Framed Multi30K: A Frame-Based Multimodal-Multilingual Dataset

Marcelo Viridiano[1], Arthur Lorenzi Almeida[1], Tiago Timponi Torrent[1,2], Ely Edison da Silva Matos[1], Adriana Silvina Pagano[2,3], Natália Sathler Sigiliano[1], Maucha Gamonal[1,3], Helen de Andrade Abreu[1], Lívia Vicente Dutra[1,4], Mairon Samagaio[1], Mariane Carvalho[1], Franciany Campos[1], Gabrielly Azalim[1], Bruna Mazzei[1], Mateus Oliveira[1], Ana Carolina Luz[1], Livia Padua Ruiz[1], Júlia Bellei[1], Amanda Pestana[1], Josiane Costa[1], Iasmin Rabelo[3], Anna Beatriz Silva[3], Raquel Roza[3], Mariana Mota[3], Igor Oliveira[3] and Márcio Freitas[3]

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

Abstract

This paper presents Framed Multi30K (FM30K), a novel frame-based Brazilian Portuguese multimodal-multilingual dataset which i) extends the Multi30K dataset (Elliott et al., 2016) with 158,915 original Brazilian Portuguese descriptions, and 31,104 Brazilian Portuguese translations from original English descriptions; and ii) adds 4,577,122 frame and frame element labels to the 158,915 English descriptions and to the ones created for Brazilian Portuguese; (iii) extends the Flickr30k Entities dataset (Plummer et al., 2015) with 169,560 frames and Frame Elements correlations with the existing phrase-to-region correlations.

Keywords: dataset, multimodality, multilinguality, FrameNet

1. Introduction

Over the past decade, the fields of Computer Vision (CV) and Natural Language Processing (NLP) have witnessed a surge in datasets that combine image and language to solve a diverse range of tasks achievable only by leveraging information from both modalities. Some of these datasets, like MS-COCO (Lin et al., 2014), Flickr8k (Hodosh et al., 2013), and Flickr30K (Young et al., 2014), have become benchmark English image-caption datasets for tasks at the intersection of NLP and CV, such as Image Captioning and Multimodal Machine Translation (Uppal et al., 2022).

For that reason, many researchers have been expanding these image description datasets for languages other than English, such as the extensions of MS-COCO to German, (Hitschler et al., 2016), Japanese (Miyazaki and Shimizu, 2016; Yoshikawa et al., 2017), and Chinese (Li et al., 2019); the extension of Flickr8K to Chinese (Li et al., 2016); and the extensions of Flickr30K to German (Elliott et al., 2016), French (Elliott et al., 2017), Dutch (van Miltenburg et al., 2017), Chinese (Lan et al., 2017), Czech (Barrault et al., 2018), Japanese (Nakayama et al., 2020), Turkish (Citamak et al., 2020), and Ukranian (Saichyshyna et al., 2023).

However, in general, these multimodal datasets still lack the semantic representation of several contextual aspects that are only made possible by the combination of visual and textual modalities with the data from fine-grained semantic databases. Even the extensions that propose phrase-to-region correlations, linking lexical items in the description with their corresponding entities in the image – such as the Flickr30K Entities dataset (Plummer et al., 2015) – or extensions that use multimodal embeddings to represent word senses for visual disambiguation – like the VerSe extension to MS-COCO (Gella et al., 2016) – lack the means to represent fine-grained contextual information.

FrameNet (Fillmore and Baker, 2009) is one such fine-grained semantic database. In a FrameNet, the semantics of lexical items in a sentence is represented in terms of a frame: a scene featuring participants and props against which meaning is to be construed. For instance, a verb like *cook* would have its meaning in the sentence (1) relativized to the `Cooking_creation` frame in Figure 1. Therefore, the sentence (1) could be annotated for the participants – or Frame Elements (FEs) – in the `Cooking_creation` frame.

Belcavello et al. (2020) argue in favor of extending the FrameNet model to the multimodal domain, claiming that, similarly to lexical items in a sentence, visual elements may also evoke frames or work complementarily to the verbal language communicative mode in meaning making processes. Torrent et al. (2022) propose that such an exten-

LREC-COLING 2024, pages 7438–7449
20-25 May, 2024.

sion improves the capacity of the FrameNet model to represent contextual information.

Cooking_creation [@Action] [@Food] [@Lexical] [#243]

Definition

This frame describes food and meal preparation. A Cook creates a Produced_food from (raw) Ingredients. The Heating_Instrument and/or the Container may also be specified.

Core Frame Elements

FE Core:

Cook
semantic_type: @sentient — The Cook prepares the Produced_food.

Produced_food — The Produced_food is the result of a Cook's efforts.

Non-Core Frame Elements

Container
semantic_type: @container — It identifies the Container that holds the food being produced.

Degree
semantic_type: @degree — It identifies the Degree to which an event occurs.

Duration — For how long the cooking process lasts.

Heating_instrument
semantic_type: @physical_entity — It identifies the Heating_Instrument with which the Cook prepares the Produced_food.

Ingredients — It identifies the Ingredients which are altered by the Cook to create the Produced_food.

Figure 1: The `Cooking_creation` frame.

Therefore, given that (1) is a description extracted from the Flickr30K corpus for the image in Figure 2, the image could also be annotated for the FEs in the `Cooking_creation` frame.

(1) [A woman in a blue plaid shirt and white apron]$^{\text{Cook}}$ **cooks** [food]$^{\text{Produced\_food}}$ [in a pot]$^{\text{Container}}$ next to vegetables and rice.

Figure 2: Image 2661118494.jpg from the Flickr30K dataset, annotated for FEs in the `Cooking_creation` frame.

Based on these motivations, in this paper, we present Framed Multi30k (FM30K): a new frame-based multilingual multimodal dataset, which correlates frames and frame elements to entities in images. More specifically, we expand both Multi30K and Flickr30K Entities by adding:

1. 158,915 new Brazilian Portuguese original descriptions to Multi30K (Elliott et al., 2016);
2. 31,104 Brazilian Portuguese translations to the original English descriptions, aligned with the 31,014 German translations of Multi30K (Elliott et al., 2016);
3. 4,577,122 frame labels associated to the image descriptions in both English and Brazilian Portuguese;
4. 169,560 frames and FE correlations to the part of the existing co-reference chains and manually annotated bounding boxes of the Flickr30K Entities dataset (Plummer et al., 2015).

In the remainder of this paper, we discuss, in section 2, the related work and datasets used as base for building FM30K. Next, in section 3, we explain the methodology and process for compilation, translation, annotation of the corpus, and how Frame Semantics and its computational implementation – FrameNet – incorporate perspective to the core of semantic representations. Then, in section 4, we discuss information aspects of this new frame-based dataset, while in section 5 presents our conclusions and further work.

2. Related Work

In this section, we introduce the multilingual-multimodal datasets that form the basis of our new corpus.

The Multi30K (Elliott et al., 2016) and Flickr30K Entities datasets (Plummer et al., 2015) – both extensions of Flickr30K (Young et al., 2014) – are the primary sources for FM30K. Flickr30K contains 31,783 photos of everyday activities and events, each paired with five different English captions describing entities and events in each image.

The Multi30K dataset is the multilingual expansion of Flickr30K into multiple languages. For FM30K, we followed the work already carried out for the original German expansion (Elliott et al., 2016), as well as for the French (Elliott et al., 2017) and Czech (Barrault et al., 2018) expansions of Multi30K. Such expansions consist of 31,104 German, French and Czech translations of English descriptions - one per image, produced by professional translators – and 155,070 original descriptions in German, created by native speakers independently from the original English descriptions.

The Flickr30K Entities dataset (Plummer et al., 2015), in turn, extends Flickr30K by adding

image-to-text correlations, with manually annotated bounding boxes that assign region-to-phrase correspondences – each categorized into eight coarse-grained conceptual types: people, body parts, animals, clothing, instruments, vehicles, scene, and other. The dataset features 244,035 co-reference chains that link mentions to the same entities in the images to corresponding noun phrases in the five descriptions associated with each image.
Taken in combination, Multi30K and Flickr30K Entities provide the set of data exemplified in Figure 3.

EN: A boy dives into a pool near a water slide.
DE: Ein Junge taucht in der Nähe einer Wasserrutsche in ein Schwimmbecken.
FR: Un garçon plonge dans une piscine près d'un toboggan.
CS: Chlapec skáče do bazénu poblíž skluzavky.

Figure 3: Image 213216174.jpg from the Flickr30K dataset, with one of the source descriptions in English, and its translations into German, French, and Czech. Correlations between image regions and noun phrases in the English description are indicated by colors.

3. The Framed Multi30K Dataset

This section presents the steps for building the Framed Multi30K dataset. This process includes both the expansion of Multi30K into Brazilian Portuguese and the annotation of descriptions and bounding boxes in Flickr30K Entities for frames and frame elements[1].

[1]While Torrent et al. (2022) and Viridiano et al. (2022) discuss the initial rationale behind the Framed Multi30K Dataset, they omit specific details about the dataset itself. This paper marks the report on the release of FM30K (publicly available at https://github.com/FrameNetBrasil/framed-multi30k), offering comprehensive statistics regarding annotation, conducting comparative analysis of the Brazilian Portuguese original and translated descriptions – juxtaposed with existing descriptions in the Multi30K dataset, – and, finally, presenting the frames and Frame Element correlations with Flickr30K Entities.

3.1. Dataset Preparation Process

The first step in the creation of FM30K was the collection and organization of the datasets that we expand upon. All of the original Flickr30K contents were loaded into a database for this project, linking the 158,915 English sentences to the 31,783 images they describe and assigning the image *names+sentence* numbers as sentence IDs. The next step was to include the 31,014 German translations from Multi30K and link them to their English counterparts. To achieve that, we used the data splits from Multi30K to obtain triples containing the sentences in each language and the image that they describe. We then queried the original Flickr30K for the sentence and linked the German sentence to the English sentence ID. With this part of the database, we flagged the English sentences that had a German translation as references for our translation task.
For the frame and FE annotation tasks, data from Flickr30K Entities was imported into the same structure. Each bounding box was linked to a single image and to up to five sentences, depending on whether it was grounded or not in a noun phrase contained in the sentence. This relation between a sentence and a bounding box always specifies where in the sentence the image entity is being referred to. This structure allowed us to have FEs annotated on top of a (bounding box, sentence span) tuple or only over bounding boxes.

3.2. Brazilian Portuguese Translations

English to Brazilian Portuguese translations (PTT) were done by a total of 28 university student annotators with advanced English proficiency, split into two groups. The permanent annotation group comprised 12 annotators hired for this task and was responsible for creating 23,074 PTT descriptions out of the 31,014, or 74.3% of the total. The remainder of the captions – 25.7% of the PTT descriptions – were created by a group of 16 students who enrolled in hands-on linguistic annotation workshops in exchange for academnic credit hours. Both groups underwent training in the annotation task during a 15-hour workshop, conducted by some of the authors, where they were instructed in the use of the annotation tool, engaging in hands-on practice sessions with test subsets of the original corpus. Additionally, annotators also participated in weekly alignment meetings with the authors, providing them with a platform to ask questions and seek clarification on any issues that arose during the annotation process. The quality of annotations was evaluated

both manually – by periodical checks of subsets of the annotations – and with automated methods – for issues such as typos, missing periods, incorrect spacing and casing, adjuncts inside parenthesis, forward slashes indicating conjunctions. Best-performing annotators were offered permanent positions in the annotation team. This was made so as to ensure that high-quality annotations compose most of the dataset. Low-performing annotators received additional feedback during the weekly meetings. In case annotations still did not meet the guidelines, those annotations were discarded and annotators were removed from the task.

To ensure the alignment between Multi30K German, French and Czech translations and Brazilian Portuguese translations, we selected the same subset of 31,014 English original descriptions used by Multi30K for the German translation task. Brazilian Portuguese translators also followed the same methodology used in the translation task of the Multi30K dataset – while seeing the image and the English original description, annotators were asked to produce a correct and fluent translation of the image description into Brazilian Portuguese.

3.3. Brazilian Portuguese Original Descriptions

The original descriptions in Brazilian Portuguese (PTO) were created by a total of 148 university students, once again split in the same configuration of groups. The permanent annotation team featured 22 students and produced 81,834 (51.5%) of the 158,915 PTO descriptions – five per image of the Flickr30K dataset. The permanent annotation team was compensated equally to that of the PTT creation task. The remainder of the captions were created by 126 students enrolled in hands-on annotation workshops granting academic credits, with an average of 612 PTO descriptions created per student. Once again, both groups were instructed in the task for 15 hours before engaging with the work.

Once again following the same methodology used for the creation of German, French and Czech descriptions in the Multi30K dataset, students were presented with a translated version of the data collection interface originally developed by (Hodosh et al., 2013). The instructions for the sentence creation task were translated from English to Brazilian Portuguese by one of the authors. To prevent fatigue and ensure the quality of descriptions, each student was assigned a weekly quota of approximately fifty sentences per hour of work.

To ensure the quality of the new original descriptions, both manual and automated inspection methods were used. First, we looked for any duplicate descriptions of the same image and replaced them with a new sentence, created by another annotator. This was the case of 76 sentences of the total (less than .1%). Another group of sentences that were replaced were those with less than 4 words. It was comprised of 1,635 sentences, which was slightly over 1% of the total. Finally, 168 other sentences were inspect because they contained special characters. These included issues such as typos, adjuncts in parenthesis and forward slashes indicating conjunctions. These sentences were either re-annotated or edited, when the change would not alter its meaning. Other issues were automatically fixed, such as missing periods and incorrect spacing and casing.

The creation of PTT and PTO descriptions adds another language to the Multi30K dataset and is the first contribution of the FM30K dataset. We now turn to the second contribution: the enrichment of both images and descriptions with FrameNet-like semantic annotation.

3.4. Automatic Frame Semantic Role Labeling of Image Descriptions

To enrich the FM30K dataset sentences with explicit semantic information, we opted to use pretrained LOME (Xia et al., 2021). LOME (Large Ontology Multilingual Extraction) is a system developed for mulitingual information extraction that uses a FrameNet parser to identify textual entities and events. It also performs co-reference resolution, fine-grained entity typing and temporal relation prediction of events. For this release, we have included frame and frame element information.

One of the main advantages of LOME over similar systems is that it performs full parsing, instead of only labeling the frame elements of a given frame in a sentence. The FrameNet parser being trained over XLM-R (Conneau et al., 2020) representations also allows the model to learn representations in one language and extrapolate it to others. For this work, we used a version of LOME that was trained not only on Berkeley FrameNet 1.7 *fulltext* annotations, but also in 8558 *fulltext* annotations from FrameNet Brasil. Because of that, it include FrameNet 1.7 and FrameNet Brasil frame and FE labels.

The FrameNet parser in LOME works by first encoding input sentences into a list of vectors where each position represents a token. Each vector represents different parts of an annotation, such as token XLM-R embeddings, span boundaries, labels and parent indices (required by FEs, to identify the token evoking their frame). The vector list is then fed to a BIO tagger that identifies trigger spans. These trigger spans are labeled by a typing module. This process is ran twice for every sentence: in the first iteration, the tagger identifies frame evoking spans and labels their frames;

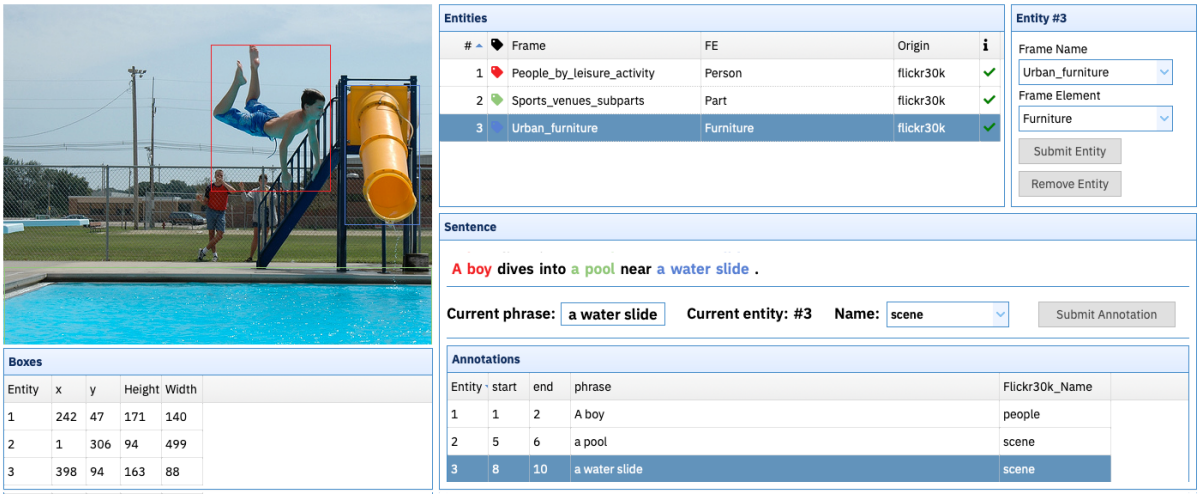


Figure 4: Interface of the annotation tool used to assign frames and frame elements to the phrase-to-region data of the Flickr30K Entities dataset.

the second iteration does the same thing, but conditioned to the already predicted labels, to predict FEs. Since FEs belong to a single frame, during training, this constraint is enforced by maximizing loss every time the model predicts an FE with an invalid parent frame.

FM30K includes frame parsing with LOME for all the 158,915 English original descriptions (ENO) in Flickr30K, as well as for the 31,104 PTT and the 158,915 PTO descriptions added to Multi30K by the research effort reported on in this paper. The number of frame and FE labels associated to each group of descriptions is shown in Table 1.

	frames/FEs	**avg. p/ sent**
ENO	2,073,114	13.0
PTO	2,131,036	13.4
PTT	372,972	12.0
total	4,577,122	-

Table 1: Counts of frame labels included for each split of the FM30K dataset and the averages per sentence.

3.5. Human Annotation of Images for Frames and Frame Elements

As pointed out in section 2, the Flickr30K Entities dataset associates bounding boxes to the 31,783 images on Flickr30K based on the entities referred to in the ENO descriptions by noun phrases. Because each image is associated with five different ENO descriptions, the number of bounding boxes correlated to NPs in the description varies. For the task of assigning frames and FEs to the manually annotated phrase-to-region correlations from the Flickr30k Entities dataset, we selected 29,920 sets of bounding boxes derived from the 31,104 ENO descriptions which were used as source for the creation of the PTT descriptions.

The image annotation task was carried out using Charon (Belcavello et al., 2022), a tool for annotating multimodal corpora with FrameNet categories. Figure 4 shows the interface for the static image annotation mode, used to annotate image-text pairings.

The top left corner of the annotation interface displays the reference image from Flickr30K. The 'Boxes' panel shows the coordinates for the manually annotated bounding boxes – obtained from the Flickr30K Entities dataset – and the number associated with each entity being annotated for that image. The 'Entities' panel displays the correlations between each entity in the image, its correspondent noun phrase descriptor – color coded with the description seen on the 'Sentence' panel –, and the frame and frame element assigned to that image-text pairing by the annotator, using the 'Entity' panel on the top right corner of the interface.

Figure 4 also shows the resulting annotation for the sentence "A boy dives into a pool near a water slide". In that sentence, the noun phrases "A boy", "a pool", and "a water slide" are co-referenced with three distinct entities in the image. For the noun phrase "A boy", corresponding to Entity 1, the annotator assigned the frame `People_by_leisure_activity` and the FE Person. For the noun phrase "a pool", corresponding to Entity 2, the annotator selected the frame `Sports_venues_subparts` and the FE Part. Finally, for the noun phrase "a water slide", corresponding to Entity 3, the annotator chose the frame `Urban_furniture` and the FE Furniture.

A slightly modified version of the annotation inter-

face in Figure 4 was also used. In this version, the 'Sentence' panel was not displayed to annotators. The two versions of the annotation interface allowed for the production of frame and FE annotation for bounding boxes in the images under two conditions: with the presence of the description associated to the image and without it. The motivation behind this annotation design was to evaluate the influence of the description on the semantic representation of the images resulting from the annotation. Preliminary experimental results reported on by Viridiano et al. (2022) indicate that such an influence is statistically relevant.

For both annotation conditions, annotators were instructed to associate a frame and a FE to each of the bounding boxes listed seen on the 'Entities' panel, provided that they were visible in the image. In the regular cases, each entity is associated to a set of coordinates – shown on the 'Boxes' panel – delimiting a bounding box. However, in some cases, the entity to be annotated was the whole background of the picture.

Annotators were explicitly instructed to not annotate entities which were not visible in the image. For example, in the case of Figure 5, which is paired with the description " A man standing on a stage playing a guitar and a harmonica waving to the crowd.", the Flickr30K Entities dataset attributes a region-to-phrase correlation between "the crowd" and an entity that is not visible in the image. In cases like this, annotators were explicitly instructed to not assign frames and FEs to those correlations. For those phrase-to-region correlations where the entity is shown in the image – "A man", "a stage", "a guitar", and "harmonica" – frames and frame elements were assigned.

EN: A man standing on a stage playing a guitar and harmonica waving to the crowd.

Figure 5: Image 4827958485.jpg paired with the description " A man standing on a stage playing a guitar and a harmonica waving to the crowd." in the Flickr30K Entities dataset, where the noun phrase "the crowd" is correlated to an entity that is not shown in the image.

Annotators were also instructed to not annotate for NPs describing events. Hence, they should not seek to annotate the entities for the NPs in bold in (2), for example.

(2) A group of people attending either **a concert** or **a party**.

Each setup of the visual annotation task – with the presence of the description (VWC) and without it (VNC) – was carried out for 29,920 out of the 31,104 sets of bounding boxes associated to the ENO descriptions used as source for creating the PTT sentences. In the end, 29,831 sets had at least one annotation. Not all of the 29,920 sets of bounding boxes were annotated because some of them had issues such as no entity bounding boxes, incorrect bounding boxes or no adequate frame for the entity.

Both conditions of the image annotation task were carried out by the same setup of university student teams of annotators described in sections 3.2 and 3.3. The permanent team for this task featured 16 students compensated in the same manner already described. The pool of students enrolled in the hands-on annotation workshops was composed of 32 students. The total number of annotations generated from this task in each condition is given in Tables 2 and 3. The total number of frame and FE correlations annotated was of 169,560.

	frames/FEs	**avg. p/ student**
Permanent	49,475	2,248.9
Workshop	38,063	302.0
total	87,538	-

Table 2: Counts of frame and FE labels annotated for the images in FM30K dataset by each annotation team and the averages per student annotator under the VWC condition.

	frames/FEs	**avg. p/ student**
Permanent	65,538	2,979.0
Workshop	16,484	130.8
total	82,022	-

Table 3: Counts of frame and FE labels annotated for the images in FM30K dataset by each annotation team and the averages per student annotator under the VNC condition.

We now turn to the discussion of FM30K in relation to both the original datasets it expands upon and the semantic representations built.

4. Discussion

As it is, FM30K further expands data that is already applicable to a multitude of NLP tasks. It adds semantic information to existing bounding boxes that could improve object detection tasks, as well as text in Brazilian Portuguese that could be used for machine translation, automatic-image description and many other tasks. Instead of discussing these possibilities, we highlight how different languages, modes and annotation procedures impact the existing information in a dataset.

4.1. Translated vs. Independent Descriptions

Table 4 presents token and type counts, as well as other statistics, for FM30K's PTO and PTT sentences, compared to those of the other languages included in the Multi30k dataset.

Translation sentences can be organized into two different groups: (i) Brazilian Portuguese, English and French, and (ii) German and Czech, both sharing among themselves similar counts for tokens, types and singletons. The first group contains corpora with more tokens, longer sentences, but less diverse types. The second one has the opposite attributes. The PTT sentences from FM30K are on average longer than all other translations in terms of words, with the exception of French. The Brazilian Portuguese translations also have 23% more singletons than English, but less than half of Czech. In comparison to English and French, PTT sentences have good token variety and adequate lengths.

In regards to the original descriptions, the ones in Brazilian Portuguese are longer than those in English and German, in terms of both characters and words. On average, there is a 9% increase in sentence length in number of words, and a 21.6% in terms of characters, when compared to the ENO dataset. These numbers are 39.5% and 27.7% in comparison to German. When looking at type and singleton counts, PTO and ENO are very close to each other, while German has more than double the number of types and more than tree times the number of singletons.

These numbers not only show language differences, but also how original descriptions can be quite different from translations. For instance, the difference in average sentence length between the two in Brazilian Portuguese and German is of more than one word.

4.2. English vs. Brazilian Portuguese

To compare semantic representations across different languages and different modes, we follow the same methodology as in Viridiano et al. (2022). It consists of calculating cosine similarities between data pairs (sentences or images) using vectors derived from the frames annotated on each data point. These vectors are not embeddings like the ones obtained from language or image models. Instead, they are computed using a Spreading Activation (SA) algorithm over the frame network. First, the FrameNet digraph is built using frames and their relations. Then, for each sentence or image, the annotated frames serve as activation nodes in the graph. These starting nodes receive a maximum energy value that act as a weight for that frame. The algorithm works in iterations, spreading the energy from active nodes to other, related nodes and, in turn, activating then. Because in each iteration energy is decayed, the algorithm eventually ends when there is no energy left and each frame that was activate receives a weight. These weights are used to build a frame vector that is used to compute cosine similarities.

When comparing sentences in different languages, a relevant step is the selection of comparison pairs. In the case of PTT, since these sentences are translations, they are always compared to their English originals, resulting in 31,014 pairs and cosine similarity scores. For PTO, we used an heuristic approach to determine pairs. For each image in Flickr30K, the five sentence sets in each language were aligned, considering the combinations that would minimize the average difference in number of words in each pair. This was done to avoid the comparison between long sentences, that can evoke more frames, and shorter ones. The average difference in number of words for the pairs computed using this method was 2.3, with 47% of the pairs having only one extra word or the same length. After running the SA algorithm for each pair, 158,915 cosine similarity scores were obtained.

Table 5 shows the average normalized similarities for PTT and PTO when compared to ENO. As expected, translated sentences are considerable closer to their sources in a semantic space. This effect can also be observed in the distributions in Figure 6, where the PTT $\times$ ENO distribution is shifted, when compared to the more neutral ENO $\times$ PTO chart. Considering the similar variances between the distributions and their approximation to a Normal distribution, the significance of the differences was assessed using Student's t-test. The Brazilian Portuguese translations (PTT), when compared against ENO, have significantly higher similarities (M = 0.77, SD = 0.14) than the original annotations created with only the image reference (PTO) (M = 0.53, SD = 0.16), with test statistic $t(317828) = -216.41$, $p < 0.001$.

This information combined shows that semantic representations of image descriptions can change considerably depending on whether the description is a translation or not. This highlights the im-

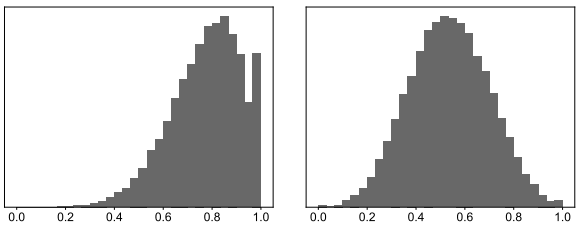
	Tokens	Types	Characters	Avg. length	Singletons
Translations (31,014 sentences)					
Brazilian Portuguese	374,097	12,212	1,749,365	12.0	5,426
English	369,848	10,500	1,523,855	11.9	4,406
German	345,326	19,363	1,834,937	11.1	11,226
French	387,536	12,129	1,796,038	12.4	5,232
Czech	281,138	23,166	1,346,020	9.0	12,369
Descriptions (158,915 sentences)					
Brazilian Portuguese	2,127,452	19,135	9,798,114	13.4	7,660
English	1,950,410	20,278	8,057,457	12.3	7,788
German*	1,482,389	44,033	7,669,557	9.6	25,229

Table 4: Corpus statistics for Translations and Descriptions in Brazilian Portuguese compared to other languages in Multi30k. * German original descriptions corpus has 155,070 sentences

portance of having both types of sentences in multilingual scenarios: the translations are more easily related to their sources, but having original sentences increases the variety of ways in which the same scene can be represented.

	ENO	
	avg. sim	stdev
PTT	0.77	0.14
PTO	0.53	0.16

Table 5: Similarity for image frame annotation setups with and without captions present.



(a) ENO × PTT (b) ENO × PTO

Figure 6: Distribution of similarity values between ENO, PTT and PTO.

4.3. Image vs. Description

Another important evaluation for the FM30K dataset is the comparison between distinct types of visual annotation. As discussed in subsection 3.5, frame and FE labels were assigned to entities under two different conditions, one in the presence of the image description (VWC) and another without it (VNC). For each of those two scenarios, the semantic vectors of each entity set from a image were computed using the same SA method. Since the entity set linked to an image is also linked to an ENO sentence, that sentence is used to compute two sets of 29,831 similarity scores. Their averages are shown in Table 6.

When compared to PTT and PTO, the difference between VWC (M = 0.49, SD = 0.17) and VNC (M = 0.42, SD = 0.18) is considerably smaller, despite also being statistically relevant, with Student's t-test statistic $t(59660) = 25.24$ and $p < 0.001$. These results are also apparent in Figure 7, where the distributions are more similar (in constrast with 6).

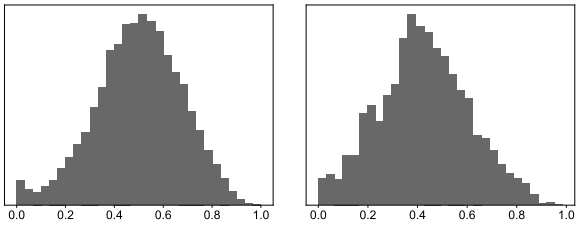
The gap between VWC and VNC similarities shows that there are shifts in semantic representation when annotators are shown sentences together with an image. Similarly to PTT, the inclusion of two modes restricts the number of possible construals and approximates the semantic representations to those of the original descriptions. Once again, it shows the importance of knowing how an annotation was created and of using data produced in different contexts to accommodate multiple perspectives.

	ENO	
	avg. sim	stdev
VWC	0.49	0.17
VNC	0.42	0.18

Table 6: Similarity for image frame annotation setups with and without captions present.

5. Conclusion(s) and Further Work

Expanding Multi30K to include Brazilian Portuguese introduces one of the world's top ten



(a) ENO × VWC (b) ENO × VNC

Figure 7: Distribution of similarity values between ENO, VWC and VNC.

most spoken languages to the dataset family, thereby addressing the under-representation of Brazilian Portuguese in the field of Natural Language Processing. This expansion also stimulates multilingual-multimodal research, opening new possibilities for shared tasks on Multimodal Machine Translation, and fostering a more diverse research landscape in NLP.

As future work for a second release, we plan on including FE data generated by the semantic parser to all English and Brazilian Portuguese sentences. Another relevant piece of information that could further enrich FM30K is the inclusion of annotations of events on images. These would further link entities within the same image as participants of the same scene and improve the semantic descriptions of the situation being presented.

We are also in the initial stages of developing a novel annotation task aimed at expanding the existing phrase-to-region correlations provided by Flicker30K Entities to Brazilian Portuguese. In this new round of annotation, annotators are invited to assigned one or more eventive frames to each image and annotate the bounding boxes in such images not with FEs indicating the entities they represent, but the role they play in the eventive frame(s) chosen for annotation. Moreover, by aligning the noun phrases from the English original captions with their respective Brazilian Portuguese translations, we expect to be able to use their manually annotated bounding boxes to also assign frames and frame elements to image-text correlations in our translated descriptions.

6. Acknowledgements

Authors acknowledge the support of the Graduate Program in Linguistics at the Federal University of Juiz de Fora. Research presented in this paper is developed by ReINVenTA - Research and Innovation Network for Vision and Text Analysis of Multimodal Objects. ReINVenTA is funded by FAPEMIG grant RED 00106/21, and CNPq grants 408269/2021-9 and 420945/2022-9. Viridiano's research was funded by CAPES PROBRAL PhD exchange grant 88887.628830/2021-00 and CAPES PROEX PhD Grant 88887.816219/2023-00. Lorenzi's research was funded by CAPES PROBRAL PhD exchange grant 88887.628831/2021-00 and CAPES PROEX PhD Grant 88887.816228/2023-00. Torrent is an awardee of the CNPq Research Productivity Grant number 315749/2021-0. Pagano is an awardee of the CNPq Research Productivity Grant number 313103/2021-6. Gamonal's research was funded by CAPES/PRINT grant 88887.936139/2024-00.

7. Ethical Considerations and Limitations

Addressing the ethical considerations that are inherent to the design and development of peer-sourced datasets is an essential step to ensure the responsible development of NLP research (Bender and Friedman, 2018). In this section, we outline the strategies we employed, during the design and development of Framed Multi30K, to keep track of factors that may have an impact on the resulting dataset.

Curation Rationale: The choice of the foundational datasets upon which FM30K expands was due to the fact that the Flickr30K family of datasets is a benchmark for many multimodal NLP tasks (Uppal et al., 2022).

In an effort to incorporate a perspectivized annotation of both images and descriptions to those datasets, we chose to augment them with frame and FE annotations. The expansion of the descriptions into Brazilian Portuguese aims to integrate one of the most spoken – yet still low resourced – of the world's languages in the Flickr30K family of datasets.

Language Variety: PTT and PTO descriptions were all produced by native speakers of Brazilian Portuguese. All of them are, at least, pursuing an academic degree. Therefore, the variety of Brazilian Portuguese which is prevalent in the dataset is the one identified with urban highly educated speakers in a monitored communicative setting.

Annotator Demographics: This dataset contains annotations from annotators recruited among the undergraduate students from the BA in Language and Linguistics (54.9%), students with a completed BA in Language and Linguistics (22.5%), students pursuing a MA in Linguistics (12.7%), and students pursuing a PhD in Linguistics (9.9%).

Among the participants, 43.7% fell within the 18-24 age group, 18.3% between 29 and 36 years old, 16.9% between 25 and 28 years old, 1.3% between 37 and 45 years old, and 9.9% above the age of 46 years old.

In terms of gender, the group was constituted by 63.4% of annotators identifying themselves as women, 33.8% identifying as men, and 2.8% as non-binary.
Regarding ethnicity, 69% self-declared as white or Caucasian, 25.3% identified as black or brown, and 1.4% as indigenous. Two annotators – representing approximately 4.3% of the group – chose not to disclose their racial or ethnic background.
Concerning the proficiency in English, 52.1% self-declared to be proficient or having advanced level of proficiency, while 25.4% declared to be at an intermediate level, and 15.5% to have a basic or beginner level of English proficiency. Only 7% reported having no proficiency in the language.

Annotators Well-Being & Compensation: In addition to defining annotation quotas aimed at accommodating the personal and academic commitments of annotators, we also ensured that annotators received a fair compensation for their contributions by paying a monthly stipend of 700.00 BRL (approximately 140.00 USD) per 20 hours of work per week. This value is 1.66% above the national minimum wage and is defined by Brazilian research funding agencies.
Before starting each task, annotators took part in a series of training sessions aimed at familiarizing them with the annotation tools, and received comprehensive written materials that provided detailed instructions and guidelines pertaining to the task. Graduate students were available to provide assistance and guidance to annotators during the whole annotation task not only in the lab, but also through an online platform – Slack – which also facilitated communication and collaboration among annotators, serving as a channel for annotators to interact, ask questions, and seek clarification on any aspect of the task.

8. Bibliographical References

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Frederico Belcavello, Marcelo Viridiano, Alexandre Diniz da Costa, Ely Edison da Silva Matos, and Tiago Timponi Torrent. 2020. Frame-based annotation of multimodal corpora: Tracking (a) synchronies in meaning construction. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 23–30.

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

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Begum Citamak, Ozan Caglayan, Menekse Kuyu, Erkut Erdem, Aykut Erdem, Pranava Madhyastha, and Lucia Specia. 2020. Msvd-turkish: A comprehensive multimodal dataset for integrated vision and language research in turkish. *Machine Translation*, pages 265–288.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. *arXiv preprint arXiv:1710.07177*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Charles J. Fillmore and Collin Baker. 2009. A frames approach to semantic analysis. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 791–816. Oxford University Press, Oxford, UK.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings.

In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–192, San Diego, California. Association for Computational Linguistics.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. pages 2399–2409.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-guided cross-lingual image captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1549–1557.

Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016. Adding chinese captions to images. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval*, pages 271–275.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790.

Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4204–4210.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Andrii Babii, and Olena Turuta. 2023. Extension Multi30K: Multimodal dataset for integrated vision and language research in Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 54–61, Dubrovnik, Croatia. Association for Computational Linguistics.

Tiago Timponi Torrent, Ely Edison da Silva Matos, Frederico Belcavello, Marcelo Viridiano, Maucha Andrade Gamonal, Alexandre Diniz da Costa, and Mateus Coutinho Marim. 2022. Representing context in framenet: A multidimensional, multimodal approach. *Frontiers in Psychology*, 13.

Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. 2022. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain. Association for Computational Linguistics.

Marcelo Viridiano, Tiago Timponi Torrent, Oliver Czulo, Arthur Lorenzi, Ely Matos, and Frederico Belcavello. 2022. The case for perspective in multimodal datasets. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 108–116, Marseille, France. European Language Resources Association.

Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. LOME: Large ontology multilingual extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, Online. Association for Computational Linguistics.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.