Measuring encyclopedic content in silent gesture: Gesture not as ambiguous as once thought

Keywords: gesture, sign language emergence, semantic similarity, vector space modeling

Background: Gesture contains a wealth of imagistic, yet ambiguous information. Studies show that non-signers are poor at determining the encyclopedic content of gestures (van Nispen et al., 2017) and signs from natural sign languages (Klima & Bellugi, 1979; Sehyr & Emmorey, 2019), which calls into question how collocutors converge on a shared interpretation in novel communicative settings (e.g., an emerging sign system). However, these studies adopted a strict definition of 'accuracy,' where a guess and the gloss of the sign/gesture must be string identical (e.g., 'brush' and 'comb' are *not* considered a match). We argue that this underestimates the information contained within the visual signal, by not taking into consideration the similarity between guess and gloss. To this end, we conducted action- and silent gesturing-labeling experiments, and compared the similarity of labels using a computational approach to semantic similarity. We show that non-signing participants converge on a certain range of interpretations when assigning meaning to gestures that correspond to the meaning of the actions these gestures represent. This suggests the pervasiveness and usefulness of form-meaning correspondences in comprehension.

Methods: We produced vignettes of 69 unique events (e.g., *break*, *walk*, *hammer*). Using these action vignettes, we elicited silent gestures from 6 non-signing participants (6 * 69 = 413 silent gestures, with one gesture discarded). For each video, we elicited one-sentence descriptions of what the gesture intended to convey using Amazon Mechanical Turk (30 sentences per action vignette, and 20 per gesture). We extracted the verbs from the sentences and scored them according to their Semantic Distance (SD). For example, the verbs in the set {*eat, dine, drink*} are more similar to each other than {*eat, think, drip*}, which can be represented numerically (i.e., SD(*eat, dine, drink*) < SD(*eat, think, drip*)). Specifically, we obtained 300-dimensional word-representation vectors from GloVe (Pennington et al., 2014), which characterize words based on their co-occurrence with other words (frequently co-occurring words are more likely to be semantically related). For each video, we calculated the average pairwise semantic distance between each verb by computing the angular distance between each verb's word-representation vector (lower score = words more closely related).

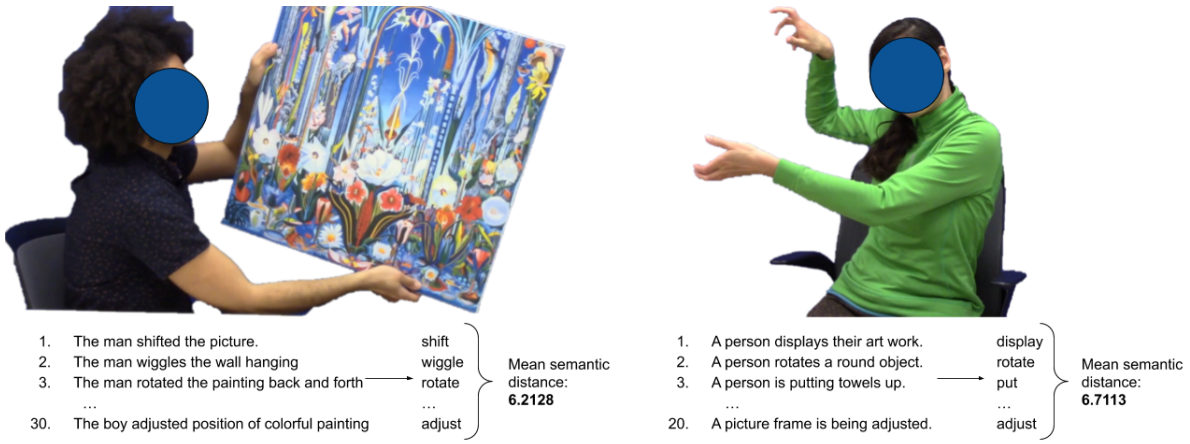

Figure 1: **Experimental design:** Turkers annotated live action videos (left) or videos of silent gestures (right). For each video, verbs were compared for similarity (semantic distance) and annotated for transitivity.

We assessed the consistency of the perception of semantic content (a) within verbs generated from viewing action videos (SD(action verbs)), (b) within verbs generated from viewing gestures (SD(gesture verbs)), and (c) between verbs generated from both tasks (SD(action verbs,gesture

verbs)). To compute a baseline (chance) measure, we randomly drew 8,260 verbs from the GloVe corpus and computed semantic distance scores for 413 sets of 20 verbs. If the encyclopedic meaning of gestures is truly ambiguous, we would expect that verbs elicited from gesture videos would not correspond with each other greater than chance (SD(gesture verbs) = SD(random verbs)). Further, we would expect no correspondence between verbs elicited from action videos and those from gesture videos (SD(action verbs,gesture verbs) = SD(action verbs,random verbs)).

Results, interpretation: Verbs generated in response to action videos were significantly more consistent with each other than verbs generated in response to gestures ($t(412) = -13.51$, $p<0.001$). However, the latter group was more internally consistent than random words ($t(412) = -41.69$, $p<0.001$; Fig. 1a). Verbs generated from action videos were significantly more consistent with those generated from gesture videos than with randomly generated verbs ($t(135) = -16.78$, $p<0.001$; Fig. 1b). Thus, despite the reported ambiguity of silent gesture, we found that non-signers consider only a certain range of interpretations of silent gestures: The visual properties of a gesture constrain its possible interpretations, which can be further constrained by environmental or discursive context. These results not only inform the nature of the signal (e.g., silent gesture contains ambiguous, but iconically constrained information), but inform discussions on how meaning is assigned to new signs in emerging sign languages.

(a) (b)

Figure 2: (a) Verbs produced in response to gesture videos were less similar to each other than those produced in response to action videos (SD(action verbs) $<$ SD(gesture verbs)) . However, both sets of verbs were more internally consistent than randomly selected verbs (SD(action verbs), SD(gesture verbs) $<$ SD(random verbs)). The red line represents the mean of SD(random verbs); (b) Verbs produced in response to gesture videos and action videos were more similar to each other than randomly selected verbs (SD(action videos,gesture verbs)) $<$ (SD(action verbs,random verbs)). The red line represents the mean of SD(action verbs,random verbs).

References

Klima, E. S., & Bellugi, U. (1979). *The signs of language*. Harvard University Press.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Sehyr, Z. S., & Emmorey, K. (2019). The perceived mapping between form and meaning in ASL depends on linguistic knowledge and task. *Language and Cognition*, *11*(2), 208–234.

van Nispen, K., van de Sandt-Koenderman, W. M. E., & Krahmer, E. (2017). Production and comprehension of pantomimes used to depict objects. *Frontiers in Psychology*, *8*, 1095.