linguistics approach we built on Cohn’s (2016a) systematization of the semantic investigation in multimodal data, according to the grammaticality of the modalities involved. It was used as a first reference to evaluate the relation expressed by spoken audio and video images in the selected corpus. Moreover, in a later section, we present FrameNet Brasil as a semantic representation compatible with such a grammar.

### 3.1 VISUAL AND FILMIC NARRATIVE GRAMMARS

Cohn (2016a) proposes an expansion of Jackendoff’s (2002) parallel architecture of language, which relies on the assumption that “language has multiple parallel sources of combinatoriality, each of which creates its own characteristic type of structure” (JACKENDOFF, 2002, p. 107). Cohn (2016a) focuses on how grammar and meaning coalesce in multimodal interactions, going beyond the semantic taxonomies typically discussed within the domain of text–image relations. Figure 6 shows a schematic representation of this framework. He thus classifies the relations between text and image in visual narratives, evaluating the presence or absence of grammar – syntax for written, verbal or sign language and narrative for visual language – in the process of structuring each of the modalities.

Figure 6 – The parallel architecture expanded to allow for multimodal interactions



Source: COHN, 2016a, p. 309

To analyze the presence of grammar in visual narratives, Cohn (2013) developed the theory of Visual Narrative Grammar (VNG). The main concept of VNG is that the meaningful information of a visual narrative sequence is organized by a narrative grammar in a way analogous to how the semantic information of a sentence is organized by the syntactic structure. Another key aspect of the theory is that it emphasizes the separation between narrative and meaning for sequential images: “while event structure is the knowledge of meaning, narrative structure organizes this meaning into expressible form” (COHN, 2013, p. 416).

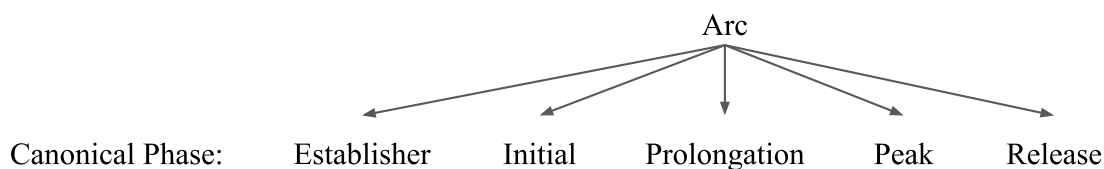
VNG is similar to previous approaches comparing narrative to syntax, such as story grammars of discourse (MANDLER; JOHNSON, 1977; RUMELHART, 1975). However, these older models used procedural phrase structure rules based on generative-transformational

grammars (CHOMSKY, 1965), and had imprecise distinctions between narrative and semantics. In contrast, VNG is modeled after construction grammars (CULICOVER; JACKENDOFF, 2005; GOLDBERG, 1995), which posit that sequencing schemas are entrenched in long-term memory with mappings to an explicitly separate semantics (see COHN, 2013b, 2015).

Based on Loschky et al.'s (2018) Scene Perception and Event Comprehension Theory, the semantic processing for event structure comprehension is described by Cohn (2019) as divided into two general domains: front-end processes and back-end processes.

Thus, processing the meaning of visual narratives breaks down into front-end processes for negotiating information in the visual modality, and back-end processes for constructing a situation model. Front-end processes use attentional selection and information extraction to feed into backend processes activating semantic memory to then construct a progressively updating situation model. Such processes involve both forward-looking expectancies on the basis of activated information, and backward-looking updating to reconcile those expectations with incoming information. Taken together, these processes are consistent with established theories for the processing of discourse in the verbal domain (Graesser, Millis, & Zwaan, 1997; McNamara & Magliano, 2009; van Dijk & Kintsch, 1983), yet adapted to the unique affordances of the visual-graphic modality. (COHN: 2019, p. 105-106).

The narrative level of representation, described by Cohn (2013) in VNG, runs in parallel with semantic processing. Panels are the basic units of visual narrative and fit into five core narrative categories: Establisher, Initial, Prolongation, Peak and Release. These categories are combined to form constituency phases, which correspond, then, to coherent pieces of a structure. The phases are part of an Arc in the narrative and the canonical constituency can be represented in Figure 7:



Source: adapted from COHN (2013).

This schema shows the canonical phase with the occurrence of all five categories in the prototypical order. It is possible to have some phases that do not feature all categories, but the Peak is almost always indispensable<sup>8</sup>. It is the category that motivates a sequence. When considering a narrative as a whole, VNG expands the canonical arc into substructures. Cohn (2016b) points out that similarly to human language, VNG does not use basic abstract combinatory schemas. On the contrary, VNG allows for constructional patterns beyond the basic schemas or related to the canonical patterns.

Cohn (2013a) states that this general idea of structure, development and processing are governed by a guiding Principle of Equivalence. This means that, given modality-specific constraints, it is expected to perceive the mind/brain treating expressive capacities in similar ways:

That is, from the perspective of cognition, different modalities like language, music, and visual narratives should share in their processing resources. However, their differences should be motivated by the affordances of the modalities themselves, either with processing of that modality or with how that modality subsequently facilitates cognitive mechanisms (COHN, 2013a, p. 195).

Therefore, because we are interested in the characterization of audiovisual narratives, we pay attention to the similarities present in the way static visual narratives and films are perceived. Cohn (2016b) introduces the postulation of a Filmic Narrative Grammar (FNG) based on VNG. The author believes that VNG overlaps with notions of film theories and can thus be applied to films. Considering that films have been the foundation for all other audiovisual combination media, we can exploit similarities to audiovisual productions in general, as it would be the case of the TV Travel Series that composes the corpus used in this dissertation.

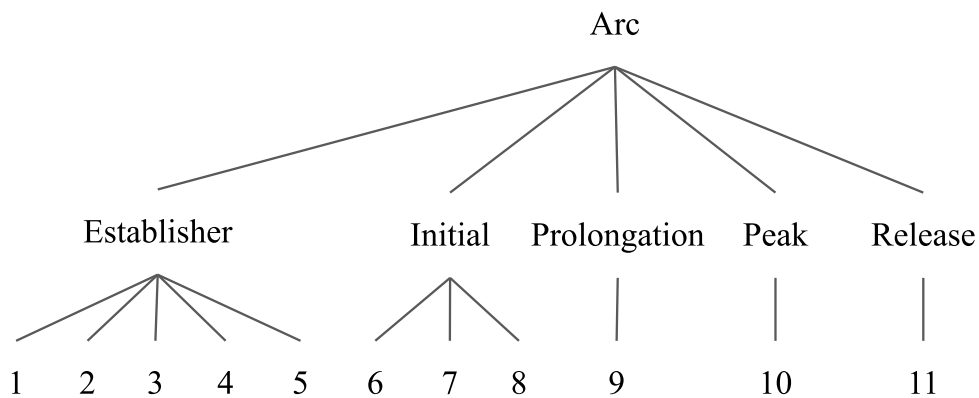
The foundational similarity of VNG and FNG is that both are theoretic models that account for visual narrative sequences. Then, when Cohn (2016b) builds FNG as an adaptation of VNG, he recalls (i) the categorical roles played by units and constituents, (ii) the hierarchic structures which allow connections across distances, (iii) modifiers that expand on basic sequencing and (iv) the storage of these elements as constructional patterns in memory.

In Figure 8 we apply narrative grammar analysis as proposed by both VNG and FNG to a sequence from “*Pedro pelo Mundo*”. In this sequence, Pedro wakes up in his hotel room and leaves for having breakfast at the hotel restaurant.

---

<sup>8</sup> Peaks can be omitted felicitously under certain constrained, inference-generating conditions (COHN; KUTAS, 2015; MAGLIANO et al., 2015 apud COHN 2016b).



Figure 8 – Narrative grammar applied to a sequence from “*Pedro pelo Mundo*”

Source: elaborated by the author.

The sequence presented opens with five conjoined Establisher shots (1-5), which locate the action in the bucolic scenery of Edinburgh, where we see the hotel building. Then we have the Initial phase (conjoined shots 6-8), in which we see Pedro inside the hotel room, looking through the window to the bucolic scenery before walking away. After that we have a Prolongation shot (9) that shows a detail of the hotel room while Pedro has already left it. The

climax is reached at shot 10, the Peak, that shows Pedro walking into the restaurant, looking to the buffet and moving in the direction of a table. We see Pedro at the table in shot 11, which is the Release of this sequence.

The representation of the shots as static panels shows not only that it is possible to adapt VNG to FNG, but also highlights the differences. Cohn's (2016b) hypothesis is that the main structural differences between VNG and FNG came from the differences of static image sequences and moving image sequences, which are nothing more than a direct result of the differences between the modalities themselves. In fact, Figure 8 exemplifies this, once we are not able to reproduce Pedro walking in shots 8 and 10, as well as we cannot represent the camera moving in shots 7 and 9. In Figure 9 we reproduce Cohn's table of gross differences between static and moving narratives.

Figure 9 – Gross differences in dimensions between prototypical cases of drawn and filmed

<b>Static, drawn narratives</b>	<b>Moving, filmic narratives</b>
Production is a biologically based human ability (drawing)	Production is technologically mediated (nonnatural)
Uses patterned graphic schema for both iconic and symbolic elements (i.e., stored lines and shapes in a visual vocabulary)	Uses general perception (not a patterned visual vocabulary, with the exception of animation)
Static content in images	Moving content in film
Static depictions in images	Moving camera in film (panning, zooming)
Ambiguous temporality between units unless otherwise depicted	Pervasive sense of temporality between units because of ongoing temporality of motion
Spatial juxtaposition of units (in page layout) requiring non-content based navigational rules	Temporal juxtaposition of units unfurling on a screen. When spatially juxtaposed frames appear on a screen, they involve no independent navigational rules.

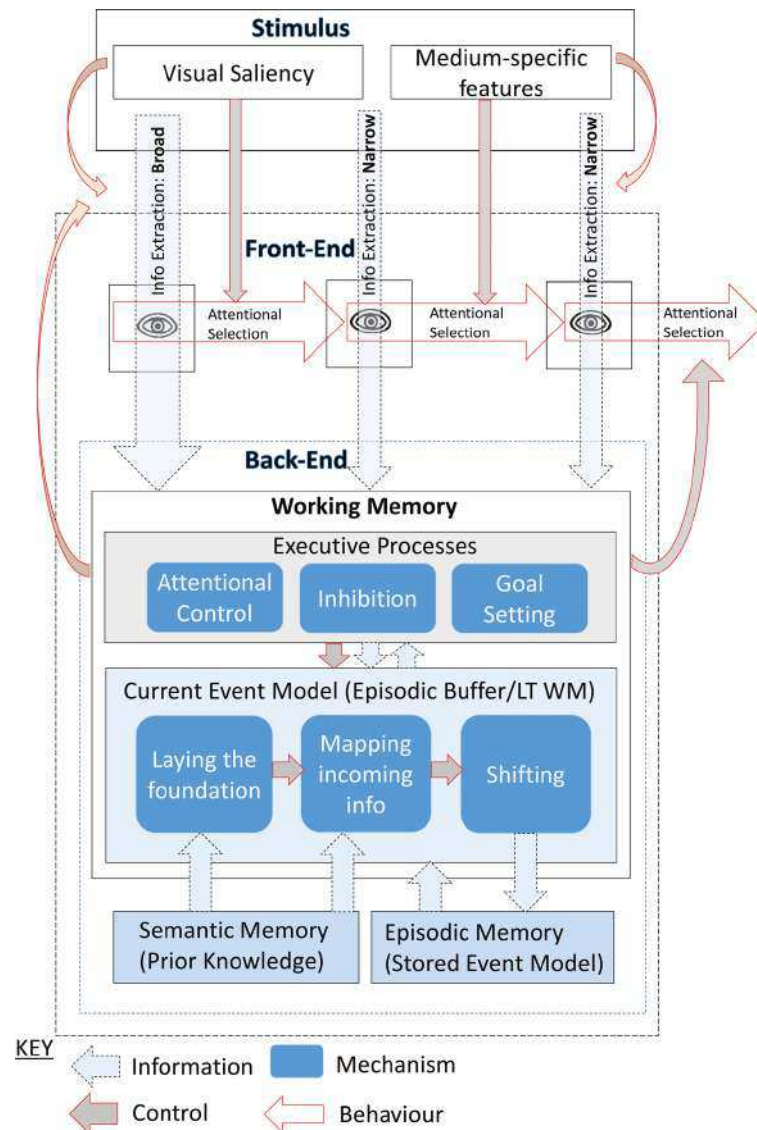
Source: COHN: 2016b, p. 18.

In general, these are very important differences and each of them should be considered, depending on the type of analysis proposed.

Both VNG and FNG can be interpreted in the context of cognitive science theories and models, an example of which is the Scene Perception & Event Comprehension Theory (SPECT), proposed by Loschky et al. (2018; 2020) to investigate the cognitive processing of visual narratives. SPECT is a framework built on the assumption that two different cognitive processes, front-end processes and back-end processes go in action when people perceive and

understand visual narratives. The framework, then, explores the relationships between front- and back-end processes, as well as some specificities in the processes related to the medium (e.g., comics and film). Front-end processes occur in the interface between the medium and the sensorimotor system, in the activities of eye fixation, attentional selection and information extraction, what makes them related to perception. Back-end processes refer to the activities of comprehension, occurring across multiple fixations and related to the construction of event models. The understanding of each part of the narrative is stored in a working memory and the elements that are taken over the course of the entire narrative are stored in long-term episodic memory. Losckhy et al. (2020) argue that SPECT sheds light on the “interplay between these levels of processing”. Figure 10 presents a schematic overview of the theoretical framework.

Figure 10 – Model of the Scene Perception & Event Theory (SPECT) theoretical framework



Source: LOSCHKY et al. 2020, p. 316.

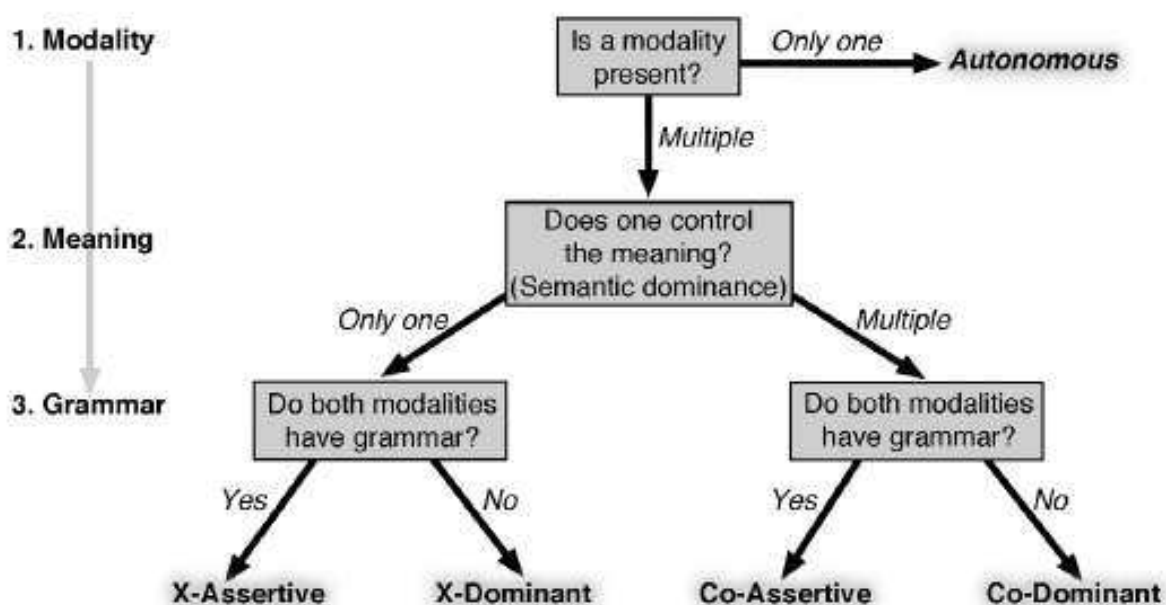
For the purposes of this dissertation, we should highlight some aspects of the theory and the framework. In the same way that Cohn (2016b) describes differences between static visual narratives and filmic narratives, Losckhy et al. (2020) point out specificities of film within this framework. When watching films, viewers have their attention guided not only by framing, but also by camera movements, cuts, cues and the pace of the moving images: “the combination of medium-agnostic and medium-specific stimulus features shape what potential information is available to the viewer, and likely influence how front-end and back-end processes interact in processing this information.” (LOSCKHY et al.: 2020, p. 315). When reading comics, viewers move their eyes through the page layout from panel to panel, in a Z-path – mainly from left to right and top to bottom, with some cultural exceptions. On the other hand, according to the authors, film viewers show a pattern oriented toward the center of the screen. Moreover, the predetermined pace of films does not give many chances for viewers to look around and/or reexamine the screen – not considering situations with digital video, when viewers can pause, rewind and watch again. Comic readers, instead, can determine how much time they spend in each panel and also go back whenever they want. These differences affect both information extraction and attentional selection and, as a result, the subsequent processing in front-end, working memory, comprehension and explicit long-term memory for events.

The event model construction is another aspect we should pay attention to. Losckhy et al. (2020, p. 317) state that “an event model is a particular type of mental model that captures a sequenced event”. The construction of an event model is operated in the back-end from the outputs of front-end processes which activated semantic representations that are sent to the working memory. This representation is encoded into episodic long-term memory and calls a stored event model, from where event schemas can be derived, making it possible to understand the event. Losckhy and colleagues, however, do not explicitly define the structures and/or the dynamics of these event models. We propose that FrameNet can provide what is missing in terms of detailing the nature of such models, because (i) semantic frames, as proposed by Fillmore (1982), are schemas of memory structure evoked or invoked in the process of meaning construction; (ii) FrameNet provides a computational model for those memory structures which can be used as an analytical tool, and (iii) such a model allows for encoding inferential processes. Details are presented in the next section.

Before that, however, we should return to an aspect of multimodal analyses, which is relevant to the analyses to be pursued in this dissertation. The theoretical framework developed by Cohn (2016a) focuses on determining not only the presence of grammar in each modality, but also on how the modalities relate to each other in terms of dominance: does one of the

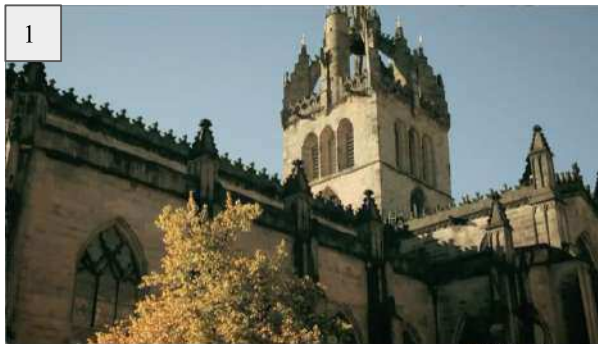
modalities play a preponderant role in determining the meaning expressed by the media? If the answer is yes, there will be a relation of assertiveness or dominance. If the answer is no, the relation will be of co-assertiveness or co-dominance. In some way, this idea of dominance and assertiveness evokes the Barthes (1976 [1964]) relay/anchorage functions of the linguistic message. Figure 11 shows a step-by-step method to evaluate the multimodal interactions.

Figure 11 – Step-by-step method for the analysis of multimodal interactions



Source: COHN: 2016a, p. 320.

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. Our working hypothesis – as we present further – is that, throughout the TV show, spoken audio may play a controlling role in establishing meaning. Most of the sequences have voiceover/off-camera narration, talking heads, standups or interviews (e.g. Figure 12). For these cases in the corpus, we hypothesize that the spoken audio (which can be transcribed into text) is the modality that has grammar and controls meaning. Most of the time, visual sequences are descriptive illustrations for the spoken audio, although there are some visual narrative sequences – e.g. Figure 8 – in which we can perceive narrative grammar.

Figure 12 – Audio guided sequence from “*Pedro pelo Mundo*”

1  
Quando a gente pensa na Escócia, a primeira coisa que vem à mente é  
*When we think of Scotland, the first thing that come to mind is*



2  
homem de saia, uísque  
*man in skirt, whisky*



3  
escocês, gaita  
*Scotch, bag-*



4  
de fole. Mas, na verdade,  
*-pipe. But, in fact,*



5  
Edimburgo é uma meca pra ciência,  
*Edinburgh has been a Mecca for Science.*



6  
arquitetura e filosofia, desde o século dezoito  
*Architecture and Philosophy since the eighteenth century*

Source: elaborated by the author.

Figure 12 shows an example of a sequence in which spoken audio may be framed as dominant in relation to video images, which are illustrations of what is synchronically said. The sequence is the first segment of “*Pedro pelo Mundo*” Edinburgh episode. It starts with Pedro (voice over) talking about the general idea people might have about Scotland, listing some prototypical Scottish culture images. Then he contrasts these elements with a fact about the prominent role of Edinburgh in different fields. Shots 2, 3 and 4 try to directly illustrate what is said, by depicting whisky bottles and a man wearing a kilt and playing a bagpipe. Shots 1 and 5 show what might be perceived as random instances of Edinburgh’s historical architecture,

landmarks or exemplars of the city's beauty. In shot 6 Pedro appears on screen, walking in an open field, talking directly to the viewers, closing his initial statement. Therefore, , in a first analysis, no substantial visual grammar, in terms of narrative structure, is present in this sequence.

Although Cohn's (2016a) 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 dissertation aims to tackle the issue of meaning construction in multimodal settings, specifically on what concerns the interaction between spoken audio (verbal expression transcribed into text) and video (visual objects components of a shot), based on a principled structured model of human semantic cognition: FrameNet.

## 3.2 FRAMENET BRASIL: A FRAME-BASED MEANING REPRESENTATION ENRICHED WITH QUALIA STRUCTURE

In this section we present FrameNet as a model of linguistic cognition. First, we describe some FrameNet basics. Next, we present the implementations developed by FrameNet Brasil so as to enrich the FrameNet model with additional layers of semantics analyses.

### 3.2.1 FrameNet Basics

A FrameNet is a computational implementation of Frame Semantics, as proposed by Charles J. Fillmore (1982, p.111): "a research program in empirical semantics and a descriptive framework for presenting the results of such research". In Frame Semantics, words are understood relative to the broader conceptual scenes they evoke (FILLMORE, 1977). This means that to understand the meaning of a word it is necessary to understand its semantic structure in terms of the properties of the schematized scene in which the word occurs. As an example, consider the signs in Figure 13:

Figure 13 – Child safe and tornado safe signs



Source: <https://www.thesignmaker.co.nz> and <https://www.smartsign.com>.

The expression *child safe area* in Figure 13, for example, is understood only in the context of a scene in which an Asset (the child) is exposed to some potentially Harmful event (vehicles in high speed, for example). In a different way, the expression *tornado safe area*, although very similar to the previous one in terms of surface structure, has a different perspective and points to the Harmful event (the tornado) as causing a Risky situation. Both contexts, however, are related to the broader scenario of Risk. These contexts could be, as conceptual scenes, considered what Fillmore called a frame.

By the term 'frame' I have in mind any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits; when one of the things in such a structure is introduced into a text, or into a conversation, all of the others are automatically made available. I intend the word 'frame' as used here to be a general cover term for the set of concepts variously known, in the literature on natural language understanding, as 'schema', 'script', 'scenario', 'ideational scaffolding', 'cognitive model', or 'folk theory'. (FILLMORE, 1982, p. 111).

Frames are, then, the pivot structures for Frame Semantics and, therefore, for FrameNet. 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.



Berkeley FrameNet started in 1997, as a project based in the International Computer Science Institute (ICSI). Fillmore and collaborators proposed building a frame-based lexicon to cover the general vocabulary of English whose aim was “to demonstrate the usefulness of the database as a lexical resource for its application to speech and language technology” (FILLMORE, JOHNSON, PETRUCK, 2003, p. 242). The principles and foundations of Berkeley FrameNet are reported in many papers and compiled by Ruppenhofer et al. (2016) in what is called “The Book<sup>9</sup>”, which works as a handbook for FrameNet theory and practice. Figure 14 shows an example of a Frame in the Berkeley FrameNet database.

Figure 14 – Risk\_scenario frame in Berkeley FrameNet

## Risk\_scenario

[Lexical Unit Index](#)

### Definition:

An Asset is in a particular Situation, which has a likelihood of leading to or inviting a Harmful\_event which will negatively affect the Asset.

**Semantic Type:** Non-Lexical Frame, Non-perspectivalized\_frame

### FEs:

#### Core:

<b>Asset</b> [ass]	Something judged to be desirable or valuable which might be lost or damaged.
<b>Harmful_event</b> [har]	An event that may occur or a state which may hold which could result in the loss of or damage to the <b>Asset</b> .
<b>Situation</b> [sit]	The <b>Situation</b> under which the <b>Asset</b> is safe or unsafe.

#### Non-Core:

<b>Degree</b> [deg]	A modifier expressing the deviation of the actual level of security from the expected value given the <b>Asset</b> , the <b>Situation</b> , and state indicated by the target itself.
<b>Place</b> [pla]	The <b>Place</b> at which the degree of safeness holds.
<b>Time</b> [tim]	The <b>Time</b> during which the degree of safeness holds.

### Frame-frame Relations:

Inherits from:  
 Is Inherited by:  
 Perspective on:  
 Is Perspectivalized in: [Being at risk](#), [Risky situation](#), [Run risk](#)  
 Uses:  
 Is Used by: [Rescuing](#)  
 Subframe of:  
 Has Subframe(s):  
 Precedes:  
 Is Preceded by:  
 Is Inchoative of:  
 Is Causative of:  
 See also:

### Lexical Units:

Created by RLG on 07/19/2006 03:31:03 PDT Wed

Source: [framenet.icsi.berkeley.edu/fndrupal/frameIndex](http://framenet.icsi.berkeley.edu/fndrupal/frameIndex)

<sup>9</sup> [framenet.icsi.berkeley.edu/fndrupal/the\\_book](http://framenet.icsi.berkeley.edu/fndrupal/the_book)

To demonstrate the structure of a frame, we detail the parts depicted in in Figure 14. At the top we can see name of the frame which is written with the first word capitalized and attached to the second word by the underscore sign: `Risk_scenario`.<sup>10</sup> The name is followed by the Definition, which is elaborated considering the evaluation of the properties needed to schematically represent an event, state, attribute, relation or entity.

The subsequent field is Semantic Type, which currently records information related to the type of the frame in terms of its perspective and the possibility of it being evoked by some lexical item. In the `Risk_scenario` frame, the types `Non-lexical frame` and `Non-perspectivalized frame` are present.

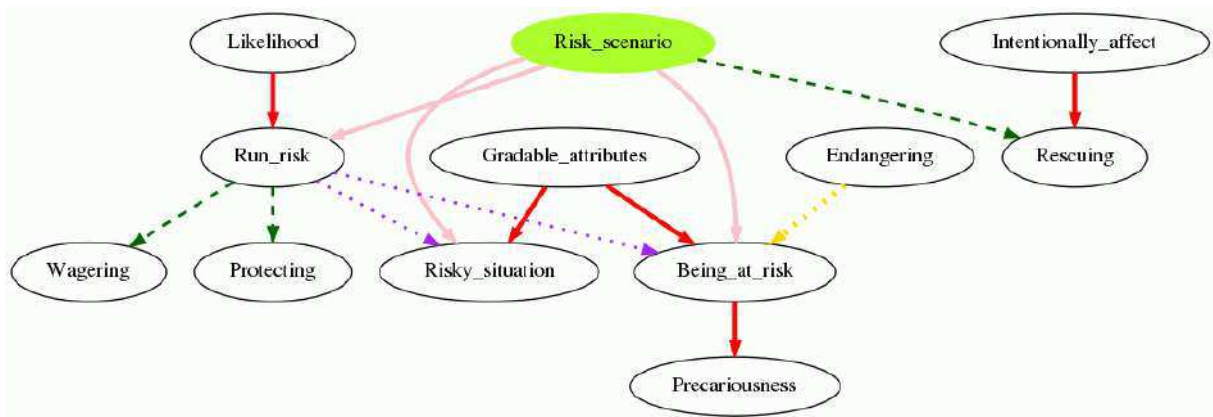
The FEs segment of the frame report refers to the Frame Elements, which are the component elements of the frame and model the semantic roles that constitute the scene description. There are three types of Frame Elements: (i) core, which are indispensable concepts for the frame to be instantiated; (ii) peripheral, which provide additional characteristics to the circumstances in which the scenes systematized by the frames occurs; (iii) extra-thematic, which amplifies the context of the scene, incorporating extra information from attributes of other frames. The `Risk_scenario` frame in Figure 14 shows three core Frame Elements (`ASSET`; `HARMFUL_EVENT` and `SITUATION`) which are all part of the frame definition. The non-core FEs (`DEGREE`, `PLACE` and `TIME`), on the other hand, are not often mentioned in the definition, once their presence is optional when the frame is evoked.

The subsequent field of the frame description – Frame-to-Frame Relations – refers to relations between frames. Berkeley FrameNet (as well as all FrameNets) is composed of frames and their associated roles in a network of typed relations. The `Risk_scenario` frame alluded to above, for example, is an umbrella frame for more specific perspectivalized frames such as `Being_at_risk` (in which the `ASSET` is exposed to a risky situation), `Run_risk` (in which a `PROTAGONIST` puts an `ASSET` at risk voluntarily) and `Risky_Situation` (in which a particular `SITUATION` is likely or unlikely to result in a harmful event). In Figure 15, a graph shows the relations between `Risk_scenario` and other frames in FrameNet.

---

<sup>10</sup> Following the convention usually adopted in FrameNet-related literature, frame names will be presented in this dissertation in `Courier new`, while Frame Element names are presented in `SMALL CAPS`.

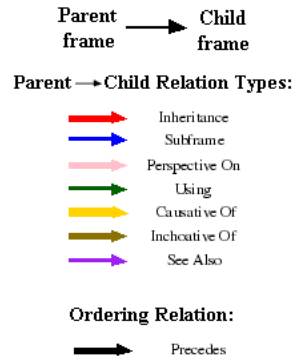
Figure 15 – Frame to frame relations for the Risk\_scenario



Source: <https://framenet.icsi.berkeley.edu/fndrupal/FrameGrapher>

The relations depicted in Figure 15 follow the legend in Figure 16. It shows the *Perspective On* relation mentioned previously with the pink arrow coming from the *Parent frame* Risk\_scenario to the *Children frames* Run\_risk, Risky\_situation and Being\_at\_risk. The *Inheritance* relation (red arrow) is the strongest relation between frames and determines that each semantic fact about the parent must be projected to the child frame in an equally specific or more specific fashion. For instance, in Figure 15, Precariousness inherits from Being\_at\_risk, projecting the attributes of the core ASSET FE in the latter to the also core THEME FE in the first. The *Using* relation (green arrow) refers to situation in which a part of the scene evoked by the child refers to the parent frame, as it is the case of the relation held between Rescuing and Risk\_scenario. The *Causative\_of* relation (yellow arrow) express situations in which the scene of the parent frame causes the scene of the child frame as a consequence, which is the case of Endangering and Being\_at\_risk. In a similar way, the *Inchoative\_of* relation (brown arrow, not present in Figure 15) connects a parent frame to a child frame that holds as a consequence of change of state. Finally, the *See also* relation (purple arrow) connects frames that are very similar and should be carefully differentiated, compared and contrasted, when an annotator is to choose the one that is evoked by a given LU, as it is the case between Run\_risk and Risky\_situation or Run\_risk and Being\_at\_risk.

Figure 16 – Frame-to-frame relations legend



Returning to the perspectivized daughters of the `Risk_scenario` frame, 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`. `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; ATKINS 1992). At last, `Risky_situation` is also evoked by nouns such as `risk.n` and adjectives such as `safe.a`; adverbs such as `safely.adv`, but not by any verb. A list of Lexical Units that can evoke each frame is at the bottom of the frame description – in Figure 14 there are no Lexical Units because the frame is Non-lexical. All that said, in the case of Figure 13, it seems adequate to associate the evocation of `Being_at_risk` to the Lexical Unit `safe.a` in the Child Safe Area sign. On the other hand, in the Tornado Safe Area the Lexical Unit `safe.a` evokes the `Risky_situation` frame. The database structure of Berkeley FrameNet also features annotated sentences to attest the use of a given word in the target frame. Figure 17 shows sentences in the `Risky_situation` frame annotated for the LU `safe.a`.

Figure 17 – Lexical Annotations in Berkeley FrameNet

## Annotation

[Lexical Entry](#) [Risky\\_situation](#)

### safe.a

Frame Element	Core Type
Asset	Core
Circumstances	Peripheral
Dangerous_entity	Core
Degree	Peripheral
Domain	Extra-Thematic
Frequency	Extra-Thematic
Place	Peripheral
Situation	Core
Time	Peripheral

#### [Turn Colors Off](#)

- added

1. Westinghouse Anniston , its subcontractors and the Anniston Chemical Agent Disposal Facility have more than 8.7 million **SAFE work** hours , 2,214 safe work days and a recordable injury rate of 0.49 per 200,000 hours worked as of the end of May .
2. Westinghouse Anniston , its subcontractors and the Anniston Chemical Agent Disposal Facility have more than 8.7 million safe work hours , 2,214 **SAFE work** days and a recordable injury rate of 0.49 per 200,000 hours worked as of the end of May .
3. This makes working at the chemical weapons destruction facility in Anniston one of the **SAFEST places in the nation to work** , according to statistics from the National Safety Council . ( U.S. Army Chemical Materials Agency , 23Jun06 , Washington Group International ) ( Link )
4. ANCDF [ Anniston Chemical Agent Disposal Facility ] receives third consecutive **SAFE Operating Facility** of the Year Award
5. Westinghouse Anniston and the Anniston Chemical Agent Disposal Facility ( ANCDF ) received its third consecutive **SAFE Operating Facility** of the Year Award from Washington Group International this week .
6. `` Awards like this reflect the **SAFE work** the employees do each and every day as they protect themselves , the environment and the community .
7. **No personal computer** , not even the one on a chief executive 's desk , is **SAFE** , this speaker noted .**CNI**
8. So far it seems to me that we are on **SAFE ground** .
9. I was to see that the doors were open and the signal of a green or white light in a window which faced the drive was to give notice if **all** was **SAFE** or if the attempt had better be postponed .**DNI**

**Annotator ID(s):** EGp, JKs, KmG, RLG

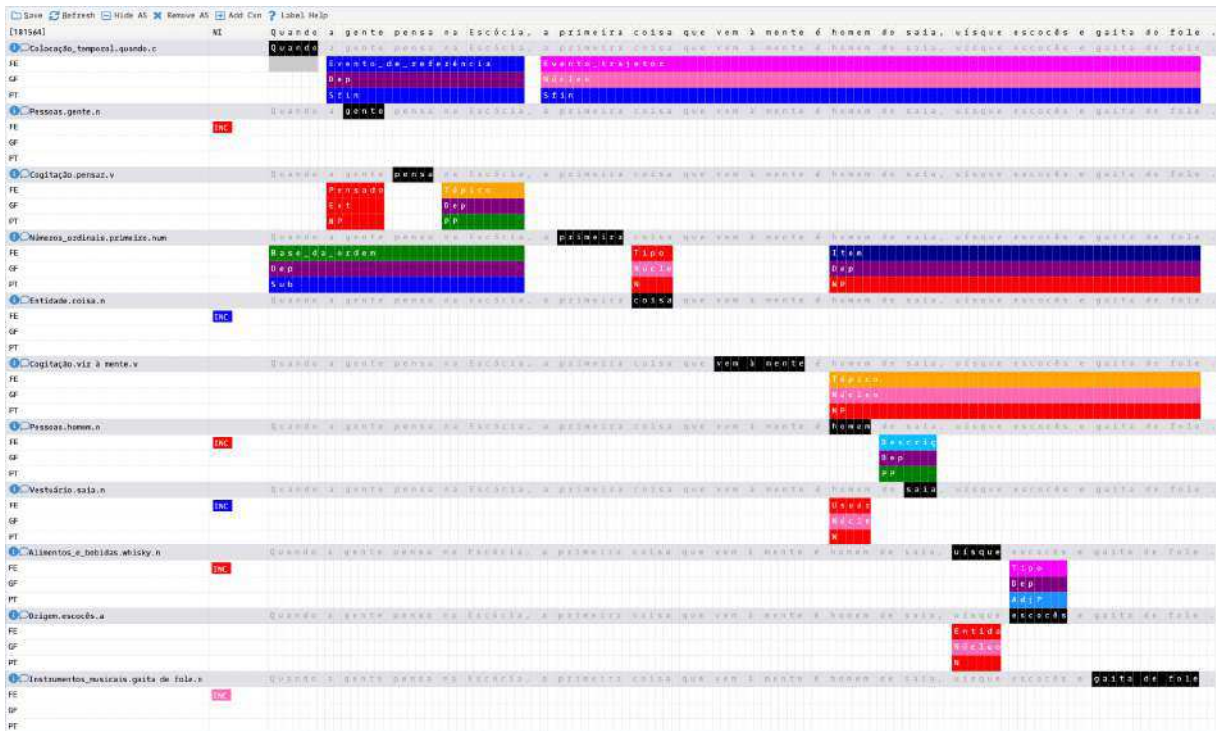
Source: [framenet.icsi.berkeley.edu/fndrupal/frameIndex](http://framenet.icsi.berkeley.edu/fndrupal/frameIndex) .

The sentences in Figure 17 are examples of lexicographic annotation, one of the two methods of FrameNet’s annotation process. Ruppenhofer et al. (2016) states that in lexicographic annotation the focus is on “recording the range of semantic and syntactic combinatory possibilities (valences) of each word in each of its senses.”. It means that sentences are extracted from different texts in a corpus. All sentences contain a predetermined target LU. A selection of the extracted sentences is annotated in respect to that particular LU. Full-text annotation is the other method. In this case the sentences chosen to be annotated are components of a full text. The annotator analyzes sentence by sentence, word by word, selecting one frame

for each word – or multiword expression – as a target – highlighted in black –, as the example in sentence (1) and in Figure 18.

- (1) **Quando**<sup>Temporal\_collocation</sup> a gente **pensa**<sup>Cogitation</sup> na Escócia, a **primeira**<sup>Ordinal\_numbers</sup> **coisa**<sup>Entity</sup> que **vem à mente**<sup>Cogitation</sup> é: **homem**<sup>People</sup> de **saia**<sup>Clothing</sup>, **uísque**<sup>Food</sup> **escocês**<sup>Origin</sup> e **gaita de fole**<sup>Musical\_instruments</sup>.<sup>11</sup>

Figure 18 – Full-text annotation in the FrameNet Brasil database



Source: [webtool.framenetbr.ufjf.br](http://webtool.framenetbr.ufjf.br)

In both lexicographic and full-text methods, the annotation is computationally organized by sets of annotation layers, comprising, at least, the following: frame element (FE), grammatical function (GF) and phrase type (PT)<sup>12</sup> – see Figure 19.

<sup>11</sup> When<sup>Temporal\_collocation</sup> we think<sup>Cogitation</sup> of Scotland, the first<sup>Ordinal\_numbers</sup> things<sup>Entity</sup> that come to mind<sup>Cogitation</sup> are: man<sup>People</sup> in skirts<sup>Clothing</sup>, Scotch<sup>Origin</sup> whisky<sup>Food</sup> and bagpipe<sup>Musical\_instruments</sup>.

<sup>12</sup> Grammatical functions e phrase types are both language specific e defined by each frame. For the sets used in English see Ruppenhoffer et al (2016). For the sets used in Brazilian Portuguese see Torrent and Ellsworth (2013).

Figure 19 – Example of annotation layers in the FrameNet Brasil WebTool



Source: <http://webtool.framenetbr.ufjf.br>

In Figure 19, we see a part of sentence annotated in the full-text method in the FrameNet Brasil database, one of Berkeley FrameNet 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. There is also an international multilingual initiative: Global FrameNet<sup>13</sup>.

In this sample, the target LU is *pensa.v* (think.v) evoking the *Cogitation* frame. The Frame Elements annotated in the first layer are the core ones for this frame: ‘*a gente*’ (we) as COGNIZER and ‘*na Escócia*’ (of Scotland) as TOPIC. The second layer shows the Gramatical Function annotations for the same phrases annotated in the first layer in relation to the target LU *pensa.v* (think.v), which are, in the case, External argument (Ext) for ‘*a gente*’ (we) and Dependent (Dep) for ‘*na Escócia*’ (of Scotland). The third layer is dedicated to labeling the phrases for Part of Speech, which in the sample are Noun phrase (NP) for ‘*a gente*’ (we) and Prepositional Phrase (PP) for ‘*na Escócia*’ (of Scotland).

Next section details the specificities of structure and methods of the Brazilian version of FrameNet.

### 3.2.2 FrameNet Brasil

FrameNet Brasil is the Brazilian branch of FrameNet. On top of expanding FrameNet into Brazilian Portuguese, it has been implementing additional semantic structure to the FrameNet model aimed at enriching the database structure and amplifying the granularity of the semantic representations (TORRENT et al., 2022).

The first additional implementation is the *Frame element to frame* relation, which links FEs to the frames licensing the lexical items that typically instantiate those elements. Gamonal (2017) presents the modeling of this relation and its implementation in FrameNet Brasil as a way of both enriching the database and providing a model of semantic specification for the FEs that could be more efficient than semantic types. In this sense, the relation was applied to FEs

<sup>13</sup> [globalframenet.org](http://globalframenet.org)

of an event frame that are semantically specified by an entity frame – through its core FEs. Hence, the FE `Tourist`, in the `Touring` frame, for instance, is linked via and *FE-Frame* relation to the `People_by_leisure_activity` frame. Gamonal (2017) also points out that the *FE-Frame* relation can occur between entity frames – e.g., FE `ORIGIN` in the `People` frame, which is linked to the `Political_locale` frame. Currently, there are 3,582 instances of this relation in FrameNet Brasil. A total of 1,198 out of the 1,306 frames in FrameNet Brasil – 91.7% – have at least one instance of the *FE-Frame* relation, and the average count of *FE-Frame* relation instances per frame is 2.98. Therefore, this new relation represents a sensible increase in the density of the database, compared to the original Berkeley FrameNet model, which only features *Frame-to-frame* relations. The number of instances of *Frame-to-frame* relations in the FrameNet Brasil database is 1,846, a little more than half of the total instances of *FE-Frame* relation instances (TORRENT et al., 2022).

Another relation added by FrameNet Brasil connects core FEs to non-core FEs in the same frame. Gamonal (2017) shows this implementation for cases in which the linguistic material that instantiates an FE can be defined in terms of another FE metonymically. In other words, the non-core FE can act as metonymic substitute for the core FE. For instance, in the `Business` frame the FE `PLACE` can act as a metonymic substitute for the core FE `BUSINESS`.

The most recently added group of relations developed by FrameNet Brasil are the *frame-mediated ternary qualia relations* (TORRENT et al., 2022; forthcoming). These relations hold between LUs and are inspired by the *Qualia Structure* (or *Qualia Roles*) categorization, as postulated by Pustejovsky (1995) in the Generative Lexicon Theory (GLT). GLT arises as an approach to lexical semantics which focuses on the combinatorial and denotational properties of words. It also pays attention to peculiar aspects of the lexicon such as polysemy and type coercion. In this sense, qualia roles emerged as characteristics or different possible context predication modes of a lexical item, in a context of 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. Pustejovsky and Jezek (2016, p.3) argue that a *Quale* “indicate[s] 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”. Pustejovsky (1995) defined four qualia roles as the aspects of a word structural meaning:

- The Formal quale (*Formal\_of*) 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.



- The Telic quale (*Telic\_of*) 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).
- The Constitutive quale (*Constitutive\_of*) is established between an object and its constituents and the material involved in its production.
- The Agentive quale (*Agentive\_of*) 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 natural type and a causal chain.

Pustejovsky and Jezek (2016) summarize the structure and the roles of a lexical item  $\alpha$  as follows:

Figure 20 – Qualia structure representation

$$\left[ \begin{array}{l} \alpha \\ \text{QUALIA} = \left[ \begin{array}{l} F = \text{what } \alpha \text{ is} \\ C = \text{what } \alpha \text{ is made of} \\ T = \text{function of } \alpha \\ A = \text{origin of } \alpha \end{array} \right] \end{array} \right]$$

Source: PUSTEJOVSKY; JEZEK, 2016, p. 8.

Based on this, a qualia structure representation for the word *pizza.n* can be built as follows:

Figure 21 – Qualia roles for *pizza.n*

$$\left[ \begin{array}{l} \text{pizza.n} \\ \text{QUALIA} \left[ \begin{array}{l} F = \text{food.n} \\ T = \text{eat.v} \\ C = \text{flour.n} \\ A = \text{cook.n, pizza restaurant.n} \end{array} \right] \end{array} \right]$$

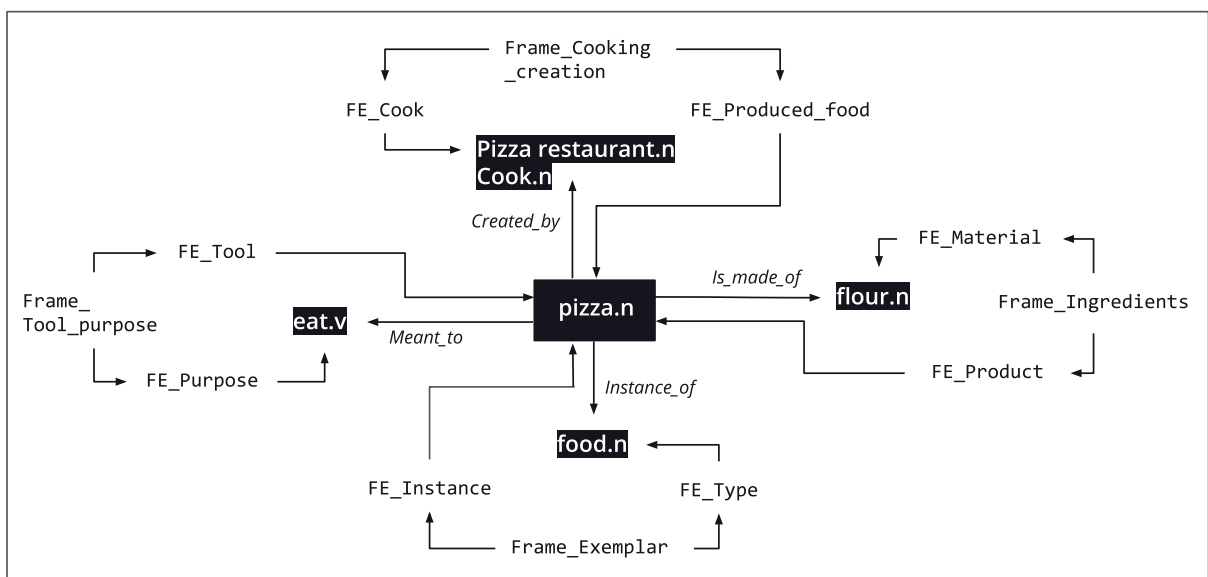
Source: BELCAVELLO et al., 2020, p. 25.

In Figure 21, we see that *food.n* is represented as *formal\_of* *pizza.n*, being a more general category to which *pizza* belongs, which means that *pizza* is a specific type of food. The word *eat.v* is *telic\_of* *pizza.n* since a *pizza* has the purpose of being made to be eaten. Once it is an ingredient used in it, *flour.n* is *constitutive\_of* *pizza.n*. *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, both matching the origin aspect of it.

Working with *Qualia*, however, usually leads to one recurrent issue: although these four aspectual meaning relations are conceived as subparts of the meaning of a word, they are still too general. Therefore, the attempts to improve the specificity have led to the proposal of long lists of subtypes for each relation, such as the SIMPLE (LENCI et al., 2000) and the Brandeis (PUSTEJOVSKY et al., 2006) ontologies.

FrameNet Brasil, however, instead of incorporating another list of relations to the database, tackles the granularity issue in a different way. Costa (2020) proposes modeling a more specific and granular subtype of qualia relations: the *frame-mediated ternary qualia*. In this innovative type of ternary relation, two LUs, 1 and 2, are linked to each other via a given quale which uses specific frames as background, 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. This modeling addresses both the lack of direct links between LUs in the FrameNet model and the poor specificity of qualia relations. An example of this implementation is shown in Figure 22:

Figure 22 – Frame-mediated ternary qualia relations for pizza.n in FN-Br



Source: BELCAVELLO et al. 2020.

As Figure 22 shows, the LU `pizza.n` is related with five other LUs via qualia in the FrameNet Brasil database:

- Agentive relation (`created_by`) with `pizza restaurant.n` mediated by the `Cooking_creation` frame, which relates `pizza.n` to the FE `PRODUCED_FOOD` and `pizza restaurant.n` the FE `COOK`;<sup>14</sup>
- Also an Agentive relation (`created_by`) with `cook.n` mediated by the `Cooking_creation` frame, which relates `pizza.n` to the FE `PRODUCED_FOOD` and `cook.n` to the FE `COOK`;
- 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`;
- 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.

The degree of generality of frames determines two criteria that orient how frame are recruited to mediate Qualia: (i) frames should be as general as possible, provided that they do not conflict with or overgeneralize the quale and (ii) frames should be as specific as necessary. Moreover, only core – and core unexpressed – FEs can be recruited as ternary qualia mediators, because only core FEs are absolutely frame-specific, hence, they are the only ones that actually differentiate one frame from another (TORRENT et al., forthcoming).

Taking the `Tool_purpose` frame in Figure 22 as an example, there are two more general frames in the inheritance chain leading to it 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 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.

---

<sup>14</sup> Because we also implement metonymy relations between FEs, the peripheral FE `PLACE` can stand for the core FE `COOK` in the `Cooking_creation` frame.

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.

Currently, the FrameNet Brasil database has 7,642 instances of ternary qualia relations holding between pairs of LUs in Brazilian Portuguese, 17,256 instances for LU pairs in English, 400 in Spanish and 148 in French. Those instances are distributed across 43 types of relations. The current instances of ternary qualia relations were mostly modeled for the domains of tourism and sports, and the number of instances continues to grow as new domains are included in and/or refined by FrameNet Brasil.

Qualia, FE-FE and FE-Frame relations, once implemented, allowed not only for the modeling of relations between frames, but also for connections on other levels of the FrameNet database structure, such as Frame Elements and Lexical Units. This was key to make FrameNet Brasil able to represent aspects of meaning and context not captured by the original Berkeley FrameNet database structure (TORRENT et al., 2022).

Thus, it is based on this set of efforts to extend the capability to represent meaning and context that FrameNet Brasil has directed itself to the implementation of a multimodal approach. In this sense, we present in Chapter 4 the process of building a multimodal corpus and the methods used to define an annotation methodology for generating a substantially robust dataset in terms of representing the multimodal grammar and semantics present in the phenomenon taken for the corpus.

## 4 CORPUS AND METHODS

This chapter details and motivates the corpus chosen for this research, describes the annotation pipeline proposed and the development of a specific tool to implement it. Moreover, it presents the experiment carried out in order to validate the annotation methodology proposed. Finally, the chapter describes the corpus annotation task.

### 4.1 CORPUS

The guidelines to build the corpus for this research were (i) being in the tourism domain – because FrameNet Brasil has it as one of its main areas of interest and development and, hence, more coverage – and (ii) provide rich examples of audio and video combination in terms of meaning making with Brazilian Portuguese content. We found those features in the TV Travel Series “*Pedro pelo Mundo*”.

The show premiered in 2016 on GNT, a cable channel dedicated to entertainment and lifestyle productions. Four seasons of “*Pedro pelo Mundo*” were aired until 2019. There were 40 episodes in total. 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 dissertation, the corpus will be limited to the 10 episodes of the first season.

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. Therefore, most of the episodes focus on a specific city, like the already mentioned example of Edinburgh, but some propose a broad view of a country, like Iran, Colombia and Iceland. There is also one episode specifically dedicated to a neighborhood: Brooklyn.

The format of the show combines stand ups, voice-over sequences, short interviews and video clip sequences. It offers, then, rich material as an exemplar of complex composition of audio and video for meaning making.

For each 23-minute-episode, the audio transcription generates approximately 200 sentences, which means 2,000 sentences for the first 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 (BELCAVELLO et al., 2020). For the video, the annotation average calculated during the pilot

study reported by Belcavello et al. (2020) is of 500 visual objects per episode, which means 5,000 visual objects in total (BELCAVELLO et al., 2022; TORRENT et al., 2022).

## 4.2 THE CORPUS IMPORT PIPELINE

FrameNet Brasil has its own tool to annotate textual data within the framework of frame semantics, the FN-Br Annotation WebTool (TORRENT et al, forthcoming). It is a web-based database management and annotation software used by both local FrameNet projects in Brazil, Sweden, Croatia and Japan, and in the Global FrameNet Shared Annotation Task (TORRENT et al., 2018) Figures Figure 18 and Figure 19 show examples of text annotation using FN-Br Annotation WebTool.

Therefore, the process of developing a multimodal annotation scheme compatible with the FN-Br database would depend on the integration of a multimodal module to the WebTool. We, then, explored the possibility of building a tool capable of annotating visual objects, correlate them with the synchronic textual data (transcription of every verbal data) and label frames and frame elements evoked by them.

With all that in mind, we started the development of Charon<sup>15</sup> (BELCAVELLO et al., 2022), the multimodal annotation tool and database management application, as a project for Google Summer of Code 2020<sup>16</sup>, which, in turn, was derived from another project for Google Summer of Code 2019<sup>17</sup>. Charon is a semi-automatic, human-in-the-loop tool for annotating static and dynamic images for semantic frames. Charon was developed to meet the following key requirements: (i) compatibility with existing FrameNet software; (ii) annotation of image with FrameNet categories; (iii) linkage of image and textual annotations.

Another goal was making this tool capable of preprocessing videos to extract verbal data and make it ready for being annotated in the FN-Br WebTool. In parallel, the visual data would be submitted to computer vision processes so as visual objects could be identified and made available for annotation. This goal was accomplished by the construction of the Corpus Building Module of Charon. Figure 23 shows a schematic representation of the pipeline designed for data processing.

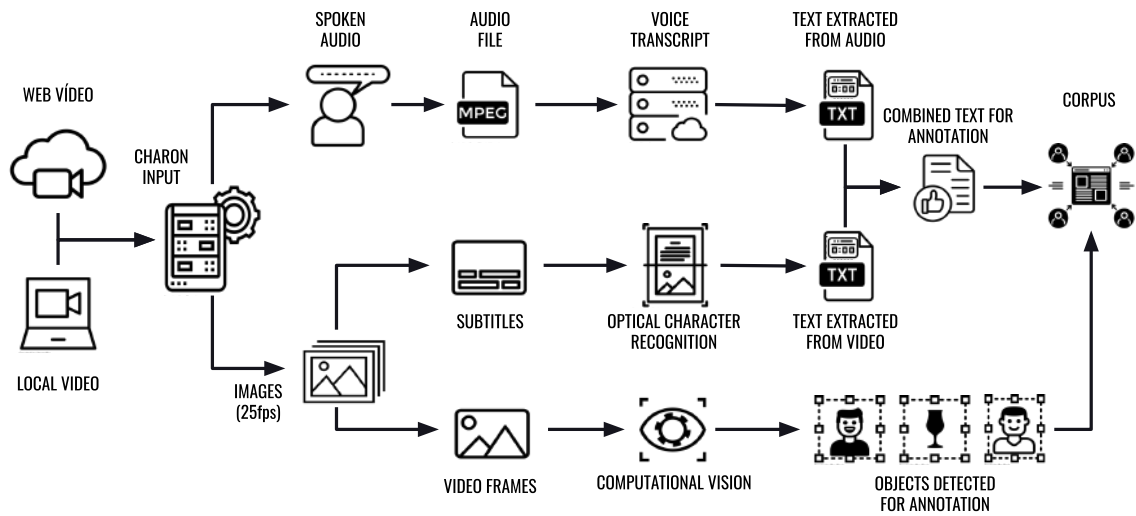
---

<sup>15</sup> charon.frame.net.br .

<sup>16</sup> <https://summerofcode.withgoogle.com/archive/2020/projects/4857286331203584/> .

<sup>17</sup> <https://summerofcode.withgoogle.com/archive/2019/projects/5902293138931712/> .

Figure 23 – Multimodal Corpus Import Pipeline



Source: the author.

The pipeline designed for corpus import and video preprocessing starts with the selection of the video input, either from a web address or a local file. Then the tool imports, pre-processes and separates audio data and image data to proceed with parallel tasks in two data flows: one for the audio and another for the images. The extracted spoken audio runs through a speech-to-text cloud service, which detects word by word what is said throughout the video. Each word receives time stamps indicating the time span during which they are spoken.

From the image flow, subtitles are extracted using an optical character recognition software. They are going to integrate the textual data, combined with the content extracted from the spoken audio. This is necessary because the subtitles found in the corpus bring the Portuguese translation for all the non-Portuguese spoken content. As a TV show produced for the Brazilian audience, Pedro speaks Portuguese when talking to the camera or narrating off screen. But when he interviews people, most of the times they speak English – sometimes the interviewees are Brazilians or Portuguese speakers who live in the destination, and sometimes other languages are spoken and translated on the flow of action. The subtitles extracted are timestamped and then merged to the text corpus with the output of the speech-to-text software. Words and sentences extracted then go through a human-in-the-loop stage, where users can build sentences from the words, edit them, as well as check and adjust time stamps. Finally, the textual part of the corpus is saved and sent to the FN-Br WebTool for annotation.

In the third segment of the pipeline, Charon also processes non-verbal visual data. The images extracted at a 25 frames per second rate are stamped for both time (in seconds) and

video frame (in sequential numbers). They run through a computer vision algorithm, which automatically tags objects in each frame, associating a bounding box and a category to them. At the end of the pipeline, the data is available for the following annotation methods:

- independent text annotation.
- independent visual annotation.
- text-orientated multimodal annotation.
- visual-oriented multimodal annotation.

To decide which method should be adopted, we designed a spoken audio dominance investigation experiment, which is presented next.

### 4.3 SPOKEN AUDIO DOMINANCE INVESTIGATION EXPERIMENT

Since the first pilot experiment we conducted (BELCAVELLO, 2020), the working hypothesis that emerged was that we were facing a corpus in which the construction of meaning of the messages expressed in each excerpt and in the program as a whole is organized from the spoken audio. Therefore, once the annotation pipeline gives us different possibilities of annotation, we had to define which one would be the most appropriate for the corpus and the goals we have established, considering what the data indicates. In this sense, the aim was to determine the starting point and the order of the steps of the annotation process, having in mind that this definition is part of the development of a frame-based annotation methodology for audiovisual corpora.

As presented previously through the example in Figure 12 and based on Cohn's (2016a) step-by-step method for analysis of multimodal interactions (Figure 11), we hypothesized that the corpus was composed by *audio-dominant* material, with some sequences of *co-assertiveness* and sporadic occurrences of *visual-dominance* (e.g. Figure 8). This would indicate that the appropriate annotation methodology should be text-oriented. To investigate the validity of this claim, we designed an eye-tracking experiment (DUCHOWSKI, 2017; CONKLIN, PELLICER-SANCHEZ, CARROL, 2018; FONSECA, MAIA, 2022) aimed at measuring to what extent a semiotic mode – the verbal expressions present in spoken audio – acts in directing the audience's gaze to certain elements or regions of the images – another semiotic mode – presented simultaneously on screen. In other words, we designed the experiment to investigate the hypothesis that, in certain multimodal audiovisual objects, the



verbal semiotic mode – the spoken audio<sup>18</sup> – guides the audience's gaze to specific visual elements or specific regions of the image (visual semiotic mode) presented on screen.

The experiment was designed as follows: we selected 6 excerpts from episode 6 – Edinburgh – of “*Pedro pelo Mundo*”’s first season. The excerpts were assembled as a single 10-minute-long video, which includes, in addition to the excerpts, the written instructions on screen for the participants. The assembled video had two different versions: a complete one, intended for the control group; and a modified one, from which all the spoken audio was extracted, intended for the experimental group.

To participate in the research, we recruited 39 native Brazilian Portuguese speakers who were literate and had no recorded limitations in their auditory and/or visual abilities. Each participant was randomly assigned to one of the groups: group A, control, and group B, experimental. Group A consisted of 20 participants and Group B, 19 participants. Each participant watched the assigned group's version of the 10-minute video in a booth equipped with eye-tracking equipment. Immediately after each recording, we checked that at least 80 percent acuity was achieved in eye gaze capture. The recorded data were saved for later analysis.

The 6 excerpts<sup>19</sup> vary in length and themes as follows:

- Extract 1 (1min 33s): Opening – rethinking Scottish identity in Edinburgh.
- Extract 2 (2min 14s): Haggis for breakfast.
- Extract 3 (55s): Kilt.
- Extract 4 (1min 2s): Pub and beer.
- Extract 5 (50s): Walking and whisky.
- Extract 6 (1min 46s): Final remarks.

The guiding questions to interpret the results of the experiment are:

- (i) Are there regular and significant differences in points of fixation between the group exposed to the complete version and the group exposed to the modified version?
- (ii) If there are differences, can we associate the points of fixation of the group who watched the complete version with the spoken audio guidance?

---

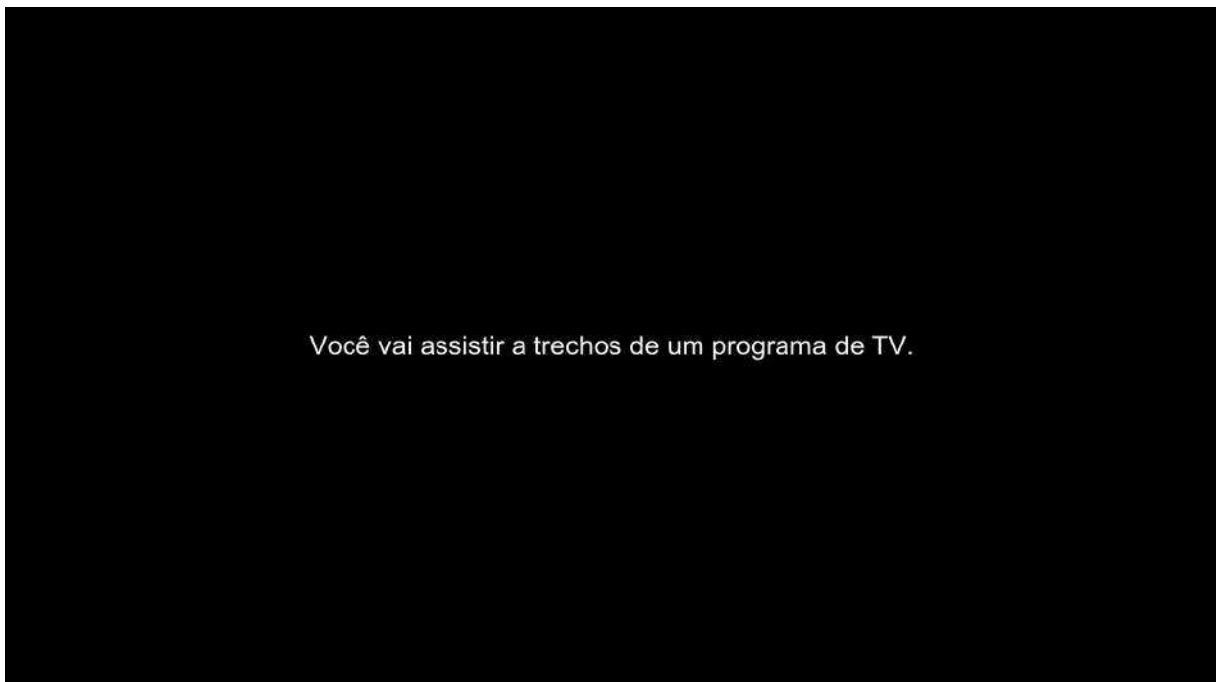
<sup>18</sup> We consider that subtitles can also be seen as components of the verbal semiotic mode, but in terms of media and screen, these elements manifest themselves visually and are thus perceived by sight and not by hearing. Thus, in an eye-tracking experiment, it is empirical to predict that subtitles attract gazes to themselves, which would not be adequate for the verification of image-sound pairing proposed here.

<sup>19</sup> The 12 videos are available at [https://drive.google.com/drive/folders/11iNBHiKnFEa8HND06UPJXhnz8dax1CkE?usp=share\\_link](https://drive.google.com/drive/folders/11iNBHiKnFEa8HND06UPJXhnz8dax1CkE?usp=share_link).

Positive answers to these two questions can corroborate the indication that the annotation process should start with text annotation and also that the visual annotation should follow the former, looking for correspondences or complementarity. Results are presented and discussed in the next chapter.

Some of the choices made in the process of selecting the video clips is worth mentioning. First, all the excerpts correspond to sequences of the program where there are only speeches in Portuguese and no subtitles – since we know that subtitles would impact on the capturing of the participants' gaze. Second, the video clips that were altered with the removal of the speech had the soundtracks kept as the original, as well as any sound effects. Third, all the excerpts start with a short sequence of images with music, and only a few seconds later the speech begins. We proceeded that way to avoid data loss in the first frames or in the first takes, creating an accommodation or warm-up space for the participants' gaze. Likewise, the instructions were written on a black background (Figure 24), in the center of the screen, with clear indication of the intervals between video excerpts and a countdown warning that a new excerpt was about to begin. Finally, the choice of clips lasting between 50 seconds and 2 minutes, interspersed with instruction and rest screens, sought to avoid generating boredom or fatigue for each participant during the viewing event.

Figure 24 – Example of instructional sentence on screen



Source: Print screen from a Tobii Studio project. The sentence reads “You will watch extracts from a TV program”.

About the event itself, we took care to make the participants as comfortable as possible. Each one was instructed to try to watch the video as if they were watching a TV program at home. In this sense, the choice of a reasonably comfortable chair to avoid tense postures, the assurance of silence and inviolability in the booth were principles carefully observed. The characteristics of the equipment used, Tobii TX300, which does not require the use of glasses or static positioning, also contributed to bring the participants' experience closer to a routine, everyday experience. The experiment was conducted at the Center for Studies in Language Acquisition and Psycholinguistics at (NEALP), which owns the equipment and the expertise on designing, conducting, and analyzing eye-tracking experiment.

The experiment was submitted to UFJF's Ethics in Human Research Committee and approved under protocol number CAAE 48300921.4.0000.5147 on August 30, 2021. Data was collected between November 10, 2022, and January 11, 2023.

#### 4.4 CORPUS ANNOTATION

Regardless of the result of the spoken audio dominance investigation experiment, the annotation process was devised as divided into two parts: text annotation and image annotation. Annotation was carried out by undergraduate students trained in the task by the FrameNet Brasil research team. 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 students hired by the FrameNet Brasil Lab to perform several annotation tasks, including the one described here. They receive 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. Stipends are funded by the National Council for Scientific and Technological Development (CNPq), by the Minas Gerais State Foundation for Research Support (FAPEMIG) and by the Federal University of Juiz de Fora (UFJF). A total of 12 undergraduate students took part in the permanent annotation team.

The temporary annotation teams were assembled among the students enrolled in the undergraduate division courses offered by the FrameNet Brasil team of researchers every semester: the Language Technology Workshop. Each workshop is composed of 45 hours of academic work, comprising tutoring and annotation practice. Two classes of the Language Technology Workshop contributed to the annotation of the dataset presented in this dissertation: one in October/November of 2022 and one in February/March of 2023. A total of 32 undergraduate students were part of the temporary annotation team.

The teams conducted the annotation using both the WebTool – for text annotation – and Charon’s dynamic mode – for image annotation. Next chapter presents how these tools were used to generate the annotation methodology proposed in this dissertation.

## 5 REFINING THE ANNOTATION METHODOLOGY

This chapter presents the results of the spoken audio dominance investigation experiment and their impacts on the guidelines for annotation; It also details Charon – the annotation tool in use for both for corpus building and video annotation – and proposes a methodology for multimodal annotation of videos using FrameNet categories.

### 5.1 RESULTS OF THE AUDIO-DOMINANCE VERIFICATION EXPERIMENT

As mentioned previously in section 4.3, there were two guiding questions to interpret the results of the experiment:

- (i) Are there regular and remarkable differences in points of fixation between the group exposed to the complete version and the group exposed to the modified version?
- (ii) If there are differences, can we associate the points of fixation of the group who watched the complete version with the audio guidance?

To answer these questions, it was necessary to determine in which parts of the video extracts we would look for responses. So, it is worth saying that the six video extracts were chosen after preliminary analysis that pointed out the possibility of having Lexical Units triggering meaning construction and/or directing gaze to specific visual elements on screen. In that sense, the analysis of the results of the experiment focused on looking for differences in points of fixation in shots related to some sentences.

Therefore, in Extract 1 – Opening we analyzed the behavior of participants in group A when listening to sentence (1) reproduced in (2):

- (2) *Quando a gente pensa na Escócia, a primeira coisa que vem à mente é: homem de saia, uísque escocês e gaita de fole.*<sup>20</sup>

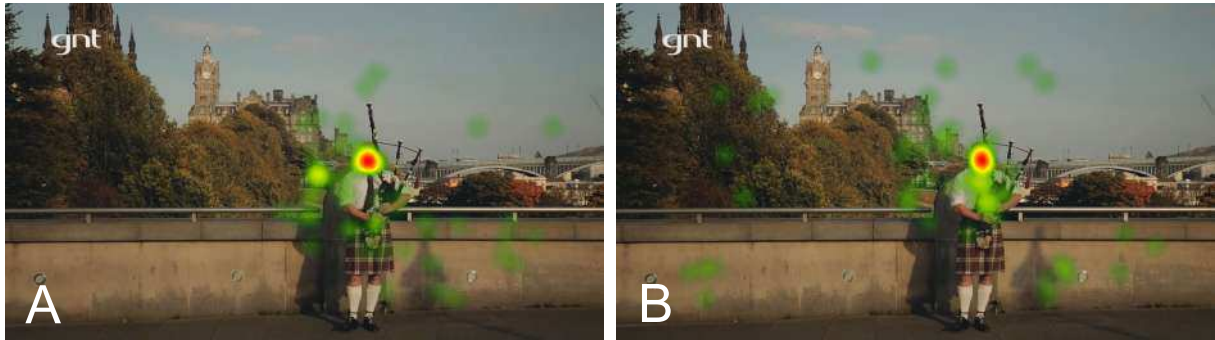
Figure 12 shows what participants in group A saw, when listening to sentence (2). Our first question analyzed was: once participants in group A listen to ‘homem de saia’ (man in skirt), do they look particularly at the kilt showed on screen? The answer is no. Figure 25 shows

---

<sup>20</sup> When we think of Scotland, the first things that come to mind are: man in skirts, Scotch whisky and bagpipe.

the answer: no, it is not possible to say people in group A look significantly at the kilt more than people in group B. There is some, but minor, fixation at the kilt in both groups. It also shows that most people in both groups pay attention to the man's face most of the time.

Figure 25 – Heat map comparison for ‘homem de saia’



Source: the author.

The subsequent shot (Figure 26) is a medium shot and brings a closer view of the referred ‘man in skirt’ – which is actually a kilt – playing the bagpipe. At this point, viewers in group A have listened to Pedro saying ‘gaita de fole’ (bagpipe). We then decided to check whether the participants in group A looked at the bagpipe first. Figure 26 shows that they do. However, it is important to note that the region that concentrates most gaze fixations coincide with the region of the previous shot where the man's face was. Hence, it seems a bit far-fetched to be adamant and assure that the gazes are there because of the mention of the lexical unit, having this other very relevant factor. But why, then, do participants in group B present a different behavior? Possibly, because, lacking one of the factors, the attraction to the face has been more quickly imposed. Notice, however, that the participants in group B are not looking more at the face than at the bagpipe. In fact, the intensity of the fixations on these two points is quite similar.

Figure 26 – Heat map comparison for ‘gaita de fole’



Source: the author.

Add to this the fact that this situation lasts only about half a second. At the end of the shot, considering the fixations throughout its duration, the gaze plot map of the two groups is quite similar, dividing the fixations between the face region and the middle region – the tenor drones – of the bagpipe, in general.

Figure 27 – Gaze plot for ‘gaita de fole’



Source: the author.

In Extract 2, ‘Haggis for breakfast’, Pedro seats at a table in a restaurant to have a meal. This sequence is preceded by the one described in Figure 8, which is very important for making the viewers infer that the meal is a breakfast, not lunch or dinner. When seated, Pedro says sentence (3):

- (3) *Resolvi pedir um café da manhã tradicionalíssimo por aqui chamado haggis, que vem a ser um bucho de carneiro recheado de vísceras.*<sup>21</sup>

This sentence makes clear for participants in group A that it is a breakfast, and that Pedro is waiting to be served. Pedro goes on describing what he knows about the dish. All this creates a certain expectation about the dish that will arrive and then puts the participants in group A on alert. What we examine here is whether participants in group A look for the food and look at the plate sooner than participants in group B, since they are aware that a dish described as exotic is on the way. Figure 28 shows the moment when some people in group A are looking at the food while only two people in group B are spotting the dish area. The difference is not major, nonetheless, it is present.

<sup>21</sup> I decided to order a very traditional breakfast here called haggis, which is a sheep's stomach stuffed with entrails.

Figure 28 – Gaze plot for the haggis arrival



Source: the author.

When analyzing the shot throughout its entire duration, Figure 29 shows that there are two important areas that concentrate the gazes: Pedro's face and the dish. There are some registers beyond these two areas, but they are very light.

Figure 29 – Heat map for the dish arrival shot



Source: the author.

In Extract 3 – Kilt, Pedro visits a traditional kilt manufacturer and store and presents its owner, Gordon Nicolson, the kilt specialist. For participants in group A, Nicolson appears on screen for the first time after Pedro pronounces sentences (4) and (5):

- (4) *Um dos grandes símbolos da Escócia é o kilt, ou aquela saia xadrez que os homens usam há séculos.*<sup>22</sup>
- (5) *Gordon Nicolson é referência local no assunto.*<sup>23</sup>

<sup>22</sup> One of the great symbols of Scotland is the kilt, or that plaid skirt that men have worn for centuries.

<sup>23</sup> Gordon Nicolson is a local reference on the subject.



It is a shot with camera movement: a left to right pan. It clearly directs gaze from Pedro to Nicolson, revealing Nicolson to viewers. However, Figure 30 shows that all the 20 participants in group A look at Nicolson face, as soon as he is framed, while around half of the participants in group B continue exploring the ambient and don't fixate at Nicolson's face in the first opportunity.

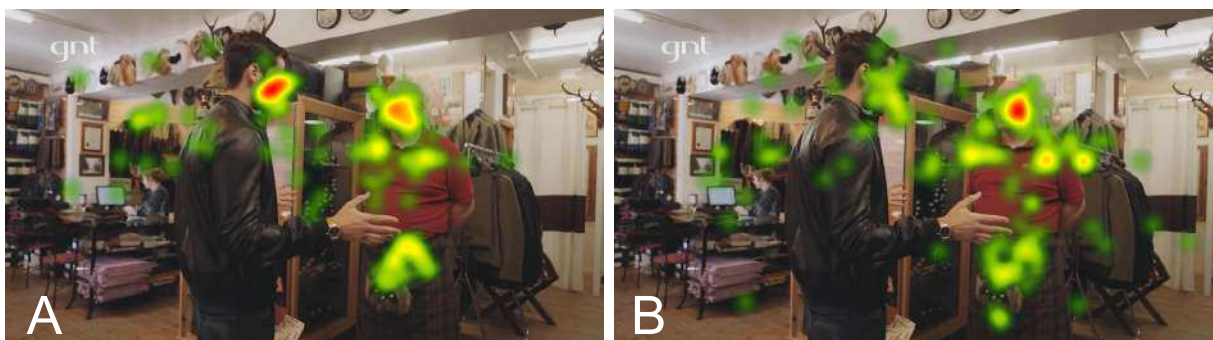
Figure 30 – Gaze plot for Gordon Nicolson revelation



Source: the author.

The subsequent shot is a medium wild shot which reveals that Nicolson is wearing a kilt. Once it is said – then, explicit – for participants in group A that the subject of the extract is the kilt, we evaluated if they pay more attention at the kilt than participants in group B. Figure 31 shows that the concentration of fixations is very similar. However, it is interesting to observe that participants in group B had looked less at Pedro's face than participants in group A.

Figure 31 – Heat map for Nicolson and his kilt



Source: the author.

In the final half of extract 3 viewers see Pedro walking through the streets of Edinburgh wearing a complete Scott costume, which includes a kilt, the main subject of the extract. In this segment we evaluated participants' gaze in three different shots.

The first shot is a back view of the kilt – Figure 32. The shot initiates in a wild shot and moves with a dolly in until an insert close up shot. The heat map of the entire shot shows that participants in group A had gaze very concentrated at the kilt, while participants in group B explored the shot a little more beyond the kilt – it happened specially at the beginning of the shot, before the dolly in. Considering that group B do not have the explicit information that Pedro has the kilt as the main subject of the extract, we can infer that participants tried to explore the wild shot. On the other hand, guided by the lexical unit ‘kilt’ throughout the extract, participants in group A were, then, oriented to pay attention at the kilt. It is important to say that this shot is a good example of how much framing and camera movement work actively on gaze orientation. The kilt is seen at the very center of the image. The angle is lightly low. The dolly in keeps the kilt in the center and makes it fill most of the screen. It is, then, natural that the shot attracts gaze to the kilt.

Figure 32 – Initial framing and heat map for the back view of the kilt



Source: the author.

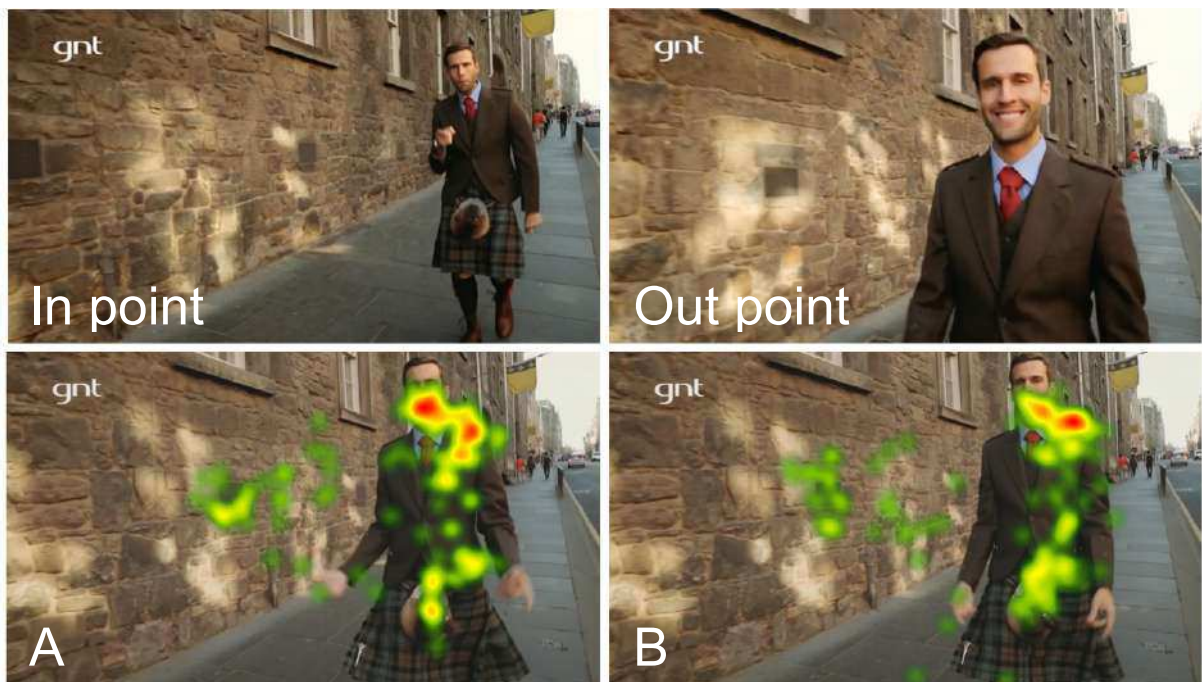
The subsequent shot (Figure 33) starts in a wild shot with Pedro framed in the right third of the screen looking at the camera and saying sentence (6):

(6) *Pedro pelo mundo e de saia, na Escócia!*<sup>24</sup>

<sup>24</sup> Peter around the world and in a skirt in Scotland!

Figure 33 shows the initial video frame and the last video frame of this shot. It also shows the heat map<sup>25</sup> of fixations counted from the initial frame until the intermediate video frame showed in Figure 33, which is the last video frame in which it is possible to see the entire kilt. The heat map shows that Pedro's face is the area with more fixations, followed by the kilt area, especially the region where the kilt is overlaid by the sporran, the Scottish traditional purse or pouch. Here we evaluated if the lexical unit '*saia*' (skirt), when pronounced, would guide participants in group A's gaze to the kilt, generating more fixations. However, although there is a small increase of concentration, it is not possible to say that people looked at the kilt directed by the LU, once participants in group B also looked at the kilt and the sporran.

Figure 33 – In point, out point and heat map for 'Pedro de saia'



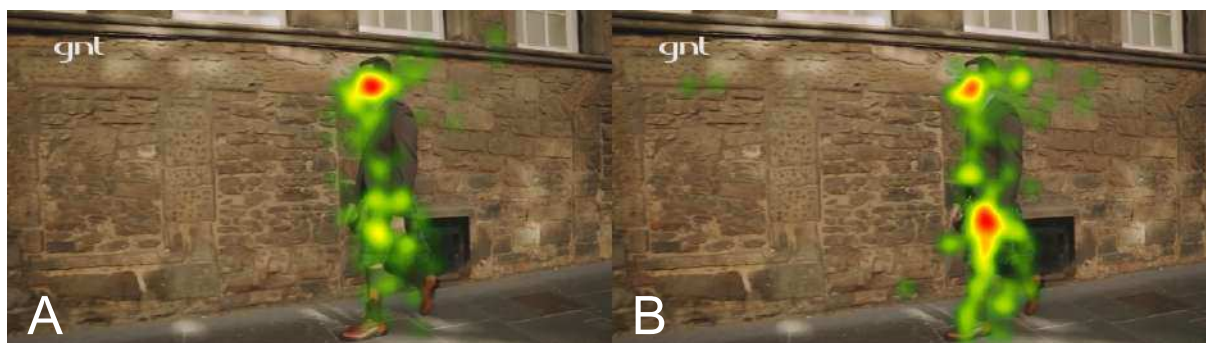
Source: the author.

The shot presented in Figure 34 follows the one presented in Figure 33. So, it is seen after Pedro pronounces the lexical unit '*saia*'. In this case, once more, we evaluate the possibility of the lexical unit '*saia*' make impact on group A participants, making them look at the kilt. The heat map in Figure 34, however, surprised us with the information that participants

<sup>25</sup> This heat map is also illustrative of the reminiscence of fixations from the previous shot. The fixations in the center of the screen are basically the fixations recorded at the beginning of the shot and correspond to the area where viewers were seeing the kilt in the previous shot, as demonstrated in Figure 32. It takes then some milliseconds before each viewer moves gaze from the area where she/he was looking at in the previous shot to the elements she/he is interested in or attracted by in the following shot.

in group B paid more attention to the kilt, then people in group A. It also shows that participants in group A focused more on Pedro's face, while participants in group B had two different focal areas with almost the same weight: the face and the kilt.

Figure 34 – Pedro in a skirt heat map 2



Source: the author.

Extract 4 – Pub and beer brings Pedro to visiting an Edinburgh's pub and trying a local beer. It is a night sequence, so there is eventually more contrast between light and dark in the image. This is a factor that often impacts the gaze fixation points in a video sequence. Then we analyzed the shot and heat maps presented in Figure 35. Participants in group A see the shot after hearing Pedro saying sentence (7):

(7) *Pedi uma cerveja escocesa, ele falou que é a melhor que eles têm aqui no pub.*<sup>26</sup>

With this sentence and this shot we evaluated if viewers in group A, after hearing that Pedro is talking about the beer, would concentrate their attention in the beer tulip which Pedro holds. The heat map, however, shows that their gaze fixations are concentrated at Pedro's face. Viewers in group B did not hear the lexical unit '*cerveja*' (beer) and have Pedro's face as their primary focal point too. The curiosity here is that some people in group B paid a lot of attention at the barman's face, who is in the background, but in the brightest region. Moreover, there are more fixations from group B in the tulip than in group A. This means that we cannot say that the lexical unit '*cerveja*' (beer) is playing a role of directing viewers' gaze.

<sup>26</sup> I asked for a Scottish beer, he said it's the best they have here in the pub.

Figure 35 – Heat map for the beer in the pub



Source: the author.

Further in this extract, Pedro explains how much the pubs are relevant to the Scottish culture and that, specifically, the pub he is visiting was where separatists had their meetings during the campaign for the 2014 Scottish independence referendum. Participants in group A get to know this by listening to Pedro saying sentences (8) and (9):

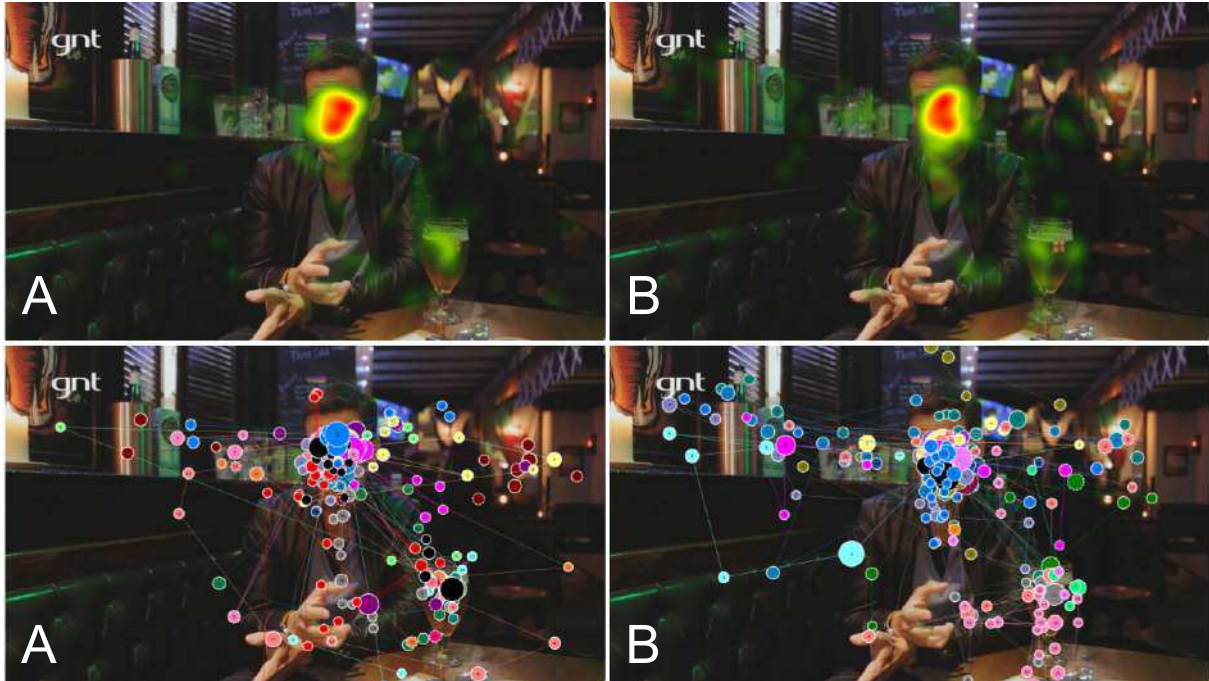
- (8) *Os escoceses, eles se encontram no pub mesmo.*<sup>27</sup>
- (9) *E esse pub onde eu estou é super famoso porque naquele momento onde a Escócia estava escolhendo se ela ia fazer parte, permanecer no Reino Unido ou ser independente, os separatistas se encontravam aqui.*<sup>28</sup>

Here we explored the possibility that, once Pedro talks about the pub, participants in group A would explore the ambient more than participants in group B who are not listening to him. The heat map in Figure 36, however, shows that fixations for both groups are very similar. Moreover, the gaze plot in Figure 36 also indicates that the behavior of participants in both groups is close, with some participants in group B exploring a little more on the left side of the screen.

<sup>27</sup> The Scots, they meet in the pub actually.

<sup>28</sup> And this pub that I'm in is super famous because at that time where Scotland was choosing whether it was going to be part of, remain in, or be independent, the separatists were meeting here.

Figure 36 – Heat map and gaze plot for the pub



Source: the author.

In extract 5 – Walking and whisky, viewers see a sequence of images of Pedro walking through the center of Edinburgh, contemplating some of the iconic urban landscapes. He says that the best way to get to know a city is on foot. And then pronounces sentence (10):

(10) *Andando pelas ruas de Edimburgo, é fácil ver algumas semelhanças com o interior da Inglaterra, mas o que não falta aqui é brilho próprio.*<sup>29</sup>

Figure 37 shows the heat maps for the shots participants in group A see while and just after hear sentence (10). What we evaluated here was if they pay more attention to the red phone booths and to the clock in the tower, once they could act in their process of meaning making as instances of English references – the clock in the tower making reference to the Big Ben. What the heat map shows is that in the case of the booths participants in both groups had spread fixations, mostly focused on the booths. Group A had fewer red spots and group B more red spots. But in the end, although different, the fixations work similar in terms of matching the booths. Thus, it is very hard to find any connection of these fixation patterns and mention of England in the audio. In the case of the clock in the tower, the heat map shows that both groups

<sup>29</sup> Walking through the streets of Edinburgh, it is easy to see some similarities with England's countryside, but what is not lacking here is its own brilliance.

have fixations strongly focused on the clock. Evaluating in detail, it is possible to say that group B, even without the sound cue, has the fixations a little more concentrated than group A.

Figure 37 – Heat maps for England similarities



Source: the author.

For extract 6 – Final remarks, we did not identify combinations of lexical units and visual objects that could apply as candidates for us to evaluate as examples of possible audio-dominance. Throughout this extract, Pedro makes his remarks about his visit to Edinburgh. Then the sentences have a reflexive character and do not look very anchored at images.

After analyzing the data collected with the experiment, we are then able to answer the guiding questions presented in the beginning of this chapter:

- (i) Are there regular and remarkable differences in points of fixation between the group exposed to the complete version and the group exposed to the modified version?

As demonstrated in the analysis of the extracts, heat maps and gaze plots show some differences between the fixations of participants in group A and B. However, differences do not occur in all the analyzed shots and, when differences occurred, regularity and remarkability are more light than strong features.

- (ii) If there are differences, can we associate the points of fixation of the group who watched the complete version with the audio guidance?

Considering the analysis, we cannot associate fixation with audio guidance because in many cases there is great similarity between fixations of both groups. Another reason is that there are some counterexamples in which participants in group B, without listening to a single lexical unit, have fixations associated with visual objects we previewed would be objects of attention of the listeners. Moreover, sequences where Pedro is speaking to the audience show increased fixation of participants in group A on his face, which may be an indication that phenomena typically located towards the pragmatic end of the Semantics-Pragmatics continuum – such as discourse organization and turn taking – may also play a role in gaze fixation.

For the refining of the annotation methodology these results and answers impacted in the way we understand the idea of audio-dominance. Despite our initial working hypothesis (see Figure 12) considers that the verbal content plays a controlling role in establishing meaning and organizing the video message, the experiment showed that in this TV show genre lexical units do not have a direct and absolute impact in guiding gaze. Therefore, we came to the conclusion that an adequate methodology for video annotation should not overweight the role of text by itself, but on the contrary consider it in context with image.

Hence, the results in the experiment led us to the proposition of two major annotation guidelines:

- (i) when annotating text using FrameNet Annotation WebTool for audiovisual corpora, annotators should always watch the video and see sentences in its multimodal context.
- (ii) in the same way, when annotating image using Charon's dynamic mode, annotators should always listen to the spoken audio and should also read its transcribed sentences made available in the video annotation workspace.

Following these statements still configures adopting the idea of a text-oriented annotation methodology. However, it is mandatory that, for a multimodal approach, text is always be considered in its multisemiotic context. We believe that this is the first impactful contribution of this work to FrameNet and to Frame Semantics theory.

Adopting a text-oriented annotation methodology is the best way of dealing with evocation, in the way Fillmore (1982) has defined it. Looking for the evocation patterns in text



and looking for correspondence and/or complementarity in visual objects makes frames instantiation tangible for the model. We do believe that a methodology for annotating image independently from text is possible to be developed, but (i) it would not be a multimodal approach for audiovisual corpora if it considers only visual objects – and not multiple modes in a visual media or support – and (ii) it would rely mostly on invocation (FILLMORE, 1985), instead of evocation, once it would not be anchored in lexical units.

Moreover, although the spoken audio dominance investigation experiment did not indicate a consistent position of dominance, it demonstrated that, most of the time, the points of fixation match the entities of interest for annotation, as we present and detail ahead, in section 5.3.

To the best of our knowledge, this is the first work that combines multimodal approach and Frame Semantics for video annotation, and so it will not encompass all possible ways of combining modes for meaning making. On the contrary, its purpose is to inspire and instigate the development of other possibilities, once the foundational methodology and tools are in place.

Next section presents the annotation tools and the detailed methodology proposed for annotating multimodal data in an audiovisual corpus.

## 5.2 FRAME-BASED ANNOTATION METHODOLOGY OF AUDIOVISUAL MULTIMODAL CORPORA

Frame-based multimodal annotation of audiovisual corpora is performed using two tools: the FrameNet Brasil WebTool (TORRENT et al., forthcoming) – for text annotation – and Charon (BELCAVELLO et al., 2022) – for image annotation. Adopting the text-oriented methodology, the annotation process starts with the FN-Br WebTool following FrameNet’s guidelines (RUPPENHOFER, 2016) for full-text annotation, as previously depicted in Figure 18. Going sentence by sentence in the corpus, annotators create Annotation Sets (AS) for each word for which there is a Lexical Unit in FrameNet – if annotators identify a word that is not a LU in the database, they can ask for the creation of the LU; once created, annotators can go back and complete the annotation.

For demonstrating the process, we present sentence (11), which belongs to “*Pedro pelo Mundo*” episode 2 – Iceland/Reykjavik:

(11) *Bom que aqui a gente bebe e vai esquentando, né?*<sup>30</sup>

Figure 38 demonstrates how the full-text annotation of sentence (11) is displayed on the FN-Br WebTool screen. In (11) the word forms *bom*, *aqui*, *bebe* and *esquentando*, highlighted in black in Figure 38, are the annotation targets. Note that, for each of them, there are three layers of annotation: Frame Element (FE), Grammatical Function (GF) and Phrase Type (PT). The column NI is used for indicating that core FEs are not instantiated in the sentence, but can be inferred. The frames *Desirability*, *Locative\_relation*, *Ingestion* and *Change\_of\_temperature* are indicated in the gray lines, in the left column. An annotation set is the combination of all the data related to a LU target.

Figure 38 – Full-text annotation example 2

[165820]	NI	Bom	que	aqui	a	gente	bebe	e	vai	esquentando,	né?
Desirability.bom.a		Bom	que	aqui	a	gente	bebe	e	vai	esquentando,	né?
FE	INI						Evaluatee				
GF							Dep				
PT							Sfin				
Locative_relation.aqui.adv		Bom	que	aqui	a	gente	bebe	e	vai	esquentando,	né?
FE	INC						Figure				
GF							Ext				
PT							NP				
Ingestion.beber.v		Bom	que	aqui	a	gente	bebe	e	vai	esquentando,	né?
FE	INI						Ingesto				
GF							Ext				
PT							NP				
Change_of_temperature.esquentar.v		Bom	que	aqui	a	gente	bebe	e	vai	esquentando,	né?
FE	CNI INI										
GF											
PT											

Source: [webtool.framenetbr.ufjf.br](http://webtool.framenetbr.ufjf.br)

In the methodology proposed here, there are two possibilities to carry out the full-text annotation of a corpus: (i) annotators exhaust the sentences of a corpus first, and then start annotating images; or (ii) annotators complete the annotation of the sentences that correspond to a sequence<sup>31</sup>, then annotate image in the respective sequence, and go back to the sentences of the following sequence.

To annotate the image of the correspondent sequence in the video, annotators use Charon's dynamic mode. We should remember that, as previously mentioned and also pointed

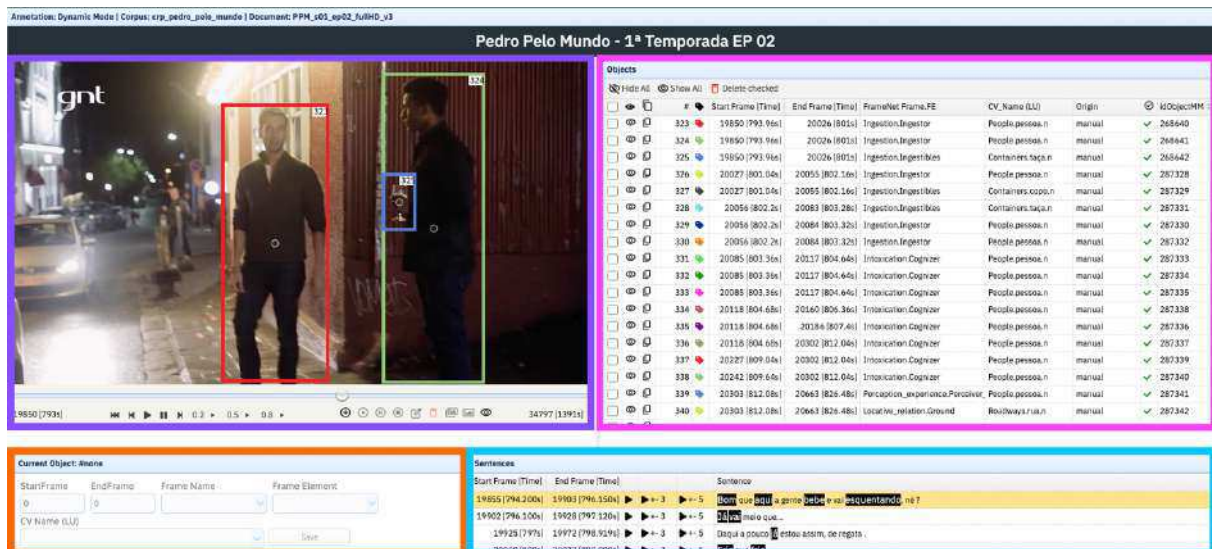
<sup>30</sup> It is good that here we drink and warm ourselves up, innit?

<sup>31</sup> 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. For instance, sentence (11) is part of the 'Reykjavik's night life' sequence that corresponds to 4 minutes and 21 seconds and is the last part of the second block of the episode.

out by Belcavello et. al (2022), Charon provides a myriad of possibilities for video annotation by human users, in terms of both methodologies and goals. So far, using the text-oriented methodology, Charon’s dynamic mode has been used to annotate and compare semantic frames evoked by visual objects with those evoked by LUs in sentences. This is why the dynamic annotation module features not only the annotation tools for tagging images, but also the visualization of the sentences annotated in the FN-Br WebTool for the same corpus.

Figure 39 shows Charon’s dynamic mode annotation workspace. Highlighted in purple is the ‘*video panel*’ in which annotators control video playback, draw and edit bounding boxes. Below it, highlighted in orange, is the ‘*annotation panel*’ in which annotators see the indication of start and end video frames of the object, associate a Frame with the object, select the Frame Element and indicate a Computer Vision name for the object. Moving right, highlighted in pink, is the ‘*objects panel*’ which presents the list of objects created – both manually and automatically – and its associated data. In the bottom right corner is the ‘*sentences panel*’, which shows the sentences annotated in the FN-Br WebTool, associated with its timestamps and playback controls for the ‘*video panel*’ to visualize the sentences in action in the episode.

Figure 39 – Charon’s dynamic mode annotation workspace



Source: charon.frame.net.br

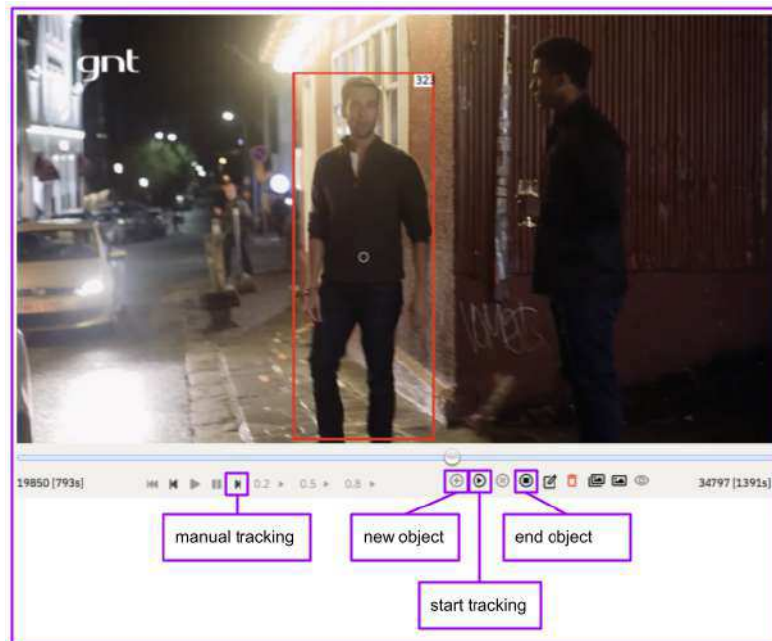
The image annotation proposed here refers to the selection of part of the screen by using a bounding box understanding this selection as a correspondent visual demonstration of a Frame Element in a Frame. In this sense, a *visual object* is defined as a set of bounding boxes in a time

interval that is associated with a Frame Element. For instance, in Figure 39 looking at the *video panel* and at the *objects panel*, object 323 – highlighted in red – stores the information that:

- (i) that portion of the image refers to the INGESTOR FE in the Ingestion frame.
- (ii) the bounding box list starts at frame 19850 – which is also correspondent to second 793.96 – and ends at frame 20026 – second 801.
- (iii) it is also associated to the LU *pessoa.n* (person) in the People frame for the Computer Vision Name (CV Name) categorization<sup>32</sup>.

To create a new object, in the ‘*video panel*’, (see Figure 40) annotators use the new object button, draw the bounding box over the object they want to detect, then start tracking it. Tracking can be executed manually, frame by frame, or automatically, using the start tracking button. In both cases, annotators determine the end point for the bounding box in the end object button, when the object is not visible anymore or there is a cut (or any editing transition) ending that shot – hence, each object has its maximum duration limited by the shot duration.

Figure 40 – Video panel detail for creating new objects



Source: charon.frame.net.br

Next, annotators must attribute a Semantic Frame and a FE to the object, in the ‘*annotation panel*’ (see Figure 41). They choose the frame from the list under the *Frame Name*

<sup>32</sup> The CV Name categorization was created for matching the object with pre-trained computer vision categories. At first, the CV Name field is to be filled in from the automatic labeling of the visual objects using The Open Images Dataset v6 classes (<https://github.com/DmitryRyumin/OIDv6/blob/master/oidv6/classes.txt>).

field. Once the frame is chosen, a list of its FEs is loaded in the *Frame Element* field. Annotators should also assign a Computer Vision name to the object. This category associates one LU with the object, considering its value as an entity recognizable by computer vision tools or algorithms. In the *CV Name (LU)* field, users may choose from any LU in the FrameNet database they are using. By using FrameNet LUs as categories similar to the ones used in benchmark multimodal datasets such as MS COCO (LIN et al., 2014) and Open Images (KUZNETSOVA et al., 2020), the annotation methodology devised here adds yet another layer of density to the resulting dataset, since the frames evoked by such LUs, their relations with other frames and other types of structure in the enriched FrameNet Brasil database also become associated with the bounding box.

Figure 41 – Annotation panel detail for labeling new objects

StartFrame	EndFrame	Frame Name	Frame Element
19850	20026	Ingestion	Ingestor

CV Name (LU)  
People.pessoa.n

Save

Source: charon.frame.net.br

The multimodal text-oriented approach for this annotation can be explained as follows. When looking for correspondences between text and image, objects 323 and 324 (Figure 39) were annotated as the *INGESTORS* for the *Ingestion* frame. On the other hand, as what is visually recognizable are two human figures, the CV Name chosen for each object was *pessoa.n* (person.n) in the *People* frame. Object 325 was annotated as the *INGESTIBLES* in the *Ingestion* frame and as *taça.n* (glass.n) in the *Container* frame for the CV Name. What is interesting here is that in sentence (11) – Figure 38 – Full-text annotation example 2 – the *INGESTIBLES* FE is not instantiated by any of the sentence’s components in the annotation of the *Ingestion* frame for *bebe.v* (drink.v) as target. Actually, the *INGESTIBLES* FE is annotated as a null instantiation (INI), once the sentence does not point out what the *INGESTORS* are having. By looking at the image, however, it is empirically noticeable that the *INGESTIBLES* mentioned is what fills the container – the glass – one of the human figures holds in his hands. Although there is no Lexical Unit evoking the *Container* Frame, it is a straightforward cognitive task to identify the container in the image. Thus, the annotator can add this information to the database by using the CV Name field in the ‘*annotation panel*’. Therefore,

this example shows how meaning layers and granularity can be added to the FrameNet semantic representation by annotating visual data in correspondence with textual data in a corpus.

We have also developed a semi-automatic annotation process, as it was mentioned in the description of the Multimodal Corpus Import Pipeline – Figure 23 – Multimodal Corpus Import Pipeline. In that case, bounding boxes with the duration of one video frame would be automatically drawn and CV Name labels would be automatically assigned. The human annotation process would start with reviewing the objects automatically detected by the computer vision software. If annotators agreed with a bounding box automatically generated, they would first select the object in the ‘*objects panel*’, then use the edit tracking button in the ‘*video panel*’ to finish the creation of the bounding box set. After that the process is manual and proceeds just as it was described previously. If annotators do not agree with the bounding box automatically generated, they can select the object in the ‘*objects panel*’ and delete it. Similarly, if they agree with the CV Name label assigned, they let it stay. If not, they would change it. So far, the automatic recognition and bounding box creation is running using YOLOv6<sup>33</sup> (LI, 2022) trained on the MS COCO data set<sup>34</sup> (LIN et al., 2014). During the annotation task developed for this dissertation, we considered that the results of the automatic annotation were not satisfactory. The detected objects often did not correspond to the objects one wanted to annotate both in terms of their position on the screen and in terms of the labels assigned to the CV Name. Furthermore, the fact that in our implementation, so far, the generated objects have a duration of only one frame makes the amount of manual work still very large. In other words, semi-automation did not translate into effort saving as an advantage and proved limited in terms of identifiable categories. We believe that after we train a model with the categories we are using manually, the automatization process can be improved and then taken as the primary way of proceeding with annotation.

On top of the two general annotation guidelines stated at the beginning of this section, the following annotation methodology guidelines were also proposed:

- 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.
- 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.

---

<sup>33</sup> <https://github.com/meituan/YOLOv6> .

<sup>34</sup> <https://cocodataset.org/> .

- One visual object can be duplicated as many times as necessary, if it instantiates different FEs – both in one same frame or in different frames.
- The limit of asynchrony 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 or ahead the LU pronunciation, if they are both located within the same video sequence and/or if there is not a better connection with a closer LU.
- 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.
- CV Names should always be chosen taking the most empirical and concrete LU to designate what is seen on screen.

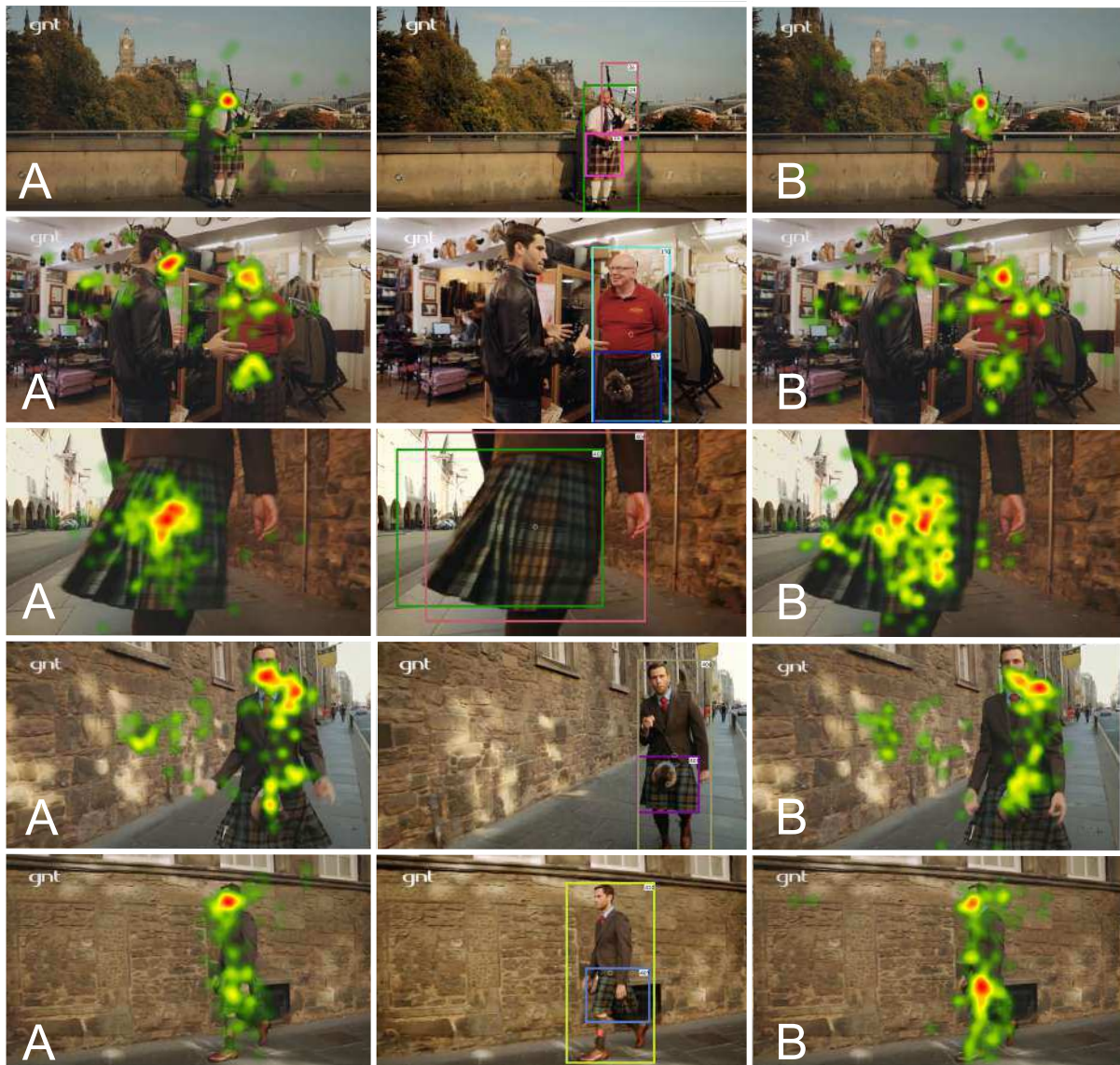
### 5.3 CONTRASTING ANNOTATIONS AND FIXATIONS

As previously mentioned, the results of the spoken audio dominance investigation did not indicate that lexical units, phrases, or sentences are determinant to define what were the participants’ points or areas of gaze fixation. However, the results did identify which regions of each image in each sequence attracted more attention from the participants in both experimental conditions. Therefore, in general, it was not a great challenge to infer entities on screen candidates to be the target of annotations, but this data was kept apart during the annotation task as a way of avoiding biases in the annotators’ interpretation of the corpus. Nonetheless, it was later revisited to analyze the annotated data and evaluate the task. The ReINVenTA methodology proposes that the annotated datasets are submitted to psycholinguistic validation of the bounding boxes after the annotation is concluded. The episodes of the first season of “*Pedro pelo Mundo*” will go through this protocol soon and the results will be published.

Heat maps and gaze plots show some consistency in terms of concentrating fixations. It happens even in the case of shots in which the results were more spread. When contrasted with the annotations, it is usually possible see with reasonable clearness that the areas of interest for the viewers match those where bounding boxes were drawn and, as a consequence, there is a visual object. presents a sample of comparison between heat maps and bounding boxes. It shows the heat map of points of fixation for group A – control – of the experiment in the left column; the bounding boxes of objects created during image annotation in the central column,

and the heat map of points of fixation for group B – experimental – of the experiment in the right column.

Figure 42 – Comparative shot board of image annotation and gaze fixations



Source: the author.

The shot of the Scot bagpiper wearing a kilt is depicted in the first line of the board in . As previously pointed out, both heat maps show the most relevant concentration of fixations in the man's face. The bounding boxes for the correspondent shot refer to the man, the kilt and the bagpipe. Therefore, the region of the screen with most fixations is also a region where there are bounding boxes.

The second line shows the encounter between Pedro and Gordon Nicolson. The bounding boxes created objects correspondent to Nicolson and his kilt. The heat map of group



A's fixations shows one red spot at Pedro's face and an orange spot at Nicolson's face. Here there is some difference for group B's heat map in which the major concentration of gaze is at Nicolson's face. There is, then, a match at Nicolson, but not at Pedro. Because Pedro is the host, annotators tend not to tag him when he is presenting something or someone in the video, since the audio in those fragments is composed of sequences where the elements being presented are described by Pedro. On the other hand, because he is the one talking in the scene, viewers – most prominently those in group A – tend to concentrate the gaze on his face.

The third line looks very straightforward once the kilt fills much of the screen. There is no surprise, then, that both heatmaps and bounding boxes refer to it. One detail is that the pink bound box annotates the user of the garment, not the garment itself. But they are mostly overlapping.

Both the fourth and the fifth lines show shots of Pedro walking in a sidewalk wearing a full Scottish costume. In both cases heatmaps of both groups show concentration of gaze over Pedro's face and garments. There is just a little more concentration at the kilt for group B in the last line. Either way, the bounding boxes drawn by the annotators in both cases match Pedro and the kilt.

The analysis of these samples indicates that the annotations performed following the guidelines established by the methodology account for the detection of objects relevant to the audience. It is important to remember that, although we have highlighted in this section the visual aspects of both annotation and eye-tracking, these bounding boxes were drawn taking into account the spoken audio in the respective sequences. This is what denotes the multimodal character of the proposed approach. That is, to determine which entities, areas or visual objects would be worthy of annotation, the spoken audio played an active role, standing out as a balancing parameter in the interpretation of the meaning produced by its combination with the image.

By placing the combination between the semiotic modes and their acting together more at the center of the question, we understand that we set out for a model that moves somewhat away from the initial approaches of VNG (COHN, 2013, 2016a, 2019) and FNG (COHN, 2016b). Despite the important influence of these models for the maturation of our approach on narrative sequences, we believe it was necessary to abandon the notions of dominance and assertiveness and opt for a model that was structured less in a parallel architecture and more in an intersectional architecture. This means looking more at: how to establish modellable links between what is heard and what is seen? How to represent the link between Lexical Units and

Visual Objects? How to account for the intrinsicity with which one semiotic mode affects the other in terms of meaning making?

In that sense, we believe that the FrameNet Brasil model enriched by intramodal relations and multimodal data aligns more with the principles stated by Bateman, Wildfeuer and Hiippala (2017) on developing a methodological approach that accounts not only for each semiotic mode, but primarily for their interactions.

Therefore, it can be stated at this point of the work that, even if it is necessary to execute separately the annotation of the spoken audio – in the form of transcribed text – and the image, the interpretation of meaning that guides the two annotations must derive from the multimodal perception of the combination between what is heard and what is seen. This was, precisely, the foundation that drove the process of building the Frame<sup>2</sup> dataset, which we present and discuss in the next chapter.

## 6 THE FRAME<sup>2</sup> DATASET

The Frame<sup>2</sup> dataset is composed by the multimodal objects, i.e., the annotated data, both for text and image, in the “*Pedro pelo Mundo*” corpus and the relations between the annotated data as mediated by the structure modeled in the FrameNet Brasil database. It was built to serve as a gold standard dataset for a variety of NLP based tasks.<sup>35</sup> It brings 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, forming a video. It offers data that reflects audio and video combination possibilities in terms of frames, as in the example shown in Figures Figure 38, Figure 39, Figure 40 and Figure 41. The first data release of Frame<sup>2</sup> will comprise the annotation of all 10 episodes of the show’s first season. This means 11,796 annotation sets (AS) for text and 6,841 visual objects (VO) for image. Table 1 shows the annotation totals.

Table 1 – Corpus annotation totals

EPISODE	ANNOTATION SETS	SENTENCES	VISUAL OBJECTS
<b>01 – Cairo</b>	1164	226	593
<b>02 – Reykjavik</b>	890	205	805
<b>03 – Atenas</b>	1029	208	638
<b>04 – Myanmar</b>	1011	199	562
<b>05 – Copenhagen</b>	1385	248	657
<b>06 – Edinburgh</b>	1191	226	503
<b>07 – Havana</b>	1087	218	698
<b>08 – Seattle</b>	1373	227	545
<b>09 – Singapore</b>	1403	215	779
<b>10 – Oman</b>	1263	223	1061
<b>TOTAL</b>	<b>11,796</b>	<b>2,195</b>	<b>6,841</b>

Source: the author.

These totals are very close to the numbers estimated after conducting the pilot study (BELCAVELLO, 2020) of 12,000 annotation sets and 2,000 sentences. The variations were 1,7% less annotation sets and 9.75% more sentences. For the visual objects, however, the result

<sup>35</sup> The notion of gold standard dataset should be regarded here with caution. As it is the case for every dataset produced within the scope of the ReINVenTA initiative, the annotations in the Frame<sup>2</sup> dataset are meant to represent possible perspectives on the meaning construction processes that may be triggered by the combination of different communicative modes. Therefore, more than one set of annotations is possible and even different annotation methodologies can be proposed so as 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** (<http://pdai.info/>) and Basile et al. (2021).

of 6,841 VOs compared to the estimative was of 5,000 (BELCAVELLO et al., 2022; TORRENT et al., 2022) represents an increase of 36.82%.

The average of annotations per sentence ranged from 4.34 – the lowest value – to 6.53 – the highest value – as shown in Table 2.

Table 2 – Corpus annotation averages

EPISODE	ANNOTATION SETS	SENTENCES	ANNOTATION AVERAGE
<b>01 – Cairo</b>	1164	226	5.1504
<b>02 – Reykjavik</b>	890	205	4.3415
<b>03 – Atenas</b>	1029	208	4.9471
<b>04 – Myanmar</b>	1011	199	5.0804
<b>05 – Copenhagen</b>	1385	248	5.5847
<b>06 – Edinburgh</b>	1191	226	5.2699
<b>07 – Havana</b>	1087	218	4.9862
<b>08 – Seattle</b>	1373	227	6.0485
<b>09 – Singapore</b>	1403	215	6.5256
<b>10 – Oman</b>	1263	223	5.6637
<b>TOTAL/AVERAGE</b>	<b>11,796</b>	<b>2,195</b>	<b>5.3598</b>

Source: the author.

Table 2 also informs that the corpus average of annotation sets per sentence was 5.3598. This result is below the FN-Br full-text annotation average of 6.1 AS per sentence (BELCAVELLO et al., 2020). We can empirically associate this 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 amount of greetings and other more pragmatic level operators that are not yet covered by FrameNet frames.

Concerning the variability of the corpus, Table 3 shows how many discrete frames were used in each episode and in the corpus.

Table 3 – Numbers of discrete frames annotated

EPISODE	DISCRETE FRAMES IN TEXT	DISCRETE FRAMES IN IMAGE	DISCRETE FRAMES IN IMAGE – CV NAME
<b>01 – Cairo</b>	279	163	42
<b>02 – Reykjavik</b>	256	91	29
<b>03 – Atenas</b>	243	110	55
<b>04 – Myanmar</b>	257	89	31
<b>05 – Copenhagen</b>	284	103	33
<b>06 – Edinburgh</b>	278	110	30
<b>07 – Havana</b>	265	123	24
<b>08 – Seattle</b>	298	141	39
<b>09 – Singapore</b>	291	106	28
<b>10 – Oman</b>	292	136	51
<b>THE CORPUS</b>	<b>611</b>	<b>393</b>	<b>129</b>

Source: the author.

The numbers in Table 3 are robust in showing that the rate of variability in the use of frames in textual annotation is much higher than the rate in the use of frames evoked by visual objects and the ones evoked by LUs annotated as CV Name. 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 Video Object discrete frame per AS discrete frame is 0.64 – higher than the VOs/ASs value –, while the ratio of CV Name discrete frame per AS discrete frame is 0.21 – much lower than the VOs/ASs value. Finding the explanation for this difference may be a goal for future research, but we have empirical elements to believe that it may be related to the predominance of entities annotated for CV Name and also to the high rate of repetition of some frames during annotation.

Table 4 presents the discrete number of LUs used as CV Name per episode.

Table 4 – Discrete LUs as CV Name per episode

<b>EPISODE</b>	<b>DISCRETE LUs USED AS CV Name</b>
<b>01 – Cairo</b>	91
<b>02 – Reykjavik</b>	88
<b>03 – Atenas</b>	93
<b>04 – Myanmar</b>	52
<b>05 – Copenhagen</b>	73
<b>06 – Edinburgh</b>	55
<b>07 – Havana</b>	53
<b>08 – Seattle</b>	82
<b>09 – Singapore</b>	49
<b>10 – Oman</b>	81
<b>THE CORPUS</b>	<b>478</b>

Source: the author.

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 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<sup>36</sup> (KUSNETZOVA et al., 2020), but with the difference that the Frame<sup>2</sup>'s classes are not hierarchized, but organized in a more complex net of concepts that is FrameNet, using 129 different frames, as presented in Table 3.

Table 5 presents another aspect that accounts for improved granularity of the Frame<sup>2</sup> dataset: the matching ratio of only 1.61 between the frames used for the Visual Object

<sup>36</sup> [https://storage.googleapis.com/openimages/web/factsfigures\\_v7.html#class-definitions](https://storage.googleapis.com/openimages/web/factsfigures_v7.html#class-definitions)

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 net.

Table 5 – Matching ratio of frames annotated for images

<b>EPISODE</b>	<b>VO FRAME TO CV NAME FRAME MATCHING RATIO</b>
<b>01 – Cairo</b>	3.54
<b>02 – Reykjavik</b>	2.88
<b>03 – Atenas</b>	6.76
<b>04 – Myanmar</b>	4.92
<b>05 – Copenhagen</b>	4.42
<b>06 – Edinburgh</b>	4.14
<b>07 – Havana</b>	2.9
<b>08 – Seattle</b>	5.32
<b>09 – Singapore</b>	2.58
<b>10 – Oman</b>	3.28
<b>THE CORPUS</b>	<b>1.61</b>

Source: the author.

Because the annotation of video sequences in correlation with linguistic data adds a whole new dimension of meaning construction possibilities, the Frame<sup>2</sup> dataset was planned as a domain-specific dataset for the Tourism domain. The modeling of this domain in the FN-Br database counts with all the additional dimensions described in section 3.2.

Next, we present some data samples and discuss how they add granularity to the model.

## 6.1 IMAGE SPECIFIES TEXT

Sentence (1) – reproduced as (12) –, from episode 6 – Edinburgh, is a good example of data in which image specifies text, and, so, the annotation of a visual object specifies the annotation of a lexical unit, adding a layer of meaning that can only be perceived multimodally. It was previously explored in the pilot study (BELCAVELLO, 2020). The full annotation of (12) yielded ten lexical annotation sets, whose targets are highlighted in black and respective frames annotated are superscripted next to them.

- (12) Quando<sup>Temporal\_collocation</sup> a gente pensa<sup>Cogitation</sup> na Escócia, a primeira<sup>Ordinal\_numbers</sup> coisa<sup>Entity</sup> que vem à mente<sup>Cogitation</sup> é: homem<sup>People</sup> de saia<sup>Clothing</sup>, uísque<sup>Food</sup> escocês<sup>Origin</sup> e gaita de fole<sup>Musical\_instruments</sup>.<sup>37</sup>

Following the guidelines for the text-oriented frame-based multimodal methodology annotation of images, at the shot presented in Figure 43, the annotator chose the *People\_by\_origin* frame for the visual object 24 – green bounding box. This VO refers to the *homem.n* (man.n) LU in sentence (12), which was annotated for the *People* frame.

Figure 43 – Image annotation for *People\_by\_origin*



#	📌	Start Frame [Time]	End Frame [Time]	FrameNet Frame.FE	CV_Name (LU)
24	🟢	911 [36.4s]	936 [37.4s]	People_by_origin.Person	People_by_origin.escocês.n
25	🟡	911 [36.4s]	936 [37.4s]	Clothing.Garment	Clothing.kilt.n
26	🔴	911 [36.4s]	936 [37.4s]	Musical_instruments.Musical_instrument	Musical_instruments.gaita de fole.n

Source: charon.frame.net.br

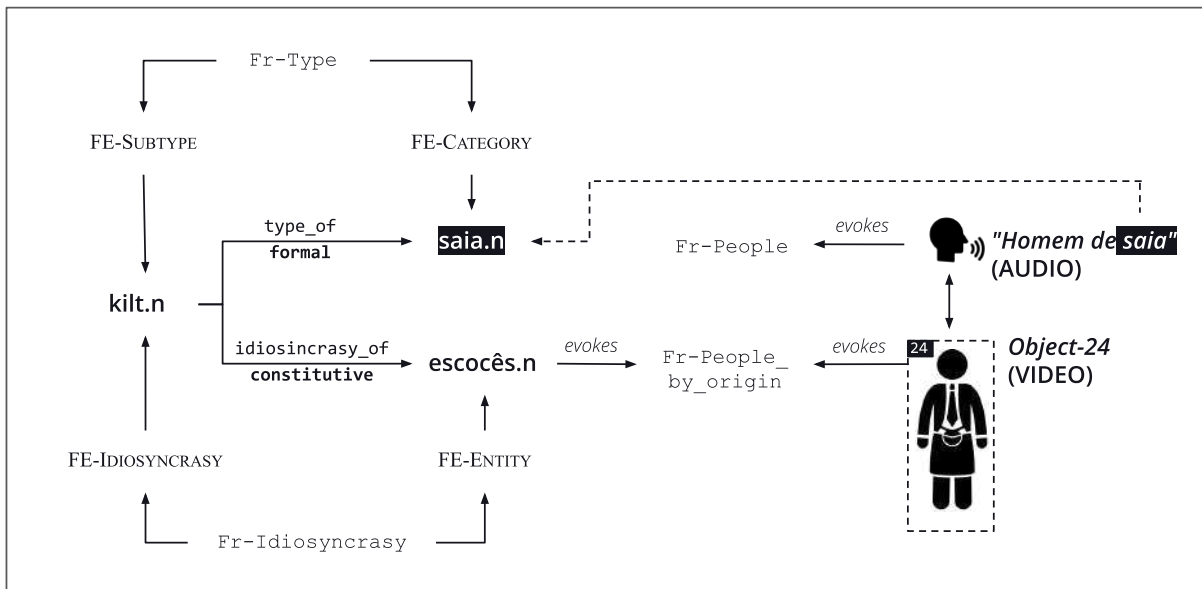
The reason behind this choice is the fact that the man depicted in the video right after the spoken audio mentions ‘*homem de saia*’ (man in skirt) is wearing a kilt and playing a bagpipe, which are typical clothing and musical instrument of Scotland, respectively. This combination of factors makes it very likely to infer that what we see is a Scot person. Therefore,

<sup>37</sup> When<sup>Temporal\_collocation</sup> we think<sup>Cogitation</sup> of Scotland, the first<sup>Ordinal\_numbers</sup> things<sup>Entity</sup> that come to mind<sup>Cogitation</sup> are: man<sup>People</sup> in skirts<sup>Clothing</sup>, Scotch<sup>Origin</sup> whisky<sup>Food</sup> and bagpipe<sup>Musical\_instruments</sup>.

it makes possible for the annotator to choose the `People_by_origin` frame instead of the `People` frame. The CV Name reinforces this possibility by attaching the LU `escocês.n` in the same frame `People_by_origin` to the Object 24.

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. Frame-mediated ternary qualia relations provide the answer to both. Figure 44 shows a schematic representation of the relations in action.

Figure 44 – Frame-mediated ternary qualia relations for *homem de saia.n* and object 24



Source: the author.

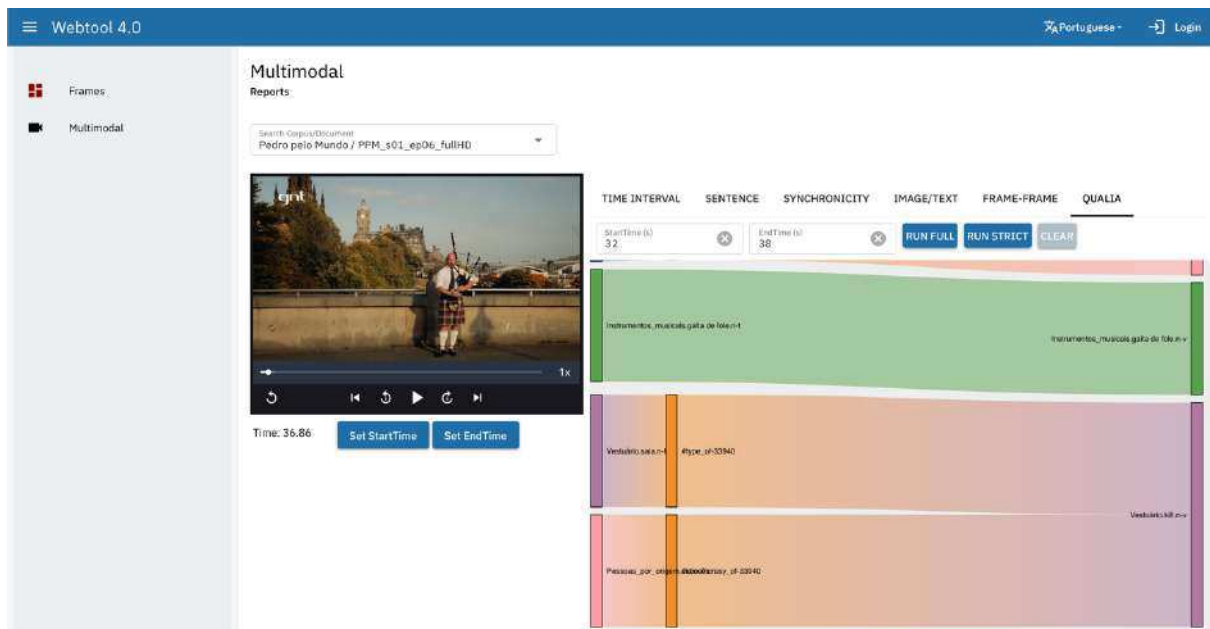
First, a subtype of the **formal** quale, mediated by the `Type` frame connects the LUs `kilt.n` and `saia.n` (skirt) in FrameNet Brasil database. Second, a subtype of the **constitutive** quale mediated by the `Idiosyncrasy` frame connects the LU `kilt.n`, instantiating the FE `IDIOSYNCRASY` to the LU `escocês.n` (in Portuguese it is the same word form for Scottish.a for the whisky and `Scot.n` for the person)', 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 24 (as shown in Figure 43).

Those connections between the modes annotated, as mediated by the semantic structure modeled in the FrameNet Brasil database, can be seen in the Sankey diagram generated by the FN-Br WebTool Report Module devised for the exploring the dataset and depicted in Figure



45. The purple bar on the left of the diagram represents the LU *saia.n* in the *Clothing* frame. The purple bar on the right of the diagram represents the LU *kilt.n* also in the *Clothing* frame. These LUs are connected by the orange bar between the other bars, which represents the **formal** quale mediated by the *Type* frame, and establishes that *saia.n* is a *type\_of* *kilt.n*. The diagram also shows that the *kilt.n*'s purple bar on the right is connected by another orange bar in the center with the pink bar on the left. The pink bar represents the LU *escocês.n* and the orange bar represents the **constitutive** quale mediated by the *Idiosyncrasy* frame and establishes that *kilt.n* is an *idiosyncrasy\_of* *escocês.n*.

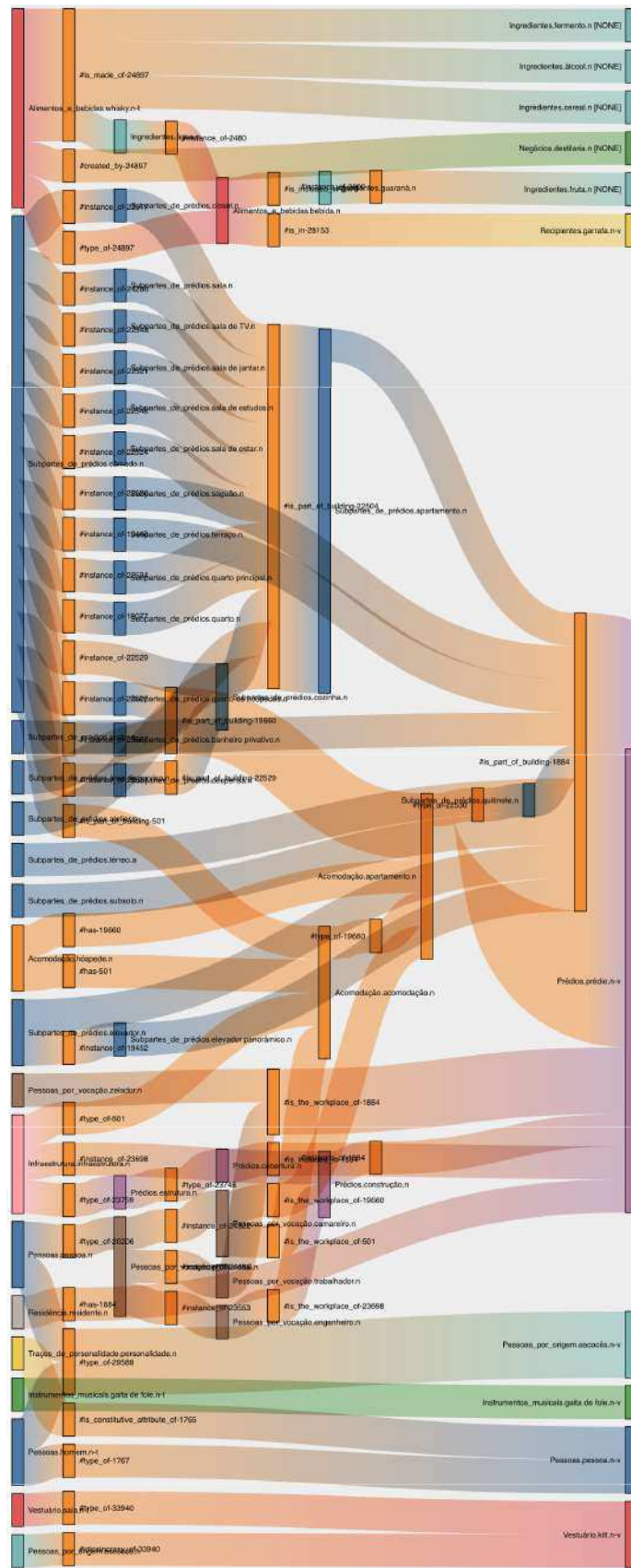
Figure 45 – Sankey diagram for qualia relation between *saia.n* and *kilt.n* at FN-Br WebTool Report Module



Source: <http://webtool.frame.net.br/reportMultimodal>.

Figure 45 presents a portion of the Strict Qualia Report, which means that it informs relations established only between LUs that are annotated in the text and LUs that are annotated in the image in the specified time interval. The Full Qualia Report, on the other hand, presents relations of the LUs annotated with all the others non-annotated to which each one has a relation. This functionality was designed with the purpose of enhancing the visualization of other potential qualia relations and, therefore, inferences, to be established from the annotated LUs. For instance, Figure 46 presents the Sankey diagram for potential qualia relations established from each LU annotated in text or in image throughout the whole time span – 32 seconds to 43.592 seconds – correspondent to the appearance of sentence (2) in the video.

Figure 46 – Sankey diagram for potential qualia relations derive from LUs in a time interval



Source: <http://webtool.frame.net.br/reportMultimodal>.

FN-Br WebTool Report Module allows not only to look for qualia relations, but to explore other types of relations established between LUs, frames, text-image etc. In Figure 47, the menu ‘Time interval’ presents two lists of the LUs annotated in the time interval set: (i) the list of LUs annotated for text in sentences that are totally or partially circumscribed in the interval, and (ii) the list of LUs annotated as CV Names for Visual Objects that occur on screen totally or partially circumscribed in the time interval. Occurrences are presented at maximum 5 records per page. For the interval correspondent to sentence (2) the report presented 18 LUs annotated for text and 25 LUs annotated for image.

Figure 47 – Report example of the 'Time interval' menu in the FN-Br WebTool Report Module

The screenshot displays the FN-Br WebTool Report Module interface. On the left is a video player showing a man walking on a path. Below the video, the current time is 43.588, and there are buttons for 'Set StartTime' and 'Set EndTime'. On the right, the 'TIME INTERVAL' menu is selected, showing a start time of 32 and an end time of 43.592, with a 'RUN' button. Below this, there are two tables of LU annotations.

**Textual annotation table:**

LU	Frame	FE
Clothing.saia.n	Clothing	Garment
Clothing.saia.n	Clothing	Wearer
Cogitation.pensar.v	Cogitation	Cognizer
Cogitation.pensar.v	Cogitation	Topic
Cogitation.vir à mente.v	Cogitation	Topic

Records per page: 5 | 1-5 of 18 | < > >>

**Video annotation: from [32] to [43.592]**

LU	Frame	FE
Clothing.kilt.n	Clothing	Garment
Containers.garrafa.n	Cogitation	Topic
Containers.garrafa.n	Cogitation	Topic
Containers.garrafa.n	Cogitation	Topic
Containers.garrafa.n	Cogitation	Topic

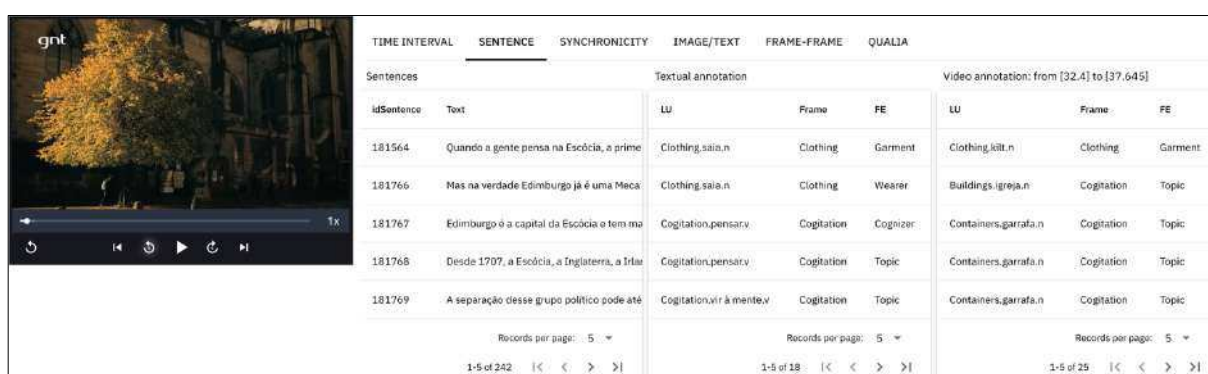
Records per page: 5 | 1-5 of 25 | < > >>

Source: <http://webtool.frame.net.br/reportMultimodal>.

The ‘Sentence menu’ – depicted in Figure 48– presents the lists of LUs annotated for both text and image in the time span of the presence of a sentence in a video. This feature allows for the analysis of synchronicity, taking a sentence as the referent point. Thus, for the text, the list will be organized by the target LUs in the annotation, while for the images, the LUs listed will be the ones annotated as CV Name for visual objects that occur fully circumscribed into

the sentence time interval or that occur partially in the sentence time interval – the object starts before the sentence starts, but ends while the sentence is still being said; or the object starts after the sentence starts, while the sentence is still being said, and ends after it; or the object starts before the beginning of the sentence and ends after the sentence; or the object starts and ends within the sentence appearance interval. The table on the left shows the list of sentences that form the episode, each of them specified by its number of identification in the database – the ‘idSentence’. The central table shows the list of LUs with its correspondent frame and FE in the same line. The table on the right shows the list of LUs annotated for the image in the sentence with its correspondent frame and FE in the same line.

Figure 48 – Report example of the 'Sentence' menu in the FN-Br WebTool Report Module



Sentences		Textual annotation			Video annotation: from [32.4] to [37.645]		
idSentence	Text	LU	Frame	FE	LU	Frame	FE
181564	Quando a gente pensa na Escócia, a prime	Clothing,saia.n	Clothing	Garment	Clothing,kilt.n	Clothing	Garment
181766	Mas na verdade Edimburgo já é uma Meca	Clothing,saia.n	Clothing	Wearer	Buildings,igreja.n	Cogitation	Topic
181767	Edimburgo é a capital da Escócia e tem mo	Cogitation,pensar.v	Cogitation	Cogizer	Containers,garrafa.n	Cogitation	Topic
181768	Desde 1707, a Escócia, a Inglaterra, a Ir	Cogitation,pensar.v	Cogitation	Topic	Containers,garrafa.n	Cogitation	Topic
181769	A separação desse grupo político pode até	Cogitation,vir à mente.v	Cogitation	Topic	Containers,garrafa.n	Cogitation	Topic

Source: <http://webtool.frame.net.br/reportMultimodal>.

The ‘Synchronicity menu’ reports annotations connected to a specific video frame. The LU list of annotations shows the LUs annotated as CV Names that occur in the specific video frame set as the ‘Start time’. The LU list of textual annotation retrieves the LUs annotated in the sentence detected as occurring in that specific video frame. As an example, Figure 49 presents the lists of annotations – 22 LUs for text and 3 LUs for image – related to the video frame located at second 38.588 of Episode 6 – Edinburgh. The list also shows for each LU the respective frame and the Frame Element instantiated in the text annotated or in the visual object.

Figure 49 – Report example of the 'Synchronicity menu' in the FN-Br WebTool Report Module

The screenshot displays the FN-Br WebTool Report Module interface. On the left, a video player shows a man playing a bagpipe. Below the video, the time is set to 38.588, and there is a 'Set StartTime' button. The main area on the right is titled 'SYNCHRONICITY' and contains a table of annotations. The table is organized into two sections: 'Textual annotation' and 'Video annotation'. Each section has columns for 'LU' (Lexical Unit), 'Frame', and 'FE' (Frame Element). The 'Textual annotation' section lists several LU-Frame-FE relationships, such as 'Calendric\_unit.século.n' linked to 'Calendric\_unit' and 'Name'. The 'Video annotation' section lists LU-Frame-FE relationships for visual objects, such as 'Noise\_makers.gaita de fole.n' linked to 'Cogitation' and 'Topic'. At the bottom of the table, there are pagination controls showing 'Records per page: 5' and '1-5 of 22'.

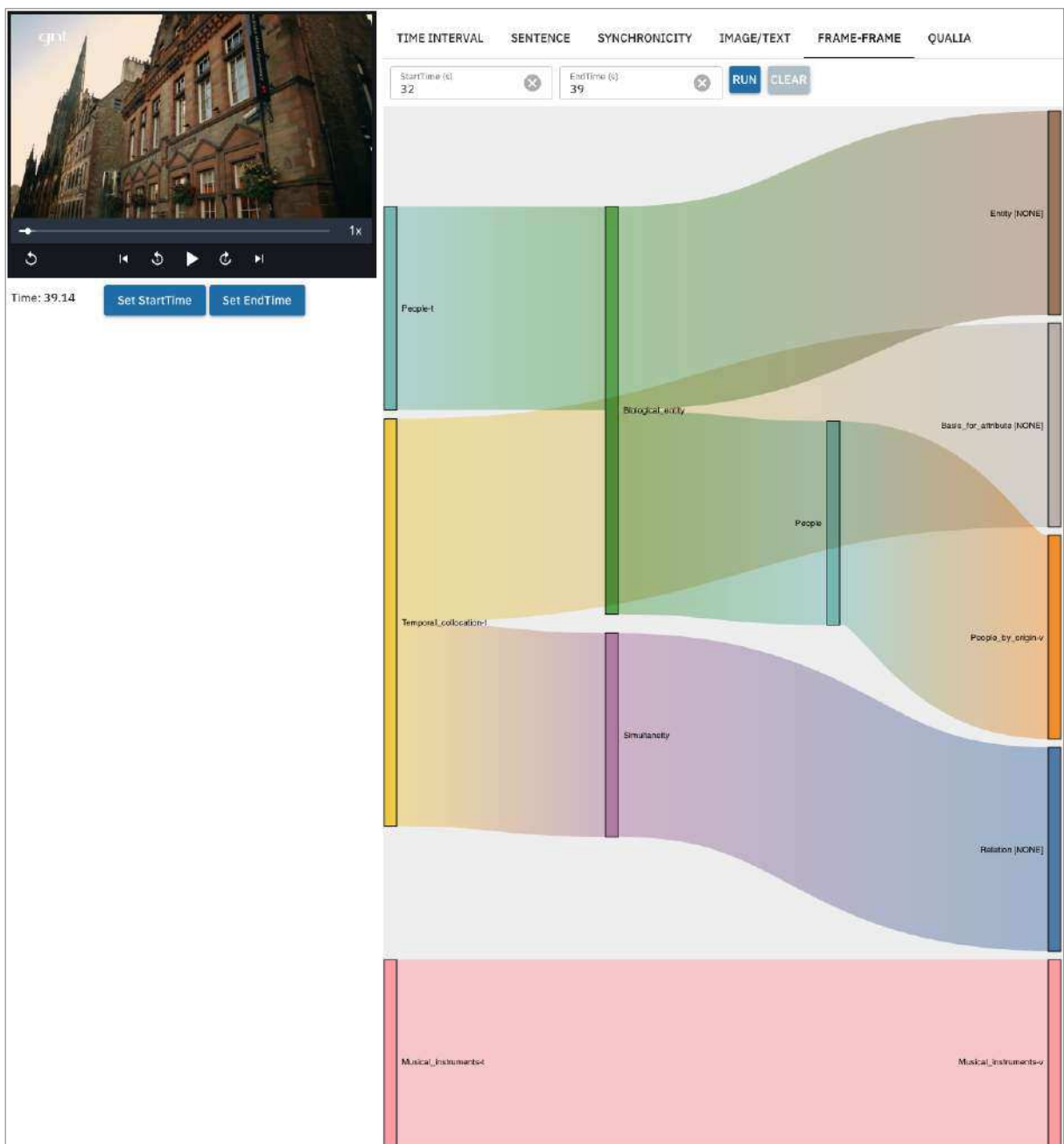
LU	Frame	FE
Calendric_unit.século.n	Calendric_unit	Name
Calendric_unit.século.n	Calendric_unit	Unit
Concessive.mas.c	Concessive	Conceded_state_of_affairs
Concessive.mas.c	Concessive	Main_assertion
Correctness.verdade.n	Correctness	Information
Records per page: 5 1-5 of 22 < > >		
LU	Frame	FE
Noise_makers.gaita de fole.n	Cogitation	Topic
Musical_instruments.gaita de fole.n	Musical_instruments	Musical_instrument
People_by_origin.escocês.n	People_by_origin	Person
Records per page: 5 1-3 of 3		

Source: <http://webtool.frame.net.br/reportMultimodal>.

The 'Frame-frame menu' was designed as a tool to present a report of the relations between the frames that are evoked in a portion of the video by the LUs in the text and the visual objects in the image. The report is presented in the form of a Sankey diagram. On the left side of the diagram, each vertical-colored bar refers to a frame evoked by an LUs in the text. They are listed in alphabetical order. On the right side, the vertical-colored bars denote the frames evoked by visual objects. The bars in the central region of the diagram demonstrate account for frames that are part of the net to which each frame is connected. Some frames listed on the left or on the right do not represent evoked frames – the ones tagged with [NONE] –, but frames that are at the end of a chain initiated in the opposite side of the diagram. Figure 50 shows, for example, that the *People* frame – pale green bar – evoked in the text is connected to the *Biological\_entity* – medium green bar – which is connected to the *Entity* frame (brown bar on the right) which is not evoked by a visual object, but is a top-level frame for this chain. On the other hand, the *People\_by\_origin* frame (orange bar) evoked by a visual object – the bar is in the right end and have a 'v' (for video) attached to the frame name – is connected to the *People* frame – pale green bar in the center of the diagram – which is

connected to the `Biological_entity` – medium green bar in the center of the diagram – which, by its turn, is connected to the `People` frame – pale green bar on top left – evoked in the text. There are also situations of the same frame evoked by LUs in the text and visual objects in the image, as it is the case of the `Musical_instruments` frame – pink bar in the bottom – for which both ends of the table are straightly connected with no other frame in between.

Figure 50 – Report example of the 'Frame-frame menu' in the FN-Br WebTool Report Module



Source: <http://webtool.frame.net.br/reportMultimodal>.

## 6.2 VISUAL OBJECT ANNOTATED FOR ONE FRAME INSTANTIATES ANOTHER FRAME EVOKED BY THE LEXICAL UNIT IN THE TEXT

There is a fencing sequence in episode 6 – Edinburgh. It shows Pedro exploring the highlanders’ way of fencing as a tradition kept by Scots nowadays. Pedro 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 Pedro a practical fencing lesson. The sequence begins with sentence (13) which was annotated as follows:

(13) É **impossível**<sup>Likelihood</sup> **não**<sup>Negation</sup> **associar**<sup>Cause\_to\_amalgamate</sup> esse **lugar**<sup>Locale</sup>  
à **esgrima**<sup>Custom</sup>.<sup>38</sup>

Later, when viewers see Pedro and McDonald fencing with ancient swords, the subtitles translate the spoken audio into Portuguese in (14).

(14) **Sempre**<sup>frequency</sup> fomos **ligados**<sup>Social\_connection</sup> às nossas **tradições**<sup>Custom</sup>.<sup>39</sup>

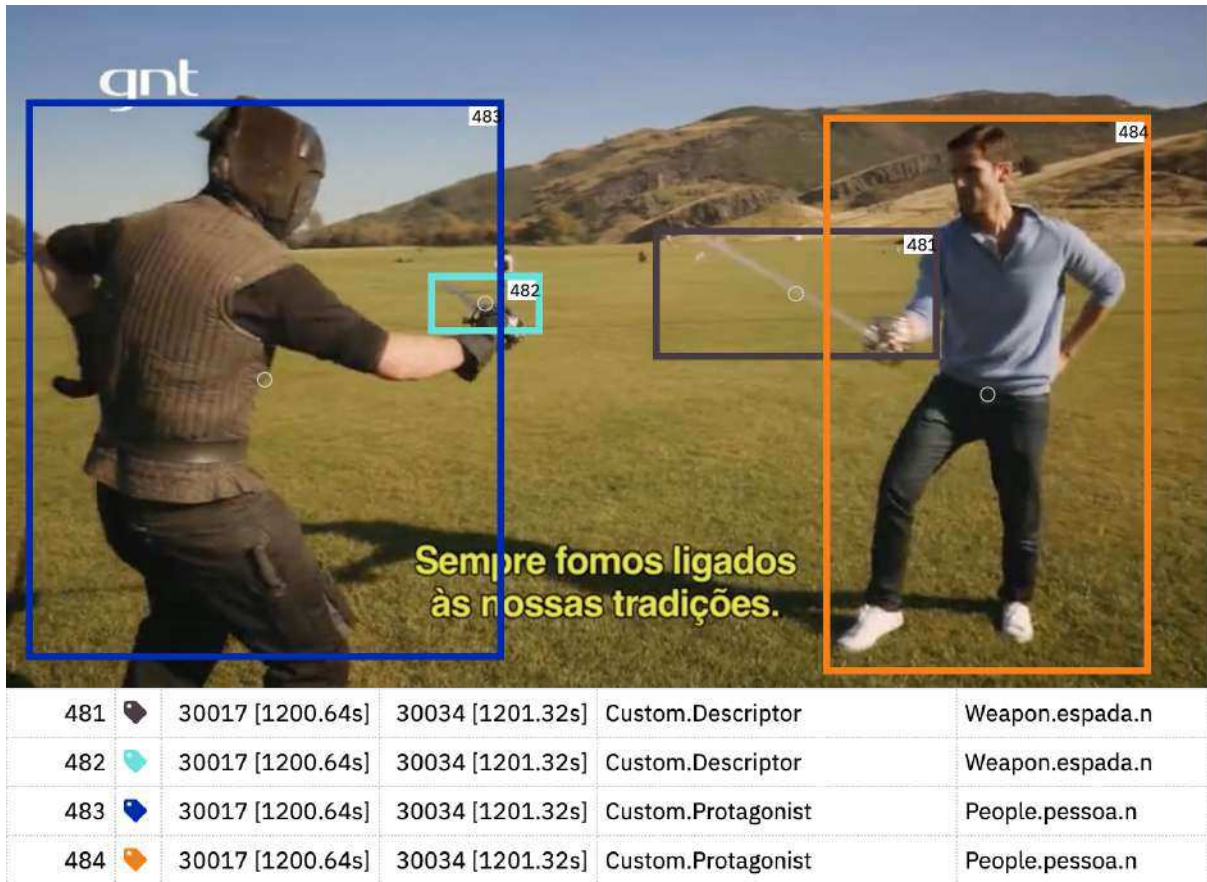
In sentence (14), *tradições.n* (traditions.n) is annotated for the *Custom*<sup>40</sup> frame – see Figure 51. The FE BEHAVIOR is incorporated in the LU, while it’s PROTAGONIST is annotated in the video – objects 483 and 484. Objects 481 and 481 are annotated for the FE WEAPON in the *Weapon* frame and designated as *espada.n* also in the *Weapon* frame for the CV Name. See Figure 51.

<sup>38</sup> It is impossible<sup>Likelihood</sup> not<sup>Negation</sup> to associate<sup>Cause\_to\_amalgamate</sup> this place<sup>Locale</sup> with fencing<sup>Custom</sup>.

<sup>39</sup> We have always<sup>frequency</sup> been connected<sup>Social\_connection</sup> to our traditions<sup>Custom</sup>.

<sup>40</sup> The *Custom* frame is defined in the FrameNet database as: “A Behavior is classified as entrenched for a Protagonist or a Society. The Behavior may be associated with a particular Domain of experience and described with a Descriptor. This indicates that the Behavior is commonly performed by the Protagonist or members of the Society.”

Figure 51 – Visual objects annotation of (14)

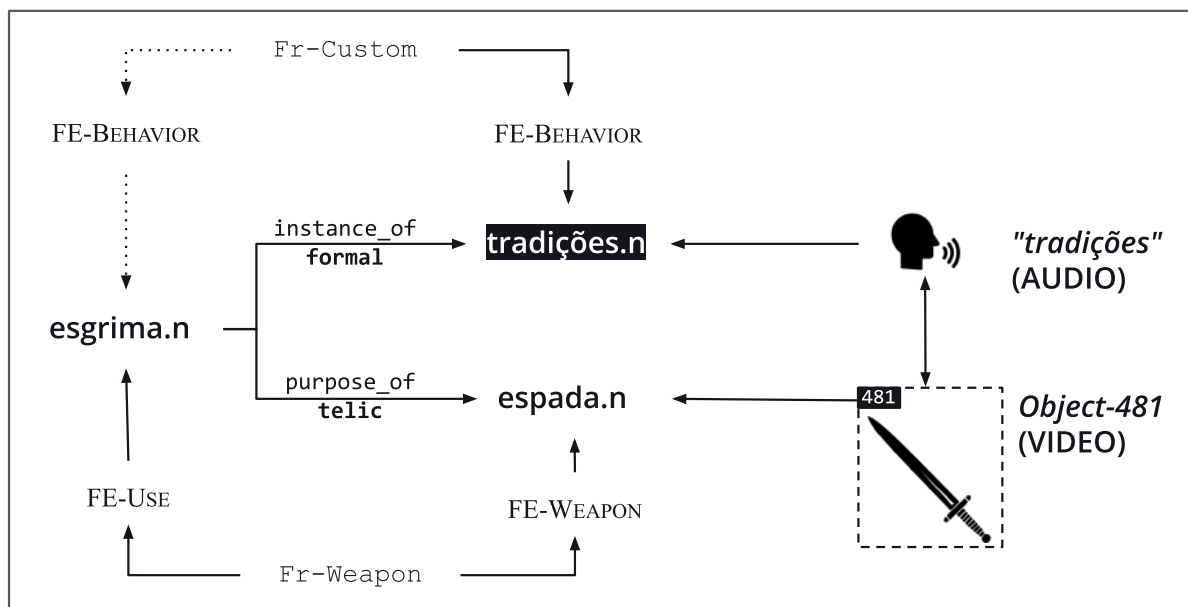


Source: charon.frame.net.br.

The arising issue would be how to represent the connection between the art of fencing, previously mentioned in sentence (13) and triggered in the shot by the sword – objects 481 and 482 – combined with the *Custom* frame, evoked by *tradições.n*. A ternary qualia relation once again solves the problem – see Figure 52.



Figure 52 – Ternary qualia relations in the multimodal annotation of sentence (14)



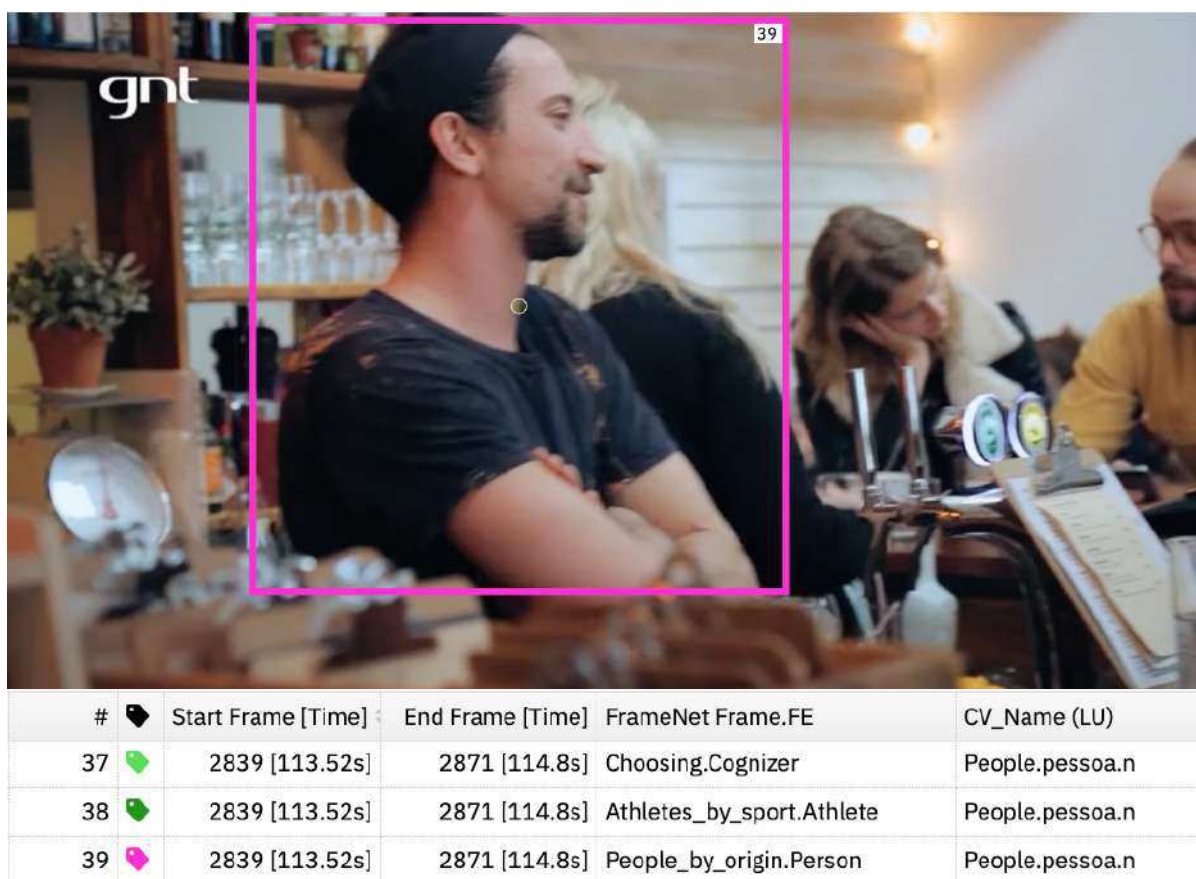
Source: the author.

Note that the existence of a ternary qualia relation mediated by the `Exemplar` frame connects *esgrima.n* (fencing.n) to *tradição.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.

### 6.3 DUPLICATED VISUAL OBJECT FOR MATCHING DIFFERENT INSTANCES OF DIFFERENT LEXICAL UNITS

The `Frame2` dataset has many examples of duplicated visual objects for matching two different instances of two lexical units. This happens frequently when there is an entity on screen which is referred to by more than one noun in the correlated audio, and also when this referred entity is an instance of an FE in a frame evoked by a verb. Look at the example from Episode 2 – Reykjavik in Figure 53:

Figure 53 – Multiple coincident visual objects



Source: charon.frame.net.br.

On the screen we can see object 39 only. This happens because objects 37 and 38 have exactly the same coordinates for their bounding boxes. The three different objects refer to the person wearing a black hat and a black shirt. Therefore, the three different objects select the same portion of the screen. However, each of them refers to a different lexical unit spoken in the audio and annotated in sentence (15): object 37 instantiates the COGNIZER FE in the Choosing frame; object 38 instantiates the ATHLETE FE in the Athletes\_by\_sport frame; and object 39 instantiates the PERSON FE in the People\_by\_origin frame.

(15) O frio de Reykjavik não parece uma **escolha**<sup>Choosing</sup> óbvia para um **skatista**<sup>Athletes\_by\_sport</sup> **californiano**<sup>People\_by\_origin</sup>.<sup>41</sup>

<sup>41</sup> The cold of Reykjavik does not seem an obvious choice<sup>Choosing</sup> for a Californian<sup>People\_by\_origin</sup> skateboarder<sup>Athletes\_by\_sport</sup>.

The interesting aspect of this structure is that it blends in the same entity – the human figure in the foreground of the screen – the attributes of the COGNIZER, the ATHLETE and the PERSON. Moreover, it exemplifies the perspectivized nature (BASILE et al., 2021) of the dataset. Because Frame<sup>2</sup> is annotated for Frame Semantics, it is, by default, capable of (re)framing one same element multiple times as the discourse proceeds.

#### 6.4 PERSPECTIVE AND TEXT-ORIENTATION SHIFTS VISUAL OBJECT ANNOTATION

Episode 2 – Reykjavik offers an example for combination of perspectives. At a certain point of the episode, viewers see Pedro walking through the city center and listen to sentence (16) listing three things that make it famous for:

- (16) O centro de Reykjavik é famoso pela arte de rua, pelas casas coloridas e por algumas atrações turísticas<sup>Attraction\_tourism</sup>.<sup>42</sup>

The annotator chose the *Attraction\_tourism* frame evoked by the multiword expression *atrações turísticas.n* (tourist attractions.n) for sentence (16). However, in the shot synchronically presented – Figure 54 – the annotator chose the *Attracting\_tourists* frame, which, in turn, is a perspective of the *Attraction\_tourism* one. The latter is an unperspectivized frame modeling, in a general fashion, the part of the *Tourism\_scenario* in which the tourism industry is built around attractions – and not events, for example. On the other hand, the *Attracting\_tourists* frame adopts the perspective of the *Attraction* FE, and is usually evoked by verbs such as *attract.v* and *lure.v*, as opposed to what happens in the *Touring* frame, which adopts the perspective of the *Tourist* FE and is evoked by LUs such as *visit.v* and *tour.n*.

That happened because the viewer sees only the Cathedral, that is, an *Attraction* as a place that brings people from different origins interested in enjoying its features. Then, the focus on the *Attraction* makes the agentive process of attracting the perspective adopted for annotation.

---

<sup>42</sup> Reykjavik's center is famous for street art, colored sidewalks, and some tourist attractions<sup>Attraction\_tourism</sup>.

Figure 54 – The Cathedral annotated as ATTRACTION in Attracting\_tourists



Source: charon.frame.net.br.

In the subsequent shot (Figure 55), however, the same cathedral was annotated as an FE of the *Attraction\_tourism* frame. In this shot, viewers see Pedro walking in a street in the foreground and the cathedral in the background. Annotation has Object 114 instantiating the PLACE FE, covering the whole screen, Object 115, the cathedral, instantiating the ATTRACTION FE, and Object 116, the host, instantiating the TOURIST FE.

Figure 55 – Image annotation for *Attraction\_tourism*

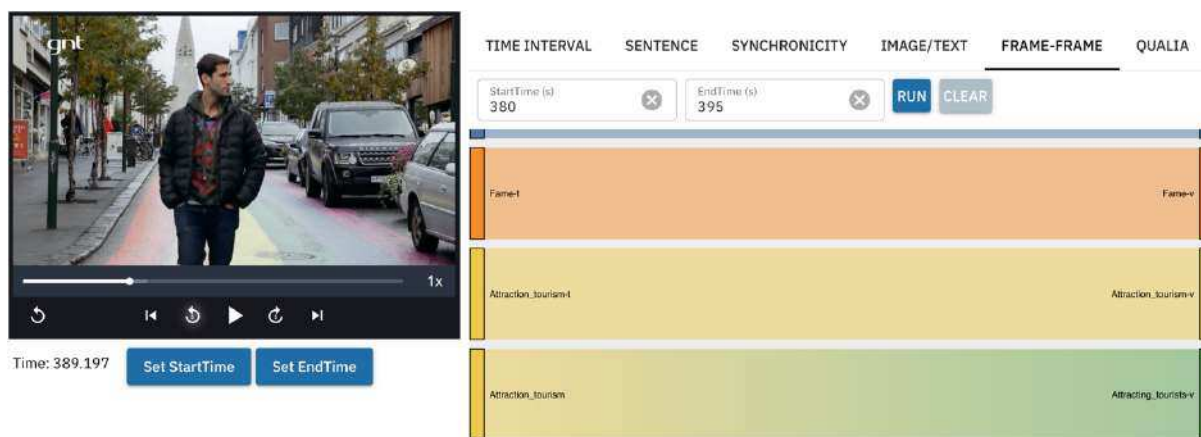


Source: charon.frame.net.br.

Consider, then, the situational context. When a viewer sees the shot depicted in Figure 55 it is natural to recognize the cathedral shown in the previous shot. This second appearance of the cathedral, now in the background, instantiates the *ATTRACTION* FE. At the same time, in the foreground we have the host figure, which instantiates the *TOURIST* FE with these two FEs annotated, we can say we have the *Attraction\_tourism* frame in the image matching the *Attraction\_tourism* frame evoked by the multiword expression *atrações turísticas.n* (tourist attractions.n).

Once again, the report system shows the relations between the frames evoked as a sankey diagram. However, this time, the links in the diagram are not ternary qualia, but frame-to-frame relations, as it can be seen in Figure 56. The bottom line of the diagram shows the *Attraction\_tourism* frame, as the yellow vertical bar, connected straightly to the green vertical bar, which refers to the *Attracting\_tourists*.

Figure 56 – Sankey diagram report on frame-to-frame relation between *Attraction\_tourism* and *Attracting\_tourists* frames



Source: <http://webtool.frame.net.br/reportMultimodal>.

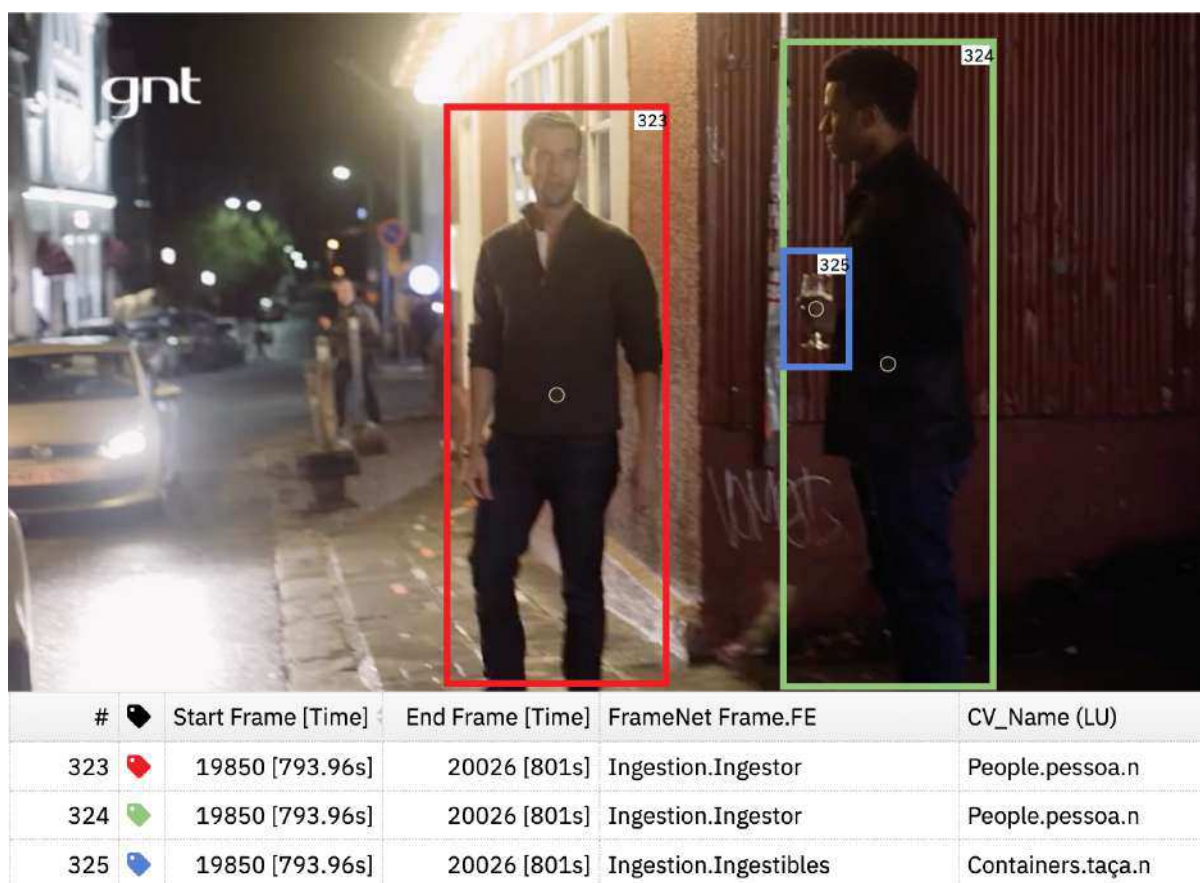
## 6.5 VISUAL OBJECT INSTANTIATES WHAT IS A NULL INSTANTIATION IN TEXT

For this sample we go back to the ‘Night life’ sequence in the Episode 2 – Reykjavik, mentioned previously in sentence (11) and in Figure 38 and Figure 39. In this sequence, Pedro is exploring Iceland’s capital night live accompanied by a Brazilian who owns a pub in the city. At some point, when they are walking outdoors, Pedro comments that it is cold, but their drinks make them warm. The sentence was annotated as shown in (17):

- (17) **Bom**<sup>Desirability</sup> que **aqui**<sup>Locative\_relation</sup> a gente **bebe**<sup>Ingestion</sup> e vai **esquentando**<sup>Change\_of\_temperature</sup>, né?<sup>43</sup>

The focal point here is the *Ingestion* frame. It has *INGESTIBLES* and *INGESTORS* as core FEs. However, in sentence (17), the *INGESTIBLES* are not instantiated and annotated as an Indefinite Null Instantiation (INI). Once sentence (17) is synchronically aligned with the shot depicted in Figure 57 – Image annotation for the *Ingestion* frame it is actually possible to see what they are drinking in object 325.

Figure 57 – Image annotation for the *Ingestion* frame



Source: charon.frame.net.br.

Object 325 was annotated as the *INGESTIBLES* in the *Ingestion* frame. Actually what is visible in the man's hand is a glass, and the annotator pointed this by choosing *taça.n* (glass.n)

<sup>43</sup> It is good<sup>Desirability</sup> that here<sup>Locative\_relation</sup> we drink<sup>Ingestion</sup> and warm<sup>Change\_of\_temperature</sup> ourselves up, innit?

in the `Containers` frame for the CV Name of the object. This kind of connection could be modeled in the FrameNet Brasil database, although it currently does not exist. A subtype of the telic ternary qualia relation mediated by the `Tool_purpose` frame could be posited to model the relations between different LUs in the `Food_and_beverage` frame and LUs in the `Container` and `Utensils` frames. That is another contribution of the multimodal annotation methodology devised in this dissertation: it allows for the inference of new instances of ternary qualia relations that should be added to the FrameNet Brasil database

## 6.6 MULTIPLE VISUAL OBJECTS MATCHING A SINGLE LEXICAL UNIT

This is a recurrent situation in the Frame<sup>2</sup> dataset: we have one lexical unit in a sentence that evokes a frame and have some Frame Elements instantiated which are associated with multiple visual objects on the image. From Episode 9 – Singapore, Figure 58 shows six different objects – 540, 542, 544, 546, 548 and 557 – annotated for the FE PERSON in the `People` frame. All these objects are instances of the FE PERSON instantiated by the LU *todo mundo.n* (everybody.n) evoking the `People` frame in sentence (18), which is synchronically aligned with the shot in Figure 58. Here, the CV Names also duplicate the information, once all these objects were labeled as *pessoa.n* (person.n) in the `People` frame.

Figure 58 – Example of multiple objects for a single lexical unit



Source: charon.frame.net.br.

(18) Está **todo mundo**<sup>People</sup> aqui para apreciar esta estátua atrás de mim: o Merlion ou Merlión.<sup>44</sup>

<sup>44</sup> Everyone<sup>People</sup> is here to appreciate this statue behind me: the Merlion or Merlión.

One detail here is that in this example we also have the duplication of objects described in section 6.3. Objects 540, 542, 544, 546, 548 and 557 match objects 541, 543, 545, 547 and 549 which instantiate the FE TOURIST in the `Touring` frame, evoked in another sentence.

This multiplication also happens in the previously mentioned sentence (1), in which the LU *uisque.n* (whisky.n) evokes the `Food_and_beverages` frame and incorporates the `FOOD_OR_BEVERAGE` FE. This FE is also annotated and visible in multiple visual objects – 6, 9, 11, 13, 15, 17, 19 and 21 – annotated in the synchronically aligned shot shown in Figure 59.

Figure 59 – Multiple objects instantiating `FOOD_OR_BEVERAGE` FE



#	Start Frame [Time]	End Frame [Time]	FrameNet.Frame.FE	CV_Name (LU)
6	877 [35.04s]	910 [36.36s]	Food_and_beverage.Food_or_beverage	Containers.garrafa.n
7	877 [35.04s]	910 [36.36s]	Cogitation.Topic	Containers.garrafa.n
8	877 [35.04s]	910 [36.36s]	Cogitation.Topic	Containers.garrafa.n
9	877 [35.04s]	910 [36.36s]	Food_and_beverage.Food_or_beverage	Containers.garrafa.n
10	877 [35.04s]	910 [36.36s]	Cogitation.Topic	Containers.garrafa.n
11	877 [35.04s]	910 [36.36s]	Food_and_beverage.Food_or_beverage	Containers.garrafa.n
12	877 [35.04s]	910 [36.36s]	Cogitation.Topic	Containers.garrafa.n
13	877 [35.04s]	910 [36.36s]	Food_and_beverage.Food_or_beverage	Containers.garrafa.n
14	877 [35.04s]	910 [36.36s]	Cogitation.Topic	Containers.garrafa.n
15	877 [35.04s]	910 [36.36s]	Food_and_beverage.Food_or_beverage	Containers.garrafa.n
16	877 [35.04s]	910 [36.36s]	Cogitation.Topic	Containers.garrafa.n
17	877 [35.04s]	910 [36.36s]	Food_and_beverage.Food_or_beverage	Containers.garrafa.n
18	877 [35.04s]	910 [36.36s]	Cogitation.Topic	Containers.garrafa.n
19	877 [35.04s]	910 [36.36s]	Food_and_beverage.Food_or_beverage	Containers.garrafa.n
20	877 [35.04s]	910 [36.36s]	Cogitation.Topic	Containers.garrafa.n
21	877 [35.04s]	910 [36.36s]	Food_and_beverage.Food_or_beverage	Containers.garrafa.n

Source: charon.frame.net.br.

In terms of CV Names, in this sample they add extra information, once objects received the label of *garrafa.n* (bottle.n) in the `Containers` frame. This means we have once again the qualia relation associating *garrafa.n* (bottle.n) in the `Containers` frame to *uisque.n* in the `Food_and_beverages` frame would hold.

This is also another example of the duplication of objects described in section 6.3. Here Objects 6, 9, 11, 13, 15, 17, 19 and 21 were duplicated as objects 7, 8, 10, 12, 14, 16, 18 and 20. In the second group, each of them instantiate the FE TOPIC in the `Cogitation` frame, once they fill this role for the evocation of the frame established by *vem à mente.v* (come to mind.v) in sentence (1).

## 6.7 BLENDING ENTITY FROM VISUAL OBJECT TO INSTANTIATE FRAME ELEMENT IN TEXT



A very rich example of image-text combination comes from Episode 2 – Reykjavik. In this sequence, Pedro’s interviewee describes how people in the city bounced back from the economic crisis of 2008. Sentence (19) appears in this context as a subtitle on screen.

(19) *O povo voltou a ser criativo*<sup>Mental\_property</sup>.<sup>45</sup>

For sentence (19), the annotator chose the *Mental\_property* frame, evoked by the LU *criativo.a* (*creative.a*). The BEHAVIOR FE is instantiated incorporated in the LU. Although both frame and FE indicate intangible concepts – a mental property and a behavior –, Figure 60 shows that the annotator found an entity on screen that embodies those properties.

Figure 60 – Visual object embodies a FE



Object 63 in Figure 60 is the one the annotator chose as the instance of the BEHAVIOR FE in the *Mental\_property* frame on screen, which matches the text annotation. For the CV Name, the annotator chose the *grafite.n* (*graffiti.n*) which suitably express the materiality of the painting in a wall depicted in the image. The prominent factor in this example is that the direct association of the specific graffiti piece of artwork shown with the general idea of the behavior of non-specific people being creative is a metaphor and possible only because of the

<sup>45</sup> People started being creative<sup>Mental\_property</sup> again.

multimodal approach proposed in this dissertation. Using ternary qualia relations, it is possible to establish that *grafite.n* is an instance\_of *criativo.a*, mediated by the Instance frame.

## 6.8 REMARKS ON THE FRAME<sup>2</sup> DATASET

The Frame<sup>2</sup> dataset exploits the complexity of the enriched model of FN-Br to create meaningful connections also enhanced between semiotic modes, between communicative modalities. The dataset offers the means for the FN-Br database 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 deal with the dataset keeps the linguistic anchorage to the way its elements may be analyzed, explored, and used. However, the research conducted to culminate in this dissertation shows that the path to approach image in processes of meaning-making is broad and offers other ways to be explored. This means that the new connections proposed for the model now will also serve to further enrich the basis of FN-Br, as they uncover new sources of modeling.

For the next stage of the Frame<sup>2</sup> dataset development includes its usage by the Research and Innovation Network for Visual and Textual Analysis of Multimodal Objects (ReINVenTA) to train a model for improving and actually start to automatically identify and tag Visual Objects for LUs and frames in the CV Name label. In parallel, we are sure that other ways of annotating other elements of visual composition will arise. Also, we believe that the dataset is full of other ways to foresee the perception of texts and images combinations.

## 7 CONCLUSION

This dissertation reports on the first research on the encounter between Frame Semantics and FrameNet with the Multimodal approach. It consolidates what has been developed in the last five years and reported on by Belcavello et al. (2020; 2022) and Torrent et al. (2022). The hypothesis was that visual elements in a video sequence may (i) evoke frames and instantiate Frame Elements, similarly to the way in which words in a sentence do, and (ii) combine with the spoken audios' words and phrases in sentences to offer a broader sense in the frame evocation patterns which provide different profiling and perspective options for meaning construction, while also exploring alternative connections between concepts in the FrameNet Brasil model.

To test this hypothesis we, first, conducted a pilot study (BELCAVELLO et al. 2020), which indicated a promising path while pointing out aspects necessary to achieve the main goal of developing a video annotation methodology capable of adding to the FrameNet Brasil model the multimodal perspective foreseen in the hypothesis.

The theoretical foundations for promoting the multimodal shift of FrameNet begun with a deep study of the concept of Multimodality. We reviewed theories and frameworks that sustain the multimodal approach, as well as its connections with genre studies and computational applications. This provided the basis for us to present the TV Travel Series as the multimodal genre in which we looked for the narrative grammar in the path for identifying the semantic elements and patterns for the combination of spoken audio and image that sustained the proposal for verifying the hypothesis.

On the way for achieving the necessary goals of this research, we built and developed resources that stand as the dissertation's contributions. The first one is the creation and development of Charon (BELCAVELLO et al., 2022), the multimodal corpus builder, annotation tool and database management application integrated with the FrameNet Brasil Web Annotation Tool. Charon was used to build the second contribution, which is the "*Pedro pelo Mundo*" corpus, composed by 230 minutes of video, divided in ten documents – one for each episode –, and the 2,195 sentences generated. The third contribution is the methodology developed, as it was defined and described in section 5.2. Finally, the fourth contribution is the Frame<sup>2</sup> dataset composed by the multimodal objects, i.e., the annotated data, both for text and image, in the "*Pedro pelo Mundo*" corpus and the relations between the annotated data as modeled in the FrameNet Brasil database. All these contributions are available to be used in new research or tasks, and to be enhanced by new contributions and/or future developments.

The academic production resulting from this research includes this dissertation itself, as well as 14 oral or poster presentations in national and international conferences and the following publications:

- Journal paper:
  - TORRENT, Tiago Timponi; MATOS, Ely E. d. S.; BELCAVELLO, Frederico; VIRIDIANO, Marcelo; GAMONAL, Maucha A; COSTA, Alexandre D. d.; and Marim, Mateus C. Representing context in FrameNet: A multidimensional, multimodal approach. **Frontiers in Psychology**, 13, 2022.
  
- Full papers in conference proceedings:
  - BELCAVELLO, Frederico; VIRIDIANO, Marcelo; MATOS, Ely E. d. S; TORRENT, Tiago T. . Charon: a FrameNet Annotation Tool for Multimodal Corpora. In: Sameer Pradhan; Sandra Kübler. (Org.). Proceedings of the 16<sup>th</sup> Linguistic Annotation Workshop (LAW-XVI) within LREC2022. 1ed.Marseille: European Language Resources Association (ELRA), 2022. P. 91-96.
  - TORRENT, Tiago T. ; ALMEIDA, Arthur L. ; MATOS, Ely E. d. S. ; BELCAVELLO, Frederico ; VIRIDIANO, Marcelo ; GAMONAL, Maucha A. . Lutma: a Frame-Making Tool for Collaborative FrameNet Development. In: Gavin Abercrombie; Valerio Basile; Sara Tonelli; Verena Rieser; Alexandra Uma. (Org.). Proceedings of the 1<sup>st</sup> Workshop on Perspectivist Approaches to NLP @LREC2022. 1ed.Paris: European Language Resources Association (ELRA), 2022. P. 100-107.
  - VIRIDIANO, Marcelo; TORRENT, Tiago T. ; CZULO, Oliver ; ALMEIDA, Arthur L. ; MATOS, Ely E. d. S. ; BELCAVELLO, Frederico . The Case for Perspective in Multimodal Datasets. In: Gavin Abercrombie; Valerio Basile; Sara Tonelli; Verena Rieser; Alexandra Uma. (Org.). Proceedings of the 1<sup>st</sup> Workshop on Perspectivist Approaches to NLP @LREC2022. 1ed.Paris: European Language Resources Association (ELRA), 2022. P. 108-116.

- BELCAVELLO, Frederico; VIRIDIANO, Marcelo; COSTA, Alexandre Diniz da; MATOS, Ely E. d. S. ; TORRENT, Tiago T. . Frame-Based Annotation of Multimodal Corpora: Tracking (A)Synchronies in Meaning Construction. In: TORRENT, Tiago T.; BAKER, Collin F.; CZULO, Oliver; OHARA, Kyoko; PETRUCK, Miriam R. L. (Org.). Proceedings of the LREC International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet. 1ed. Paris: European Language Resources Association (ELRA), 2020, v. 1, p. 23-30.

For future developments we foresee the application of the methodology proposed for new research and tasks, as it is already being done by the Research and Innovation Network for Visual and Textual Analysis of Multimodal Objects (ReINVenTA) for the Audition corpus. We also believe that the “*Pedro pelo Mundo*” corpus can be used to perform other types of annotation and devise other methodologies, e. g. ones that account for other visual aspects such as camera angles, camera movements or gestures, which were not in the scope of this dissertation. Finally, as a next step of this research, we project the use of the Frame<sup>2</sup> dataset as a gold standard resource for training models and develop semi-automatic detection of visual objects and frame labeling.

## REFERENCES

- ADAM, Jean-Michel. **A Linguística Textual: Introdução à análise textual dos discursos**. São Paulo: Cortez, 2011.
- ADAMI, Elisabetta; KRESS, Gunther. Introduction: Multimodality, meaning making, and the issue of “text”. **Text & Talk**, v. 34, n. 3, p. 231-237, 2014.
- ARONCHI DE SOUZA, José Carlos. **Gêneros e formatos na televisão brasileira**. São Paulo: Summus, 2004.
- AKSOY, Eren Erdal et al. Unsupervised linking of visual features to textual descriptions in long manipulation activities. **IEEE Robotics and Automation Letters**, v. 2, n. 3, p. 1397-1404, 2017.
- BAKHTIN, Mikhail. **Speech genres and Other Later Essays**. Austin: University of Texas Press, 1986 [1952-1953].
- BAKHTIN, Mikhail. **Problemas da poética de Dostoiévski** (1929). 2. ed. Trad. Paulo Bezerra. Rio de Janeiro: Forense Universitária, 1997. 276p.
- BARBER, X. Theodore. The roots of travel cinema: John L. Stoddard, E. Burton Holmes and the nineteenth-century illustrated travel lecture. **Film History**, v. 5, n. 1, p. 68-84, 1993.
- BARTHES, Roland. The Photographic Message. In: BARTHES, Roland (ed). **Image-Music-text**. London: Fontana, 1977[1961] p. 15-31.
- BARTHES, Roland. Rhetoric of the Image. In: BARTHES, Roland (ed.) **Image-Music-text**. London: Fontana, 1977[1964] p. 33-51.
- BASILE, Valerio; CABITZA, Federico; CAMPAGNER, Andrea; and FELL, Michael. Toward a perspectivist turn in ground truthing for predictive computing. **arXiv preprint arXiv:2109.04270**, 2021.
- BATEMAN, John; WILDFEUER, Janina; HIIPPALA, Tuomo. **Multimodality: Foundations, research and analysis—A problem-oriented introduction**. Berlin/Boston: Walter de Gruyter GmbH & Co KG, 2017.
- BATEMAN, John A.. **Multimodality and genre: A foundation for the systematic analysis of multimodal documents**. Hampshire: Palgrave Macmillan, 2008.
- BATEMAN, John A.. **Text and Image: a critical introduction to the visual/verbal divide**. London and New York: Routledge, 2014.
- BATRA, Vishwash; HE, Yulan; VOGIATZIS, George. Neural caption generation for news images. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki: ELRA, 2018.
- BELCAVELLO, Frederico; VIRIDIANO, Marcelo; COSTA, Alexandre Diniz da; MATOS, Ely E. S.; TORRENT, Tiago T.. Frame-Based Annotation of Multimodal Corpora: Tracking

(A) Synchronies in Meaning Construction. In: **Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet**. Marseille: ELRA, 2020. p. 23-30.

BELCAVELLO, Frederico; VIRIDIANO, Marcelo; MATOS, Ely E. S.; TORRENT, Tiago T.. Charon: a FrameNet Annotation Tool for Multimodal Corpora. In: **Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022**. Marseille: ELRA, 2022. p. 91–96.

BÉRARD, Alexandre et al. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In: **Proceedings of the 30th Neural Information Processing Systems - NIPS Workshop on end-to-end learning for speech and audio processing**. Barcelona, 2016.

CALIXTO, Iacer; ELLIOTT, Desmond; FRANK, Stella. DCU-UvA multimodal MT system report. In: **Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers**. 2016. p. 634-638.

CHALABY, Jean K.. The making of an entertainment revolution: How the TV format trade became a global industry. **European Journal of Communication**, v. 26, n. 4, p. 293-309, 2011.

CHALABY, Jean K.. At the origin of a global industry: The TV format trade as an Anglo-American invention. **Media, Culture & Society**, v. 34, n. 1, p. 36-52, 2012.

COHN, Neil. Visual narrative structure. **Cognitive science**, v. 37, n. 3, p. 413-452, 2013.

COHN, Neil. A multimodal parallel architecture: A cognitive framework for multimodal interactions. **Cognition**, v. 146, 2016a, p. 304-323.

COHN, Neil. From Visual Narrative Grammar to Filmic Narrative Grammar: The narrative structure of static and moving images. **Film text analysis: New perspectives on the analysis of filmic meaning**, 2016b, p. 94-117.

COHN, Neil. Visual narratives and the mind: Comprehension, cognition, and learning. In: **Psychology of learning and motivation**. Academic Press, 2019. p. 97-127.

CONKLIN, Kathy; PELLICER-SÁNCHEZ, Ana; CARROL, Gareth. **Eye-tracking: A guide for applied linguistics research**. New York: Cambridge University Press, 2018.

COSTA, Alexandre Diniz da. **A tradução por máquina enriquecida semanticamente com frames e papéis qualia**. Tese (Doutorado em Linguística) – Programa de Pós Graduação em Linguística, Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora, 2020.

CREEBER, Glen (Ed.). **The television genre book**. 3. ed. London: Bloomsbury Publishing, 2015.

DAMKJAER, Maja Sonne; WAADE, Anne Marit. Armchair tourism: The travel series as a hybrid genre. In: **Travel Journalism**. Palgrave Macmillan, London, 2014. P. 39-59.

- DEVLIN, Jacob et al. Language Models for Image Captioning: The Quirks and What Works. In: **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**. Beijing, 2015. P. 100-105.
- DIAS, Anair Valênia Martins et. Al. Minicontos multimodais: reescrevendo imagens cotidianas. In: ROJO, Roxane; MOURA, Eduardo. **Multiletramentos na escola**. São Paulo: Parábola Editorial, 2012. P. 95-122.
- DIAS, Anair Valênia Martins. Hipercontos multissemióticos: para a promoção dos multiletramentos. In: ROJO, Roxane; MOURA, Eduardo. **Multiletramentos na escola**. São Paulo: Parábola Editorial, 2012. P. 75-94.
- DOLZ, Joaquim; SCHNEUWLY, Bernard. Gêneros e progressão em expressão oral e escrita – elementos para reflexões sobre uma experiência suíça (francófona). In: ROJO, Roxane; CORDEIRO, Gláís Sales. (trad. Org.). **Gêneros orais e escritos na escola**. Campinas: Mercado de Letras, 2004[1996]. P. 41-70.
- DUCHOWSKI, Andrew T.; DUCHOWSKI, Andrew T. **Eye tracking methodology: Theory and practice**. Cham: Springer, 2017.
- ELLIOTT, D.; FRANK, S.; HASLER, E. Multi-language image description with neural sequence models. CoRR. **arXiv preprint arXiv:1510.04709**, 2015.
- ESCALERA, Sergio et al. Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. In: **Proceedings of the 15th ACM on International conference on multimodal interaction**. 2013. p. 365-368.
- FANG, Hao et al. From captions to visual concepts and back. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. Boston, 2015. p. 1473-1482.
- FECHINE, Yvana. Gêneros televisuais: a dinâmica dos formatos. In: **Revista Symposium** – Universidade Católica de Pernambuco. Recife, v. 5, n. 1, 2001. p. 14-26.
- FILLMORE, Charles J. . An alternative to checklist theories of meaning. In **Proceedings of the First Annual Meeting of the Berkeley Linguistics Society**. Berkeley: BLS, 1975. p. 123-131.
- FILLMORE, Charles J. . Frame semantics and the nature of language. In **Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech**, Volume 280. New York: New York Academy of Sciences, 1976. p. 20-32.
- FILLMORE, Charles J. . Scenes-and-frames semantics, Linguistic Structures Processing. In ZAMPOLLI, Antonio (Ed.): **Fundamental Studies in Computer Science**, No. 59, Amsterdam: North Holland Publishing, 1977a. p. 55-88.
- FILLMORE, Charles J. . The need for a frame semantics in linguistics. In KARLGREN, Hans (Ed.). **Statistical Methods in Linguistics**. v. 12. Stockholm: Skriptor, 1977b. p. 5-29.



FILLMORE, Charles J. The case for case reopened. In: **Grammatical relations**. Brill, 1977c. p. 59-81.

FILLMORE, Charles J. Frame semantics. In: **Linguistics in the morning calm**. Linguistics Society of Korea. Seoul: Hanshin, 1982. p. 111-137.

FILLMORE, Charles J. Frames and the semantics of understanding. In: **Quaderni di semantica**. v. VI, n. 2. Bologna: Società editrice il Mulino, 1985. p. 222-254.

FILLMORE, Charles J.; ATKINS, Beryl T. Toward a frame-based lexicon: The semantics of RISK and its neighbors. In: LEHER, Adrienne; KITTAY, Eva Feder. (ed). **Frames, fields and contrasts: New essays in semantic and lexical organization**. Lawrence Erlbaum Associates, 1992. p. 75-102.

FILLMORE, Charles J.; JOHNSON, Christopher R.; PETRUCK, Miriam R. L.. Background to framenet. **International journal of lexicography**, v. 16, n. 3, p. 235-250, 2003.

FONSECA, Aline; MAIA, Marcus. Na trilha do processamento da linguagem: o uso de rastreadores oculares na análise de dados linguísticos. In: OLIVEIRA, Cândido Samuel Fonseca de; SÁ, Thaís Maira Machado de; org. **Métodos experimentais em psicolinguística**. São Paulo: Pá de Palavra, 2022.

FORCEVILLE, Charles. Non-verbal and multimodal metaphor in a cognitivist framework: Agendas for research. **Applications of Cognitive Linguistics**, v. 1, p. 379, 2006.

FORCEVILLE, Charles J. **Multimodality: A Social Semiotic Approach to Contemporary Communication**: Gunther Kress, Routledge, London, 2010, 212 pp., 45 b/w illustrations+ 15 colour plates, ISBN 13: 978-0-415-32061-0 (pbk). 2011.

FRAPA (THE FORMAT RECOGNITION AND PROTECTION ASSOCIATION). **What is a format?** Naarden, 2020. <https://frapa.org/>.

FUSCO, Serafina; PERROTTA, Marta. Rethinking the Format as a Theoretical Object in the Age of Media Convergence. **Observatorio (OBS\*)**, v. 2, n. 4, 2008. p. 89-102.

GAMONAL, Maucha Andrade. **Modelagem linguístico-computacional de metonímias na Base de Conhecimento Multilíngue (m.knob) da FrameNet Brasil**. Tese (Doutorado em Linguística) – Programa de Pós Graduação em Linguística, Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora, 2017.

GILAKJANI, Abbas Pourhossein; ISMAIL, Hairul Nizam; AHMADI, Seyedeh Masoumeh. The effect of multimodal learning models on language teaching and learning. **Theory & Practice in Language Studies**, v. 1, n. 10, 2011.

GOMES, Itânia Maria Mota. A Noção de Gênero Televisivo como Estratégia de Interação: o Diálogo entre os Cultural Studies e os Estudos da Linguagem. **Revista Fronteiras – estudos midiáticos**, Unisinos/São Leopoldo, v. IV, n. 2, Dez. 2002. p. 165-185.

GOMES, Itânia Maria Mota. Metodologia de análise de telejornalismo. In: GOMES, Itânia Maria Mota (org.) **Gênero televisivo e modo de endereçamento no telejornalismo**. Salvador: EDUFBA, 2011.

GOODFELLOW, Ian J. et al. Generative Adversarial Nets. In: **Advances in neural information processing systems** – Proceedings of 28th Annual Conference on Neural Information Processing Systems 2014, v. 1050, p. 2672-2680, 2014.

GRANSTRÖM, Björn; HOUSE, David; KARLSSON, Inger (Ed.). **Multimodality in language and speech systems**. Springer Science & Business Media, 2013.

HALLIDAY, Michael AK. 1994. **An introduction to functional grammar**, v. 2, 1985.

HIIPPALA, Tuomo. 11 Multimodal Genre Analysis. **Interactions, images and texts: A reader in multimodality**, v. 11, p. 111, 2014.

HIIPPALA, Tuomo. An overview of research within the Genre and Multimodality framework. **Discourse, context & media**, v. 20, p. 276-284, 2017.

HODGE, Robert; KRESS, Gunther. **Social Semiotics**. Cambridge: Polity, 1988.

JACKENDOFF, Ray; JACKENDOFF, Ray S. **Foundations of language: Brain, meaning, grammar, evolution**. New York: Oxford University Press, 2002.

JEWITT, Carey Ed. **The Routledge handbook of multimodal analysis**. Routledge/Taylor & Francis Group, 2011.

JEWITT, Carey; BEZEMER, Jeff; O'HALLORAN, Kay. **Introducing multimodality**. Routledge, 2016.

JEWITT, Carey; KRESS, Gunther R. (Ed.). **Multimodal literacy**. New York: Lang, 2003.

JOO, Jungseock; STEEN, Francis F.; TURNER, Mark. Red Hen Lab: Dataset and tools for multimodal human communication research. **KI-Künstliche Intelligenz**, v. 31, n. 4, p. 357-361, 2017.

KATSAMANIS, Athanasios et al. Multimodal gesture recognition. In: OVIATT, S. et al. (Eds.) **The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations**. Volume 1. ACM Books / Morgan-Claypool Publishers: San Rafael, 2017. p. 449-487.

KISLOVA, Larisa Sergeevna; ERTNER, Daria Evgenievna. The Revival of the Genre: Key Trends in the Contemporary Travelogue Development. **Philological Class**. 2019. № 3 (57), p. 127-133, 2019.

KRESS, Gunther R. **Multimodality: A social semiotic approach to contemporary communication**. Taylor & Francis, 2010.

KRESS, Gunther; VAN LEEUWEN, Theo. **Reading Images: The Grammar of Visual Design**. 2. ed. London/New York: Routledge, 2006.

KUZNETSOVA, Alina et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. **International Journal of Computer Vision**, v. 128, n. 7, p. 1956-1981, 2020.

LENCI, Alessandro et al. SIMPLE: A general framework for the development of multilingual lexicons. **International Journal of Lexicography**, v. 13, n. 4, p. 249-263, 2000.

LIN, Tsung-Yi et al. Microsoft coco: Common objects in context. In: **Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13**. Springer International Publishing, 2014. p. 740-755.

LI, Chuyi et al. YOLOv6: A single-stage object detection framework for industrial applications. **arXiv preprint arXiv:2209.02976**, 2022.

LO BIANCO, Joseph. Multiliteracies and multilingualism. In: COPE, Bill; KALANTZIS, Mary (Ed.). **Multiliteracies: Literacy learning and the design of social futures**. Psychology Press, 2000.

LORENZI, Gislaine Cristina Correr; DE PÁDUA, Tainá-Rekã Wanderley. *Blog nos anos iniciais do fundamental I: a reconstrução de sentido de um clássico infantil*. In: ROJO, Roxane; MOURA, Eduardo. **Multiletramentos na escola**. São Paulo: Parábola Editorial, 2012. p. 35-54.

LOSCHKY, Lester C. et al. Viewing static visual narratives through the lens of the Scene Perception and Event Comprehension Theory (SPECT). **Empirical comics research: Digital, multimodal, and cognitive methods**, 2018. p. 217-238.

LOSCHKY, Lester C. et al. The scene perception & event comprehension theory (SPECT) applied to visual narratives. **Topics in cognitive science**, v. 12, n. 1, 2020. p. 311-351.

MACHADO, Arlindo. **A televisão levada a sério**. 4.ed. São Paulo: Editora Senac São Paulo, 2005 [2000].

MARCUSCHI, Luiz Antônio et al. Gêneros textuais: definição e funcionalidade. In: DIONÍSIO, Ângela Paiva; MACHADO, Anna Rachel; BEZERRA, Maria Auxiliadora. (Org.). **Gêneros textuais e ensino**. Rio de Janeiro: Lucerna, v. 20, 2002. p. 19-36.

MARTINEC, Radan; SALWAY, Andrew. A system for image–text relations in new (and old) media. **Visual communication**, v. 4, n. 3, p. 337-371, 2005.

MARTÍNEZ LIROLA, María. Teaching visual grammar and social issues in an English language course: an example using multimodal texts on immigrant minors from a Spanish newspaper. In: PÉREZ, Francisco Javier Díaz et al. (ed) **Global issues in the teaching of language, literature and linguistics**. Bern : Peter Lang, 2013. ISBN 978-3-0343-1255-4, p. 195-215.

MCKEVITT, Paul. MultiModal semantic representation. In: **First Working Meeting of the SIGSEM Working Group on the Representation of MultiModal Semantic Information**. Tilburg University, 2003. p. 1-16.

- MELO, José Marques de; ASSIS, Francisco de. Gêneros e formatos jornalísticos: um modelo classificatório. **Intercom: Revista Brasileira de Ciências da Comunicação**, v. 39, n. 1, p. 39-56, 2016.
- MINSKY, Marvin. **A Framework for Representing Knowledge**. Cambridge: Massachusetts Institute of Technology, 1974.
- MORAN, Albert. The pie and the crust: television program formats. In: ALLEN, Robert Clyde; HILL, Annette. (ed). **The television studies reader**. London and New York: Routledge, 2004. p. 258-266.
- MORAN, Albert. Television formats in the world / the world of television formats. In: KEANE, Michael; MORAN, Albert (ed.). **Television across Asia: TV industries, programme formats and globalisation**. London and New York: RoutledgeCurzon, 2004b.
- MORAN, Albert; MALBON, Justin. **Understanding the global TV format**. Bristol and Portland: Intellect Books, 2006.
- NIKOLAUS, Mitja et al. Compositional Generalization in Image Captioning. In: **Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)**. 2019. p. 87-98.
- NORRIS, Sigrid. **Analyzing multimodal interaction: A methodological framework**. Routledge, 2004.
- OZOLA, Diāna. Theoretical aspects on travelogue in modern literature. **Journal of Comparative Studies/Komparatovistikas Almanahs**, n. 6, Daugavpils University, Daugavpils, 2014.
- PEIRCE, Charles Sanders; WELBY, Lady Victoria. In: HARDWICK, Charles S. (ed) **Semiotic and signifiacs: the correspondence between Charles S. Peirce and Lady Victoria Welby**. Indiana University Press, 1977.
- PENNOCK-SPECK, Barry; DEL SAZ-RUBIO, María Milagros (Ed.). **The multimodal analysis of television commercials**. Publicacions de la Universitat de València, 2013.
- PITSIKALIS, Vassilis et al. Multimodal gesture recognition via multiple hypotheses rescoring. In: **Gesture Recognition**. Springer, Cham, 2017. p. 467-496.
- PUSTEJOVSKY, James. **The Generative Lexicon**. Cambridge, Estados Unidos: MIT Press, 1995.
- PUSTEJOVSKY, James. Type construction and the logic of concepts. In: BOUILLON, P.; BUSA, F. **The language of word meaning**. New York: Cambridge University Press, 2001. p. 91-123.
- PUSTEJOVSKY, James et al. Towards a Generative Lexical Resource: The Brandeis Semantic Ontology. In: **Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)**. 2006.

PUSTEJOVSKY, James; JEZEK, Elisabetta. Integrating Generative Lexicon and Lexical Semantic Resources. In: **10th Language Resources and Evaluation Conference**, n. 10, 2016, Portorož, Eslovênia. Anais [...]. Portorož: LREC, 2016, p. 3-139.

RADFORD, Alec; KIM, Jong W.; HALLACY, Chris; RAMESH, Aditya; GOH, Gabriel; AGARWAL, Sandhini; SASTRY, Girish; ASKELL, Amanda; MISHKIN, Pamela; CLARK, Jack; KRUEGER, Gretchen; SUTSKEVER, Ilya. Learning transferable visual models from natural language supervision. In: **International conference on machine learning**. PMLR, 2021, p. 8748-8763.

RAMESH, Aditya; PAVLOV, Mikhail; GOH, Gabriel; GRAY, Scott; VOSS, Chelsea; RADFORD, Alec; CHEN, Mark; SUTSKEVER, Ilya. Zero-shot text-to-image generation. In: **International Conference on Machine Learning**. PMLR, 2021. p. 8821-8831.

RAMESH, Aditya; DHARIWAL, Prafulla; NICHOL, Alex; CHU, Casey; CHEN, Mark. Hierarchical text-conditional image generation with clip latents. **arXiv preprint arXiv:2204.06125**. 2022

ROUSELL, J.; COLLIER, D. R. Researching multimodality in language and education. **Research Methods in Language and Education**. (3. ed) Dordrecht: Springer, 2017.

ROUSELL, Jennifer; WALSH, Maureen. Rethinking literacy education in new times: Multimodality, multiliteracies & new literacies. **Brock Education**, Volume 21, No. 1, Fall 2011, 53-62.

ROJO, Roxane; MOURA, Eduardo. **Multiletramentos na escola**. São Paulo: Parábola Editorial, 2012.

RUPPENHOFER, Josef.; ELLSWORTH, Michael; PETRUCK, Miriam R. L.; JOHNSON, Christopher R.; BAKER, Collin F.; SCHEFFCZYK, Jan. **FrameNet II: Extended Theory and Practice**. Berkeley: International Computer Science Institute, 2016.

SAUSSURE, F. de. **Course in general linguistics** (W. Baskin, Trans.). New York: Philosophical Library, 1959[1916].

SANABRIA, Ramon et al. How2: A Large-scale Dataset for Multimodal Language Understanding. In: **NeurIPS**. 2018.

SCHNEUWLY, Bernard. Gêneros e tipos de discurso: considerações psicológicas e ontogenéticas. In: ROJO, Roxane; CORDEIRO, Gláís Sales. (trad. org.). **Gêneros orais e escritos na escola**. Campinas: Mercado de Letras, 2004 [1994]. p. 21-40.

SPECIA, Lucia et al. A shared task on multimodal machine translation and crosslingual image description. In: **Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers**. 2016. p. 543-553.

STEEN, Francis; TURNER, Mark B. Multimodal construction grammar. **Language and the Creative Mind**. Borkent, Michael, Barbara Dancygier, and Jennifer Hinnell, editors. Stanford, CA: CSLI Publications, 2013.

STEEN, Francis F. et al. Toward an infrastructure for data-driven multimodal communication research. **Linguistics Vanguard**, v. 1, n. open-issue, 2018.

SULUBACAK, Umut et al. Multimodal machine translation through visuals and speech. **Machine Translation**, v. 34, n. 2, p. 97-147, 2020.

SUN, Chen et al. Videobert: A joint model for video and language representation learning. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision**. 2019. p. 7464-7473.

TEIXEIRA, Denise de Oliveira; MOURA, Eudardo. Chapeuzinho Vermelho na cibercultura: por uma educação linguística com multiletramentos. In: ROJO, Roxane; MOURA, Eduardo. **Multiletramentos na escola**. São Paulo: Parábola Editorial, 2012. p. 75-94.

THOMSON, Clive. Bakhtin's "Theory" of Genre. **Studies in 20th & 21st Century Literature**, v. 9, n. 1, p. 4, 1984.

TIMBERG, Bernard M. History of television talk: defining a genre. In: TIMBERG, Bernard M. **Television Talk**. Austin: University of Texas Press, 2002. p. 1-18.

TORRENT, Tiago Timponi; ELLSWORTH, Michael. Behind the Labels: criteria for defining analytical categories in FrameNet Brasil. **Veredas-Revista de Estudos Linguísticos**, v. 17, n. 1, p. 44-66, 2013.

TORRENT, Tiago Timponi, ELLSWORTH, Michael, BAKER, Collin, and MATOS, Ely E. d. S. The Multilingual FrameNet Shared Annotation Task: a Preliminary Report. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Paris: European Language Resources Association (ELRA), 2018

TORRENT, Tiago Timponi; MATOS, Ely E. d. S.; BELCAVELLO, Frederico; VIRIDIANO, Marcelo; GAMONAL, Maucha A; COSTA, Alexandre D. d.; and Marim, Mateus C. Representing context in FrameNet: A multidimensional, multimodal approach. **Frontiers in Psychology**, 13, 2022.

TORRENT, Tiago Timponi; MATOS, Ely Edison da Silva; COSTA, Alexandre Diniz da; GAMONAL, Maucha Andrade; PERON-CORRÊA, Simone; PAIVA, Vanessa Maria Ramos Lopes. A Flexible Tool for a Qualia-Enriched FrameNet: The FrameNet Brasil WebTool. **Language Resources and Evaluation**, forthcoming.

TURCHYN, Sergiy et al. Gesture Annotation with a Visual Search Engine for Multimodal Communication Research. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. 2018.

WADE, Anne Marit. Travel Series as TV Entertainment: Genre characteristics and touristic views on foreign countries. **MedieKultur: Journal of media and communication research**, v. 25, n. 46, p. 17 p.-17 p., 2009.

WANG, Chengyi et al. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. 2020. p. 9161-9168.

WEISS, Ron J. et al. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. **Proc. Interspeech 2017**, p. 2625-2629, Stockholm, 2017.

WU, Di et al. Deep dynamic neural networks for multimodal gesture segmentation and recognition. **IEEE transactions on pattern analysis and machine intelligence**, v. 38, n. 8, p. 1583-1597, 2016.

XU, Tao et al. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2018. p. 1316-1324.

YANG, Zhishen; OKAZAKI, Naoaki. Image Caption Generation for News Articles. In: **Proceedings of the 28th International Conference on Computational Linguistics**. 2020. p. 1941-1951.

ZHANG, Han et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. **IEEE transactions on pattern analysis and machine intelligence**, v. 41, n. 8, p. 1947-1962, 2018.

ZHENG, Dongping; NEWGARDEN, Kristi; YOUNG, Michael F. Multimodal analysis of language learning in World of Warcraft play: Languaging as values-realizing. **ReCALL**, v. 24, n. 3, p. 339-360, 2012.

ZHENG, Renjie et al. Ensemble Sequence Level Training for Multimodal MT: OSU-Baidu WMT18 Multimodal Machine Translation System Report. In: **Proceedings of the Third Conference on Machine Translation: Shared Task Papers**. 2018. p. 632-636.

ZHU, Minfeng et al. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. 2019. p. 5802-5810.