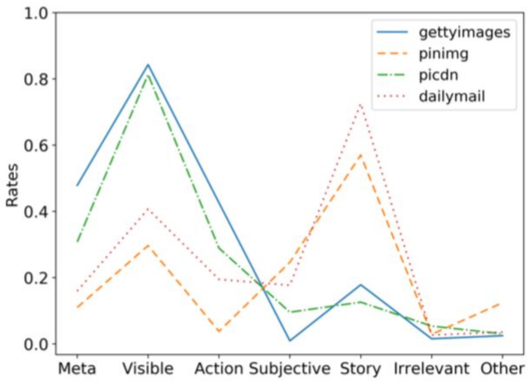
Intention and Attention in Image-Text Presentations: A Coherence Approach

Image-text presentations are widely available on the internet, in captioned images, social media posts and web pages; at the same time, they provide a valuable proxy for situated language, enabling indirect inferences about face-to-face conversation, the primary setting for language learning and language use. In this work, we analyze a corpus of image-text presentations to characterize the communicative goals and context-dependence that speakers exploit in describing the world around them. Our framework is coherence theory, as pioneered by researchers such as Nicholas Asher, Jerry Hobbs, Andy Kehler and Alex Lascarides. Drawing on annotated coherence relations between text and imagery in our corpus, along with annotated information about pronoun use, we show that coherence offers a powerful abstraction for linking speaker intentions and attention to the semantics of utterances.

In another paper that is currently in submission, we have established an annotation protocol for analyzing the relations between images and their captions. The relations that we have identified are *Visible*, *Subjective*, *Action*, *Story*, and *Meta*.

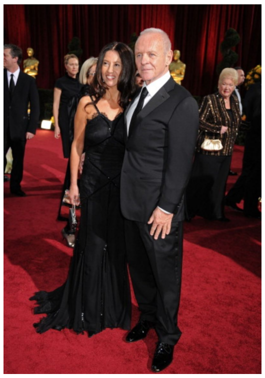
- *Visible*: the caption presents information which is intended to recognizably characterize what is depicted in the image, analogous to *Restatement* relations in text.
- *Subjective*: the caption describes the speaker's reaction to or evaluation of what is depicted in the image, analogous to *Evaluation* relations in text.
- *Action*: the text describes an extended and dynamic process of an action of which the image captures a representative snapshot, analogous to *Elaboration* relations in text.
- *Story*: the text is understood as providing a description of the circumstances depicted in the image, analogous to Hobbs's *Occasion* but inclusive of instructional, explanatory and other background relations.
- *Meta:* the text allows the reader to draw inferences not just about the scene depicted in the image but about the production and presentation of the image itself, analogous to *Meta-talk* relations in text.

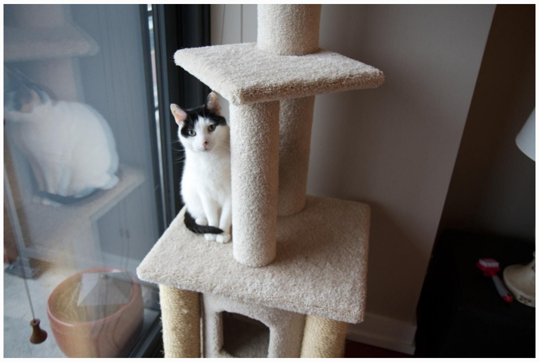
This taxonomy differs from information science approaches to the communicative functions of text and images in multi-modal documents (e.g., Marsh and Domas White, J Documentation 2003) in characterizing specific content-level inferences. Our taxonomy mirrors coherence frameworks for discourse semantics. As in text, multiple relations can hold at once. Using this protocol, we have annotated 10,000 image-caption pairs, including naturally occurring examples from the Conceptual Captions (Sharma et al., ACL 2018) and Open Images (Kuznetsova et al., arXiv:1811.00982 2018) datasets. Our team assessed our inter-rater agreement using Cohen's κ, resulting in a κ coefficient of 0.81.



In our first analysis, we look at how the distribution of different coherence relations varies across the publication domain of image-text pairs, using four case studies: *Getty Images,* a distributor for photos of current events; *Pinterest* (identified by domain pinimg), a social network for style and living where users build curated collections of web photos; *Shutterstock* (identified by domain picdn), a source for commercial stock photography; and *Daily Mail,* a British tabloid newspaper*.* The full distributions are shown in the graph at left. Text descriptions in Getty

Images and Shutterstock overwhelmingly exhibit the *Visible* relation, indicating that the text is generally limited to content that a viewer of the image can be expected to recognize. By contrast, the majority of Pinterest and Daily Mail descriptions display the *Story* relation, indicating that the text provides independent information about the situations captured in the imagery. It's not surprising, of course, that different authors and publishers use different styles and different content. Nevertheless, the disparity across collections shows that different situations and communicative goals can be reliably correlated with different coherence relations. In such cases, context enables reliable inferences about how a text might be interpreted—inferences that can guide semantic processing and inform language learning. The full paper gives the theoretical and experimental context describing these implications.





Left, an image with a *Story/Action* caption "actor and guest arriving at the premiere". Right, an image with a *Story/Subjective* caption "He's mostly a good kitty", which is an example of pronoun with a deictic reference to an entity in the image.

Different relations also lead to different ways to refer to objects in imagery. A key case concerns the use of pronouns, which in image-text presentations can refer deictically to entities from the image. To study the correlations of frequencies of pronouns and the coherence relations between images and text, we ran a pilot study where we annotated 1000 image-text pairs from the *makeup* and *animal* subgroups of the Reddit dataset (Hessel et al, AVinDH SIG 2017). The rates of the pronouns in pairs with the *Meta*, *Visible*, *Action*, *Subjective* and *Story* relations are respectively 10.12%, 9.23%, 14.16%, 25.28% and 28.41%. The difference between the rates demonstrates the potential effect of coherence in licensing different kinds of descriptions. For example, although the *Visible* relation directly characterizes entities in the image, speakers who use this kind of coherence relation seems not to take it for granted that those entities are at the center of attention in the discourse and so describe them with full noun phrases. By contrast, speakers who use the *Subjective* or *Story* relations to describe their opinion about an image seem much more likely to draw on the prominence of entities in the image in formulating their utterance. This is seen in the image above, on right. Next, we plan to annotate 10,000 image-text pairs from different subgroups of the Reddit dataset with an upgraded annotation protocol that accommodates coreference resolution in this context. The full paper describes the implications of these findings for the grammar of context-sensitivity.