We collect datasets that describe videos with a sequence of captions, one for each action, forming the story of the video; and their corresponding timestamps. We also (i) annotate each video segment with 34 labels; (ii) show the benefits of training on story-like data; (iii) establish human evaluation of video stories; and (iv) re-affirm the need for better automatic metrics for video generation.