Chapter 2

How people describe images

2.1 Introduction

The first part of this thesis is dedicated to the question: how do people describe images? This chapter provides the theoretical background to this question, and presents an overview of different linguistic phenomena in image description data. Although some of these linguistic phenomena are quantified, the main claims of this chapter rest on existence arguments. As discussed in §1.8, the point of an existence argument is to describe and illustrate different phenomena that exist in the data. If the goal for automatic image description systems is indeed to mimic human image description behavior, then any complete solution to this problem must be able to account for the phenomena described in this chapter. Specifically, they should be able to exhibit the same level of variation in the use of different labels, and they should be able to reason about the situation depicted in a given image.

Image description data also presents us with some phenomena that we may not want systems to exhibit. We will observe how image descriptions are subjective, and may reflect stereotypes and biases held by the speaker. Furthermore, descriptions of other people may make reference to properties that could be considered inappropriate. Having established that these phenomena exist, one might also decide to limit the kinds of descriptions that a system should produce. In other words: to establish guidelines for what proper descriptions should look like. But a prerequisite of image description guidelines is that we have a clear idea of what descriptions could look like, i.e. that we understand the full range of variation, before we make a selection from the rich palette of human image descriptions. This chapter provides the foundations for such an understanding.

2.1.1 Contents of this chapter

The first chapter introduced the concept of a semantic gap between human and machine performance in image recognition, and we argued that image description also requires us to look at how people choose to talk about images (the pragmatic level). This chapter provides a broader theoretical background, and gives an overview of the different pragmatic phenomena that we may find in image description data.

Theoretical background

Section 2.2 relates the semantic gap to different theories of image understanding. We will discuss Panofsky’s (1939) meaning hierarchy, along with Shatford’s (1986) contributions to image indexing (based on Panofsky’s work). Following Ørnager (1997), we note that there are parallels between this body of literature and the work of Barthes (1957, 1961, 1978). Closing off this section, we show how these theories may inform our thinking about automatic image understanding, and how they may lead to hypotheses about system performance (§2.2.3).

Section 2.3 extends the discussion of the pragmatic level from the first chapter. We provide a short introduction to Gricean pragmatics (Grice, 1975), and show how we might apply Gricean analyses to image description data. These analyses put the speaker at the center stage.
We show how different descriptions for the same image may be the result of differences in knowledge about the world, or a different weighing of the Gricean Maxims.

Section 2.4 explains how the Flickr30K and MS COCO datasets were developed, followed by a final discussion of image description as perspective-taking (§2.5). Difference in perspectives on an image may lead to different descriptions of that image. The rest of the chapter explores this variation from several different angles.

### Empirical data

Section 2.6 presents two ways to explore the labels used to refer to different entities in the Flickr30K Entities dataset. First, we explain how we can organize these labels using a graph-clustering approach. Each cluster of labels shows us the different ways people refer to similar entities. Second, we present a manual categorization of labels used to refer to people. We will see that these labels are based on a wide range of properties. But humans never describe other people by listing all of their properties. (This would make communication very inefficient.) Rather, they make a selection of the properties that are somehow relevant to mention. Variation in image descriptions arises when different participants select different properties to make reference to.

Following the discussion of variation in entity labels, we will discuss stereotyping and bias in image descriptions, and show how the descriptions reflect different participants’ perspectives on the world. We will look at three phenomena: 1. unwarranted inferences, where participants provide speculative descriptions (§2.7); 2. linguistic bias in the use of adjectives (also called reporting bias, Misra et al. 2016) (§2.9); 3. linguistic bias and evidence of world knowledge in the use of negations (§2.10). Together, these phenomena show us that image descriptions are the result of a reasoning process based on world knowledge and (generalizations over) past experiences.

#### 2.1.2 Publications

This chapter was edited from the following publications:


2.2 Levels of interpretation

The previous chapter discussed the idea of a *semantic gap* between image recognition systems and humans with respect to their ability to interpret images (Smellders et al., 2000; Hare et al., 2006). The concept of a semantic gap implies that there are different levels of understanding that we can have of a picture. This idea is in line with previous research in image description and image categorization. A good place to start is Erwin Panofsky’s (1939) meaning hierarchy, which defines three levels of understanding in the context of renaissance paintings:

1. **Pre-iconography** giving a low-level description of the contents of a picture (factual description), and the mood it conveys (expressional description).

2. **Iconography** giving a more specific description of the image, also using information about the historical and cultural context in which the image was produced.

3. **Iconology** interpreting the image, establishing its cultural and intellectual significance.

The more we move up through the hierarchy (from level 1 to 3), the more (world) knowledge is required.\(^1\) Panofsky’s hierarchy was used by Markey (1983), Shatford (Shatford, 1986; Layne, 1994) and Jaimes and Chang (1999) as a theoretical framework to index image libraries. Shatford’s work, in particular, has been very influential, because she proposed an intuitive distinction between what a picture is *Of*, and what a picture is *About*. She also adapted Panofsky’s framework to a more practical scheme for indexing images (commonly referred to as the Shatford/Panofsky matrix; see e.g. Enser 1995; Stewart 2010; Ørnanger and Lund 2018).

### 2.2.1 The Of/About distinction

Shatford (1986) argues that the Panofsky’s first two levels consist of two aspects: *Of* and *About*. At the pre-iconographic level, *Of* corresponds to the factual properties of the image, and *About* corresponds to the expressional properties. At the iconographic level, we can say that an image is *Of* specific objects and events (possibly using their proper names), and *About* mythical beings and symbolic meanings.

Shatford proposes to analyze the subjects of a picture in terms of three aspects: Specific *Of* (at the iconographic level), Generic *Of* (at the pre-iconographic level), and *About* (for which she argues that “aside from mythical beings and locales, *About* words describe emotions and abstract concepts, and may be thought of as inherently generic (p. 47).”). Having established three different *aspects* of a picture (Specific *Of*, Generic *Of*, and *About*), Shatford introduces four *facets*: Who, What, Where, When. If we want to fully analyze the subject of a picture, we should look at all combinations of these facets and aspects. These combinations are commonly presented in a matrix, as in Table 2.1. This matrix may be used as a practical guide to systematically index collections of images. Following Shatford’s work, different researchers have proposed modifications or additional features to supplement the Shatford/Panofsky matrix. See Stewart 2010 for an overview.

---

\(^1\)But, as Christensen (2017) notes, Panofsky’s hierarchy is not meant to interpret images in a bottom-up process. Rather, the interpretation of images is a more circular, hermeneutic process in which answers at ‘higher’ levels may also inform us about the interpretation of images at a ‘lower’ level.
Chapter 2 How people describe images

<table>
<thead>
<tr>
<th>Panofsky</th>
<th>Iconography</th>
<th>Pre-Iconography</th>
<th>(See caption)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shatford</td>
<td>Specific Of</td>
<td>Generic Of</td>
<td>About</td>
</tr>
</tbody>
</table>

| Who      | Named entities | Kinds of entities | Abstractions and mythical beings |
| What     | Named events   | Actions, conditions | Emotions and abstractions |
| Where    | Named locations | Kind of place      | Place as symbol, Symbol as place |
| When     | Linear time    | Cyclical time      | Time as symbol   |

Table 2.1 The Shatford/Panofsky matrix, but with the top right corner unspecified. For Shatford (1986), the About-aspect seems to cover both Pre-iconography and Iconography (to the extent that mythical beings are relevant for the indexation of pictures), and she explicitly excludes Panofsky’s Iconology level from the practice of indexation because “it cannot be indexed with any degree of consistency” (p. 45). Others, tracing back at least to Enser (1995), equate the About-aspect with Iconology.

2.2.2 Barthes’ Denotation and Connotation

Ørnager (1997) argues that Panofsky’s hierarchy and the Shatford/Panofsky matrix can be tied to Roland Barthes’ levels of understanding images (Barthes, 1957, 1961, 1978). Barthes was a literary theorist and semiotician who studied (among many other things) the meaning of photographs and advertisements. According to Barthes, a photograph can be said to convey meaning at two levels: Denotation and Connotation. The former corresponds to the objective contents of the image, while the latter corresponds to our associations with the image, and the implicit message behind the image. Ørnager equates Barthes’ Denotation and Connotation with Shatford’s Of and About-aspects, respectively.²

2.2.3 Understanding the semantic gap

Shatford’s work has been referenced by Hodosh et al. (2013) as a source for the three kinds of image descriptions defined earlier in Section 1.3 (conceptual, perceptual, and non-visual descriptions). They argue that automatic image description systems should aim to generate conceptual descriptions, that provide concrete information about the depicted scene and entities. This goal roughly corresponds to Panofsky’s first two levels, and to Shatford’s Of and Barthes’ Denotation aspects.

Theories about different levels of interpretation may help us reflect on the information that a picture may convey, and hypothesize about the nature of the semantic gap. For example, one possible hypothesis might be that image description systems are better at identifying what a picture is Of than what it is About, since the latter typically requires a higher level of abstraction. A naive version of this hypothesis might be illustrated as in Figure 2.1.

We could also take our cue from the multifaceted approach of Shatford (1986). Instead of a single dimension from zero to full comprehension, we can also consider image understanding as the complex ability to understand Who and What are depicted, and Where and When the

²Next to these two levels, Barthes also proposes a third level of meaning: the linguistic message, corresponding to the “textual matter in, under, or around the image” and what that textual matter refers to (Barthes, 1978). The linguistic message is important for advertisements (Barthes, 1978) and pictures in newspapers (Barthes, 1961), because it affects how the images are interpreted. In this context, Barthes also talks about Anchorage and Relay. Text can help anchor the meaning of an image; i.e. help us understand how an image should be interpreted. And text can also serve as a relay in that it can help communicate messages that are hard or impossible to convey through images alone. We will not look into this, as this thesis focuses on decontextualized images.
2.3 Pragmatic factors in image description

The semantic gap has been defined by Smeulders et al. (2000) and Hare et al. (2006) in terms of image understanding: identifying the components of an image and how they relate to each other. The goal is to understand the semantics of an image (what the image denotes, in Barthes’ terminology). One important difference between image description and full image understanding is that people are usually not exhaustive in their descriptions, simply because they consider some parts to be irrelevant to report (as we discussed in §1.1). This does not mean that image description is easier than identifying all the contents of an image. Rather, image description comes with the additional challenge of identifying which parts of the image are actually relevant to mention. This behavior does not fit into earlier characterizations of the semantic gap, because it goes beyond the level of semantics. For image description, we need to modify Hare et al.’s (2006) proposal as in Figure 2.3 to add an additional, pragmatic level.

In its broadest sense, pragmatics is the study of language use (Levinson, 1983). A central figure in pragmatics is the philosopher H.P. Grice (1913-1988), who argued that in normal conversations, speakers typically follow the Cooperative Principle: “Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or
Figure 2.3 Update to Hare et al.’s (2006) proposal. We added a pragmatic level on top of the semantic level, to account for the fact that people may only report a selection of the information contained in an image.

direction of the talk exchange in which you are engaged” (Grice, 1975). This principle can be divided into four conversational maxims (cited from Grice 1975):

**Quantity** Make your contribution as informative as is required (for the current purposes of the exchange). Do not make your contribution more informative than is required.

**Quality** Try to make your contribution one that is true. (1) Do not say what you believe to be false. (2) Do not say that for which you lack adequate evidence.

**Relation** Be relevant.

**Manner** Be perspicuous: (1) Avoid obscurity of expression. (2) Avoid ambiguity. (3) Be brief (avoid unnecessary prolixity). (4) Be orderly.

Grice’s conversational maxims are reasonable assumptions about how people tend to behave in cooperative conversation. Once we assume that a speaker is cooperative, we can use these maxims to reason about the intended meaning of their utterances. For example, consider the following exchange (again due to Grice):

(1) **Context:** Marten is standing next to his immobilized car.
   Marten: I am out of petrol.
   Filip: There’s a garage round the corner.
   ⇒ You may be able to get some petrol there.

   If we assume Filip to be helpful, their utterance should be relevant to Marten’s utterance. Even though Filip did not say so explicitly, Marten may reasonably conclude that Filip thinks the garage is likely to be open, and that it has petrol to sell. (Or at least that Filip does not have any reason to believe otherwise.) Another example concerns the use of quantifiers, such as some, most, all. Consider the next exchange (adapted from Van Tiel 2014).

(2) **Piek:** Was the exam difficult?
   Hennie: Most of the students failed.
   ⇒ Not all of the students failed

   From Hennie’s statement, we may conclude (through the maxim of Relevance) that the exam was difficult. But we may also infer that not all students failed the exam, through the maxim of Quantity: if it were the case that all students failed, Hennie could have been more...
informative by saying so. Because he did not, we may conclude that at least some students passed the exam. Examples like these are also called scalar implicatures (Horn, 1972). The idea is that sets of expressions like some, most, all can be represented on a scale from least to most informative. The use of a less informative term tends to implicate that, according to the speaker, the stronger, more informative term does not hold. Some examples of scales are given in (3, adapted from Levinson 1983).³

(3) a. \{some, most, all\}  
    b. \{or, and\}  
    c. \{1,2,3,4,5,…,n\}  
    d. \{lukewarm, warm, hot, scalding\}  
    e. \{sometimes, often, always\}  
    f. \{like, love\}

As can be seen from the examples above, pragmatic reasoning often uses the concept of alternative utterances: things the speaker could also have said in the same situation, but for some reason chose not to say. Often this comes in the form of “If the speaker believed that X instead of Y, then they should have said so.” The inferred reason for making a particular utterance adds a new layer of meaning to that utterance. Especially in the first part of this thesis, we will also employ this kind of pragmatic reasoning to better understand the data in image description corpora like Flickr30K or MS COCO. One interesting aspect of these corpora is that they already contain multiple descriptions, so we can directly compare each utterance with what other people have said in the same situation. Consider the toy example below, with the image in Figure 2.4 and two sets of descriptions in (4) and (5).

(4) a. Two strange animals next to the river.  
    b. Looks like two duck-billed otters.

(5) a. Two platypuses at the riverside.  
    b. One platypus is about to swim, while the other looks at him.

The subject of the picture is quite clear to the informal viewer: two platypuses. But the descriptions in (4) do not refer to them as such. These two descriptions implicitly signal, through their avoidance of the term platypus, that the authors do not know what kind of animals these are exactly. The two descriptions also show two strategies for handling unfamiliar entities: either use a more general term (animals), or describe their general characteristics (duck-billed, otter-like). Knowledge of these strategies is part of the pragmatic level.

The descriptions in (5) capture different aspects of the image. Which one is better depends on the context.⁴ The former (5a) describes what the picture shows, while the latter (5b) describes what the two platypuses are doing. The second description is also more speculative; while it is reasonable to expect that one of the platypuses is about to swim, there is no way

³Though not all scalar expressions give rise to an implicature at the same rate (Van Tiel et al., 2016).
⁴More specifically, the Question Under Discussion (QUD), see Roberts 1996; Benz and Jasinskaja 2017.
for us to know for sure. From a Gricean point of view, we might say that there is a trade-off here between Quantity (how informative we’d like to be) and Quality (how much evidence is required before we make any claims). Different situations may call for a different balance between the two. Being able to assess the situation and make that judgment is also part of the pragmatic level.

2.4 Image description datasets

Experiments in linguistics and psychology have traditionally been fairly small. For example, Marszalek et al. (2011) found that the median sample size for psychology experiments between 1977 and 2006 is between 32 and 60 participants. With the advent of crowdsourcing, it has become possible to carry out experiments on a much larger scale. In Natural Language Processing (NLP), many experiments are carried out under the guise of ‘data collection’ or ‘annotation’. We will focus on one such experiment: what happens if you ask a large group of crowd-workers to describe an even larger collection of images? This chapter explores one of the largest datasets of described images (Flickr30K, Young et al. 2014), and uses a data-driven approach to show the richness and subjectivity of crowd-sourced image descriptions.

The Flickr30K dataset contains over 30,000 images, with 5 English descriptions per image. These descriptions were collected via a relatively uncontrolled elicitation task, posted on Amazon Mechanical Turk. After passing a qualification test (to check their English skills), participants were able to enlist in the image description task. In this task, participants are shown some example images and descriptions, and provided with the following instructions (from the appendix of Hodosh et al. 2013, edited for brevity):

1. Describe the image in one complete but simple sentence.
2. Provide an explicit description of prominent entities.
3. Do not make unfounded assumptions about what is occurring.
4. Only talk about entities that appear in the image.
5. Provide an accurate description of the activities, people, animals and objects you see depicted in the image.
6. Each description must be a single sentence under 100 characters.

Participants are then asked to describe five images, in return for $0.10. Each image prompt is presented as in Figure 2.5.

Having finished the task, participants may annotate more batches of five images, for $0.10 per batch. Rashtchian et al. (2010) and Hodosh et al. (2013, in the Appendix) provide more details. The procedure for MS COCO is very similar (Lin et al., 2014; Chen et al., 2015). One of the main differences between the two is that the MS COCO instructions ask participants not to start their descriptions with there is . . . , which may lead them to use different syntactic constructions, but otherwise the instructions are practically identical. We may refer to this format as the canonical image description task. This chapter provides a characterization of the descriptions that were elicited using this task. Later chapters explore how these descriptions are affected by modifying the task, specifically the language of the task (Chapter 3) and the modality of the task (Chapter 5).

---

5This means that workers on Mechanical Turk should describe 365 images per hour (roughly 6 per minute) to be able to earn the current US minimum wage of $7.25. Low wages like these are common on Mechanical Turk, but more and more researchers are calling for fairer treatment of crowd-workers. See e.g. (Fort et al., 2011).
2.5 Image description as perspective-taking

Whenever you are asked to describe an image, you have to choose what to describe, and how to describe it. Levelt (1999) notes that, when you have decided what to say, there may be countless ways of expressing that information. Consider Figure 2.6:

Figure 2.6 A tree and a house, image composited from two clipart images (both public domain) by users rdevries (the house) and talekids (the tree) on Openclipart.org. This image is based on the drawing in Levelt 1999, page 92 (his figure 4.3).

Levelt notes that we may describe this image as in (6):

(6) a. There is a house with a tree to the left of it.
   b. There is a tree with a house to the right of it.

Both are valid descriptions of the scene in Figure 2.6, but the first description focuses on the house (orienting the tree with respect to the house), while the second description focuses on the tree. Levelt calls this perspective-taking, and notes that perspective-taking is at the core of all conceptual preparation for speech. We can also find it in the use of kinship terms (another of Levelt’s examples). Both sentences in (7) express the same relation:
(7) a. John is Peter’s father.
  b. Peter is John’s son.

More examples can be found in the work of Clark (1997), who argues that children are taught to handle multiple perspectives from a young age. Adults use different terms to refer to the same entities all the time (e.g. the dog, our pet, that animal). From these different uses, children may also infer pragmatic information about when to use them.

As the Flickr30K and MS COCO data contain multiple descriptions for the same images, from different crowd-workers, each annotated image comes with a set of different perspectives on the same situation. The next section explores the variation in how the same (or similar) entities are described.

2.6 Variation

Looking at the image descriptions in Flickr30K and MS COCO, we can see that there is a high degree of variation, both at the phrase level and at the sentence level. We explore the former now, and leave the latter for the next chapter. The goal of this section is to get a sense of the range of different expressions used by crowd workers in their descriptions.

2.6.1 Clustering entity labels

The Flickr30K dataset has been enriched with links between the descriptions and the images (Plummer et al., 2015). Each entity label (a phrase describing a person or object) is linked with a bounding box marking the relevant entity in the image. Figure 2.7 provides an example. Because each image has 5 different descriptions, each bounding box may be linked with multiple entity labels (unless only one description makes reference to the relevant entity). If we find different labels that refer to the same bounding box, we know that these are alternative ways to refer to the same entity. We can use this information to find clusters of labels that refer to similar entities. We used the Louvain method for this.

![Figure 2.7](image126594141.png)

**Figure 2.7** Image with bounding boxes indicating the entities referred to in the description, along with three sets of similar labels that would be extracted by the proposed algorithm. Data from the Flickr30k Entities dataset, visualization from the online dataset browser. Original picture by mayamoose (CC BY-NC-SA) on Flickr.com

The Louvain method is a graph clustering algorithm that is designed to optimize the modularity of each of the clusters (Blondel et al., 2008). In other words, it tries to find groups
of nodes (points in a network), such that the nodes within those groups are well-connected to each other, but only sparsely connected to nodes in other groups (if they are connected at all). Figure 2.8 provides an example of a clustered graph.

![Figure 2.8 Example of a modular graph, where modules are colored after clustering the nodes using the Louvain method. Image generated using Gephi (Bastian et al., 2009).](image)

To use the Louvain method, we need to translate the task of finding similar entity labels into a graph clustering problem. This is a natural fit, because the entity labels in the Flickr30K-Entities dataset are already linked to each other through the bounding boxes they are associated with. We can translate the Flickr30K-Entities data into a graph by representing each entity label as a node. Whenever two labels co-refer to the same bounding box, we say that there is a connection between them. This way, similar entity labels will be connected to each other, and we end up with a graph (or multiple separate graphs) of entity labels. Algorithm 2.1 provides an example implementation of the graph building code. Because the dataset was manually annotated, and may contain noise, we used a frequency threshold of 2. This means that two entity labels should co-occur at least 2 times before we make a connection between them.

After applying this algorithm to the Flickr30K Entities dataset, the \texttt{label\_graph} object contains many but not all labels from the annotated data. Labels that never co-occur twice with another label are not included. We refer to these labels as ‘orphans’ as they do not have any attachment to other labels. To remedy this situation, we first clustered \texttt{label\_graph}, generating lists of similar labels. Following this, we added the ‘orphaned’ labels to the list with the highest count of labels co-occurring with them in the Flickr30K-Entities data. Using this approach, we obtained 749 clusters. Inspecting the clusters, we can see that they capture a wide range of terms to refer to similar entities. For example, here is a cluster of different ways to refer to beards, moustaches, etc.

<table>
<thead>
<tr>
<th>beard</th>
<th>white beard</th>
<th>long brown beard</th>
<th>large white beard</th>
<th>goatee</th>
<th>red beard</th>
<th>flaming red beard</th>
<th>thick beard</th>
<th>beard and mustache</th>
<th>braided beard</th>
<th>gray beard</th>
<th>gray braided beard</th>
<th>long, white beard</th>
<th>short beard</th>
<th>neatly trimmed beard</th>
<th>scruffy beard</th>
<th>black beard</th>
</tr>
</thead>
</table>

These terms include references to the kind of hair (beard, goatee, mustache), the color (gray, black, white), length (long, short), size (big, large), orderliness (neatly trimmed, scruffy), and presentation (braided). This means that, when asked, people consider at least six different variables just to describe male facial hair. Furthermore, it is worth pausing to think about the situations when one would use these kinds of descriptions. To take just one example,
def build_graph(images, threshold = 2):
    """
    Function that takes a set of annotated images, and returns a graph
    where co-referring expressions are linked.
    """
    link_counts = defaultdict(int)
    for image in images:
        label_index = defaultdict(list)  # reset for every image.
        # Loop over descriptions and collect referring expressions:
        for description in image.descriptions:
            annotations = get_annotations(description)
            for bounding_box_id, label in annotations:
                label_index[bounding_box_id].append(label)
            # Update the counts for combinations of labels.
            for list_of_labels in label_index.values():
                for pair_of_labels in combinations(list_of_labels, 2):
                    link_counts[pair_of_labels] += 1
        # Build the graph
        label_graph = Graph()
        for pair_of_labels, count in link_counts.items():
            if count >= threshold:
                label_graph.add(pair_of_labels)
        return label_graph

Algorithm 2.1 Function to produce a graph connecting similar referring expressions (code simplified for presentation).

when would it be appropriate to say that someone has red facial hair? This expression is marked (in the third sense of Haspelmath 2006, see also Horn 1984, p. 22): it is a complex expression, used while simpler, lexicalized alternatives are available (e.g. beard, moustache, goatee). When speakers are going out of their way to express themselves like this, we may infer (through Grice’s (1975) maxim of Manner) that the phrase facial hair refers to something that is not quite like a beard, moustache, or goatee (yet), but of a more undefined nature.

Appearance versus context

We also observe that some labels are more appearance-based while others are more context-dependent. For example, police officers are immediately recognizable through their uniform. On the other hand, a bystander may only be labeled as such because of external factors (e.g. an accident happened close to where they are standing). Sometimes both appearance and situation are important, as with civilians, who are only labeled in the presence of police officers or members of the military, and if they are not wearing any uniform themselves. We can express this difference in a matrix, as in Table 2.2. Alternatively, we may imagine the labels as points in between the two forces that drive the labeling process (as in Figure 2.9).

2.6.2 Describing different people

Besides clustering all entity labels, we can also create a taxonomy and manually sort them into different semantic categories. The advantage of manually sorting the labels is that we have
### Table 2.2 A categorization of labels based on whether the label is applied on the basis of someone’s appearance or the situation they are in.

<table>
<thead>
<tr>
<th>Appearance</th>
<th>Situation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
<td>Police officer, businessman, firefighter</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Civilian</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Bystander, neighbor, passerby, orphan</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2.9** Continuous scale from Appearance-based to Contextually determined labels.

...full control over the categories. This makes it possible to make more fine-grained distinctions, and to show the breadth of the label distribution. We again use the Flickr30K-Entities corpus (Plummer et al., 2015), focusing on the different ways that crowd-workers describe other people. This restriction keeps the categorization task manageable.

**Initial selection**

The starting point for our categorization is a list of labels. We compiled this list using the Flickr30K-entities annotations provided by Plummer et al. (2015), and listed all labels that were classed as **people**. After normalization, we found 19,634 unique labels, which is too much to categorize by hand. It is not possible to crowd-source our categorization task, because the categories are not known beforehand.) Hence we focus our efforts only on the 5,526 labels that end with any of the nouns girl, boy, woman, man, female, male, or any of their plural forms.

Examples of such labels are: **barefooted little girl**, **casually dressed man**, and **husky little boys**.

Our selection makes the task more manageable, but it also reduces the variation in the data because the selected labels are more homogeneous. Specifically, we ignore all noun heads except for the abovementioned gendered nouns. The list below shows the most common excluded head nouns. Nevertheless, as we will see in Section 2.6.2, we still found a broad range of variation in the labels.

| people,   | band,  | adults, | riders, | biker, | performers, |
| player,   | kids,  | teams,  | artists, | officer, | musician,  |
| children, | couple,| guys,   | musicians,| individuals, | spectators, |
| players,  | worker,| dancers,| friends, | runners, | performer, |
| team,     | crowd, | vendor, | dancer, | kid, | onlookers, |
| child,    | guy,   | ladies, | toddler, | runner, | driver, |
| person,   | baby,  | group,  | gentleman,| fans, | crew, |
| workers,  | students, | members, | officers, | parade, | skier, |
| lady,     | rider, | class,  | family, | cheerleaders, | cyclists |

---

6 We normalized the labels by lowercasing them, and removing the characters @*, &().

7 We applied the same approach to the attributes in the Visual Genome dataset (Krishna et al., 2017), but for reasons of clarity we focus on Flickr30K. Results are available online: https://github.com/evanmiltonburg/LabelingPeople
During the categorization task, we found several typing errors, and words unrelated to people-labeling. We addressed these issues by semi-automatically correcting the typing errors, and creating a list of stopwords that were automatically removed from the labels. This further reduced the number of unique labels-to-be-categorized from 5526 to 3401.

**Sorting procedure**

We manually sorted the labels into semantic categories (shown in Table 2.3). The sorting procedure works as follows.

1. Start with a set of labels to be categorized.
2. Remove task-specific stopwords and unrelated phrases (e.g. *a picture of*) from the labels. This reduces the number of unique labels.
3. Select (partial) labels from the list, add them to an existing category file, or create a new category file with those labels.
4. Match the labels with the categories. We use a context-free grammar (CFG, see Figure 2.10; implemented using the NLTK, Bird et al. 2009) because each label may consist of multiple modifiers from different categories. For example: *African-American young man* has both ethnicity and age modifiers.
5. Remove matches from the set of labels to be categorized.
6. Either stop categorization, or go to 3.

Our goal is to get an overview of the different kinds of labels used by the crowd-workers, not to achieve a perfect categorization of all labels. Thus, our stopping criterion is as follows. The sorting task is finished whenever there are no more examples matching existing categories, or warranting new categories. New categories are warranted if there are multiple labels that clearly fall under the same umbrella, but do not fit into any of the existing categories.

**Results**

We sorted the (partial) labels into 20 different categories, until we were left with only 341 labels (10%) that could not be fully matched with our categories by the CFG matcher. Examples of uncategorized labels are *birthday girl* and *blood pressure of a man*. The former could be classed as a role associated with an event, but we did not find many such examples. The latter is an artifact of the automated label categorization process for the Flickr30K Entities dataset.

Table 2.3 shows the 20 different label categories, with examples for each category. With this table, we have an empirically derived taxonomy that provides an overview of the choices
Table 2.3 Taxonomy of labels referring to other people, with selected examples for each category. All examples are (partial) labels from the Flickr30K dataset.

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ability</td>
<td>wheelchair bound, able-bodied, disabled, handicapped, blind, one-armed</td>
</tr>
<tr>
<td>Activity</td>
<td>running, chasing, waving, speaking, parachuting, roller-skating, protesting</td>
</tr>
<tr>
<td>Age</td>
<td>young, middle-aged, adult, elderly, infant, twenty-something, teen-aged</td>
</tr>
<tr>
<td>Attractiveness</td>
<td>attractive, beautiful, pretty, sexy, cute, ugly, hot, handsome, nice</td>
</tr>
<tr>
<td>Build</td>
<td>petite, muscular, slender, lanky, heavy chested, potbellied, well built, burly</td>
</tr>
<tr>
<td>Cleanliness</td>
<td>dirty, shaggy, scruffy, mussy, disheveled, well-groomed, dirty faced</td>
</tr>
<tr>
<td>Clothing – Amount</td>
<td>shirtless, topless, barefooted, scantily clad, nude, unclothed, undressed</td>
</tr>
<tr>
<td>– Color</td>
<td>green black uniformed, brightly dressed, red shirted, colorfully clothed</td>
</tr>
<tr>
<td>– Kind</td>
<td>uniformed, casually dressed, sari-garbed, leather-clad, robed, suited, kilted</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>african-american, oriental, caucasian, chinese, foreign, middle-eastern</td>
</tr>
<tr>
<td>Eyes</td>
<td>blue-eyed, brown eyed, green eyed, bespectacled, glasses-wearing</td>
</tr>
<tr>
<td>Fitness</td>
<td>physically fit, healthy fit, healthy and fit, weak looking, out-of-shape</td>
</tr>
<tr>
<td>Group</td>
<td>cast, circle, audience, crowd, ensemble, couple, team, roomful, group, trio</td>
</tr>
<tr>
<td>Hair – Color</td>
<td>blond, dark-haired, brown-haired, brunet, redheaded, fair, dark, ginger</td>
</tr>
<tr>
<td>– Facial</td>
<td>bearded, goateed, white-bearded, mustachioed, stubbled, clean-shaven</td>
</tr>
<tr>
<td>– Length</td>
<td>bald, short-haired, long-haired, balding, nearly bald, shaved head</td>
</tr>
<tr>
<td>– Style</td>
<td>curly-haired, frizzy-haired, pony-tailed, shaggy-haired, curly, dreadlocked</td>
</tr>
<tr>
<td>Height</td>
<td>tall, short, petite, taller, long, littler, tall looking, shorter, rather tall</td>
</tr>
<tr>
<td>Judgment</td>
<td>stylish, tacky looking, strange, silly, odd looking, hip, comical, flamboyant</td>
</tr>
<tr>
<td>Mood</td>
<td>happy, excited, curious, enthusiastic, tired, thoughtful, pensive, weary, sad</td>
</tr>
<tr>
<td>Occupation</td>
<td>military, navy, photographer, coast guard, executive, cooking professional</td>
</tr>
<tr>
<td>Religion</td>
<td>muslim, hindu, amish, christian, islamic, religious, jewish, mormon, hindu</td>
</tr>
<tr>
<td>Social group</td>
<td>homeless, goth, hippie, rasta, peasant, unemployed, poor looking, trash</td>
</tr>
<tr>
<td>State</td>
<td>drunk, extremely drunk, wet, bloody, pregnant, sweaty, cold, handcuffed</td>
</tr>
<tr>
<td>Weight</td>
<td>overweight, fat, slim, skinny, obese, plump, heavyset, heftier, heavy, hefty</td>
</tr>
</tbody>
</table>

that crowd-workers have to make in order to describe other people. The different categories show the diversity and breadth of the label distribution. In future work, we hope to extend the coverage of our taxonomy (ideally to all 19,634 person-labels in Flickr30K-Entities), and present statistics about the proportion of person-labels from the Flickr30K dataset that fall into each category.

Our taxonomy also provides a reference point to think about the characteristics that we would like image description systems to describe, and also about the features we would not want those systems to refer to. For example, it seems to us that automatic description of features like RELIGION, WEIGHT, or SOCIAL GROUP would probably do more harm than good. Table 2.3 also shows us what makes image description difficult. For this domain alone, to produce human-like descriptions, systems need to be able to predict 20 different kinds of features, and decide which feature values are relevant to mention. A further complication is that even after deciding which characteristics to describe, there are still within-category choices to be made. For example, when describing a game of basketball, one might choose to talk about a man playing basketball (seeing basketball-playing as a transient property), or male basketball player (seeing basketball-playing as an inherent property). See Beukeboom
Chapter 2  How people describe images

2014; Fokkens et al. 2018 for a discussion and further references relating to this issue.

Related work

This section explored how American speakers of English describe other people in the Flickr30K dataset, and what features may be used in those descriptions. It is still an open question what drives people to prefer one feature over another. One way to come closer to answering that question, is to collect more data specifically geared towards the description of other people. Gatt et al. (2018) provide such a dataset, called Face2Text, which contains face images with natural language descriptions. The dataset is provided with demographic information about the participants in the description task, and there are equally many images of male and female faces. With this kind of data, we may be able to see e.g. whether men are described differently from women, or whether the age/gender/country of origin of the participants has any effect on the descriptions.

Gatt et al. (2018) present their dataset as a resource for training image description systems to produce rich face descriptions. At the same time they note that next to physical (blonde) and emotional (happy) properties, their participants also speculate about other characteristics that the subjects in the images may have. This is problematic for systems aiming to generate factual descriptions. One way to proceed is to categorize the different kinds of properties that people may refer to in their descriptions (as we have done above), and to assess which properties can reliably be predicted from an image and, in a next step, which of those properties we would like an automatic image description system to produce.

In earlier research, Song et al. (2017) present a system that is able to predict (to varying degrees of success) perceived social attributes from faces. Human participants rated faces from a large database for their attractiveness, friendliness, familiarity, but also to what extent they thought the subjects were egotistical, emotionally stable, or responsible.\(^8\) It is important to stress again that these ratings only indicate perceived characteristics, and do not necessarily reflect the actual characters of the individuals in the dataset. The following quote by Todorov et al. (2013) is very apt (also see Agüera y Arcas et al. 2017):

\begin{quote}
The idea that the face reflects one’s personality could be found in every ancient culture, and reached its prime in 19th century physiognomy — the pseudo-science of reading personality from faces. Physiognomy has been long discredited as a science for good reasons, but physiognomists got a few things right. Firstly, people make all kinds of social judgments from faces of strangers; secondly, there is consensus in these judgments; and thirdly, these judgments matter for social interaction.

(Todorov et al., 2013, p. 373)
\end{quote}

This quote should serve as a warning that, even though people may be able to consistently ascribe a particular property to an individual, this alone does not entail that the property actually applies.

\(^8\) Song et al. (2017) list the following 20 pairs of social traits: (attractive, unattractive), (happy, unhappy), (friendly, unfriendly), (sociable, introverted), (kind, mean), (caring, cold), (calm, aggressive), (trustworthy, untrustworthy), (responsible, irresponsible), (confident, uncertain), (humble, egotistical), (emotionally stable, emotionally unstable), (normal, weird), (intelligent, unintelligent), (interesting, boring), (emotional, unemotional), (memorable, forgettable), (typical, atypical), (familiar, unfamiliar) and (common, uncommon).
2.7 Stereotyping and bias

As we mentioned in Chapter 1, a common assumption behind image description datasets is that the descriptions provide an objective indication of the contents of an image. In other words, the descriptions are based on the images, and nothing else. Here is the relevant quote from Hodosh et al. (2013), repeated for convenience:

“By asking people to describe the people, objects, scenes and activities that are shown in a picture without giving them any further information about the context in which the picture was taken, we were able to obtain conceptual descriptions that focus only on the information that can be obtained from the image alone.” (Hodosh et al., 2013, p. 859)

We referred to this as the assumption of neutrality, and noted that it is often a useful assumption to make; if the descriptions are at least somewhat predictable on the basis of visual features alone, we can try and learn a mapping between visual features and image descriptions. But what the assumption of neutrality overlooks is the amount of interpretation or recontextualization carried out by the annotators. Consider Figure 2.11.

![Figure 2.11 Image 8063007 from the Flickr30K dataset. Author and license unknown.](image)

This image comes with the five descriptions below. All but the first one contain information that cannot come from the image alone. Relevant parts are highlighted in **bold**:

1. A blond girl and a bald man with his arms crossed are standing inside looking at each other.
2. A **worker** is **being scolded** by her **boss** in a **stern lecture**.
3. A **manager talks to an employee about job performance**.
4. A **hot, blond girl getting criticized by her boss**.
5. **Sonic employees** **talking about work**.

We need to understand that the descriptions in the Flickr30K dataset are subjective descriptions of events. This can be a good thing: the descriptions tell us what are the salient parts of each image to the average human annotator. So the two humans in Figure 2.11 are relevant, but the two soap dispensers are not. But subjectivity can also result in stereotyping descriptions, in this case suggesting that the male is more likely to be the manager, and the female is more likely to be the subordinate. Rashtchian et al. (2010) do note that some descriptions are speculative in nature, which they say hurts the accuracy and the consistency of the descriptions. But the problem is not with the lack of consistency here. Quite the contrary: the problem is that stereotypes may be pervasive enough for the data to be consistently biased. And so language

---

9 The quote is about the Flickr8K dataset, a subset of Flickr30K.
models trained on this data may make incorrect inferences and propagate harmful stereotypes, such as the idea that women are less suited for leadership positions.

Next to the manager-worker inference, the annotators also speculate about the activity taking place in the image (scolding, talking, criticizing), the mood of the presumed conversation (stern, criticizing), and the topic of the conversation (work). Finally, one crowd-worker also mentions the attractiveness of the woman on the left in their description. One might consider this a form of bias, since the attractiveness of the male on the right is not discussed.

Stereotype-driven descriptions

Stereotypes are ideas about how other (groups of) people commonly behave, what properties they tend to have, and what they are likely to do. These ideas guide the way we talk about the world. We distinguish two kinds of verbal behavior that result from stereotypes: (i) unwarranted inferences and (ii) linguistic bias.

Unwarranted inferences are the result of speculation about the image; here, the annotator goes beyond what can be glanced from the image and makes use of their knowledge and expectations about the world to provide an overly specific description (van Miltenburg, 2016). Unwarranted inferences are directly identifiable as such, and in fact we have already seen four of them (descriptions 2–5) discussed earlier.

Linguistic bias is discussed in more detail by Beukeboom (2014), who defines linguistic bias as “a systematic asymmetry in word choice as a function of the social category to which the target belongs.” So this bias becomes visible through the distribution of terms used to describe entities in a particular category. Generally speaking, people tend to use more concrete or specific language when they have to describe a person that does not meet their expectations. Beukeboom (2014) lists several linguistic ‘tools’ that people use to mark individuals who deviate from the norm. We will mention two of them (examples also due to Beukeboom 2014):

**Adjectives** One well-studied example Stahlberg et al. (2007); Romaine (2001) is sexist language, where the sex of a person tends to be mentioned more frequently if their role or occupation is inconsistent with ‘traditional’ gender roles (e.g. female surgeon, male nurse). Beukeboom also notes that adjectives are used to create “more narrow labels [or subtypes] for individuals who do not fit with general social category expectations” (p. 3). E.g. tough woman makes an exception to the ‘rule’ that women aren’t considered to be tough.

**Negation** can be used when prior beliefs about a particular social category are violated, e.g. *The garbage man was not stupid*. See also Beukeboom et al. (2010).

These examples are similar in that the speaker has to put in additional effort to mark the subject for being unusual. But they differ in what we can conclude about the speaker, especially in the context of the Flickr30K data. Negations are much more overtly displaying the annotator’s prior beliefs. When one annotator writes that *A little boy is eating pie without utensils* (for the image in Figure 2.12), this immediately reveals the annotator’s normative beliefs about the world: pie should be eaten with utensils. But if another annotator would talk about *a female basketball player* for the image in Figure 2.13, this cannot be taken as an indication that the annotator is biased about the gender of basketball players; they might just be helpful by providing a detailed description. In Section 2.9 we will discuss how to establish whether or not there is any bias in the data regarding the use of adjectives.
2.8 Categorizing unwarranted inferences

Browsing through the Flickr30K corpus, one quickly notices different kinds of unwarranted inferences that are made by the crowd-workers. We carried out a pilot study to make an initial taxonomy of those different kinds of inferences, and to find examples for each of those categories. We wrote an inspection tool to browse the Flickr30K dataset and add notes about the images and their descriptions (see Appendix A). After inspecting a subset of the Flickr30K data, we have grouped the unwarranted inferences into six categories, presented below with an example for each category.

**Goal** Quite a few annotations focus on explaining the *why* of the situation. For example, in one of the images, a man is fastening his climbing harness. One of the crowd-workers noted he was doing so *in order to have some fun*.

In an extreme case, one annotator wrote about the picture on the right, showing a dancing woman, that *the school is having a special event in order to show the american culture on how other cultures are dealt with in parties*. This is reminiscent of the Stereotypic Explanatory Bias (Sekaquaptewa et al., 2003, SEB), which refers to “the tendency to provide relatively more explanations in descriptions of stereotype inconsistent, compared to consistent behavior” (Beukeboom et al., 2010).
Activity We’ve seen an example of this earlier in Section 2.7, where the ‘manager’ was said to be talking about job performance and scolding [a worker] in a stern lecture. The picture on the right shows another example, where an annotator described the three men as sitting and contemplating their next bull ride. The Flickr30K dataset also has several images that are ambiguous in the actions that are depicted, e.g. opening/closing a door, throwing/catching a ball.

Ethnicity It is almost impossible to infer someone’s ethnicity or nationality from an image alone, but crowd-workers seem to have no problem with this. Many dark-skinned individuals are called African-American regardless of whether the picture has been taken in the USA or not. And people who look Asian are called Chinese (such as the woman in the image on the right) or Japanese.

Event In the image on the right, people sitting at a gym are said to be watching a game, even though there could be any sort of event going on. But since the location is so strongly associated with sports, crowdworkers readily make the assumption.

Relation Older people with children around them are commonly seen as parents, small children as siblings (for the picture on the right), men and women as lovers, groups of young people as friends. These kinds of relations are almost impossible to verify on the basis of an image alone, although there are different shades of gray.
2.8 Categorizing unwarranted inferences

Status/occupation Annotators will often guess the status or occupation of people in an image. Sometimes these guesses are relatively general (e.g. college-aged people being called students in image 36979), but other times these are very specific. For example, one participant called the man in the picture on the right a graphics designer (presumably because it looks like the man is drawing something). In fact, according to the author, the image shows a bookbinder in his Parisian workshop.

This categorization is not meant to be exhaustive, but rather to provide empirical evidence that crowd-workers do not necessarily produce objective descriptions of the images in the Flickr30K dataset. Given this evidence, we can ask ourselves how these kinds of speculative descriptions arise. Answering this question brings us closer to an understanding of the human image description process.

2.8.1 Accounting for unwarranted inferences

The examples provided above are unexpected, because they seem to go against the task guidelines. Specifically rule number 3: do not make unfounded assumptions about what is occurring. One explanation for this rule-breaking may be that the participants just did not bother to read the rules very well. But suppose that the participants were trying to stick to the rules. How might we explain their behavior?

One way to account for the participants’ behavior is to note that the canonical image description task is very unnatural. Imagine sitting behind your computer and being asked to provide descriptions for a series of decontextualized images. Many of the images depict everyday situations that are not particularly interesting. You are not being told about the purpose of the experiment, so the question under discussion is unclear. In other words: you have no idea what to say, because you don’t know what the experiment is about. Still, there must be some purpose to the task. Left wondering how their description will be used, participants might just be providing as much information as possible. And because the images are presented in isolation, stereotypes may be used in lieu of context to fill in the gaps.

If this characterization is on the right track, then we might improve the image description task by introducing an explicit goal (what will the descriptions be used for) as well as an audience (who will be reading the descriptions). Either way, this section has shown that we cannot blindly trust image description data to be restricted to factual descriptions. Participants may go beyond the contents of the image, and into the realm of speculation.

The question under discussion (QUD) is an analytical tool to reason about the suitability or interpretation of individual utterances in a particular discourse. The basic idea is that every conversation is guided by (implicit or explicit) questions that speakers try to provide the answers to. Utterances they make can then be interpreted in terms of those questions (Roberts 1996, see also Benz and Jasinskaja 2017).

During the collection of the German descriptions for the Multi30K dataset Elliott et al. (2016), the authors found that the German crowd-workers were discussing how boring and repetitive the task was (Desmond Elliott, personal communication). Thus, another explanation for the participants’ behavior is that they were not motivated enough to provide accurate descriptions. One remedy for this might be to make the task seem more worthwhile by explaining the purpose of the task. For example: “by writing these descriptions, you are contributing to better assistive technology, helping other people.”
2.9 Detecting linguistic bias: adjectives

We have discussed earlier that the use of adjectives and negations may reflect stereotypes carried by a speaker. This section discusses the use of adjectives, and specifically the use of ethnic markers. One pattern in the Flickr30K data is that the ethnicity/race of babies doesn’t seem to be mentioned unless the baby is black or Asian. In other words: white seems to be the default, and others seem to be marked (Jakobson, 1972). This phenomenon is also called reporting bias, see e.g. Misra et al. 2016.

2.9.1 Estimating linguistic bias in image descriptions

How can we tell whether or not the data is actually biased? The Flickr30K images are not labeled by social class, and so we don’t know whether or not an entity belongs to a particular social class (or in this case: ethnic group) until it is marked as such. In this subsection, we first show a method to (roughly) estimate whether there are any differences in the way that different social groups are marked. Later, in Section 2.9.2, we will show the results of the more precise, annotation-based approach.

Approach. We first tried to approximate the proportion by looking at all the images where the annotators have used a marker (in this case: adjectives like black, white, Asian), and for those images count how many descriptions (out of five) contain a marker. This gives us an upper bound that tells us how often ethnicity is indicated by the annotators. Note that this upper bound lies somewhere between 20% (one description) and 100% (5 descriptions).

Set-up. This study is set up such that the results can easily be compared with the annotation-based approach in Section 2.9.2. Because manual annotation is a labor-intensive process, we focused my efforts on the baby-domain. In other words: we looked at all pictures with babies in them, and ignored the images with only adults and no babies. We searched the entire Flickr30K corpus for descriptions matching the pattern (Asian|white|black|African-American|skinned) baby. Then, for each image with one or more matching descriptions, we counted the number of descriptions with a racial/ethnic marker in them, discarding all false positives (where the picture does not show babies at all).

Results. Table 2.5 presents count data for the ethnic marking of babies. It includes two false positives (talking about a white baby stroller rather than a white baby). In the Asian group there is an additional complication: sometimes the mother gets marked rather than the baby. E.g. An Asian woman holds a baby girl. We have counted these occurrences as well.

The numbers in Table 2.5 are striking: there seems to be a real, systematic difference in ethnicity marking between the groups. Whenever the ethnicity of a baby is mentioned by at least one annotator, there is a greater chance of others mentioning the baby’s ethnicity as well if the baby is Asian than if the baby is White. We also observe this effect for Black versus White babies. The next section takes our analysis one step further, and looks at all the 697 pictures with the word ‘baby’ in it. We will show that there are disproportionately many white babies in the dataset, which strengthens the conclusion that the dataset is indeed biased.

2.9.2 Validation through annotation

The method presented in the previous section is very coarse-grained, because it only enables us to find images where crowd-workers applied racial/ethnic markers. The results are skewed, because we do not get to see the images where the crowd-workers did not decide to use racial/ethnic markers. In the end, what we would like to know is whether there is any bias in the use of adjectives for all pictures of members of different social groups. The only way
### Table 2.5

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Average</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td>60%</td>
<td>11</td>
</tr>
<tr>
<td>Black</td>
<td>40%</td>
<td>3</td>
</tr>
<tr>
<td>White</td>
<td>20%</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Count</th>
<th>Description</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td>Asian child/baby</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asian baby, Asian/oriental woman</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asian girl/baby, Asian/oriental woman</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>African-American (AA)/black baby</td>
<td></td>
</tr>
<tr>
<td></td>
<td>African/AA child, black baby</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dark-skinned baby</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA baby</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>White baby boy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>White baby</td>
<td></td>
</tr>
<tr>
<td></td>
<td>White baby stroller</td>
<td>FP</td>
</tr>
<tr>
<td></td>
<td>White baby stroller</td>
<td>FP</td>
</tr>
<tr>
<td></td>
<td>Fair-skinned baby</td>
<td></td>
</tr>
</tbody>
</table>

The number of times ethnicity/race was mentioned per category, per image. The average is expressed as a percentage of the number of descriptions. Counts in the last column correspond to the number of descriptions containing an ethnic/racial marker. Images were found by looking for descriptions matching (Asian|white|black|African-American|skinned) baby. We found two false positives, indicated with FP.

To answer this question is to manually annotate all images with information about the race of the depicted individuals, and then to see for each of the different groups how often their race/ethnicity is mentioned.

**Set-up.** We first selected all images from the Flickr30K dataset with descriptions containing the word ‘baby’. Using this selection, we manually categorized each of the images as either black, white, Asian, or other. To this end, we created an annotation tool that takes a list of images, presents them in turn, and lets the user assign them to particular categories. This tool is not limited to the annotation of race/ethnicity, but could in theory be used for any kind of categorization task. See Appendix A for more information.

**Results.** Using the annotation tool, we found that there are 504 white, 66 Asian, and 36 black babies. 73 images do not contain a baby, and 18 images do not fall into any of the other categories. While this does bring down the average number of times each category was marked, it also increases the contrast between white babies and Asian/black babies. If we just focus on the images, black babies are marked as such in 4/36 images (11%), while white babies are only marked as such in 5/504 images (less than 1%). Asian babies are marked as such 4.5% of the time. It is an open question whether these observations generalize to other age groups (i.e. children and adults).\(^\text{12}\)

\(^\text{12}\) All code and data are available online through: https://github.com/evanmiltonsburg/Flickr30k-Image-Viewer
### 2.9.3 Linguistic bias and the Other

The findings above indicate that there are differences in the way that *a priori* comparable groups are treated: white people aren’t typically marked as such, while black and Asian people are marked. This kind of linguistic behavior sets up white people as the default, and non-white people as the exception. In this context, researchers in the social sciences often talk about the concept of *the Other*, which Mountz (2009) defines as follows:

The term ‘other’ serves as both a noun and a verb. By placing one’s self at the centre, the ‘other’ always constitutes the outside, the person who is different. As a noun, therefore, the other is a person or group of people who are different from oneself. As a verb, other means to distinguish, label, categorize, name, identify, place and exclude those who do not fit a societal norm. (Mountz, 2009)

That is to say, to mark specific social groups as Other is to exclude them, defining people by what they are not. So even if the individual descriptions are not necessarily wrong in their use of ethnicity-related adjectives, the corpus as a whole conveys a mostly White perspective on the world, and we should be aware of that.

### 2.9.4 Takeaway

The takeaway from this section is that adjectives are not distributed equally. Rather, we find that the distribution may be skewed by ethnicity. This finding is not unique to image descriptions, as social scientists have found similar patterns in other genres of text (Beukeboom, 2014). But to find linguistic bias in the Flickr30K data is particularly troubling because this dataset is used to train image description systems. In other words, this data is supposed to set an example for how images should be described. But the descriptions are clearly not exemplary.

### 2.10 Linguistic bias and evidence of world knowledge in the use of negations

Negations are words that communicate that something is *not* the case. They are often used when there is a mismatch between what speakers expect to be the case and what is actually the case (see e.g. Leech 1983; Beukeboom et al. 2010). For example, if Queen Elizabeth of England were to appear in public wearing jeans instead of a dress, (8a) would be acceptable because she is known to wear dresses in public. But if she were to show up wearing a dress, (8b) would be unexpected.

(8) a. Queen Elizabeth isn’t wearing a dress  
    b. ??Queen Elizabeth isn’t wearing jeans

Thus the correct use of negations often requires *world knowledge*, or at least some sense of what is expected and what is not. In (van Miltenburg et al., 2016a), we carried out a study to analyze the use of negations in the Flickr30K corpus. This analysis provides an indication of the amount of world knowledge and reasoning that is needed to generate human-like image descriptions. Here we use the term ‘world knowledge’ in a broad sense, not only including facts and statistics about the world, but also normative beliefs about how the world should be. Through the use of negations, parts of this knowledge are encoded in the Flickr30K dataset.

#### 2.10.1 General statistics

We focused on two kinds of negations: non-affixal negations (Tottie, 1980) and implicit negations (also known as *inherent negations*, e.g. Horn 1989; Morante et al. 2008). Table 2.6
provides an overview of the negations used in our study. We left affixal negations for future research.\textsuperscript{13} We used a string-matching approach to see if a description contains a negation, either matching the whole word or, in the case of verbs, the start of the word to account for differences in verb endings.

<table>
<thead>
<tr>
<th>Non-affixal negation</th>
<th>Free Bound</th>
<th>Not</th>
<th>Implicit negation</th>
<th>Verb Preposition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Never, n’t, no, none, nothing, nobody, nowhere, nor, neither</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-affixal negation</td>
<td></td>
<td>Lack, omit, miss, fail</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Without, sans, minus</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\textbf{Table 2.6} Negations used in our study.

Our search yielded 896 sentences, of which 892 unique, and 31 false positives. Table 2.7 shows frequency counts for each negation term. We carried out the same analysis for the MS COCO dataset (Lin et al., 2014) to see if the proportion of negations is a constant. Our approach yielded 3339 sentences on the training and validation splits, of which 3232 unique. The presence of negations appears to be a linear function of dataset size: 0.56\% in the Flickr30K dataset, and 0.54\% in the MS COCO dataset. This suggests that the use of negations is not particular to either dataset, but rather it is a robust phenomenon across datasets.

| No | 371 | Fail | 9 |
| Not | 198 | Never | 5 |
| Without | 141 | Nowhere | 3 |
| Miss | 69 | Neither | 2 |
| N’t | 68 | Sans | 1 |
| Nothing | 16 | None | 1 |
| Lack | 9 | Nobody | 1 |

\textbf{Table 2.7} Frequency counts for each negation term.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flickr30K</td>
<td>659</td>
<td>85</td>
<td>16</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>MS COCO</td>
<td>2406</td>
<td>277</td>
<td>78</td>
<td>30</td>
<td>5</td>
</tr>
</tbody>
</table>

\textbf{Table 2.8} Distribution of the number of descriptions of an image with at least one negation term.

Table 2.8 shows the distribution of descriptions containing negations across images. In the majority of cases only one of the five descriptions contains a negation (86.25\% in Flickr30K and 72.05\% in MS COCO). Only in very exceptional cases do the five descriptions contain negations. This indicates that the use of negation is a subjective choice.

\subsection*{2.10.2 Categorizing different uses of negations}

This section provides a categorization of negation uses and assesses the amount of required background knowledge for each use. Our categorization is the result of manually inspecting all the data twice: the first time to develop a taxonomy, and the second time to apply this taxonomy to all 892 sentences. There is already a unifying explanation for why people use negations (unexpectedness, see Leech 1983; Beukeboom et al. 2010). The question here is how people use negations, what they negate, and what kind of knowledge is required to produce those negations.

\[\text{Affixal negations are words starting with any of the negative morphemes } a-, \text{ dis-}, un-, non-, un-, or ending with the morpheme } –\text{less.}\]
Our categorization is meant to provide a general description of the different uses of negation in image descriptions. This categorization may also be used as a practical guide to be of use for natural language generation: if you want your system to be able to produce human-like descriptions including negations, these are the phenomena that the system should account for. We will now first describe eight different uses of negation, before discussing the distribution of these different uses (§2.10.3).

1. **Salient absence**: The first use of negation is to indicate that something is absent:

   (9) a. A man **without** a shirt playing tennis.
   
   ✎ You are supposed to wear a shirt while playing tennis.
   
   b. A woman at graduation **without** a cap on.
   
   ✎ You are expected to be wearing a cap.

   Shirts and shoes are most commonly mentioned as being absent in the Flickr30K dataset. From examples like (9a) speaks the norm that people are supposed to be fully dressed. These examples may be common enough for a machine to learn the association between exposed chests and the phrase *without a shirt*. But there are also more difficult cases, such as (9b). To describe an image like this, one should know that students (in the USA) typically wear caps at their graduation. This example shows the importance of background knowledge for the full description of an image.

2. **Negation of action/behavior**: The second category is the use of negation to deny that an action or some kind of behavior is occurring:

   (10) a. A kid eating out of a plate **without** using his hands.
   
   ✎ You are expected to eat with utensils.
   
   b. A woman in the picture has fallen down and **no** one is stopping to help her up.
   
   ✎ You are supposed to help others when they are in trouble.

   Examples like these require an understanding of what is likely or supposed to happen, or how people are expected to behave.

3. **Negation of property**: The next use of negation is to note that an entity in the image lacks a property. In (11a), the negation does two things: it highlights that the buildings are not finished, but in its combination with *yet* suggests that they *will be* finished.

   (11) a. A man wearing a hard hat stands in front of buildings **not** yet finished being built.
   
   b. There are four boys playing soccer, but **not** all of them are on the same team [...].

   In (11b), the negated phrase also performs two roles: it communicates that there are (at least) two teams, and it denies that the four boys are all in the same team. For both examples, the negated parts (*being finished* and *being on the same team*) are properties associated with the concepts of *building* and *playing together*, and could reasonably be expected to be true of buildings and groups of boys playing soccer. The negations ensure that these expectations are cancelled.

   Example (12) shows a completely different effect of negating a property. Here, the negation is used to compare the depicted situation with a particular reference point. The implication here is that the picture is not taken in the USA.
(12) A wild animal not found in America jumping through a field.

4. Negation of attitude: The fourth use of negation concerns attitudes of entities toward actions or others. The examples in (13) illustrate that this use requires an understanding of emotions or attitudes, but also some reasoning about what those emotions are directed at.

(13) a. A man sitting on a panel not enjoying the speech.
    b. The dog in the picture doesn’t like blowing dryer.

5. Outside the frame: The most image-specific use of negation is to note that particular entities are not depicted or out of focus:

(14) a. A woman is taking a picture of something not in the shot with her phone.
    b. Several people sitting in front of a building taking pictures of a landmark not seen.

The use of negation in this category requires an understanding of the events taking place in the image, and what entities might be involved in such events. (14b) is a particularly interesting case, where the annotator specifically says that there is a landmark outside the frame. This raises the question: how does she know and how could a computer algorithm recognize this?

6. (Preventing) future events: The sixth use of negation concerns future events, generally with people preventing something from happening. Here are two examples:

(15) a. A man is riding a bucking horse trying to hold on and not get thrown off.
    b. A girl tries holding onto a vine so she won’t fall into the water.

What is interesting about these sentences is that the ability to produce them does not only require an understanding of the depicted situation (someone is holding on to a horse/vine), but also of the possibilities within that situation (they may or may not fall off/into the water), depending on the actions taken. In other words: they require reasoning about the future.

7. Quotes and Idioms: Some instances of negations are mentions rather than uses:

(16) A girl with a tattoo on her wrist that reads “no regrets” has her hand outstretched.

Other times, the use of a negation isn’t concerned with the image as much as it is with the English language. The examples in (17) illustrate this idiomatic or conventional use of negation.

(17) a. Strolling down path to nowhere.
    b. Three young boys are engaged in a game of don’t drop the melon.

8. Other: Several sentences do not fit in any of the above categories, but there aren’t enough similar examples to merit a category of their own. Two examples are given below. In (18), the negation is used to convey that it is atypical to be holding an umbrella when it is not raining.

(18) The little boy […] is smiling under the blue umbrella even though it is not raining.

In (19), the annotator recognized the intention of the toddler, and is using the negation to contrast the goals with the ability of the toddler. Though there are many other sentences where
the negation is used to contrast two parts of the sentence (see Section 2.10.3), there is just one example where an *ability* is negated.

(19) A little toddler trying to look through a scope but *can’t* reach it.

This categorization is a generalization over uses of negation in the Flickr30K dataset, but because of the limited amount of examples (892, including false positives) and the limited domain (Flickr30K images are likely not representative for all images), there may still be other uses of negation. Future research should assess the degree to which the current taxonomy is sufficient to systematically study the production of negations in image descriptions, for example by looking at negations in image descriptions for a completely different sample of images.

### 2.10.3 Annotating the Flickr30K corpus

Two annotators categorized uses of negations in the Flickr30K corpus using the categories listed above. This categorization has two goals: to validate the categories, and to develop annotation guidelines for future work. By going through all sentences with negations, we were able to identify borderline cases that could serve as examples in the final guidelines.

Using the categories defined in the previous section, we achieved an inter-annotator agreement of Cohen’s $\kappa=0.67$, with an agreement of 77%. We then looked at sentences with disagreement, and settled on categories for those sentences. Table 2.9 shows the final counts for each category, including a Meta-category for cases like *I don’t see a picture*, commenting on the original annotation task, or on the images without describing them.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salient absence</td>
<td>488</td>
</tr>
<tr>
<td>Negation of action/behavior</td>
<td>90</td>
</tr>
<tr>
<td>Quotes and idioms</td>
<td>71</td>
</tr>
<tr>
<td>Not a description/Meta</td>
<td>40</td>
</tr>
<tr>
<td>Negation of attitude</td>
<td>36</td>
</tr>
<tr>
<td>False positive</td>
<td>31</td>
</tr>
<tr>
<td>Outside the frame</td>
<td>26</td>
</tr>
<tr>
<td>Negation of property</td>
<td>25</td>
</tr>
<tr>
<td>(Preventing) future events</td>
<td>21</td>
</tr>
<tr>
<td>Other</td>
<td>66</td>
</tr>
</tbody>
</table>

*Table 2.9 Frequency count of each category.*

Orthogonal to our categorization, we found 39 examples where negations are also used to provide contrast (next to their use in terms of the categories listed above). Two examples are:

(20) a. A man shaves his neck but *not* his beard
    b. A man in a penguin suit runs with a man, *not* in a penguin suit

Such examples show how negations can be used to structure an image. Sometimes this leads to a scalar implicature (Horn, 1972), like in (21).
Three teenagers, two **without** shoes having a water gun fight with various types of guns trying to spray each other.

⇒ One teenager *is* wearing shoes.

A striking observation is that many negations pertain to pieces of clothing; for example: 282 (32%) of the negations are about people being shirtless, while 59 (7%) are about people not wearing shoes. We expect that this distribution will make it difficult for systems to learn on the basis of the Flickr30K data how to use negations that aren’t clothing-related.

### 2.10.4 Takeaway

The takeaway from this section is that negations provide evidence that image descriptions are the result of a complex reasoning process. A subset of the negation uses are based on normative beliefs of how the world should be. This section focused on negations because they are easy to detect, and it is feasible to manually categorize all of the results. And while this chapter focuses on English descriptions, the use of negations is certainly not limited to English. In the next chapter we will see that Dutch and German participants also make use of negations to signal contrast or unexpected situations.

### 2.11 Discussion: Perpetuating bias

This chapter discussed multiple forms of stereotyping and biases in image description data. The problem with these phenomena is that the data currently serves as an example for image description systems, which are evaluated by the similarity between their generated descriptions, and the descriptions in Flickr30K and MS COCO. Because the Flickr30K and MS COCO data is used as training data, there is a possibility that they pick up on the stereotypes and bias that is present in the data, and that they will eventually produce biased descriptions of their own.

#### 2.11.1 Bias in Natural Language Processing

The problem of bias in Natural Language Processing is not hypothetical. For example, other researchers have found that word embeddings derived from large text corpora are clearly biased (Bolukbasi et al., 2016; Caliskan et al., 2017). Bolukbasi et al. (2016) focus on **gender stereotypes**, and propose a debiasing strategy, to erase gender stereotypes from the embeddings and make sure that e.g. *man* and *woman* are equally close in the embedding space to *brilliant*, so that brilliance is not seen as a typically masculine property. Caliskan et al. (2017) explore bias in word embeddings through a variation of the Implicit Association Test (Greenwald et al., 1998, IAT), revealing biases along multiple axes (e.g. age, gender, race). The difference between their work and Bolukbasi et al.’s is that Caliskan et al. (2017) argue

---

14 Word embeddings are vectors that represent word meaning in a high-dimensional vector space. (Simply put, a vector is an array of numbers. They may be used as coordinates, so that e.g. (1,2) represents a point in 2D-space, and (3,2,6) represents a point in 3D-space. Although it is difficult for us to visualize, there is no upper-bound to the number of dimensions that spaces can have in mathematics.) There are many ways to construct word embeddings, (e.g. word2vec, GloVe, FastText, see Mikolov et al. 2013a; Pennington et al. 2014; Bojanowski et al. 2017) but all methods rely on the same **Distributional Hypothesis**: similar words appear in similar contexts (see Sahlgren 2008 for a discussion). So if we want to create a set of word embeddings, we take a large collection of texts, and feed it to a system that determines the meaning of each word on the basis of the contexts in which it is used. For example, the words *cat* and *dog* may often occur with the verbs *walk*, *eat*, *sleep* and the noun *pet*. From this information, we may conclude that *cat* and *dog* are more similar to each other, than to the word *microscope*, which occurs in very different contexts.
that learned representations reflect how language is used, and it is not possible to separate bias from meaning. In an earlier version of their work, Caliskan et al. made this point more explicitly by saying that “bias is meaning.” In the section titled ‘Awareness is better than blindness,’ the authors note:

[We] see debiasing as “fairness through blindness”. It has its place, but also important limits: prejudice can creep back in through proxies (although we should note that Bolukbasi et al. (2016) do consider “indirect bias” in their paper). Efforts to fight prejudice at the level of the initial representation will necessarily hurt meaning and accuracy, and will themselves be hard to adapt as societal understanding of fairness evolves. Instead, we take inspiration from the fact that humans can express behavior different from their implicit biases (Lee, 2016). (p. 12)

In sum, Caliskan et al. argue that, rather than trying to erase all biases (and thus also knowledge about the world) from the system, we can also try to control the system’s behavior, and try to make sure that it recognizes prejudice (unacceptable biases) and refrains from acting upon prejudice. Their Word Embedding Association Test (WEAT) is a step towards being able to detect unacceptable biases.

2.11.2 Bias in Vision & Language

Researchers in the Vision & Language domain have also shown popular multimodal datasets to contain biases (Misra et al., 2016; Zhao et al., 2017a; Burns et al., 2018). Misra et al. (2016) study reporting bias in human-generated image descriptions. For example, the fact that yellow bananas are often just called ‘bananas,’ not mentioning their color because bananas are usually yellow. While the banana example is fairly harmless (and we might even want to encourage systems to display this level of pragmatic competence), we need to take extra care when talking about other people. This is closely related to the linguistic biases discussed above in Section 2.7.

Zhao et al. (2017a) show for two different existing tasks (multilabel object classification and visual semantic role labeling) that the datasets contain gender bias, and that models trained on these datasets amplify that bias. So the bias is not just perpetuated, but actively made worse. The authors propose methods to prevent this amplification, using corpus-level constraints. For example, in visual semantic role labeling, the model determines the subject and object of an action taking place in an image. Zhao et al. put limits in place so that the gender ratio (how many men versus women are predicted to perform a particular action) is within a set margin. This is in line with Caliskan et al.’s (2017) argument that we should be aware of potential biases, and then work to keep the system from acting upon those biases. Finally, Burns et al. (2018) note that crowd-workers sometimes use gendered terms like man or woman without any evidence. Figure 2.14 provides an example, where the subject (a snowboarder) is referred to as a man, even though it is impossible to determine the gender of the subject. This is an example of what we called unwarranted inferences in Section 2.7.

Burns et al. (2018) propose a new type of model (called ‘the Equalizer model’) that explicitly takes the evidence into account before using gendered terms. In cases where the gender is not clear, it is better for the model to use a gender-neutral term (e.g. person, snowboarder), or at least to assign gender labels with equal probability without being skewed towards one of the two. The authors also look at whether image description models are right for the right reasons.

---

15 Furthermore, a recent paper by Gonen and Goldberg (2019) shows that it is impossible to fully remove male/female bias from word embeddings.

16 Caliskan et al. uploaded several drafts to ArXiv. We are referring to version 4: https://arxiv.org/abs/1608.07187v4
2.11 Discussion: Perpetuating bias

Description: A man jumps off a ramp on a snowboard.

Figure 2.14 Picture taken by Christian Jimenez (CC BY-NC-SA) on Flickr.com. Description from the Flickr30K dataset (Hodosh et al., 2013). The author uses the gendered term man, even though the gender of the snowboarder cannot be identified from the picture.

In other words: whether these models use the relevant variables to make their decisions. If a model produces a correctly gendered term, but does not use any gender information in their decision, then that model would be right for the wrong reasons. We cannot trust it to make the right decision, because its decision procedure is fundamentally flawed. Burns et al. find that their Equalizer model helps to focus on the right variables (physical characteristics of a person) to make any decision about gender. Thus, it is more often right for the right reasons.

2.11.3 Addressing the biases discussed in this chapter

Some of the biases discussed in this chapter are relatively easy to address. For example, we might try to counter linguistic bias by controlling the rate at which a model produces adjectival modifiers. Others, like many of the unwarranted inferences, are more difficult to deal with because of their context-specific nature. (But at the same time, this context-specific nature also makes it more difficult for any system to generalize over these examples.)

However, we believe there is also a more fundamental issue to consider. The near-endless variation produced by humans should be a cause to take a step back and think about what we would want the ideal output to look like. With a more carefully constructed set of guidelines, we might be able to avoid many of the biases that are now present in the data. At the same time, having a clear set of guidelines would also allow us to evaluate more precisely how systems perform on the image description task. Taking yet another step back, the image description task is also too broadly defined because current datasets have not been put together with a clear application in mind. Ideally, one would start from the ground up, considering:

1. The usage context:
   - How will the image descriptions be used?
   - On what kind of visual information?
   - In what kinds of situations?

2. The needs of the user:
   - What kind of descriptions would potential users like to have?
• How reliable do the descriptions need to be? (There is a trade-off between specificity and reliability of the descriptions; it is more difficult to generate more specific descriptions.)

3. The technical possibilities:

• Is it feasible for systems to reason about images, or should we focus exclusively on directly visible properties?

Some of this work (mainly 1 and 2) has been done already, in the context of supporting visually impaired users on the web and social media. Petrie et al. (2005) presents a survey among blind users on their thoughts about ALT-text, text that is provided on websites as an alternative to images, and that can be accessed through screen readers. Although the participants’ needs differ from context to context, they indicated that they would like to have information about: objects, buildings, and people; activities that are going on; (the use of) color; the goal of the image; emotion and atmosphere in the image; and the location of the events depicted in the image. Gella and Mitchell (2016) present results from another survey among blind users, on automatic image recognition and the features that they would like to see in those descriptions. Their participants indicated that this technology would be useful for social media images, and that they would like to have information about the emotion and the atmosphere in the images, as well as whether the images are humorous. Researchers at Facebook have also investigated how visually impaired users currently interact with visual content (Voykinska et al., 2016), and what they think about automatically generated ALT-text for images on Facebook (Wu et al., 2017b).

Further guidelines to develop information systems (of which automatic image description systems are an example) are provided by Friedman et al. (2013). They present an overview of the Value-Sensitive Design approach, which aims to uphold human values that are often implicated in system design, such as privacy, freedom from bias, and universal usability.

2.12 Conclusion

This chapter explored the variation in image descriptions produced by human crowd-workers. We have seen that there is a very rich vocabulary for describing images in general, and other people in particular. It is not clear how crowd-workers choose to describe other people, but it is definitely not a shallow process. The examples in this chapter show how crowd-workers reason using stereotypes and prior expectations, resulting in subjective descriptions. What implications does this chapter have for image description systems? We will highlight three topics: the near-endless variation in the descriptions, the danger of perpetuating biases in the data, and the complexity of the task.

2.12.1 Near-endless variation

In Section 2.6, we saw that participants describing an image have a large number of variables to take into account. Even describing a single person in an image becomes complicated when you consider the number of different ways in which a person could be described. The takeaway from this chapter is that image description is not a trivial procedure. Rather, producing a description of an image involves many different choices about how to frame the contents of an image. Lacking clear guidelines, this task is necessarily subjective.
If we want to automate this process, then we should not treat variation in image description corpora as noise. Instead, we should realize that the image description task (as it is currently presented in the literature) is underspecified, and perhaps even encourages people to produce subjective descriptions. If we really want systems to produce human-like descriptions, then we should ask ourselves: what should those descriptions even look like? Current image description datasets offer us a rich palette to choose from.

2.12.2 World knowledge and reasoning about the world

This chapter has repeatedly emphasized the importance of reasoning and world knowledge for generating image descriptions. This is because current image description systems model image description as a simple mapping from images to descriptions, with no knowledge or reasoning component involved. (See Chapter 6 for an overview.) By highlighting the importance of these components for several different linguistic phenomena, we have shown that world knowledge and reasoning are not just incidentally required, but that there is a pervasive need for these components in order to account for all linguistic phenomena. Hence, world knowledge and reasoning form a recurrent theme throughout this thesis.

Of the linguistic phenomena dealt with in this chapter, the need for world knowledge is perhaps most clearly illustrated by the different uses of negation in Section 2.10. What kind of knowledge is needed, and where could image description systems obtain this kind of knowledge?

- The Outside the frame category requires an understanding of human gaze within an image, which is a challenging problem in computer vision (Valenti et al., 2012). Additionally, we also need to understand the differences between scene types, both from a computational- (Oliva and Torralba, 2001) and a human perspective (Torralba et al., 2006).

- The Salient absence category provides evidence for two kinds of expectations that play a role in the use of negations: general expectations (people are supposed to wear shirts) and situation-specific expectations (students at graduation ceremonies typically wear caps). This is the same kind of distributional information that underlies reporting bias (Misra et al., 2016). Because bananas are usually yellow, people usually only mention their color when it deviates from the norm, e.g. with green bananas.

- Finally, the Negation of action/behavior category requires action recognition, which is a challenging problem in still images (Poppe, 2010). The ability to automatically recognise what people are doing in an image, and how this contrasts with what they would typically do in similar images, would greatly help with generating this use of negation. Note that knowledge of what people typically do in a particular situation also requires experience, or some other source of event frequency.

From a linguistic perspective, background knowledge could be represented by frames (Fillmore, 1976) and scripts (Schank and Abelson, 1977). There are some hand-crafted resources that contain this kind of knowledge, e.g. FrameNet (Baker et al., 1998), but they only have limited coverage. Recent work has shown, however, that it is possible to automatically learn frames (Pennacchiotti et al., 2008) and narrative chains (Chambers and Jurafsky, 2009) from text corpora. Fast et al. (2016) show how such knowledge, as well as knowledge about object affordances (Gibson, 1977), can be used to reason about visual scenes. Still, it is an open question how to use knowledge bases to produce human-like descriptions.
2.12.3 Next chapter

The observations in this chapter are based on descriptions provided by speakers of US English. Although we have no reason to think that speakers of other languages would be less subjective, it is still necessary to see if our observations generalize to other languages. In the next chapter, we will provide an overview of existing image description datasets in other languages, and compare the English descriptions from the Flickr30K dataset with their Dutch and German counterparts. We will see that the use of subjective language in image descriptions is not restricted to English; it is present in these other languages as well.