# Human-in-the-Loop Image Annotation for frame-based Multimodal Machine Translation

Marcelo Viridiano
/ Federal University of Juiz de Fora
/ FrameNet Brasil

AILC 2021 Student Sessions
June 16th, 2021

# Thanks to

Tiago Torrent

*(Supervisor and PI at FrameNet Brasil)*

Fred Belcavello

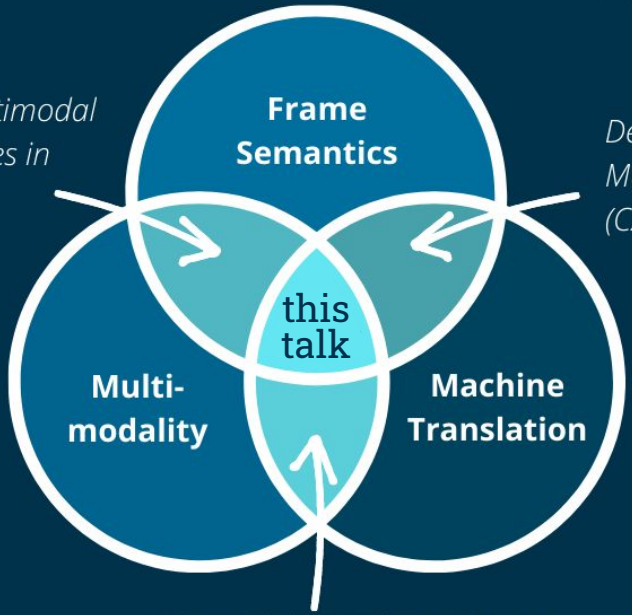
*(Senior Ph.D. student in Multimodality)*

Ely Matos

*(Researcher at FrameNet Brasil)*

# Related work

*Frame-Based Annotation of Multimodal Corpora: Tracking (A) Synchronies in Meaning Construction (Belcavello et al., 2020)*




*Designing a Frame-Semantic Machine Translation Evaluation Metric (Czulo et al., 2019)*

*Multimodal Lexical Translation (Lala and Specia, 2018)*

# What is 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.*

Fillmore (1982:111)

# What is a frame?

(1) When I was a kid, I couldn't wait till the next morning to open the presents.

This sentence invokes the Christmas frame by describing a situation that matches salient facts of Christmas practice, even though no word in it is specific to Christmas.

# What is FrameNet?

It's... a computational lexicography project started in 1997 by Charles J. Fillmore at the International Computer Science Institute, in Berkeley, CA

in the form of... a network of frames as a model of linguistic cognition that accounts for the meaning of Lexical Units
*(and, later on, other sorts of linguistic structures)*

where we... annotate corpus to attest the proposed analyses.

# Touring

## Definition

A Tourist experiences a specific Attraction, a specific Place with a unique history or other societally recognized individual character, with the goal of seeing and learning about it. Typically, the Attraction has a source of information about the Attraction like a guide, pamphlets, or displays.

## Example(s)

## Core Frame Elements

**FE Core:**

| | |
|---|---|
| Attraction [Attraction] | The societally-recognized unique location that the Tourist experiences. |
| Place [Place] | The location where the touring takes place. |
| Tourist [Tourist] | The individual that is seeking an experience at an Attraction. |

# Touring

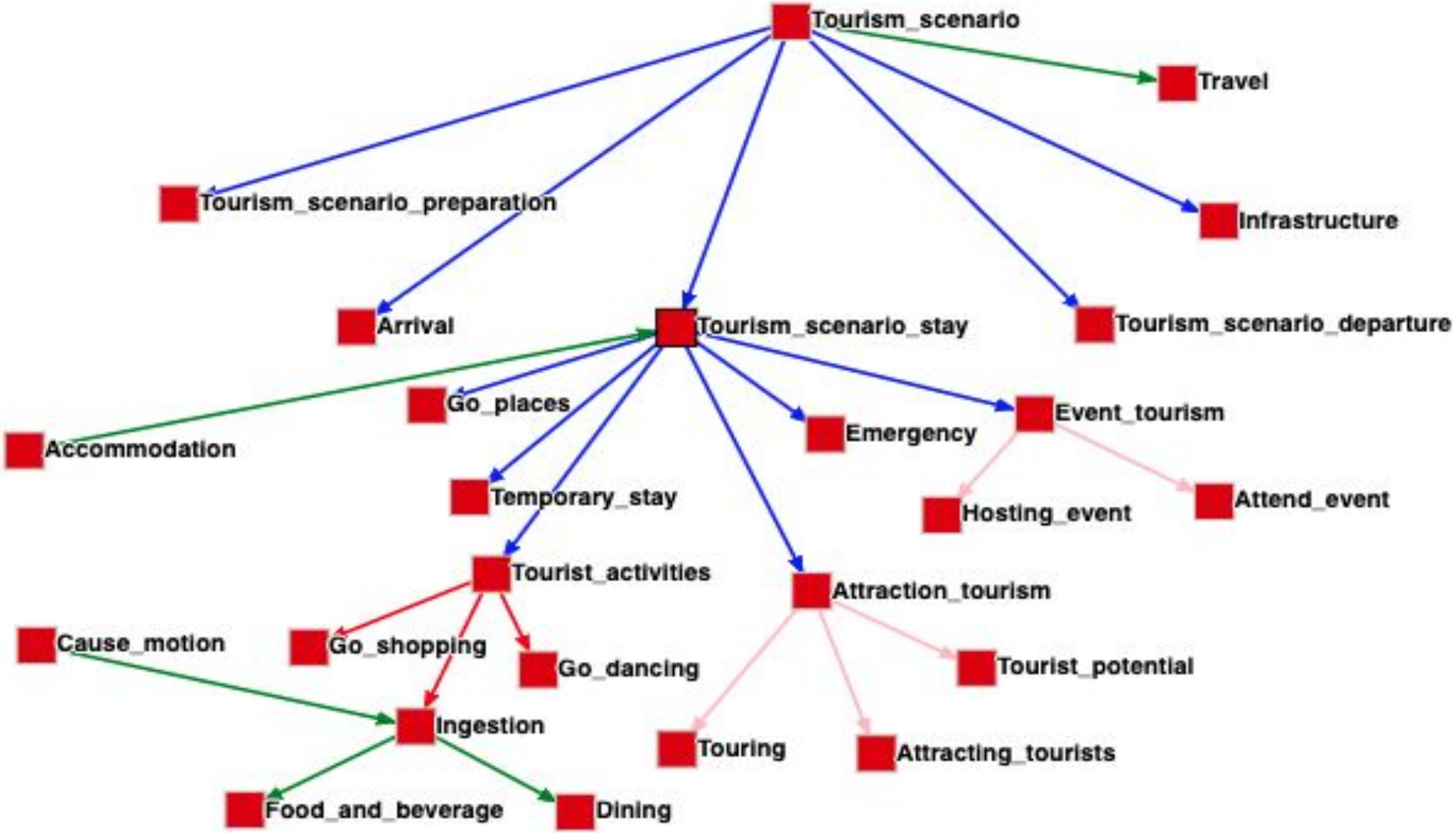
**Definition**

A Tourist experiences a specific Attraction, a specific Place with a unique history or other societally recognized individual character, with the goal of seeing and learning about it. Typically, the Attraction has a source of information about the Attraction like a guide, pamphlets, or displays.

*FE_Place* *FE_Attraction*

(2) In Rome, most foreigners do the Colosseum tour.

*FE_Turist* *Lexical Unit that evokes this frame*


Tourism_scenario
Travel
Tourism_scenario_preparation
Infrastructure
Arrival
Tourism_scenario_stay
Tourism_scenario_departure
Go_places
Accommodation
Emergency
Event_tourism
Temporary_stay
Hosting_event
Attend_event
Tourist_activities
Attraction_tourism
Cause_motion
Go_shopping
Go_dancing
Tourist_potential
Ingestion
Touring
Attracting_tourists
Food_and_beverage
Dining

# How can we improve FrameNet?

*a structure of*

Lexical Units + Frame Elements + Frame Relations

+

Qualia Relations

# Qualia

> *Indicate a single aspect of a word's meaning, defined on the basis of the relation between the concept expressed by the word and another concept that the word evokes.*

Pustejovsky and Jezek (2016)

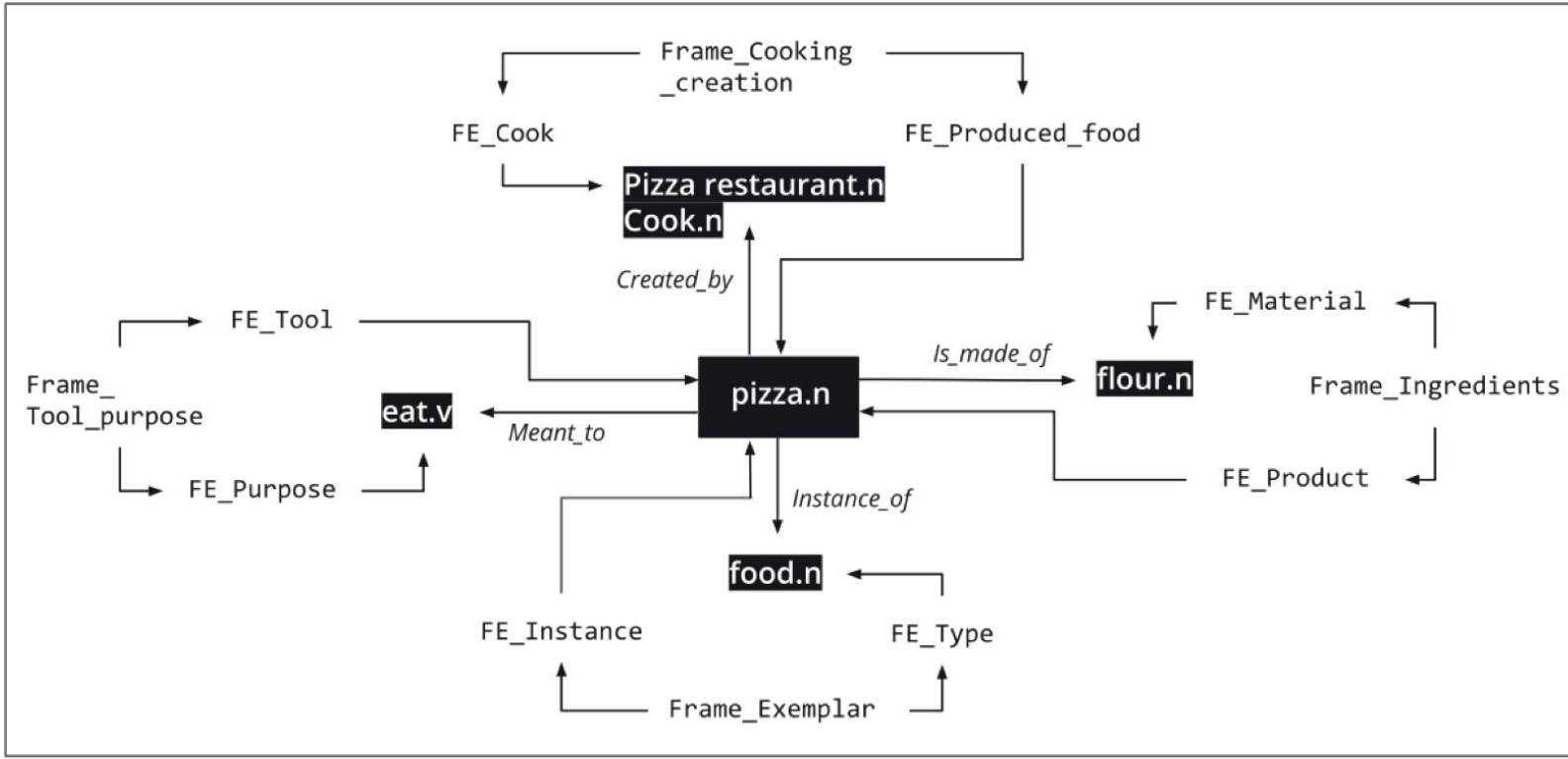
# Qualia

The Generative Lexicon Theory originally proposes four qualia roles

- Formal *(is a type of)*
- Telic *(is meant to)*
- Constitutive *(is made of)*
- Agentive *(is create by)*

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

# Ternary Qualia (mediated by Frames)



## Qualia Structure

Select Qualia Type | Search Frame

| Type | Frame | LU1 | Info | LU2 |
|---|---|---|---|---|
| Qualia Agentive | Intentionally_affect# | Agent | agentive_affect | Patient |
| Qualia Agentive | Causation# | Actor | agentive_cause | Effect |
| Qualia Agentive | Agir_intencionalmente | Agente | agentive_cause | Ação |
| Qualia Agentive | Causation# | Effect | caused_by | Cause |
| Qualia Agentive | Intentionally_create# | Created_entity | created_by | Creator |
| Qualia Agentive | Inovar | Nova_ideia | created_by | Conhecedor |
| Qualia Agentive | Cooking_creation# | Produced_food | created_by | Cook |
| Qualia Agentive | Resolve_problem# | Agent | resolves | Problem |
| Qualia Constitutive | Objective_influence# | Influencing_entity | affects | Dependent_entity |
| Qualia Constitutive | Causation# | Actor | causes_naturally | Affected |
| Qualia Constitutive | Infraestrutura | Infraestrutura | has | Usuário |
| Qualia Constitutive | Pessoas_por_origem | Pessoa | has_origin | Origem |
| Qualia Constitutive | People_by_religion# | Person | is_a_follower_of | Religion |
| Qualia Constitutive | Agir_intencionalmente | Ação | is_constitutive_activity_of | Agente |
| Qualia Constitutive | Atributos | Atributo | is_constitutive_attribute_of | Entidade |
| Qualia Constitutive | Conter | Conteúdo | is_in | Contêiner |

# What can we do with FrameNet?

Since FrameNet annotation can also take full texts as targets, we can use it for a number of computational tasks like:

- Information Extraction (Xia et al., 2020)
- Relation Extraction (Zhao et al., 2020)
- Sentence Similarity (Liu et al., 2020)
- Narrative Completion (Ou et al., 2021)
- Event Framing (Vossen et al., LREC'20)
  (Remijnse & Minemma, IFNW'20)

and also...

# Multimodal Machine Translation

> *(...) the task of translating text using information from other modalities (such as images) as auxiliary cues.*

Lala & Specia (2018)

# Multimodal Machine Translation

The motivation to combine **Multimodality** with **Frame Semantics** comes from the hypothesis that, as words in a sentence evoke frames, information from other modalities (like images) may play a role in frame evocation.

# Words in sentences can have multiple meanings

He took the airplane

Is this sentence about take.v as in "taking"...

*An Agent removes a Theme from a Source so that the Theme is in the Agent's possession*

...or is it about take.v as in "vehicle ride"?

*A Theme is moved by a Vehicle which is not directly under their power*

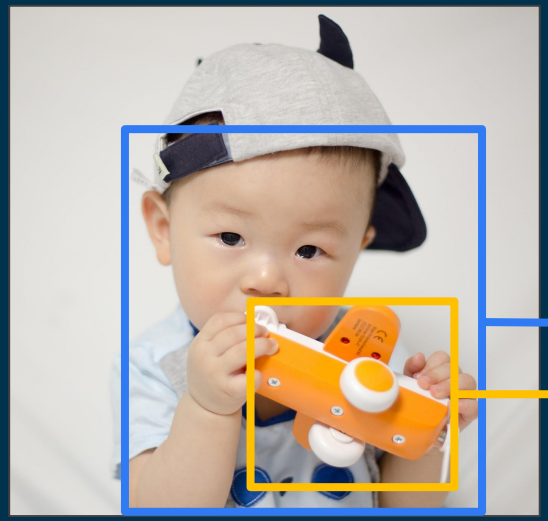



Photo by Minnie Zhou on Unsplash

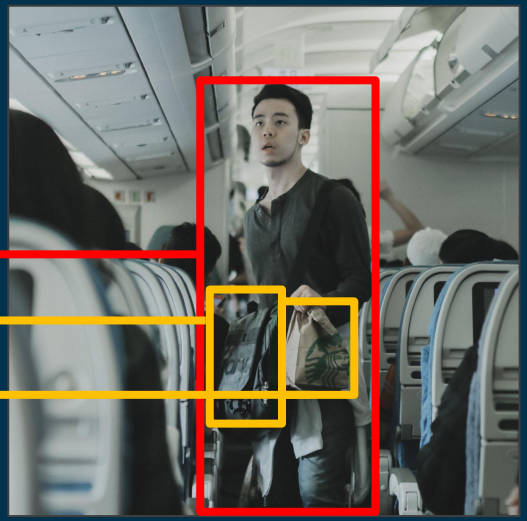



Photo by Lutfi Gaos on Unsplash

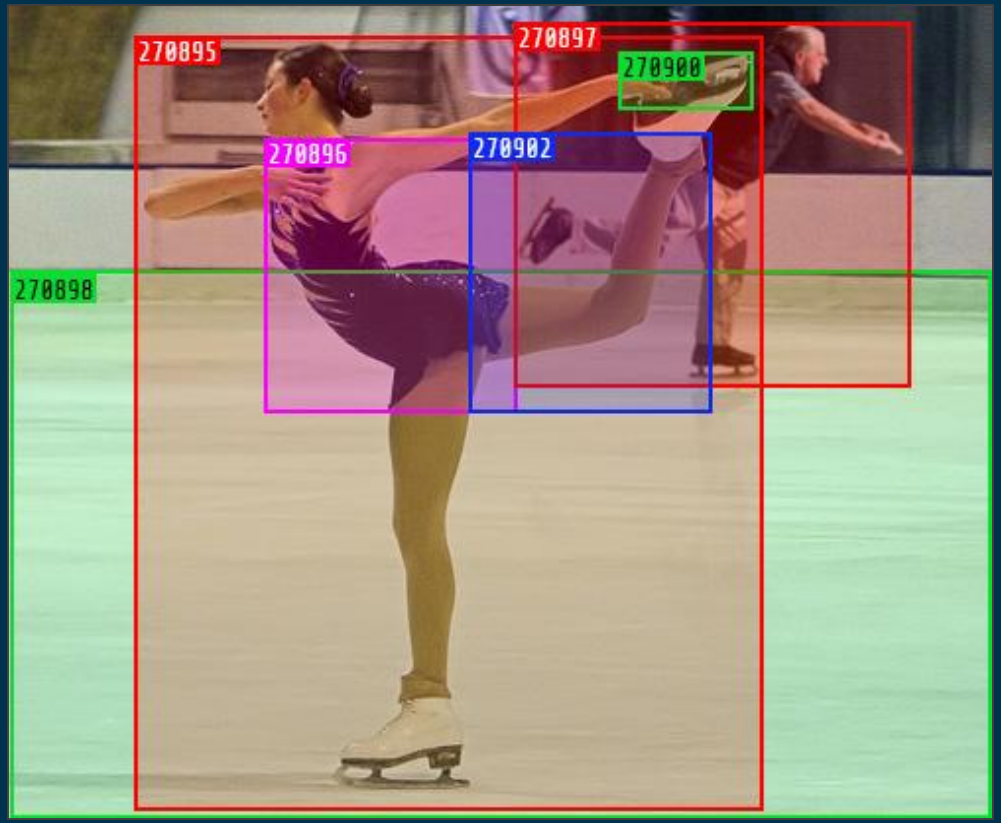
# Preliminary findings showed that words extracted from objects in images can help solve this ambiguity

How can we combine the semantic information from images with the network of semantic relations from FrameNet to help solve ambiguity problems in translations and improve Machine Translation algorithms?

In other words, how can we bring fine grained semantics into Multimodal Machine Translation?

# FIRST...

We use manually annotated images[1] to gather metadata in multilingual datasets containing image-text pairings.

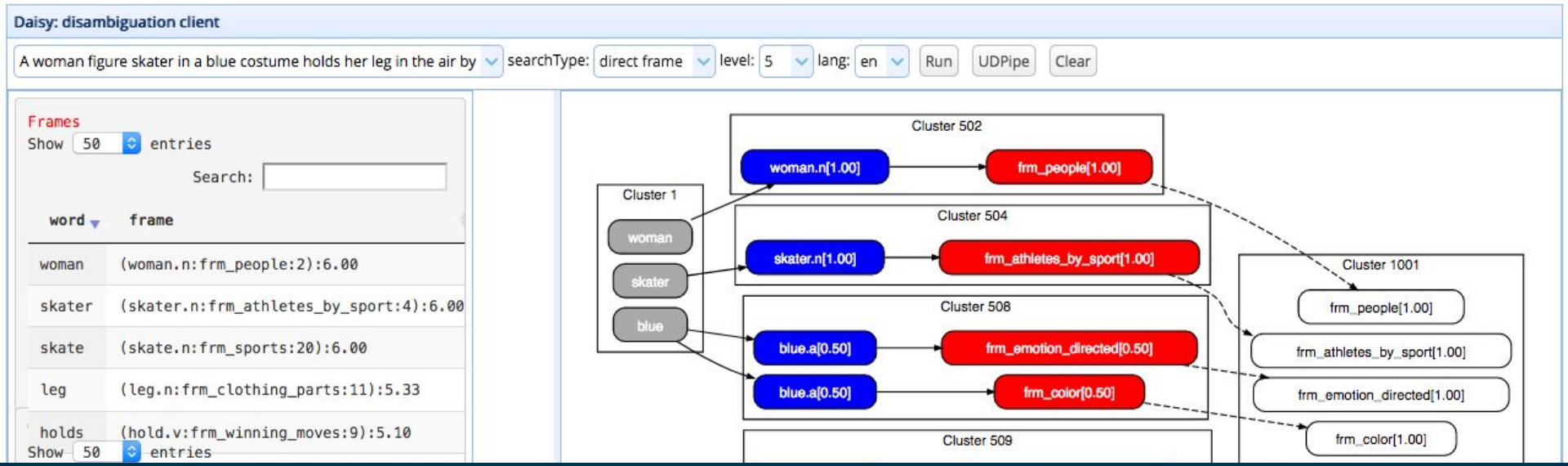



Source: Flickr 30K dataset

1. Flickr30K Entities (Plummer et al., 2015)

# ...THEN

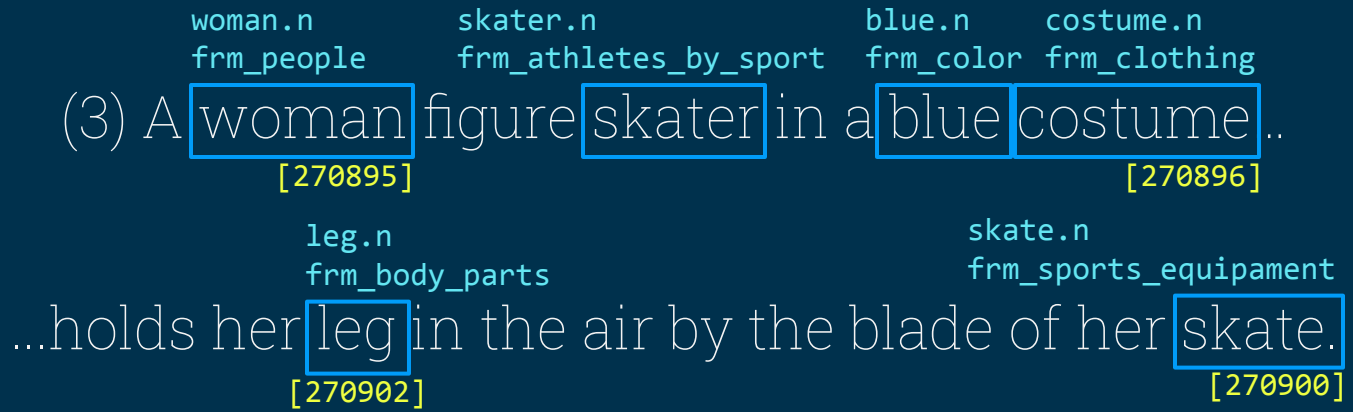
We use a semantic parser[2] to analyze image captions and check which objects in the images match FrameNet Lexical Units.




2. DAISY (Matos & Salomão, 2016)

# ...THEN

We use a semantic parser[2] to analyze image captions and check which objects in the images match FrameNet Lexical Units.




2. DAISY (Matos & Salomão, 2016)

We're currently expanding this dataset into Brazilian Portuguese, adding this language to the Multi30K corpus. We'll then align the captions in BR-Pt to the manual annotation of Flickr30K Entities and use the resulting dataset to train a frame-based Machine Translation algorithm we've been developing at FrameNet Brasil

# Thank you!

You can download this presentation at viridiano.com

globalframenet.org

ufjf
Universidade
Federal de Juiz de Fora