Anotação do Dataset Multimodal da ReINVenTA

Ana Carolina Loçasso Luz[1,2], Gabrielly Braz[1], Lívia Pádua Ruiz[1,2], Mariane de Carvalho Pinto[1], Frederico Belcavello[1], Natália Sathler Sigiliano[1], Tiago Torrent[1,2]

[1] Universidade Federal de Juiz de Fora (UFJF)
[2] Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

{livia.padua, ana.luz, gabrielly.braz, mariane.carvalho}@estudante.ufjf.br, {fred.belcavello, natalia.sigiliano, tiago.torrent}@ufjf.br

Abstract. *This paper aims to present an application of the semantic-computacional model of FrameNet Brazil to the representation of multimodal objects. Therefore, it describes the steps involved in creating a subpart of the ReINVenTA Dataset, focusing on the semantic annotation of the TV series Pedro pelo Mundo for the modalities of text and dynamic images.*

Resumo. *Este artigo tem como objetivo apresentar uma aplicação do modelo semântico-computacional da FrameNet Brasil à representação de objetos multimodais. Para tanto, descreve as etapas envolvidas na criação de uma subparte do Dataset da ReINVenTA, com foco na anotação semântica da série de TV Pedro pelo Mundo para as modalidades de texto corrido e de imagens dinâmicas.*

1. Introdução

Ao longo da última década, pesquisas relacionadas aos campos da Visão Computacional e da Linguística têm feito crescer a popularidade de conjuntos de dados que combinam informações textuais e visuais – os chamados *datasets* multimodais [UPPAL et al., 2020]. Neste artigo, apresentamos os recentes esforços desenvolvidos com o objetivo de expandir os dados da FrameNet para o domínio multimodal [BELCAVELLO et al., 2020], além de especificar as aplicações do modelo semântico-computacional. Nesse sentido, é necessário compreender que as linguagens agem conjuntamente a fim de produzir sentido em um texto e, por isso, limitar a análise semântica dos objetos multimodais apenas a texto verbal seria prejudicial aos estudos [DÁNNELS et al., 2022]. Desse modo, são feitas anotações de texto corrido e de imagens dinâmicas referentes aos corpora constituídos.

2. A FrameNet Encontra a Multimodalidade

A FrameNet é um projeto lexicográfico computacional que tem como referencial teórico a semântica de *frames* proposta por Charles J. Fillmore (1982), a qual prevê um pareamento inerente entre empirismo e linguagem ao apontar que as palavras são representações de categorias de experiências e, portanto, evocariam "cenas" (*frames*) capazes de delimitar a maneira como interpretamos seu sentido. Assim, por *frame* entendemos uma representação esquemática de "qualquer sistema de conceitos relacionados de tal forma que, para entender qualquer um deles, é necessário compreender toda a estrutura que eles se encaixam" [PETRUCK, 1996]. Tal representação é formada através da experiência humana e capaz de relativizar o sentido de uma palavra a depender do *frame* evocado.

É ancorando-se na hipótese de Fillmore (1982) que a FrameNet investiga o ato da atribuição de sentido em línguas naturais. Até pouco tempo, o projeto havia concentrado sua atenção em apenas uma modalidade de análise semântica: a textual. Entretanto, ao considerarmos o aspecto inerentemente multimodal da comunicação humana [STEEN et al., 2018], vemos que contemplar as diferentes modalidades nas tarefas de anotação nos permite realizar uma análise semântica mais completa. Assim, a FrameNet Brasil (FN-Br), preocupando-se em promover essa análise e enriquecer o seu banco de dados, passou a buscar meios de abarcar essas diferentes modalidades, o que foi impulsionado no momento em que o projeto se integrou à Rede de Pesquisa e Inovação para Visão e Análise de Texto, a ReINVenTA. A pesquisa, que investiga o processamento semântico computacional de objetos multimodais, reúne diferentes laboratórios e grupos de pesquisa mineiros que trabalham na construção e avaliação de um modelo computacional para representar objetos multimodais [BELCAVELLO, 2023]. É nesse sentido que a FN-Br conta com uma proposta de anotação multimodal a partir de ferramentas próprias, que possibilitam um estudo acerca da interação entre diferentes modalidades da linguagem humana e de seu impacto na construção de sentido.

3. O Dataset Frame2

Partindo da concepção multimodal da comunicação humana e da hipótese de que os elementos visuais em um vídeo são capazes de evocar frames ou complementar o que foi evocado pela narração [BELCAVELLO et al., 2020], a FN-Br, por meio da formação de um dataset multimodal, produzido no âmbito da iniciativa ReINVenTA, busca fornecer uma forma de correlação entre os elementos visuais e textuais de uma produção audiovisual. Assim, tem-se como objetivo investigar a interação entre os frames anotados nas tarefas de anotação de texto corrido e nas de sequências de vídeo, de modo a comparar a maneira com que os frames mobilizados para os elementos textuais interagem em combinação entre áudio e vídeo.

Os objetos multimodais selecionados para a anotação de imagens dinâmicas foram os 40 episódios da série de viagens de TV "Pedro pelo Mundo", exibida a partir de 2016 no canal GNT. O programa é apresentado por Pedro Andrade e trata de aspectos sociais, culturais e econômicos dos diferentes países nos quais os episódios se passam. A primeira temporada, objeto de anotação que compõe o Frame2, conta com 10 episódios de 23 minutos cada. Cada episódio teve suas falas transcritas automaticamente e revisadas. Depois, os anotadores[1] ocuparam-se de anotá-las manualmente usando a *Web Annotation Tool* (WebTool)[2] – vide 4.1 –, conforme as diretrizes da FN-Br, ou seja, adotando uma abordagem perspectivizada para cada anotação. Em seguida, o mesmo anotador foi responsável por anotar, por meio da ferramenta de anotação multimodal Charon[3] [BELCAVELLO et al., 2022] – vide 4.2 –, os elementos visuais presentes no

[1] O grupo de anotadores foi composto por graduandos em Letras, todos falantes nativos de português. Conforme explicitado por Belcavello (2023), 12 deles eram bolsistas do projeto, e outros 32 fizeram parte de oficinas de anotação oferecidas semestralmente pela equipe de pesquisadores da FN-Br na UFJF.

[2] Software de gerenciamento de banco de dados e anotação usado pela FrameNet.

[3] A ferramenta foi desenvolvida para auxiliar na anotação de objetos visuais, na correlação desses objetos com dados textuais e na rotulagem dos frames e elementos de frame por eles evocados.

mesmo episódio, procurando guiar-se (mas não restringir-se) pelos frames e elementos de frame (EFs), identificados em cada trecho durante a anotação de texto corrido realizada anteriormente.

4. O Passo a Passo da Anotação

A FrameNet é um modelo semântico que tem seus itens lexicais organizados em Frames [FILLMORE, BAKER, 2009]. Por exemplo, o verbo *comer* evoca o Frame `Ingestão`, que, por sua vez, pressupõe a existência de um Ingestor e de um ou mais Ingeríveis. Além disso, Frames possuem uma cadeia de relações entre si. No caso, o Frame de `Ingestão` herda do Frame de `Ingerir_substâncias` e é usado por `Alimentos_e_bebidas`.

4.1. Anotação de texto corrido

Na tarefa de anotação de texto corrido, um grupo de anotadores recebe lotes de sentenças para análise, a qual consiste, primeiramente, em atribuir um frame a cada Unidade Lexical (UL) presente na sentença. Para isso, o anotador é orientado a clicar na UL que deseja anotar, de forma que, então, é carregado um quadro com todos os frames da FN-Br associados à UL correspondente. Após a escolha do frame, é gerada uma camada de anotação de EF, na qual é possível categorizar os demais itens da sentença. A anotação pode ser feita em mais camadas, mas, para fins deste artigo, apenas os EFs serão incluídos. Um exemplo pode ser encontrado na Figura 1, em que a UL *comer* foi anotada no Frame `Ingestão` e, a partir disso, pôde-se atribuir aos elementos *eu* e *um sanduíche de porco com molho caribenho*, respectivamente, os valores de Ingestor e de Ingeríveis.

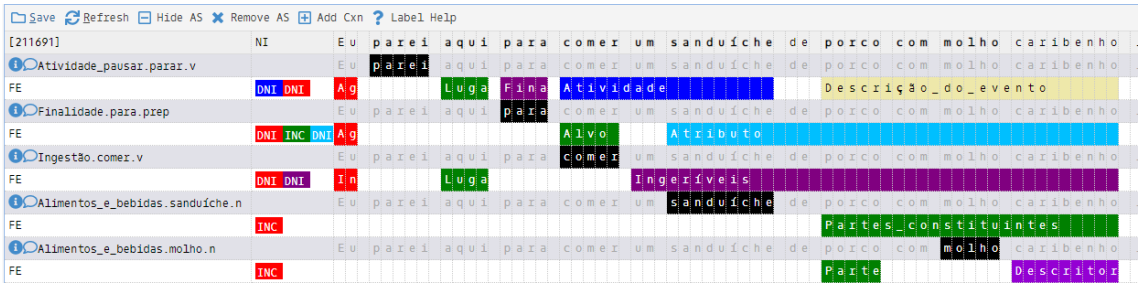

Figura 1. Anotação semântica para a Unidade Lexical *comer* no Frame `Ingestão`

4.2. Anotação de vídeo

A ferramenta de anotação de imagens dinâmicas conta com três painéis e com um arquivo de vídeo, detentor das entidades a serem anotadas, presente no canto superior esquerdo (Figura 2). Com isso, é concedida ao anotador a oportunidade de assistir ao contexto da sentença previamente anotada, o que pode ampliar, e até mesmo alterar, a sua perspectiva de anotação. Visto o vídeo, o anotador pode começar a atividade de marcação de objetos, que consiste na criação e edição de *bounding boxes* ao redor das entidades que se deseja anotar. É importante apontar que alguns objetos são criados

automaticamente pelo próprio *software* da ferramenta, mas é decisão do anotador mantê-las ou excluí-las. As atividades de criar objetos, rastreá-los, editá-los e excluí-los são feitas a partir dos respectivos botões localizados abaixo do arquivo de vídeo, junto aos botões de manipulação de reprodução do vídeo.

Após criar e editar a *bounding box* ao redor do objeto escolhido, o anotador atribui a ele um frame a partir de uma lista com todas as opções que constam na base de dados da FN-Br. Uma vez tomada essa decisão, é preciso atribuir um EF e um *Computer Vision Name* (CV Name). Essa categoria associa uma UL ao objeto delimitado, sendo essa qualquer UL do banco de dados da FN-Br que evoque um frame que estabeleça uma relação de herança com o frame de `Entidade`.

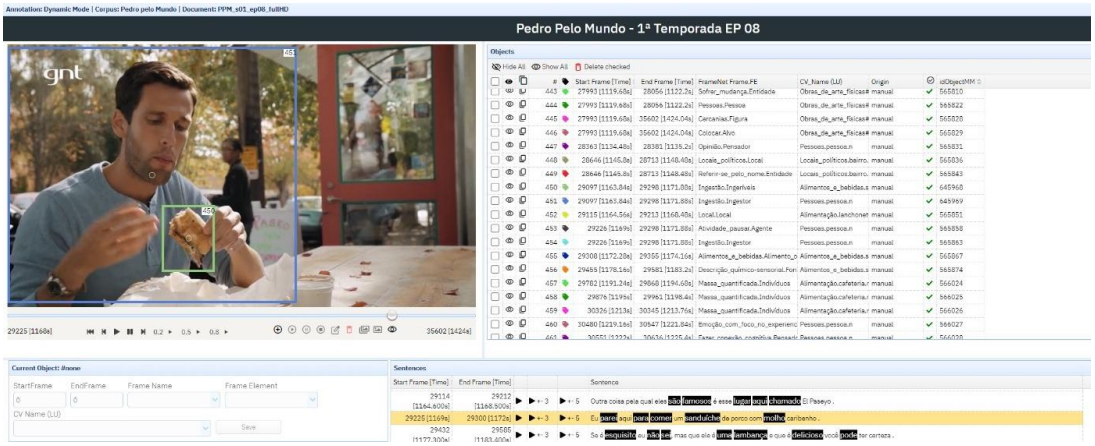

Figura 2. Interface de anotação de imagem dinâmica do corpus Pedro pelo Mundo

5. Resultados Alcançados

A partir de 2.195 sentenças, transcritas de 10 episódios do corpus Pedro Pelo Mundo, foi obtido o total de 11.796 *Annotation Sets* (AS) para texto corrido e 6.841 objetos para imagens dinâmicas. Uma vez que cada objeto visual é anotado para 3 categorias semânticas e que, em média, cada AS é anotado para 2,13 EFs, o esforço de anotação desse corpus produziu um *dataset* com 45.648 pontos de dados semânticos.

6. Considerações Finais

Neste artigo, descrevemos a proposta de anotação multimodal empregada pela FN-Br. Ao entender que a comunicação humana é um sistema multimodal, percebe-se a necessidade de se ir além de uma pesquisa que considera apenas a modalidade verbal para a construção de sentido. Assim, aponta-se a proposta de anotação da FN-Br como uma forma de realizar uma análise semântica de forma mais completa, visto que contempla, também, os aspectos visuais da comunicação.

Agradecimentos

A pesquisa apresentada neste artigo teve financiamento da FAPEMIG - processo RED-00106/21 e do CNPq - processos 408269/2021-9 e 420945/2022-9.

Referências

Baker, C. F., Fillmore, C. J. and Lowe, J. B. (1998). "The Berkeley FrameNet Project". In: COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics.

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

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

Belcavello, F. (2023). "FrameNet Annotation for Multimodal Corpora: devising a methodology for the semantic representation of text-image interactions in audiovisual productions". 135f. Tese (Doutorado em Linguística) — Faculdade de Letras, Universidade Federal de Juiz de Fora, Juiz de Fora.

Dánnels, D.; Torrent, T. T.; Sigiliano, N. S.; Dobnik, S. (2022). "Beyond Strings of Characters: Resources meet NLP – Again". In: Volodina, E.; Dánnells, D.; Berdicevskis, A.; Forsberg, M.; Virk, S. (Org.). Live and Learn: Festschrift in honor of Lars Borin (pp. 29–36). Gothenburg: Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs Universitet.

Fillmore, C. J. (1982). "Frame semantics". In.: The linguistic society of Korea. Linguistics in the morning calm. Korea: Hanshin Publishing Company.

Fillmore, C. J.; Baker, C. (2009). "A Frames Approach To Semantic Analysis". In: Heine, B.; Narrog, H. (Orgs.). The Oxford Handbook Of Linguistic Analysis (pp. 313–340). Oxford: Oxford University Press.

Steen, F., Hougaard, A., Joo, J., Olza, I., Cánovas, C., Pleshakova, A., Ray, S., Uhrig, P., Valenzuela, J., Woźny, J. and Turner, M. (2018) "Toward an infrastructure for data-driven multimodal communication research". Linguistics Vanguard, Vol. 4 (Issue 1), pp. 20170041. https://doi.org/10.1515/lingvan-2017-0041

Petruck, Miriam R. L. (1986) "Body Part Terminology in Hebrew: A Study in Lexical Semantics". Unpublished Ph.D. dissertation. University of California, Berkeley.

Salomão, M. M. M. (2009) "FrameNet Brasil: um trabalho em progresso". Calidoscópio, *[S. l.]*, v. 7, n. 3, pp. 171–182. Disponível em: <https://revistas.unisinos.br/index.php/calidoscopio/article/view/4870>. Acesso em: 6 ago. 2023.

Uppal, S., Bhagat, S., Hazarika, D., Majumder, N., Poria, S., Zimmermannz, R., & Zadeh, A. (2022). "Multimodal research in vision and language: A review of current and emerging trends". Information Fusion, 77, 149-171.