Evaluating speech and image coalescence in meaning construction for frame-based multimodal annotation

Frederico Belcavello[1]
[1]Federal University of Juiz de Fora, fred.belcavello@ufjf.br

Keywords: FrameNet, Multimodality, Eye-tracking, Annotation, Multimodal dataset.

Multimodal approaches have been gaining traction in Computational Linguistics in a myriad of datasets, model architectures and tasks (Hodosh et al., 2013; Young et al., 2014; Plummer et al., 2015; Elliott et al., 2016; Lala and Specia, 2018; Yao and Wan, 2020). This paper evaluates evidence on how speech and image coalesce in meaning construction in the experience of TV show viewers, discussing to which extent a dataset annotated following the FrameNet model (Belcavello, 2020; Belcavello, 2022; Viridiano, 2022; Torrent, 2022) can represent such meaning. We report on an eye-tracker experiment in which we compare the gaze points of interest of two different groups: one that watches the complete version of the show and another who watches a modified version, in which speech was completely removed. The hypothesis was that speech could direct gaze and, so, determine the ways image and text are combined in meaning construction. Results, however, indicate that the interference of speech in generating patterns of gaze is subtle and, in general, less effective than visual language or cinematic language expressed by camera angles, movements, framing and image composition. Such results, then, indicate that speech and text, although perceived as different modes, should be analyzed in combination with each other. In terms of Frame Semantics (Fillmore, 1982), it indicates that patterns of frame evocation should consider data as a whole, composed of both textual and visual material.

References

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 & Torrent, Tiago Timponi. 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.

Elliott, Desmond ; Frank, Stella; Sima'an, Khalil & Specia, Lucia. 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.

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

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

Lala, Chiraag & Specia, Lucia. 2018. Multimodal Lexical Translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Plummer, Bryan A.; Wang, Liwei ; Cervantes, Chris M.; Caicedo, Juan C.; Hockenmaier, Julia & Lazebnik, Svetlana. 2016. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In: *Proceedings of the IEEE International Conference on Computer Vision* (ICCV).

Torrent, Tiago Timponi, Matos, Ely Edison da Silva; Belcavello, Frederico; Viridiano, Marcelo; Gamonal, Maucha Andrade; Diniz da Costa, Alexandre & Marim, Mateus Coutinho. 2022. Representing context in framenet: A multidimensional, multimodal approach. *Frontiers in Psychology* 13 (2022): 573.

Viridiano, Marcelo; Torrent, Tiago Timponi; Czulo, Oliver; Lorenzi, Arthur; Matos, Ely & Belcavello, Frederico. 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.

Yao, Shaowei & Wan, Xiaojun. 2020. Multimodal Transformer for Multimodal Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Young, Peter; Lai, Alice; Hodosh, Micah & Hockenmaier, Julia. 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.