Intention and attention in image-text presentations: A coherence approach

Ilana Torres, Kathryn Slusarczyk, Malihe Alikhani & Matthew Stone*

Abstract. In image-text presentations from online discourse, pronouns can refer to entities depicted in images, even if these entities are not otherwise referred to in a text caption. While visual salience may be enough to allow a writer to use a pronoun to refer to a prominent entity in the image, coherence theory suggests that pronoun use is more restricted. Specifically, language users may need an appropriate coherence relation between text and imagery to license and resolve pronouns. To explore this hypothesis and better understand the relationship between image context and text interpretation, we annotated an image-text data set with coherence relations and pronoun information. We find that pronoun use reflects a complex interaction between the content of the pronoun, the grammar of the text, and the relation of text and image.

Keywords. ELM; NLP; discourse; coherence; pronoun resolution; computational linguistics; semantics; pragmatics

1. Introduction. Image-text presentations are widely available on the internet, in captioned images, social media posts, and web pages. These image-text presentations provide a valuable proxy for situated language, enabling indirect inferences about face-to-face conversation, the primary setting for language learning and language use. McCullogh (2019) surveys the linguistic significance of using online communication to study spontaneous, informal language use.

Text and imagery function together in diverse ways (Marsh & Domas White 2003). An image of a dog posted on Facebook relates to the caption, "This is my new puppy" in a way that is very unlike how an image of a model in a magazine relates to its caption "A model on a runway". One fundamental difference is the semantic relationship between text and imagery: the model caption summarizes the image while the puppy caption links the image content to further facts about the speaker. These various relations lead to different ways in which we can identify objects in imagery through the use of a caption. A key case concerns the use of pronouns, which, in image-text presentations such as in the puppy image-caption example above, can refer deictically to entities from the image.

Pronouns occur often in text and conversation; they make utterances simpler and easier to process by eliminating the need to repeat a name or other descriptive content (see e.g., Gordon and Hendrick 1998). The semantic content of pronouns contains features such as number, gender, and person which helps in clarifying who or what a pronoun is referring to (Büring 2011). However, extra-linguistic information such as real-life pointing can also be used to disambiguate a pronoun. When it comes to pronouns that are used in discourse, there is a further kind of information at hand that can be processed in order to resolve the pronoun: coherence relations (Hobbs 1979). In particular, Stojnic et al. (2013) argue that ambiguity of a pronoun in a text-image presentation can be resolved using coherence, by establishing specific inferential connections from the text to

* This research was partly supported by NSF IIS-1526723 and CCF-19349243. We thank the ELM reviewers and attendees for comments and discussion that have improved the paper. Authors: Ilana Torres, Hofstra University (itorres2@hofstra.edu), Kathryn Slusarczyk, Rutgers University (kat.slu@rutgers.edu), Malihe Alikhani, University of Pittsburgh (malihe@pitt.edu), & Matthew Stone, Rutgers University (matthew.stone@rutgers.edu).

accompanying visual information that gives the reader or listener the context needed to identify the referent. While Stojnic et al. (2013) examine video and accompanying narration, our work focuses on image-text pairs to allow for a closer analysis of the relationships between coherence relations and pronoun usage. This would mean that by processing discourse relations as we read a caption and regard the accompanying image, we are making use of relevant and important information which aids in resolving the (sometimes highly underspecified) content that can be found in captions. We can identify the referents of a pronoun by not only reading the caption but also by acknowledging what's in the image.

In previous work (Alikhani et al 2019, Alikhani et al 2020), we analyzed corpora of image-text presentations to characterize their context-dependence as well as speakers' communicative goals. In particular, for the annotation of image-text pairs in the conceptual captions data set of Sharma et al (2018), we established a protocol to select types of coherence relations. The set of coherence relations we used included: (1) *Visible*, (2) *Subjective*, (3) *Action,* (4) *Story*, (5) *Meta,* and (6) *Identification*. Examples of these relations from this dataset can be found in Figure 1. Further description of these relations from the current dataset can be found below under section 3.1., Coherence Relations. We used these coherence relations to capture how text applies to or relies on the accompanying image for information about context. This also allowed us to analyze these relations in terms of speakers' communicative goals; the type of coherence relation and context provided is influenced by, and can indicate, what kind of information speakers intend to convey.

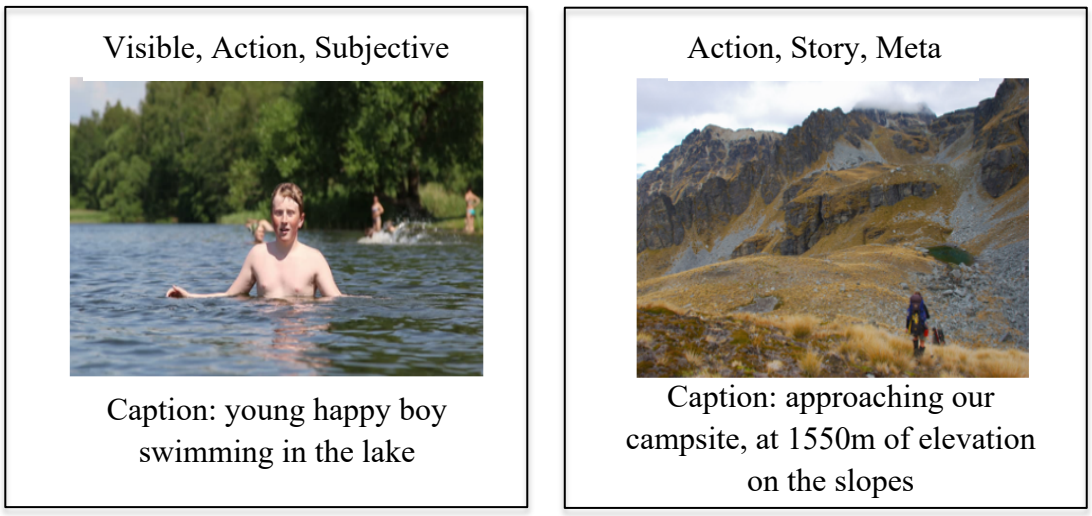

Figure 1: Images and captions from a previous Conceptual Caption dataset as an example of initial coherence relations. (Photo credits: yauhenka; Danilo Hegg)

Our previous work focused on coherence relations. Here we expand the focus to consider pronouns. This has required a change of data set, not only to make sure that images feature salient objects, animals or people, but also to make sure that captions contribute appropriate coherence relations.

Previous annotations on discourse coherence relations in image-captioning have caused us to notice that there are higher correlations of pronouns occuring in *Story* and *Subjective* relations than in other relations. This was because speakers who use the *Story* or *Subjective* relations to describe their opinion about an image seem much more likely to draw on the prominence of entities in an

image when formulating their utterance. In our current research on the usage of pronouns in image-text pairs, we aim to examine how the types and frequency of pronouns used in captions is influenced by a caption's coherence relation, and what this indicates about speaker intentions. We hypothesize that there is some pattern of correlation between image-caption discourse coherence relations and the types and frequency of pronouns within these captions. While we expect the highest frequency of all pronoun types to be in *Story* and *Subjective* type image-caption pairs, *Subjective* type pairs in particular may show a higher frequency of using indexical pronouns like *I*, whereas in *Story* relations we expect to see more examples of anaphoric pronouns. If any particular type of pronoun appears more often within certain types of coherence relations, or even in certain types of caption and utterance structures, we can draw links between image-captions, pronouns, and their references; these links may then offer insight into how speakers' intentions affect pronoun usage, and vice versa.

2. Methods. We created an interface to annotate a sample of image-text pairs. For each pronoun in the caption text, annotations were given on (1) discourse relation, (2) caption structure, and (3) pronoun type. We randomly sampled 6407 image-text pairs from the Reddit dataset that all include pronouns. Before beginning annotations, the first and second authors went through two rounds of preliminary annotations to adjust and finalize the annotation interface and establish strong inter-rater agreement. The first inter-rater agreement test we ran consisted of a set of 50 image-text pairs, with one or more pronouns per caption. This first test resulted in a low level of agreement, partially due to the inefficient first version of our caption structure types. We adjusted caption structure types to instead indicate utterance types and clarified pronoun distinctions between inter-raters. We reached a strong level of agreement with a second inter-rater agreement task and were able to continue with annotations.

3. Annotation process. The annotators were presented with an image and the accompanying text along with options for choosing coherence relations, utterance structure, and pronoun type.

3.1. Coherence Relations. In our previous work on image-text coherence relations, we had modified existing coherence relations in order to fit the relationships we saw in our annotations. These relations were based on theoretical work on discourse coherence and structure (Hobbs 1985, Roberts 2012, Webber 1999) as well as previous discourse annotation studies by Prasad et al. (2008) and previous work by Alikhani et al. (2019). As in our previous work, for each image we annotated, we chose one or more coherence relations based on the content of the text and its relation to the image. As listed above, the coherence relations were: (1) *Visible*, where the content of the caption was depicted in the image, (2) *Subjective*, where the caption was making a subjective statement about the content of the image, (3) *Action,* where the caption describes a dynamic process of an action seen in the image, (4) *Story*, where the caption provides a description of the image, or narrative-like background information, (5) *Meta*, where the caption not only describes the image but also mentions productions and presentation of the image, and (6) *Identification* where the caption uses a pronoun in order to identify a specific, salient object in the image. As mentioned, these relations are based on those previously used in text discourse; where *Visible* relations are based on *Restatement* relations, *Subjective* relations on *Evaluation* relations, *Action* relations on *Elaboration* relations, *Story* relations on *Occasion* relations, and *Meta* relations on *Meta-talk* relations (Hobbs 1985, Prasad et al. 2008). The *Identification* relation was not present in our annotation guidelines for some previous work, as conceptual captions often have content omitted for machine learning experimentation. It was added in the current work given our specific inquiry into the usage of pronouns in image-text pairs. There was also an option for (7) *Irrelevant*,

which included images where the caption was gibberish or simply did not match the image, and (8) *Other,* to indicate circumstances such as images which included text. An example of an *Irrelevant* image-caption can be found in Figure 2. Further examples of coherence relations from the specific dataset can be found in Figure 3.

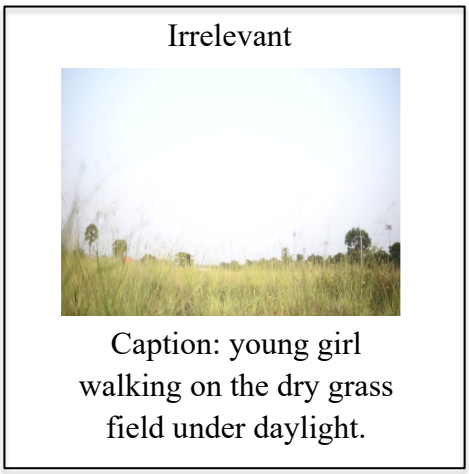

Figure 2: Example of an *Irrelevant* image-caption. (Photo credits: Andre Seale)

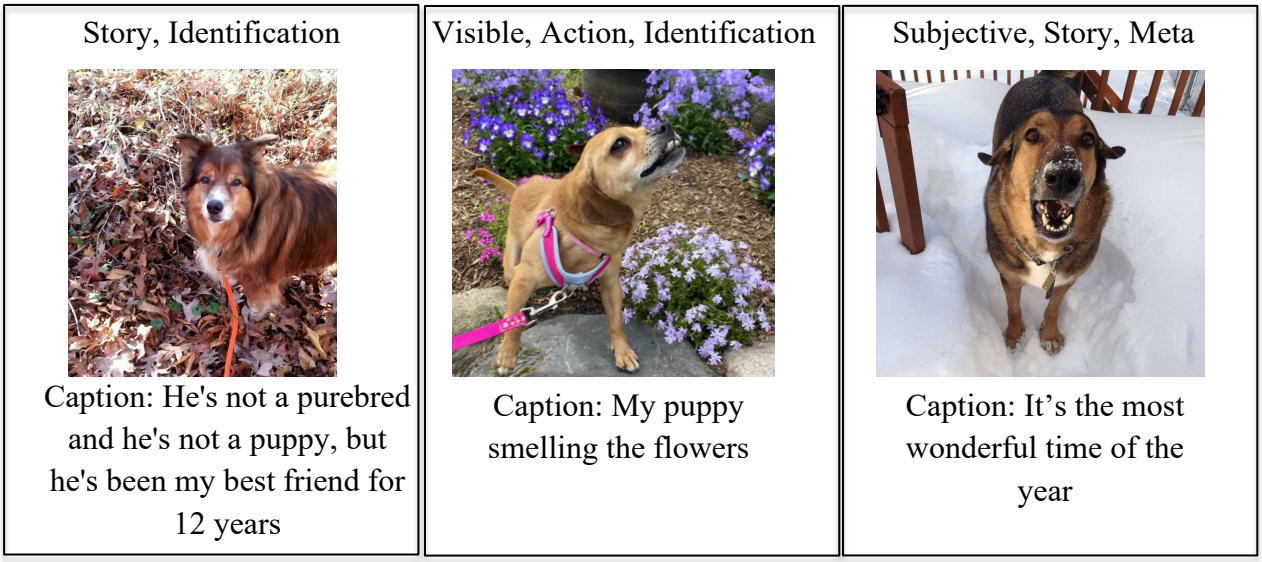

Figure 3: Examples of various coherence relations.
(Photo credits: detap_rettiwt; Ilana Torres; Alena Capil)

3.2. Utterance structures. The utterance structure type was also annotated to investigate the relationship between the structure of a caption and the frequency and types of pronouns within certain utterance structure types. With our first version of annotations for the structure of each caption, we agreed upon the following structure types; *sentence*, which indicated a full sentence regardless of punctuation; *noun phrase with an implicit topic*, with sub-categories for indicating whether the implicit topic was the image itself, the central focus of the image, or something else; and *something else*, to indicate a different structure. However, these types did not allow for

meaningful annotation of captions that were not full sentences or noun phrases, as many captions included non-finite predicates. Though an image of a kitten playing with a toy could be accompanied by the caption "my kitten is playing with her toy," the shorter caption "playing with her toy" may also be used. Annotation options were accordingly adjusted to include a wider range of structure types that appeared frequently in the dataset: (1) *simple noun phrase*, (2) *noun phrase + non-finite predicate,* (3) *non-finite predicate,* (4) *full sentence,* and (5) *other,* reserved for utterances like "ouch" that did not fall into the preceding annotation types.

The first version of this annotation system allowed submission of just one annotation for each caption, but this made it difficult to accurately capture the structure of captions that appeared to contain multiple utterances, such as captions that contained both a full sentence and a predicate. We adapted our data collection to indicate the structure of each part of a caption, or each utterance, as we have designated them. While some captions were still treated as one utterance, those with punctuation that clearly defined separate sentences, phrases, or predicates were treated as multiple utterances. For example, a caption such as "this is my new puppy" would be treated as one utterance, while a caption such as "This is my new puppy. Her name is Lucky." would be treated as two utterances, though the number of utterances within each caption was not noted. For each pronoun, we also annotated the structure of the utterance in which it appeared.

3.3. Pronoun type. Based on the definitions of pronouns in Büring (2011) and Traxler (2011), and the frequency of pronouns identified in previous analysis of coherence relations, we agreed upon the following categories for identifying pronoun type. We submitted an annotation for each pronoun in a caption. The options we agreed upon for pronoun annotations were (1) *indexical* (such as *I* and *you*), (2) *demonstrative* (such as *this* or *that*), (3) *anaphoric* (such as personal pronouns), (4) *bound* (such as bound personal pronouns), (5) *indexical/bound* (such as *my* and *your*), (6) *backwards anaphora* (such as a backward bound personal pronoun), and (7) *not actually a pronoun*, included to remove items that were mistakenly labeled as pronouns by the interface.

3.4. Annotation process outline. We will use Figure 4, below, as an example for a detailed outline of the annotation process.

Caption: He's huge and lazy but when treats are involved, this big guy'll do anything

Figure 4: (Photo credits: Annalise Burke).

- First, we identify the discourse coherence relations: *Story*, *Subjective,* and *Identification*
- Next, we identify the caption structure: one full sentence; though this example includes punctuation, this is not a necessary condition of a full sentence annotation
- Lastly, we identify the pronouns: *he* is backwards anaphoric to *this big guy*, and *this* is demonstrative

4. Results. Overall, our dataset includes 13858 image-text pairs annotated with coherence relations out of which 6407 have pronouns. Though this research is still in progress, our second inter-rater agreement task showed evidence that many of the sampled image-text pairs with pronouns fall into coherence relations of *Visible, Meta,* and *Story*, as was evidenced in previous work. Surprisingly, there were low levels of *Subjective* captions. The overall distribution of coherence relations in the dataset can be seen in Table 1. Additionally, the most frequent pronouns overall were indexical and indexical/bound pronouns, followed by anaphoric. Given that most captions were *Visible, Meta,* and *Story,* the pronouns such as *I*, *you*, and other personal pronouns appeared very frequently. The distribution of pronouns in each coherence relation can be seen in Table 2. The *Meta* relation was particularly interesting, as other pronouns such as demonstrative pronouns were often found in captions with this relation. The distribution of pronouns in fine-grained *Meta* captions can be found in Table 3. Though the distributions of each pronoun type appear to be similar across the fine-grained Meta relation types, *demonstrative* pronouns appeared less frequently in *Meta-when* relations than in *Meta-where* and *Meta-how* relations, and *bound* pronouns appeared more frequently in *Meta-how* relations than in *Meta-where* and *Meta-when* relations. Other findings include that, though not frequent, most cases of backwards anaphora appear in full sentence-annotated captions. Table 4 shows the distribution of pronoun types in specific sentence structure types. Additionally, Table 5 indicates the distribution of sentence structures in captions containing specific coherence relations. Our findings are discussed further below.

Visible	Subjective	Action	Meta	Story	Identification	Other
4014 (62.7%)	483 (7.53%)	936 (14.6%)	1998 (31.2%)	1463 (22.8%)	391 (6.1%)	785 (12.2%)

Table 1: The distribution of coherence relations in our dataset. The distribution of coherence relations for fine-grained Meta categories of *When*, *How* and *Where* are respectively 24.1%, 31.1%, and 63.3%. Note that multiple coherence relations may be present in one example which explains why the sum of this row is greater than 100%.

	Visible	Subjective	Action	Meta	Story	Identification
Indexical	29.92%	36.96%	31.75%	30.74%	34.14%	28.94%
Demonstrative	6.63%	8.26%	5.84%	10.12%	8.58%	9.13%
Anaphoric	13.72%	14.78%	13.50%	13.49%	13.47%	15.23%
Bound	8.06%	6.09%	7.66%	7.00%	5.82%	6.67%
Indexical_Bound	32.98%	25.65%	35.40%	27.11%	28.10%	31.28%
BackAnaphora	0.89%	1.30%	0.00%	1.04%	1.04%	1.36%
Other	0.04%	0.00%	0.00%	0.00%	0.05%	0.06%

Table 2: The distribution of pronouns in each category. Each percentage indicates the texts containing pronouns of the indicated type as a percentage of the texts labeled with the indicated coherence relation. For example, 29.92% of image-text pairs annotated as *Visible* contained at least one *indexical* pronoun.

	Where	When	How
Indexical	30.58%	30.28%	25.00%
Demonstrative	13.28%	7.34%	12.50%
Anaphoric	12.78%	13.99%	12.50%
Bound	7.02%	7.57%	25.00%
Indexical_Bound	25.81%	28.21%	25.00%
BackAnaphora	0.75%	1.15%	0.00%
Other	0.00%	0.00%	0.00%

Table 3: The distribution of pronouns in fine-grained Meta categories. As above, each percentage indicates the texts containing pronouns of the indicated type as a percentage of the texts labeled with the indicated fine-grained Meta category. The *notPronoun* type indicates items that were incorrectly marked as pronouns by our annotation interface and will be disregarded in the following discussion.

	Indexical	Demonstrative	Anaphoric	Bound	Back Anaphora	Indexical Bound
NP	8.51%	11.30%	13.30%	11.8%	12.00%	11.80%
Full sentence	80.30%	76.40%	77.40%	75.7%	84.00%	77.60%
NPNF Predicate	10.40%	10.48%	8.71%	11.8%	4.00%	9.70%
NF Predicate	0.20%	01.31%	0.20%	0.60%	0.00%	0.10%
Other	0.40%	0.40%	0.30%	0.00%	0.00%	0.60%

Table 4: The distribution of pronoun types in sentence structure types. Each figure indicates the utterance type containing the indicated coherence relation type as a percentage of all utterances containing the indicated coherence relation type.

	Visible	Subjective	Action	Meta	Story
NP	11.5%	7.4%	8.6%	12.7%	9.8%
Full sentence	77.5%	76.2%	81.8%	78.1%	77.9%
NPNF Predicate	10.2%	13.3%	9.1%	8.5%	11.0%
NF Predicate	0.3%	1.5%	0.0%	0.4%	0.4%
Other	0.3%	1.4%	0.2%	0.0%	0.6%

Table 5: The distribution of sentence structure types in coherence relation types.

Sentence structure type distribution for the *Identification* relation is not listed as no images with an *Identification* coherence relation have been annotated with sentence structure type yet. Sentence structure types were introduced part way into the annotation process, and *Identification* coherence relations are not very frequent, at only 9.9% of our annotated image caption pairs so far.

5. Discussion. As we continue, our hypothesis still stands; that there is some pattern of correlation between image-caption discourse coherence relations and the types and frequency of pronouns within these captions. More than the overall distribution of coherence relation types in Table 1, we are interested in the interactions of coherence relations, pronoun types, and sentence structures represented in tables 2 through 5. Table 2 indicates that the most frequent types of pronouns overall are indexical, indexical/bound, and anaphoric pronouns; while each type seems to be about evenly represented across coherence relations, some less frequent and more frequent pairings are discussed below. As mentioned, our current results confirm that many of the sampled image-text pairs with pronouns fall into coherence relations of *Visible* and *Story*. Given that pronouns in captions often refer to entities within the image, it is not surprising that *Visible* is our most frequent relation at 62.7% of the annotated data set. Of the data annotated as *Visible*, the most frequent pronoun types were *indexical/bound* at 32.98% and *indexicals* at 29.92%. When a caption refers to entities like "my dog," for example, "my" will require an *indexical/bound* annotation and "dog," as long as a dog is pictured, will require a *visible* annotation. The frequency of these annotations is expected, since the current data set is composed of user generated images and captions that aim to describe the bound indexical relationship of the image's main entity from the user's perspective. As for *Story* relations, the usage of any pronouns often give captions some element of backgrounded information that indicate their *Story* relation. Of the pronouns present in *Story* relations, indexicals were the more frequent at 34.14%, with indexical/bound pronouns slightly behind at 28.10%. Note that the most frequent and second most frequent pronoun types for *Visible* relations and *Story* relations are flipped, where images with *Visible* relations are most often annotated with *indexical/bound* pronouns and then plain *indexical* pronouns, and images with *Story* relations are most often annotated with plain *indexical* pronouns and then *indexical/bound* pronouns. *Indexical/bound* pronouns like "my" (when used to reference a user's dog, for example) can be taken as *Visible* given the image of a dog, assuming that the dog must belong to someone and "my" is not necessarily an indicator of a *Story* relation. Indexicals like "I" or "you," however, seemed to more often refer to entities that were not visibly within the image and therefore provided some information that cannot be verified for a *Visible* annotation. This may explain why *Visible* image-text pairs were slightly more often annotated with *indexical/bound* pronouns while S*tory* image-text pairs were slightly more often annotated with *indexical* pronouns.

Image-text pairs with *demonstrative* pronouns yielded some unexpected percentages. Though we annotated *demonstrative* pronouns at similar rates (between 6.6% and 9.1%) for most coherence relation types, those with *Action* coherence relations and *Meta* (of any fine-grained type) coherence relations appeared at slightly differing frequencies of 5.84% and 10.12%, respectively. The lower frequency of *demonstratives* in *Action* relations may be due to the preferred usage of *indexical*, *indexical/bound*, and *anaphoric* pronouns to refer to the entity taking action in the image. As for the higher rate of demonstratives in meta relations, we refer to the distributions in our fine-grained meta types in table 3, where *demonstrative* pronouns appeared less frequently in *Meta-when* relations (7.34%) than in *Meta-where* (13.28%) and *Meta-how* (12.50%) relations. These higher frequencies seem to be indicative of how *demonstrative* pronouns like 'this' and 'that' can be used to refer to a place or some aspect of how an image was created, such as in 'this photo' or 'that building.' Also dealing with the figures in table 3, *bound* pronouns appeared more frequently in *Meta-how* relations than in *Meta-where* and *Meta-when* relations. Captions with *Meta-how* relations often appear to be more complex, disproportionately involving further clauses with coreference.

Additionally, the data set included *Subjective* image-text pairs at a much lower frequency than we initially expected, at only about 7.53% of our data set. Given the user generated source of the data set, we expected a higher frequency of *Subjective* posts. However, the data set seemed to contain more objective *Visible* captions, or those that simply stated other background information or related *Story* captions. Within the *Subjective* image-text pairs we did have, the most frequent pronoun types were *indexical* pronouns at 36.96% and *indexical/bound* pronouns at 25.65%. Though the third most frequent pronoun type is *anaphoric* at 14.78%, the remaining pronoun types were all below 9% of the total *Subjective* image-text pairs annotated. This appears to be largely consistent with the other coherence relation types, though not all types have the same order of most frequent and second most frequent pronoun types.

As mentioned, sentence structure types were introduced part way into the annotation process, meaning that the figures reported in tables 4 and 5 represent a smaller portion of the total data set. While full sentences were most frequent in image-text pairs using any given pronoun type, they were even more frequent in image-text pairs using *backwards anaphora*, at 84% of all images annotated with *backwards anaphora*. Each of the sentence structure types have similar frequencies across the pronoun types, but *backwards anaphora* appeared relatively less frequently in *noun phrases with non finite predicates*, at only 4% of the category compared to an average of about 10% for other pronoun types. The high rate of backwards anaphoric pronouns in full sentences and lower rates in other sentence structures suggests that backwards anaphora is not efficient for captions with more truncated structures. Table 5's distribution of sentence structure types across coherence relations does not seem to show much besides a clear preference for *full sentence* type utterances; gleaning meaning from sentence structure type seems to require figures that include some information about pronoun types. Additionally, we have not yet been able to report results for the distribution of sentence structure types in *Identification* coherence relation image caption pairs yet. While the rate of *Identification* relations in our full data set is low at 6.1%, the rate of *Subjective* relations is similarly quite low at 7.53%. Our dataset is biased as it doesn't have balanced samples from each class of the relations or pronouns. We believe that a further expansion of the data set would allow us to report a distribution of sentence structure types within all coherence types.

6. Conclusion. We found that pronoun use depends on the kind of relation between the image and its caption. We saw that there is overall a high frequency of *Visible* coherence relations, and the most frequently, indexical and indexical/bound personal pronouns occurred in captions, followed by anaphoric pronouns. The kind of sentence structure type used in a caption also correlated with pronoun usage: for example, backward anaphora was most common in full sentences. Our additional annotation of utterance structures may reveal further related patterns between coherence relations, structure, and pronoun use, and allow us to analyze how image-text pairs are created according to speaker intentions. Our current research provides opportunities for future work on pronoun resolution in the context of image-captioning, and will allow the construction of more accurate and effective captioning models, which will assist in the creation of model-generated image captioning as well as better results for search engines. These models will ideally be able to create strong captions for given images, thanks to research on the content of captions and their visual referents. However thorough this research on coherence relations in the English language may be, this leaves room for research on image-text relations in various other languages. Different languages have different paradigms for usage of pronouns or a complete lack of pronouns, and similar annotations such as from this experiment would allow a better understanding of how languages process pronouns in small segments of language and especially in addition to images.

Additionally, there might be cultural differences that arise in image-caption pairs posted in different languages which could be studied as well. Our dataset is available on the project GitHub page.[1]

References

Alikhani, M., Chowdhury, S. N., de Melo, G., & Stone, M. (2019). A corpus of image-text discourse relations. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1,* 570-575.

Alikhani, M., Sharma, P., Li, S., Soricut, R., Stone, M. (2020). Cross-modal coherence modeling for caption generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*

Büring, D. (2011). Pronouns. *In von Heusinger, Maienborn and Portner (eds.): Semantics: An International Handbook of Natural Language Meaning*, *2,* 971-995.

Gordon, P. & Hendrick, R. (1998). The representation and processing of coreference in discourse. *Cognitive Science, 22*(4), 389-424.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science, 3*(1), 67-90.

Hobbs, J. R. (1985). On the coherence and structure of discourse. *Technical report, SRI INTERNATIONAL MENLO PARK CA.*

Marsh, E. E. & Domas White M. (2003). A taxonomy of relationships between images and text. *Journal of Documentation, 59*(6), 647-672.

McCulloch, G. (2020). *Because Internet*. Penguin USA.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., & Webber, B. L. (2008). The penn discourse treeBank 2.0. *In LREC.* Citeseer

Roberts, C. (2012). Information structure: towards and integrated formal theory of pragmatics. *Semantics and Pragmatics, 5*, 6-1.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1*, 2556-2565.

Shiffrin, D. (1980). Meta-talk: organizational and evaluative brackets in discourse. *Sociological Inquiry, 50*(3-4), 199-236.

Stojnic, U., Stone, M., & Lepore, E. (2013). Deixis (Even Without Pointing). (Report). *Philosophical Perspectives*, *27*(1).

Traxler, M. J. (2011). *Introduction to psycholinguistics: Understanding language science*. Wiley-Blackwell.

Webber, B., Knott, A., Stone, M., & Joshi, A. (1999). Discourse relations: a structural and presuppositional account using lexicalised TAG. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 41-48.

[1] https://github.com/malihealikhani/ELM2020-Intention-and-Attention-in-Image-Text-Presentations