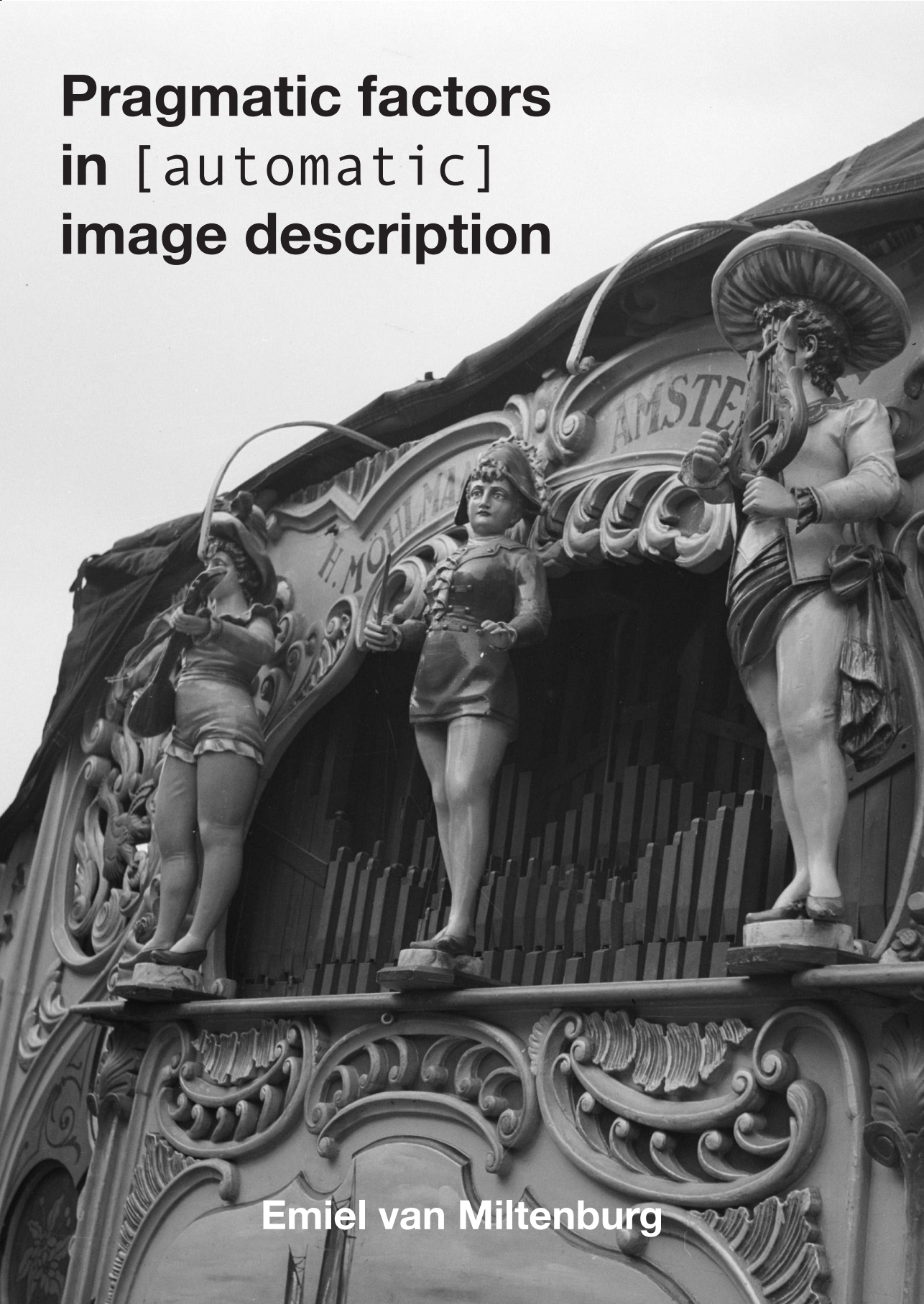
Pragmatic factors in [automatic] image description



Emiel van Miltenburg

Pragmatic factors in [automatic] image description

Emiel van Miltenburg

Promotor:	prof.dr. Piek Th.J.M. Vossen
Co-promotor:	dr. Desmond Elliott
Reading committee:	prof.dr. Antal van den Bosch
	prof.dr. Alan Cienki (chair)
	prof.dr. Kees van Deemter
	dr. Raquel Fernández
	dr. Aurélie Herbelot

SIKS Dissertation Series No. 2019-25

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Typeset using LaTeX, using the TeXGyre fonts. The small figure on the previous page comes from the `phaistos` package, which is based on the Greek Phaistos disk.

Cover photos by Willem van de Poll (1895 - 1970), licensed CC0, by *het Nationaal Archief*. Access code: 2.24.14.02. Item numbers: 254-3253 (front), 254-3252 (back)

Printed by ProefschriftMaken || www.proefschriftmaken.nl

ISBN: 9789463804899

VRIJE UNIVERSITEIT

Pragmatic factors in (automatic) image description

ACADEMISCH PROEFSCHRIFT

Ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Geesteswetenschappen
op maandag 14 oktober 2019 om 11.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Cornelis Wilhelmus Johannes van Miltenburg

geboren te Nieuwegein

promotor: prof.dr. P.T.J.M. Vossen
copromotor: dr. D. Elliott

Contents

Acknowledgments

Although only my name is on the cover of this dissertation, I could not have completed this work without the people around me. First and foremost I would like to thank my supervisors, Piek Vossen and Desmond Elliott, for all their encouragement and support. Working with Piek has taught me the meaning of visionary research, imagining long-term goals and working towards those goals despite the inevitable obstacles. There was certainly no shortage of ideas in our meetings! Having too many ideas tends to put you at the risk of drowning, but Piek always made sure I stayed afloat. Desmond has been a patient guide in the world of Vision and Language. Without him, this thesis would have looked completely different. I could not have wished for better supervisors.

Many thanks also go to the reading committee, for taking the time to read and comment on my thesis. As I am writing this, I am looking forward to the defense! At the event, I am honored to have Hennie van der Vliet and Roxane Segers as my paranymphs. Many thanks in advance.

I am also grateful to all of my co-authors. Next to Desmond and Piek, these are (in alphabetical order): Lora Aroyo, Ákos Kádar, Ruud Koolen, Emiel Krahmer, Alessandro Lopopolo, Roser Morante, Chantal van Son, and Benjamin Timmermans. Collaborating with these people has made me a better writer and researcher. If you spot a particularly good piece of writing in this thesis, there's a good chance it's theirs.

I have greatly benefited from working in a very pleasant environment, both in Amsterdam and in Edinburgh. The CLTL (*Computational Lexicology and Terminology Lab*) has felt like a second home for more than four years. Who says you cannot have two captains on one ship? Even if the waves in education became really big, Captain Hennie was always able to steer us into calmer waters. And Captain Piek made sure there was never any cause for mutiny, with regular events to keep the spirits up. Many thanks to all of the crew for all discussions, banter, and collaboration. My stay at the University of Edinburgh has been much shorter. Ten weeks is really too short a time to spend in such a nice city. Thanks to everyone at EdinburghNLP for making me feel welcome.

After almost five years working on my PhD research in Amsterdam, it was hard to imagine life after the PhD. As it turns out: it exists! There is no black hole, but a wonderful green campus in Tilburg. Thanks to all my new colleagues in Communication and Information Science for welcoming me to the department.

Finally, I would like to thank my friends and family for being there throughout my PhD. And a very big 'thank you' to Loes, for the past, the present, and the future.

Utrecht, Summer 2019

Understanding of language by machines

The research for this thesis was carried out within the context of a larger project, called *Understanding of Language by Machines* (ULM). This project is funded through the NWO Spinoza prize, awarded in 2013 to Piek Vossen. The goal of the project is:

> "…to develop computer models that can assign deeper meaning to language that approximates human understanding and to use these models to automatically read and understand text. Current approaches to natural language understanding consider language as a closed-world of relations between words. Words and text are however highly ambiguous and vague. People do not notice this ambiguity when using language within their social communicative context. This project tries to get a better understanding of the scope and complexity of this ambiguity and how to model the social communicative contexts to help resolving it."
>
> (Source: http://www.understandinglanguagebymachines.org/)

The project is led by Piek Vossen, with the help of Selene Kolman (project manager) and Paul Huygen (scientific programmer). Other members are or have been: Tommaso Caselli, Filip Ilievski, Rubén Izquierdo, Minh Lê, Alessandro Lopopolo, Roser Morante, Marten Postma, and Chantal van Son.

Notes

Language in this thesis

Research is almost impossible to carry out alone. Hence, all the content chapters from this thesis are based on collaborative work. Since this thesis is presented as a single-authored monograph, I have made the following choice. The introduction and conclusion are written from a first-person singular perspective (using *I*), but, in acknowledgment of my co-authors, all content chapters are written from a first-person plural perspective (using *we*). I remain solely responsible for any errors in this thesis.

Images and Copyright

Most of the images in this thesis originate from Flickr.com, a social image sharing platform, where amateurs and professional photographers share their work under various licenses. Many of these images are provided under a Creative Commons licence.[1] Where possible, I have tried to use images provided either under such a license, or even images that are part of the Public Domain, with the appropriate attributions.[2] Unfortunately, this was not always possible.

The research presented in this thesis focuses on image descriptions from the Flickr30K and MS COCO datasets, and some of the images from those corpora are fully copyrighted. Furthermore, some images have been deleted from Flickr.com after their publication in either Flickr30K or MS COCO. In those cases, it was not always possible to find and credit the original author (although I did try, using Google's reverse image search). I have generally tried to avoid using these images, and to look for alternative examples. In some cases, however, I have found that the copyrighted image provided the clearest example.

The use of copyrighted images is somewhat of a legal gray area. Copyright law in the US (where Flickr is based) has a Fair Use exception, that allows for the use of copyrighted images in some cases. Those cases are judged using the following four factors:[3]

The purpose and character of the use. Here, we could reasonably argue that scholarly work qualifies as 'transformative use', where we do not just copy the image, but reflect on the meaning of the image and the associated descriptions from existing image description corpora.
The nature of the copyrighted work. Here, we could argue that the images were published on Flickr.com already (meant to be seen by others), and used in existing image description datasets.
The Amount and Substantiality of the Portion Taken. Here, we need to concede that we are not just copying a portion of the image. However, this is unavoidable in discussing image descriptions, which aim to capture the heart of the work.

[1] See https://creativecommons.org.

[2] See https://fairuse.stanford.edu/overview/public-domain/welcome/

[3] See https://fairuse.stanford.edu/overview/fair-use/four-factors/

The Effect of the Use Upon the Potential Market. We do not wish to use the images for any commercial benefit, and do not foresee any effect on the potential market for the images discussed in this thesis.

Dutch law does not have a Fair Use exception. Rather, it provides for a 'Right to Quote', which arguably covers our use of the copyrighted images from Flickr.[4] After all: one cannot have a scholarly discussion of the image descriptions from MS COCO or Flickr30K without taking the images into account. Having said this, it seems to me that the current situation is not ideal. I hope that we, as a scientific community, can move toward datasets that are not limited by copyright. If this turns out to be impossible, we should at least require all new datasets to provide a list of authors to be acknowledged when citing relevant parts of that dataset.

For my part, I invite authors of any images that have gone uncredited to contact me, so that I can give credit where credit is due.

[4] The relevant Dutch juridical term for quoting images is 'beeldcitaat.' See http://www.iusmentis.com/auteursrecht/fairuse/ and http://www.iusmentis.com/auteursrecht/citeren/beeldcitaat/

Chapter 1

Introduction

1.1 Describing an image

Whenever you look at an image, you cannot help but interpret it. Take, for example, the image in Figure 1.1. If I asked you to describe this image, you might provide one of the following descriptions:[1]

- A man in a yellow waterproof jacket and his companion are on a boat in the open water.
- Two men, one in a yellow jacket and the other in a blue sweater, are on a boat.
- Two dark-haired men are sailing a fishing boat.

Figure 1.1 Picture from the Flickr30K dataset (Young et al., 2014), taken by Phillip Capper (CC-BY).

You may also have another description in mind, but it is very likely that your description will at least contain a reference to the two men, and the boat they are on. Somehow, this information is *important* for us to mention about the image (unlike the mast and the rope in the foreground). Moreover, both men are in the middle of the image, with the man on the left wearing a bright yellow coat. This makes them *visually salient* (i.e. they draw visual attention).

You may also have thought that perhaps the two men are related (e.g. father and son), even though we cannot be sure that this is true. Somehow, this information is *relevant* enough to consider. Finally, there may be differences between your description and the ones printed above. This shows us that image description is not a deterministic process; there may be several different ways to describe an image. What kind of description you eventually provide is a result of *contextual factors* and *personal preference*.

1.2 Automatic image description

What if we could make a system that could understand images and describe them for us using natural language? Such technology would surely be helpful for people to index and search the

[1] These examples are taken from a dataset of described images; the Flickr30K corpus (Young et al., 2014).

pictures on their computer or smart phone. Moreover, it would help visually impaired people to navigate their environment, both online and offline. This prospect has drawn researchers from the Computer Vision and Natural Language Processing fields to work together on the shared task of *automatic image description* (Bernardi et al., 2016). Tasks such as these cannot exist without data. Machine learning researchers need data to *train* their systems, showing the systems what they are supposed to do, and they need data to *evaluate* whether their system actually achieves that goal. This thesis is about that data. We will be studying how people describe everyday images, and what are the challenges for machines to do the same. We will also look at which properties of human-generated descriptions are desirable or undesirable for systems to reproduce.

1.3 Defining image descriptions

Hodosh et al. (2013, p. 857) distinguish three kinds of image descriptions, arguing that automatic image description systems should focus on generating conceptual descriptions:

Conceptual descriptions "identify what is depicted in the image, and while they may be *abstract* (e.g., concerning the mood a picture may convey), image understanding is mostly interested in *concrete* descriptions of the depicted scene and entities, their attributes and relations, as well as the events they participate in."

Non-visual descriptions "provide additional background information that cannot be obtained from the image alone, e.g. about the situation, time or location in which the image was taken."

Perceptual descriptions "capture low-level visual properties of images (e.g., whether it is a photograph or a drawing, or what colors or shapes dominate)."

These levels are based on earlier work by Panofsky (1939) and Shatford (1986), which I will discuss in Section 2.2. Non-visual descriptions occur in newspapers, for example, where they relate images to the contents of the article they belong to. As a matter of terminology, we will refer to this kind of descriptions as *captions*, and reserve the term *description* for conceptual descriptions, unless indicated otherwise.

1.4 Image description data

The data that we will look at was collected by image description researchers in a series of crowdsourcing tasks.[2] In these tasks, the crowd workers were presented with a small set of images, and asked to provide a 'short-but-complete' description for each of the images. The result of their efforts is a huge collection of image description data; the Flickr30K corpus (Young et al., 2014) consists of over 30 000 images with 5 descriptions per image, while the MS COCO dataset (Lin et al., 2014) contains over 160 000 images with 5 descriptions per image. We have already seen an example image with descriptions from the Flickr30K dataset at the beginning of this chapter. This data provides us with the opportunity to study human image description behavior at a much larger scale than is typical for linguistics or psychology studies. 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.

[2] Crowdsourcing tasks are small jobs (e.g. surveys, annotation tasks) that are outsourced to online *crowd workers*, through services like Mechanical Turk, Prolific, and Crowdflower. See Quinn and Bederson 2011; Wortman Vaughan 2018 for an introduction and survey of commonly used methods.

While there are some surveys providing an overview of different image description datasets (e.g. Ferraro et al. 2015b; Bernardi et al. 2016), there have been no studies to catalog the linguistic properties of image descriptions, and the implications of those properties for image description systems. This thesis aims to fill that gap.

1.5 A model of the image description process

One of the assumptions behind these datasets is that they provide objective image descriptions:

> "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)

The *assumption of neutrality* is a useful simplification: if it is more or less correct that similar images will have similar descriptions (that are not influenced by any external factors), then we can try to learn a mapping between images and descriptions. When we inspect the descriptions, however, we find that humans do not always produce objective descriptions. Rather, they frequently speculate (e.g. about relations between people in the images), or use judgmental language (e.g. regarding physical attractiveness). Figure 1.2 provides two examples. For the picture on the left, one crowd-worker for the Flickr30K dataset assumed that the image depicts a mother and a daughter, even though the image does not provide any hints as to how the two women are related. For the picture on the right, two crowd-workers commented on the looks of the woman in the image, even though attractiveness is highly subjective (and it is unclear why it would be relevant to mention in a general description of an image).

"**Mother and daughter** wearing Alice in wonderland customs are posing for a picture."

1. "A **pretty** young woman wearing a blue ruffled shirt smelling a pretty red flower."
2. "**Attractive** young woman takes a moment to stop and smell the flower."
3. "A young woman outside , smelling a red flower and smiling."

Figure 1.2 Pictures by kievcaira (CC BY-NC-ND) and antoniopringles (CC BY-NC-SA) on Flickr.com, with descriptions from the Flickr30K dataset (Young et al., 2014).

We may also note that there is a high degree of variation in the image descriptions. Indeed, Vedantam et al. (2015) found that we may collect 50 descriptions for a given image and still find meaningful variation. These findings suggest that interpretation of the image plays a big role in image description. Even when people are asked not to speculate about an image, they

cannot help but (re-)contextualize it before providing a description. And because people may differ in their backgrounds, their interpretation may also differ. As a result, their descriptions may also end up capturing different aspects of the image. Figure 1.3 provides an illustration of this process.[3]

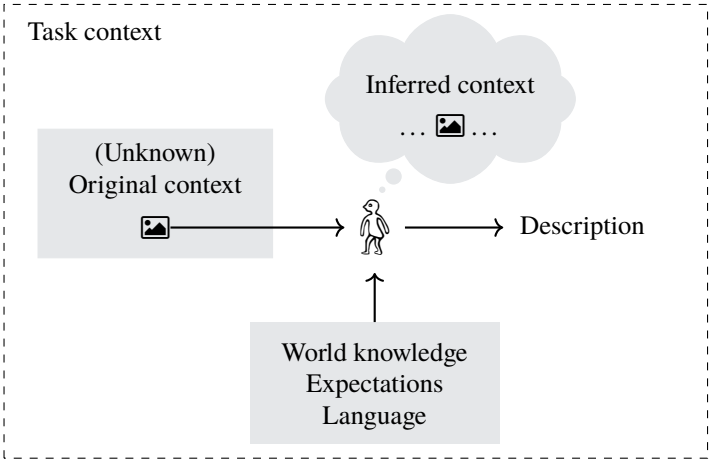


Figure 1.3 Conceptual model of description generation, modified from (van Miltenburg, 2017). Note that the original context is likely to be different from the context inferred by the subject.

In Figure 1.3, an image is taken out of context and presented to an actor who is asked to describe this image. To provide a meaningful description, the actor first has to understand what the image is about. For this, they need to rely on their world knowledge to identify the individual components of the image, and reason about what is going on. While doing so, they might fall back on their past experiences and see whether there is anything unusual about the image. This leads to a particular interpretation of the image that they have to capture in their description. Additionally, their description is limited to the vocabulary and grammatical constructions afforded by their language.

1.6 Image description systems and the semantic gap

As noted above, the image descriptions from Flickr30K and MS COCO are commonly used to train and evaluate automatic image description systems. The idea is that we can present these systems with example input (the images) and example output (the descriptions), and let them figure out how to create a mapping from visual features to sequences of words. One example of this is the system presented by Vinyals et al. (2015). I will only provide a short description of this system here, but Chapter 6 provides a more in-depth discussion of how current image description systems work.

Vinyals et al.'s system uses the pre-trained convolutional neural network (CNN) model from Ioffe and Szegedy (2015) to extract visual features from images (so that it doesn't need to learn a mapping from raw images to descriptions). Given those features, it tries to predict what are the most probable descriptions for the provided images. This simple set-up works

[3] This figure is similar to Ogden and Richards' (1923) *triangle of reference* (also known as *the semantic triangle*), in which an interpreter perceives a sign and tries to determine its referent (the meaning of the sign).

surprisingly well. It produces fluent descriptions that often seem to capture the contents of the images in the dataset. At the same time, it also makes surprising mistakes that no human would make. Figure 1.4 provides two examples. For the image on the left, the system accurately describes the man holding a tennis racket on a tennis court. But for the image on the right, the system produces a completely inaccurate description.

Accurate

Human A man with a tennis racket gets ready to swing his racket.
System A man holding a tennis racquet on a tennis court.

Inaccurate

Human A woman is stooped beside a fence, watching a polar bear.
System A couple of giraffe standing next to each other.

Figure 1.4 Accurate and inaccurate descriptions generated by Vinyals et al.'s (2015) system for images from the MS COCO dataset. Pictures taken by Spyffe (CC BY) and Ucumari (CC BY-NC-ND) on Flickr.com. Descriptions from http://nic.droppages.com

There are two important observations we can make about systems like these:

1. **Implicit standards.** There is no real standard in the image description literature for what an image description should look like, except the implicit standard that systems should try to make their descriptions as similar to human descriptions as possible. The tacit assumption here is that humans display *exemplary behavior*. As we will see in Chapter 2, this is not always the case.

2. **Naive solution.** The system does not use any external resources to reason about the provided images. There are no knowledge bases, ontologies, or reasoning systems involved in the image description process. Rather, the system just provides an end-to-end solution from images to descriptions. If Figure 1.3 provides an accurate model of the human image description process, then we may expect that systems like the one provided by Vinyals et al. (2015) will not be able to fully provide human-like image descriptions, because they lack the requisite resources.

It should be noted that the goal of automatic image description is not to model the human cognitive process. Automatic image description is an engineering challenge. If we are able to build a system that generates human-like descriptions while being cognitively implausible, that is completely fine. However, I will argue in this thesis that human-generated descriptions require more than just identifying visual features and mapping them to sequences of words; interpretation and contextualization are essential to produce human-like descriptions. There are two possible ways to resolve this issue: either we should (1) build more advanced image

description systems, or we should (2) change the (currently implicit) goal of trying to match human descriptions as closely as possible, and formulate a more restrictive standard for what image descriptions should look like.

1.6.1 The semantic gap

In the context of comparing human and machine performance, the difference between humans and machines is often referred to as *the semantic gap*. This term comes from the image retrieval literature, where it refers to the gap between machine understanding and human understanding of the content of an image. Smeulders et al. (2000) define the semantic gap as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" (p. 1353). Figure 1.5 provides an illustration, showing a scale from no understanding to full understanding of an image.[4] Machine understanding of images lags behind human understanding, and the space between the two is the semantic gap.

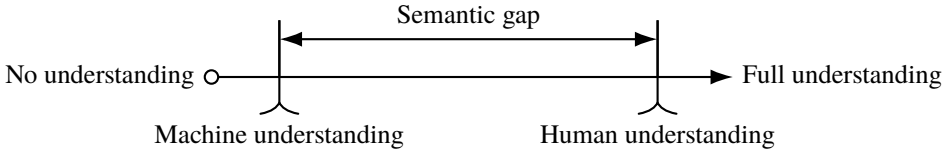


Figure 1.5 Visualization of the 'semantic gap.'

Hare et al. (2006) propose to consider the semantic gap in terms of five different levels of interpretation, illustrated in Figure 1.6. This proposal follows a long tradition in art history and information science, that I will discuss in the next chapter (§2.2). Hare et al. suggest to think of the semantic gap as consisting of two major gaps: (1) between image descriptors and object labels, and (2) between object labels and the full semantics of the image.

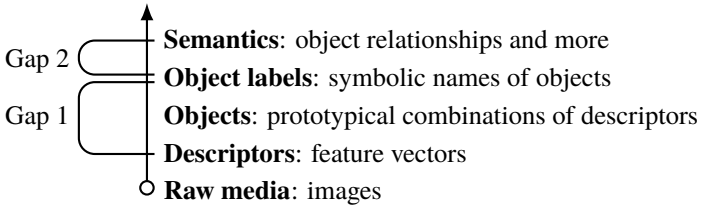


Figure 1.6 Hare's (2006) characterization of the semantic gap.

Hare's proposal predates the 'deep learning revolution' around 2012-2013 when *end-to-end* image recognition systems became mainstream research.[5] End-to-end systems are trained by

[4] Prior to their discussion of the semantic gap, Smeulders et al. also note that 2D-images may only offer us a limited understanding of the 3D-scene from which they are derived. They refer to difference between the actual scene and our understanding of an image (a mere recording of that scene) as the *sensory gap*. I will focus mainly on the semantic gap.

[5] 2012 is the year when team SuperVision won the ImageNet Large-Scale Visual Recognition Challenge, using a deep convolutional neural network, trained using a GPU (Graphics Processing Unit), which enabled them to train their model much faster than with a regular CPU (Krizhevsky et al., 2012). The year after, the majority of the entries used a similar approach (Russakovsky et al., 2015).

providing them with labeled data, and letting the system figure out relevant features to predict the right labels from the raw data. Before such systems came around, a large part of computer vision research focused on developing better *descriptors*. Descriptors are engineered feature vectors that provide low-level information about the contents of an image; examples are SIFT (Lowe, 1999) and SURF (Bay et al., 2006). We can use those descriptors to locate objects in an image, and when we have a reliable way to do this, we can try to assign labels to those objects. Each step in Figure 1.6 corresponds to a module in the classic computer vision pipeline.

Even though the classic computer vision pipeline has at least in part been superseded by newer technology, Hare's proposal is still relevant today, as it relates to different levels of understanding an image. Hare et al. note that we may want to approach the two gaps in different ways. For the first gap, we may opt for a bottom-up approach: collect a large dataset of labeled images and try to learn a mapping between images (or features extracted from those images) and object labels. This approach is exemplified by the ImageNet Large-Scale Visual Recognition Challenge (Russakovsky et al., 2015), where systems need to predict labels for unseen images, based on training data from ImageNet, a large collection of labeled images (Deng et al., 2009). This gives us a basic understanding of the entities that are depicted in the image, but not how they *relate* to each other.

For the second gap, Hare et al. propose a top-down approach using ontology-based reasoning to determine how different objects in an image may be related. But at the moment, we mostly see researchers taking the same kind of bottom-up approach for descriptions as they do for image labeling (Bernardi et al., 2016). This thesis argues that the bottom-up approach can only achieve limited success if the goal is to generate human-like image descriptions. I will show that humans often take a top-down, knowledge-rich approach to describe images, reasoning about the images that are presented to them, and using information that is external to the images themselves.

1.6.2 The pragmatic gap

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's (1978) 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. 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 1.7 to add an additional, pragmatic level.

In its broadest sense, pragmatics is the study of language use (Levinson, 1983). This thesis views image description as a reasoning process, where the speaker/writer makes choices about what to report about an image, and how to report it. During this process, the speaker/writer considers several different factors that might affect how they would describe a particular image. For example: Who is their interlocutor? What is the purpose of the description? Is there anything unusual or unexpected about the image? Is that information relevant? And so on. This thesis highlights the role of those pragmatic factors in image description.

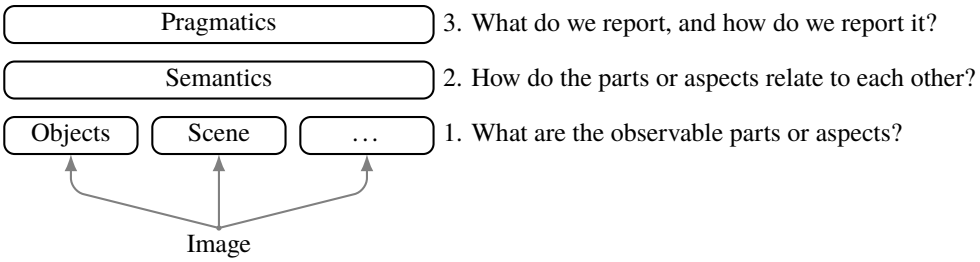


Figure 1.7 Update to Hare et al.'s (2006) proposal, including a pragmatic level.

1.7 Research questions

This thesis aims to deepen our understanding of the semantic gap between humans and automatic image description systems. I will answer the following question:

Main question To what extent are automatic image description systems able to generate human-like descriptions? This question can be split into three separate research questions:

Research Question 1 How can we characterize human image descriptions? Specifically, what does the image description process look like, what do people choose to describe, to what extent do they differ in how they describe the same images, and how objective are their descriptions?

Research Question 2 How can we characterize automatic image descriptions? Specifically, what does the image description process look like, how accurate are the automatically generated descriptions, and are they as diverse as human-generated descriptions?

Research Question 3 Should we even want to mimic humans in all respects? Specifically, are all examples in current image description datasets suitable to be generated by automatic image description systems? If not, what kinds of examples should we avoid?

To understand the semantic gap between humans and machines in automatic image description, we first need to understand what it is that people *do*. Then, when we have established the properties of human image descriptions, we can discuss which of those properties would actually be *desirable* for automatically generated image descriptions. With those goals in mind, we can start to look at the performance of automatic image description systems and see how they measure up. An important part of this process is to design automated metrics, that give us an objective measure of performance, which may be used to indicate progress in the development of better systems.

When we know how people describe images, we can also ask ourselves: to what extent do we want automatic image description systems to behave similarly? Perhaps there are also some undesirable features of human image descriptions that we should avoid. Furthermore, there may be features of human descriptions that are computationally expensive, but do not add much to the quality of the descriptions. For such features we may wonder whether they are worth the effort.

The body of this thesis consists of two parts, corresponding to the first two research questions. I will not address the third research question in the body of this thesis, but we will come back to it in the conclusion.

1.7.1 Characterizing human image descriptions

Part 1 of this thesis, titled *Humans and images*, focuses on the way people describe images. The main objective of this part is to highlight the richness and the subjectivity of human-generated image descriptions. Rich, in the sense that human language offers a virtually infinite set of different ways to describe an image. Subjective, in the sense that people will use their own knowledge and expectations to choose from all of those options how an image should be described. Research Question 1 is divided into five sub-questions:

How do people vary in their descriptions? We have already noted that different people may provide different descriptions for the same images. But we don't know the extent of this variation, and whether there may still be general tendencies in the data. We will explore this sub-question in Chapter 2, which provides an overview of different linguistic phenomena that we may observe in image descriptions. We will look at the different kinds of labels that may be used to refer to other people; the use of negations; and stereotyping and bias in image descriptions.

How objective are those image descriptions? We have also noted that people do not always produce objective descriptions. Our model in Figure 1.3 also suggests that differences in knowledge, expectations, or language may lead to differences in the descriptions that people produce. We will also explore this sub-question in Chapter 2, where I argue that image descriptions are hardly objective at all.

Do image descriptions show similar variation across different languages? We will initially only look at English image descriptions, to establish a set of linguistic phenomena that we will look at throughout this thesis. Chapter 3 discusses cross-linguistic differences and similarities in image descriptions. We will see that Dutch, English, and German image descriptions all contain the different kinds of subjective language from Chapter 2. At the same time, we will also see how cultural differences lead to differences in the descriptions.

What does the image description process look like? Most image description datasets consist of images paired with static descriptions. From this data, we cannot tell how those descriptions came about. If we want to learn more about this process, we need to record it from start to finish. Chapter 4 presents a dataset that contains this kind of dynamic data: the Dutch Image Description and Eye-tracking Corpus (DIDEC). This dataset contains spoken image descriptions along with eye-tracking data showing where participants are looking as they produce descriptions.

How does the format of the human task affect the resulting descriptions? The problem with crowdsourcing in Machine Learning is that it is typically seen as a process of 'data collection' rather than as an experiment that ought to be controlled. In Chapter 5, I argue in favor of the latter view, and show how the format of the image description task may affect the resulting descriptions. As an example, I will focus on the differences between spoken and written elicitation tasks.

1.7.2 Characterizing automatic image descriptions

Part 2, titled *Machines and images*, focuses on automatic image description systems. The main objective of this part is to provide a detailed analysis of current image description technology, and to show its limitations. Research question 2 is divided into the following subquestions:

How do automatic image description systems work? The first half of Chapter 6 (until Section 6.7) gives a short introduction to automatic image description systems. Readers experienced with natural language generation and *deep learning* may skip this part, as I will not present any new findings.

What is the quality of current automatic image description technology? The second half of Chapter 6 (Section 6.7 onwards) gives an overview of current evaluation methods, and provides a detailed error analysis of several different automatic image description systems, showing the limitations of current technology.

Do automatic image descriptions display a similar amount of variation? Having seen in Chapter 2 that humans display a high degree of variation in their descriptions, we may ask ourselves: how do automatic image descriptions compare? Chapter 7 looks at the diversity of automatically generated image descriptions. I provide an overview of existing diversity metrics, and propose several new metrics to assess the diversity of generated descriptions.

1.8 Methodology

This work relies on two types of methodology: corpus analysis and computational modeling.

1.8.1 Corpus analysis

Corpus analysis is fundamental to understand the image description task: if we don't know what the descriptions look like, we don't understand what it is that image description systems are modeling. Thus, our first task is to inspect the image descriptions, and identify linguistic phenomena that inform us about the image description process. These phenomena are found by manually inspecting the corpus. There are four kinds of arguments that we may use:

Existence If we find any amount of evidence that some linguistic phenomenon exists in the data, then we must conclude that any complete solution to the problem of automatic image description should be able to produce this phenomenon. This argument may be strengthened by frequency or cross-linguistic evidence.

Frequency If a linguistic phenomenon frequently occurs, then this is a sign of robustness: this is a feature that is *systematically* included in the descriptions, and thus enjoys some importance. We should expect automatic image description systems to be able to display this phenomenon.

Cross-linguistic evidence If a linguistic phenomenon occurs in image descriptions across different languages, then this is another sign of robustness; apparently this feature is important enough that speakers of different languages include it in their descriptions.

Systematicity If we systematically find the same linguistic phenomenon across different images sharing a particular property, then we may conclude that novel images with the same property should also elicit this phenomenon.

This dissertation frames crowdsourcing tasks to collect image descriptions as large-scale experiments, with crowd workers as the participants. This is helpful because it reminds us of (1) the role that participants have in the outcome of the experiment; (2) the potential to manipulate the task and influence the results; and (3) the need to control the experiment, to check for variables influencing the descriptions.

Corpus analysis is like a post-hoc analysis of experimental results; we observe linguistic phenomena in the data, and provide plausible explanations as to what caused the participants to describe the images in such-and-such a way. After the analysis, these explanations have the status of hypotheses: they are congruent with the data, but remain untested. New data needs to be collected to prove or refute them. In our case, we look at Dutch and German data to show that phenomena observed for English image descriptions also occur in other languages. Another role for corpus analysis is that it can be used to identify desirable or undesirable linguistic phenomena. Having observed these phenomena in the data, we can decide to alter the image description task in such a way that the participants are more (or less) likely to produce these (un)desirable phenomena.

1.8.2 Computational modeling

This thesis aims to see what is the difference between human-generated and automatically generated image descriptions. I use two different approaches for this:

Error analysis Analyze whether the output of an image description system is correct or incorrect, and categorize the mistakes. I will not look at *adequacy*, i.e. whether the descriptions are suited for any particular purpose.

Quantify behavior Determine interesting linguistic properties that might differ between human- and machine-generated descriptions, and develop automated metrics that capture those properties. This enables us to compare different systems without manually having to annotate their output.

The overall result of this is an overview of where we stand in terms of developing image description systems that can produce human-like output, and what it takes to close the semantic gap. Future research may build on these results using another computational approach:

Manipulate the model Take a basic model and add different modules that may help the model generate different kinds of output. Compare the results for different combinations of modules.

1.9 Contributions of this thesis

The field of automatic image description is still early in its development and, as such, there are no clear norms for how images should be described. Moreover, the current image description literature does not offer any framework for understanding the contents and diversity of human-generated descriptions. This thesis frames the image description task as a linguistic experiment (rather than an objective data collection procedure). I show how image descriptions may be influenced by the image description task, and provide an overview of the characteristics of human-generated image descriptions. By collecting real-time image description behavior, this thesis also offers insight in the image description process. Taken together, this thesis shows that current image description datasets are highly subjective and diverse, and that this subjectivity and diversity may be explained in terms of the model shown in Figure 1.3; the decontextualized images from the canonical image description task are re-interpreted from the perspective of the participants of the task, before they describe the images in their own words (relying on their world knowledge, general expectations, and linguistic knowledge). Furthermore, I show that this does not just hold for English, but also for Dutch and German descriptions.

Having seen how humans describe images, I analyze how automatic image description systems perform the same task. This thesis provides a summary of current research, and

assesses the quality of machine-generated descriptions. Looking at system output, this thesis shows that the vast majority of automatically generated descriptions contains at least one error. Furthermore, the descriptions are bland and generic. This genericity has been noted before, but little work has been done to quantify the (lack of) diversity of automatic image descriptions. I present different ways to measure diversity in image description data, and show how current image description systems still have plenty of room for improvement.

Datasets and Software

During this research, I published the following datasets:

The VU sound corpus is a collection of sounds from the Freesound database (Font et al., 2013), crowd-annotated with keywords (van Miltenburg et al., 2016b).

Dutch image descriptions for the Flickr30K validation and test sets (1014 + 1000 images) with 5 descriptions per image (van Miltenburg et al., 2017).

Dutch Image Description and Eye-tracking Corpus (DIDEC) for 307 images taken from MS COCO, with 16-17 descriptions per image (van Miltenburg et al., 2018a).

I also developed several annotation and inspection tools, both for these datasets and for the Flickr30K corpus. These are described in appendix A.

Publications

This dissertation is based on the research described in the following publications:

Alessandro Lopopolo and Emiel van Miltenburg. 2015. Sound-based distributional models. In *Proceedings of the 11th International Conference on Computational Semantics*. Association for Computational Linguistics, London, UK, pages 70–75

Emiel van Miltenburg. 2016. Stereotyping and bias in the flickr30k dataset. In Jens Edlund, Dirk Heylen, and Patrizia Paggio, editors, *Proceedings of Multimodal Corpora: Computer vision and language processing (MMC 2016)*. pages 1–4

Emiel van Miltenburg, Roser Morante, and Desmond Elliott. 2016a. Pragmatic factors in image description: The case of negations. In *Proceedings of the 5th Workshop on Vision and Language*. Association for Computational Linguistics, Berlin, Germany, pages 54–59

Emiel van Miltenburg, Benjamin Timmermans, and Lora Aroyo. 2016b. The vu sound corpus: Adding more fine-grained annotations to the freesound database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Portorož, Slovenia

Chantal van Son, Emiel van Miltenburg, and Roser Morante. 2016. Building a dictionary of affixal negations. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 49–56. http://aclweb.org/anthology/W16-5007

Emiel van Miltenburg. 2017. Pragmatic descriptions of perceptual stimuli. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, pages 1–10

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*. Association for Computational Linguistics, Santiago de Compostela, Spain, pages 21–30

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*

Emiel van Miltenburg, Ákos Kádar, Ruud Koolen, and Emiel Krahmer. 2018a. DIDEC: The Dutch Image Description and Eye-tracking Corpus. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*. Resource available at https://didec.uvt.nl

Emiel van Miltenburg, Ruud Koolen, and Emiel Krahmer. 2018b. Varying image description tasks: spoken versus written descriptions. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Talking about other people: an endless range of possibilities. In *Proceedings of the 11th International Conference on Natural Language Generation*. Association for Computational Linguistics, pages 415–420. http://aclweb.org/anthology/W18-6550

Part I

Humans and images

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:

Emiel van Miltenburg. 2016. Stereotyping and bias in the flickr30k dataset. In Jens Edlund, Dirk Heylen, and Patrizia Paggio, editors, *Proceedings of Multimodal Corpora: Computer vision and language processing (MMC 2016)*. pages 1–4

Emiel van Miltenburg, Roser Morante, and Desmond Elliott. 2016a. Pragmatic factors in image description: The case of negations. In *Proceedings of the 5th Workshop on Vision and Language*. Association for Computational Linguistics, Berlin, Germany, pages 54–59

Emiel van Miltenburg. 2017. Pragmatic descriptions of perceptual stimuli. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, pages 1–10

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Talking about other people: an endless range of possibilities. In *Proceedings of the 11th International Conference on Natural Language Generation*. Association for Computational Linguistics, pages 415–420. http://aclweb.org/anthology/W18-6550

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 (Smeulders 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; Ørnager 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.

Panofsky Shatford	Iconography Specific Of	Pre-Iconography Generic Of	(See caption) About
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 rougly 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

[2] Next 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.

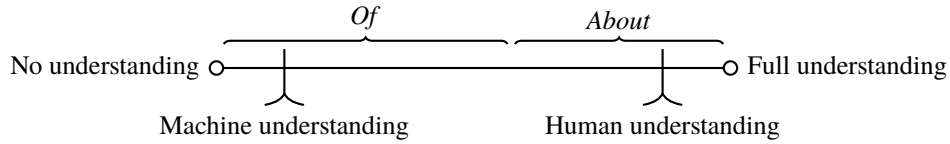


Figure 2.1 A naive interpretation of the scale from zero to full image understanding, in terms of the *Of/About*-distinction.

picture was taken. The semantic gap between humans and machines may then be illustrated as in Figure 2.2 (ignoring the *Of/About*-aspects for simplicity).

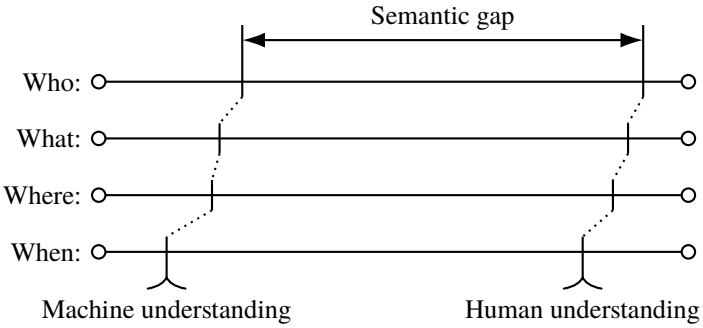


Figure 2.2 More detailed illustration of the semantic gap, using the facets from Shatford (1986). The vertical lines show the performance of machines (left) versus humans (right), and the space between these lines represents the semantic gap. The individual values on these scales are randomly chosen to illustrate the idea of having a 'multi-faceted gap' with different performance values depending on the facet under consideration.

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

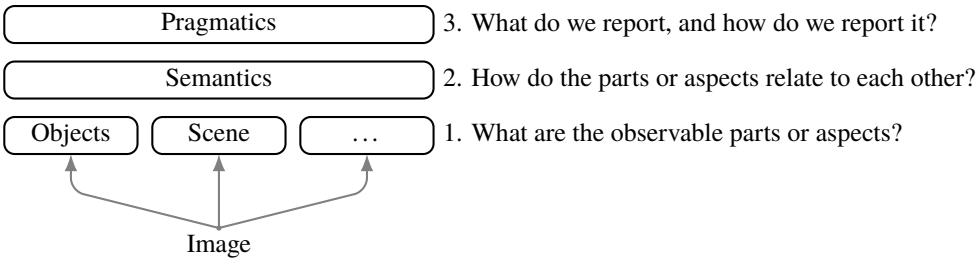


Figure 2.3 Update to Hare et al.'s (2006) proposal. We added a pragmatic level on top of the semantic level, to account for the fact that people may only report a *selection* of the information contained in an image.

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.

Figure 2.4 Painting of two platypuses by Heinrich Harder (from his *Tiere der Urwelt* series, 1916, *public domain*).

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

[3] Though not all scalar expressions give rise to an implicature at the same rate (Van Tiel et al., 2016).

[4] More 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.[5] 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).

[5] This 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).

Please describe the image in one complete but simple sentence.

Next →

Figure 2.5 Prompt for the image description task. Original picture taken by Luigi Cavasin (CC BY-NC-SA) on Flickr.com. Based on the example in (Rashtchian et al., 2010).

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.

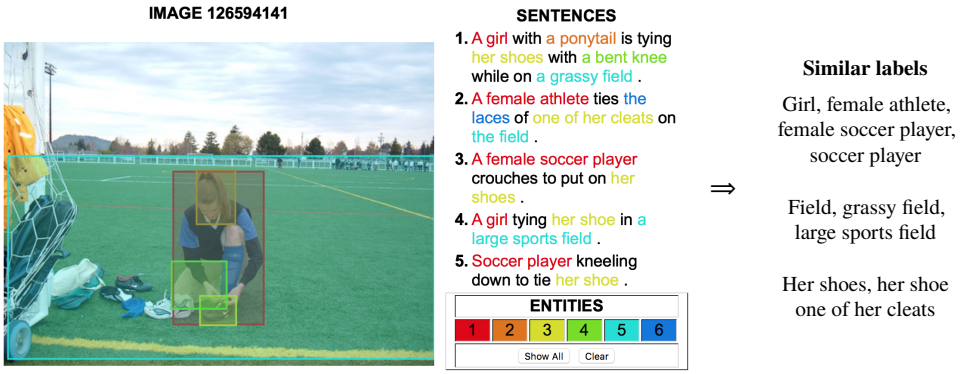


Figure 2.7 Image with bounding boxes indicating the entities referred to in the description, along with three sets of similar labels that would be extracted by the proposed algorithm. Data from the Flickr30k Entities dataset, visualization from the online dataset browser. Original picture by mayamoose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

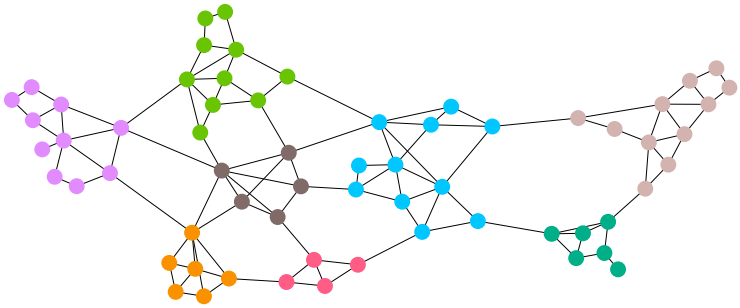


Figure 2.8 Example of a modular graph, where modules are colored after clustering the nodes using the Louvain method. Image generated using Gephi (Bastian et al., 2009).

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 `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 `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.

beard	*white beard*	*long brown beard*	*large white beard*
goatee	*red beard*	*flaming red beard*	*thick beard*
beard and mustache	*braided beard*	*big beard*	*neatly trimmed beard*
gray beard	*gray braided beard*	*short beard*	*scruffy beard*
black beard	*long, white beard*	*bubble beard*	*red facial hair*

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

Appearance	Situation	Example
Yes	No	Police officer, businessman, firefighter
Yes	Yes	Civilian
No	Yes	Bystander, neighbor, passerby, orphan
No	No	—

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.

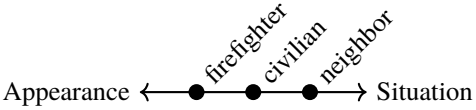


Figure 2.9 Continuous scale from Appearance-based to Contextually determined labels.

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.[6] (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.[7] 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,*	*artist,*	*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/evanmiltenburg/LabelingPeople

LABEL → MOD, GENDEREDNOUN
LABEL → MOD, MOD, GENDEREDNOUN
MOD → ABILITY | ACTIVITY | AGE | ATTRACTIVENESS | BUILD | CLEANLINESS |...
GENDEREDNOUN → woman | man | girl | boy | female | male | women | men | ...
AGE → young | old | middle-aged | adult | elderly | infant | twenty-something | ...
ETHNICITY → African-American | Asian | oriental | caucasian | Chinese | ...

Figure 2.10 Subset of our Context-Free Grammar, designed to match labels with different categories of modifiers. Production rules are based on our category files (which are updated in step 3).

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

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

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.

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

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):

> 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)

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:[9]

> "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.

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.

Figure 2.12 Original by David Gallagher (CC BY-NC-SA) on Flickr.com

Figure 2.13 Original by Mike Boening Photography (CC BY-NC-ND) on Flickr.com

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).

Picture by Caperd (CC BY-ND) on Flickr.com.

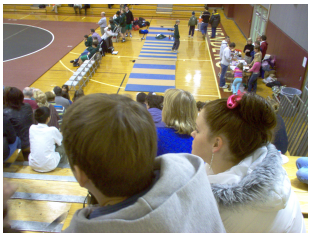
Activity We've seen an example of this earlier in Section 2.7, where the 'manager' was said to be *talking about job performance* and *scolding [a worker] in a stern lecture*. The picture on the right shows another example, where an annotator described the three men as sitting and *contemplating their 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.

Picture by Independentman (CC BY) on Flickr.com.

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.

Picture taken by Chris Palmieri (CC BY-NC-SA) on Flickr.com

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.



Picture taken by Eric Lewis (CC BY-SA) on Flickr.com.

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.

Picture by Ryan Ozawa (CC BY-NC-ND) on Flickr.com

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.

Picture by Julie Kertesz (CC BY-NC-SA) on Flickr.com.

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.[10] 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.[11]

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.

[10] 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).

[11] 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

Asian		Average 60%
2339632913	Asian child/baby	2
3208987435	Asian baby, Asian/oriental woman	3
7327356514	Asian girl/baby, Asian/oriental woman	4
Black		Average 40%
1319788022	African-American (AA)/black baby	3
149057633	African/AA child, black baby	3
3217909454	Dark-skinned baby	1
3614582606	AA baby	1
White		Average 20%
11034843	White baby boy	1
176230509	White baby boy	1
2058947638	White baby	1
3991342877	White baby	1
4592281294	White baby stroller	FP
661546153	White baby stroller	FP
442983801	Fair-skinned baby	1

Table 2.5 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).[12]

[12] All code and data are available online through: https://github.com/evanmiltenburg/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.[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.

Non-affixal negation	Free	*Not*
	Bound	*Never, n't, no, none, nothing, nobody, nowhere, nor, neither*
Implicit negation	Verb	*Lack, omit, miss, fail*
	Preposition	*Without, sans, minus*

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

Table 2.7 Frequency counts for each negation term.

Dataset	1	2	3	4	5
Flickr30K	659	85	16	1	3
MS COCO	2406	277	78	30	5

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.

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.

[13] Affixal negations are words starting with any of the negative morphemes *a–, dis–, un–, non–, un–*, or ending with the morpheme *–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 does**n'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 recognise 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 wo**n'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 do**n'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 ca**n'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 κ=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.

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

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).

(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).[14] 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.[15] In an earlier version of their work,[16] 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 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

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.

Chapter 3

Descriptions in different languages

3.1 Introduction

Do speakers of different languages differ in how they describe the same images? This chapter compares image description datasets across three different languages: US English, Dutch, and German. I show that Dutch and German speakers generally exhibit the same behavior as their American counterparts, but that they also bring their own world knowledge to bear on the image description task. This entails that one cannot simply translate image description data from one language to another, because the translated descriptions may not be suitable for the target audience. No matter how similar two languages or cultures are, we will always need some form of background knowledge to tailor the descriptions to the situation.

3.1.1 Contents of this chapter

This chapter makes three contributions. First, I provide an overview of image descriptions in different languages, and argue that these datasets are useful to compare image description behavior across different languages (§3.2), but that most existing datasets have only been used to train image description systems, and cross-linguistic comparisons have not (or at least: not systematically) been carried out (§3.3).

Second, I present a Dutch image description corpus for the validation and test images of the Flickr30K dataset (§3.4). This serves two purposes: to show how to collect image descriptions in languages other than English, and to obtain descriptions that we can compare with English and German image descriptions.

Third, I compare Dutch, English, and German descriptions and show differences and similarities in how speakers of different languages describe the same images (§3.5). Following this comparison, I look at the amount of variation between descriptions for the same images. This variation has been argued to be due to the content of the images, and Jas and Parikh (2015) refer to this idea as *image specificity*. I show that image specificity is only moderately correlated between Dutch, English, and German (§3.6).

3.1.2 Publications

This chapter was edited from the following publications:

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*. Association for Computational Linguistics, Santiago de Compostela, Spain, pages 21–30

Emiel van Miltenburg, Ákos Kádar, Ruud Koolen, and Emiel Krahmer. 2018a. DIDEC: The Dutch Image Description and Eye-tracking Corpus. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*. Resource available at https://didec.uvt.nl

3.2 Going multilingual

So far, we have only looked at English image descriptions. These are enough to train an automatic image description system, and to start exploring the linguistic properties of image descriptions. But if we were to collect descriptions for the same images in other languages as well, we would be able to find out whether there are any differences in how speakers of different languages describe images. The extent to which there are any differences informs us about the influence of language and background knowledge on the image description process. Similarities tell us which entities, objects, or properties are generally or perhaps even universally relevant to mention.

There would also be technological benefits to having image descriptions in other languages, beyond being able to train image description systems. Elliott et al. (2016) note how the availability of *multilingual multimodal data* opens up new avenues of research, such as multimodal machine translation (generating translations in a multimodal context) or multilingual image retrieval. Luckily, we do not have to speculate about the possibility of having image descriptions in other languages, as researchers have recently started to collect them. Recent years have seen a growing body of image descriptions collected for several different languages. Table 3.1 provides an overview of the different datasets that are available. We can distinguish two kinds of datasets: translations of the original source, and independently collected descriptions. The advantage of the former is that the descriptions are perfectly aligned. The advantage of the latter is that the descriptions are not influenced in any way by the original English descriptions.

Language	Source	T	I	Citation
Chinese	Flickr8K	✓	✓	Li et al. 2016b
Chinese	MS COCO*	✓	✓	Li et al. 2018
Czech	Flickr30K	✓		Barrault et al. 2018
Dutch	Flickr30K*		✓	van Miltenburg et al. 2017
Dutch	MS COCO*		✓	van Miltenburg et al. 2018a
French	MS COCO*	✓		Rajendran et al. 2016
French	Flickr30K	✓		Elliott et al. 2017
German	MS COCO*	✓		Rajendran et al. 2016
German	MS COCO*	✓		Hitschler et al. 2016
German	Flickr30K	✓	✓	Elliott et al. 2016
German	IAPR-TC12		✓	Grubinger et al. 2006
Japanese	UIUC Pascal	✓		Funaki and Nakayama 2015
Japanese	MS COCO*		✓	Miyazaki and Shimizu 2016
Japanese	MS COCO		✓	Yoshikawa et al. 2017
Spanish	IAPR-TC12	✓		Grubinger et al. 2006
Turkish	Flickr8K		✓	Unal et al. 2016

Table 3.1 Image description datasets available in languages other than English, with an indication of their source, and whether the descriptions were **T**ranslated or **I**ndependently collected. Asterisks indicate that the data is a subset of the original dataset. Flickr8K is the predecessor of Flickr30K, see Hodosh et al. 2013.

It is an open question how much speakers of different languages differ in their descriptions of the same images. Therefore, we will look at independently collected descriptions in three different languages (Dutch, English, and German), and compare them in terms of the

phenomena discussed in the previous chapter: the use of negations, racial/ethnic marking, and the presence of unwarranted inferences. I will also highlight the role of familiarity in the generation of image descriptions.

3.3 Uses of image descriptions in other languages

Work on image description in other languages generally focuses on system performance rather than cross-linguistic differences (Elliott et al., 2015; Li et al., 2016b; Miyazaki and Shimizu, 2016). Thus far, any differences have only been anecdotally described.

Li et al. (2016b) collected Chinese descriptions of images in the Flickr8K corpus Hodosh et al. (2013). They highlight the differences between Chinese and English descriptions using a picture of a woman taking a photograph. The English annotators describe the woman as *Asian*, whereas Chinese annotators describe her as *middle-aged*. The authors note that "Asian faces are probably too common to be visually salient from a Chinese point of view."

Miyazaki and Shimizu (2016) collected Japanese descriptions for a subset of the MS COCO dataset, which mostly contains pictures taken in (or by people from) Europe and the United States Lin et al. (2014). They note that in their pilot phase, the images appeared "exotic" to Japanese crowd workers, who would frequently use adjectives like *foreign* and *overseas*. The authors actively tried to combat this by modifying their guidelines to explicitly prevent crowd workers using these phrases, but the observation remains that perspective can strongly influence the nature of the descriptions.

3.4 Collecting Dutch image descriptions

Prior to this research, there was no dataset of described images for Dutch. We decided to collect Dutch descriptions to lay the foundations for the development of a Dutch image description system. This also allows us to compare Dutch, English, and German image descriptions. We used Crowdflower to annotate 2,014 images from the validation and test splits of the Flickr30K corpus with five Dutch descriptions.

Following other work, our goal is to create a comparable corpus of image descriptions, using the images as pivots. This requires us to stay as close to the original task setup as possible, thus fixing the effect of Task Design factor. We base our task on the template used by Hodosh et al. (2013) to collect English descriptions, and by Elliott et al. (2016) for German descriptions. In this design, images are annotated in batches of five images. The task for our participants is to describe each of those images "in one complete, but simple sentence." Before starting on the task, we ask participants to read the guidelines, and to study a picture with example descriptions ranging from *very good* to *very bad*. We include the instructions for our task in Appendix B.

Participants. Crowdflower does not offer the option to select Dutch participants based on their native language. Instead, we restricted our task to level 2 (experienced and reasonably accurate) workers in the Netherlands. We had to continuously monitor the task for ungrammatical descriptions in order to stop contributors from submitting low-quality responses.

Other settings. Following Elliott et al. (2016), we set a reward for $0.25 per completed task (or $0.05 per image), and required participants to spend at least 90 seconds on each task, resulting in a theoretical maximum wage of $10 per hour. We initially limited the number of judgments to 250 descriptions per participant, but due to the small size of the crowd we increased this limit to 500.

Results. A total of 72 participants provided 10,070 valid descriptions in 116 days, at a cost of $821.40. We were surprised by the number of participants who presumably used Google Translate to submit their responses. These are identifiable through their ungrammaticality, usually due to incorrectly inflected verbs. An example is given in (22), with a literal translation and original English description (verified using Google Translate).

(22) Response generated with Google Translate.

a. *Een paar kussen	(Description)
A couple of kisses	(Translation)
A couple kisses	(Original)
b. *Mensen het kopen van vis	(Description)
People the buying of fish	(Translation)
People buying fish	(Original)

Altogether, we had to remove 60 participants due to either submitting ungrammatical responses (60%), Lorum Ipsum text (12%), random combinations of characters (9%), non-Dutch responses (6%), or otherwise low-quality responses (13%).

We conclude that crowdsourcing is a feasible way to collect Dutch data, but it may still be faster to collect image descriptions in the lab (in terms of time to collect the data, not counting the time spent as an experimenter overseeing the task). For large-scale datasets, such as Flickr30K or MS COCO, the Dutch crowdsourcing population seems to be too small to collect descriptions for *all* the images in a reasonable amount of time. This is a problem; with the current data-hungry technology, low-resource languages and languages with smaller pools of crowd workers are in danger of being left behind. For example, Sprugnoli et al. (2016) note that for Flemish, an example of a *small-pool language*, they "were not able to get a sufficient response from the crowd to complete the offered transcription tasks."

3.5 Comparing Dutch, German, and English

3.5.1 General statistics

Table 3.2 shows the mean sentence length (in tokens and words) for the three languages. The English descriptions are the longest, followed by the Dutch and the German ones. However, German has the longest average word length (5.25 characters per word), followed by Dutch (4.62) and English (4.12). This difference seems due to German and Dutch compounding, which is in line with the number of word types: German has 31% more types than English (5709 versus 4355). Dutch has 19% more (5193).

	Tokens	σ	Words	σ
Dutch	11.14	4.5	10.32	4.3
English	13.60	5.6	12.48	5.3
German	9.76	4.2	8.81	3.9

Table 3.2 Mean sentence length across languages.

3.5.2 Definiteness

The five most frequent bigrams that start a description (showing the typical subjects of the images) are given in Table 3.3. The majority starts with an indefinite article, which is in line with the *familiarity theory of definiteness*: the function of definite articles is to refer to familiar referents, whereas indefinite articles are used for unfamiliar referents (Christophersen, 1939; Heim, 1982). The distribution of (in)definite articles follows from the fact that the participants have never seen the images before, nor any context for the image in which the referents could be introduced. A corollary is that systems trained on this data are more likely to produce indefinite than definite articles, and need to be told when definites should be used.

Dutch	Gloss	Count	English	Count	German	Gloss	Count
Een man	A man	517	A man	760	Ein Mann	A man	584
Een vrouw	A woman	252	A woman	367	Eine Frau	A woman	296
De man	The man	105	A young	223	Zwei Männer	Two men	120
Een jongen	A boy	92	A group	211	Ein Junge	A boy	108
Twee mannen	Two men	92	Two men	127	Der Mann	The man	93

Table 3.3 Top-5 most frequent bigrams at the start of a sentence, with their English translation.

3.5.3 Replicating findings for negation, ethnicity marking, and stereotyping

The previous chapter discussed the use of negation and ethnicity marking in English image description datasets. We now attempt to replicate these findings with the Dutch and German data, starting with the use of negations.

Negations. van Miltenburg et al. (2016a) performed a corpus study to categorize all uses of (non-affixal) negations in the Flickr30K corpus. Negations are interesting in descriptions because they describe images by saying what is *not* there. Negations may be used because something in the picture is unexpected, goes against some social norm, or because non-visible factors are relevant to describe the picture. If annotators consistently use negations, this can be seen as evidence that the negated information is part of their shared background knowledge and is a strong requirement for producing human-like descriptions. We readily found examples of negations in both the Dutch and the German data. Some examples are given in (23) and (24), respectively.

(23) Examples from the Dutch descriptions

a. De kinderen dragen **geen** kleding.
'The kids are **not** wearing any clothing.'
b. Vrouw snijdt broodje **zonder** te kijken(!)
'Woman slices a bun **without** looking(!)'

(24) Examples from the German descriptions

a. Zwei Buben **ohne** T-Shirt setzen auf der Straße.
'Two boys without T-shirt sitting on the street.'
b. Eine Ansammlung von Menschen [...] schaut auf ein Ereignis, das **nicht** im Bild ist.
'A crowd of people is watching an event not shown in the picture.'

In total, we found 11 Dutch and 20 German descriptions containing explicit negations in the corpus, while van Miltenburg et al. (2016a) found 27 in English for the same images (excluding false positives). This confirms that workers in different languages mark negations at approximately the same rate, given a sample size of 5,070 sentences. We found almost no images that consistently attracted the use of negations in all three languages: we found only four examples of co-occurring negation between languages.[1] One image is described by speakers of all three languages using a negation (a man with two prosthetic legs, described as having no legs), and there are three other images (all of shirtless individuals) where both English and German workers use negations.

Racial and ethnic marking. van Miltenburg (2016) found that the descriptions in the Flickr30K data have a skewed distribution of racial and ethnic markers: annotators used terms like *Asian* or *black* much more often than *white* or *caucasian*. If we find the same disproportionate use of ethnicity markers in Dutch and German, then we can conclude that this is not a quirk in the English data, but a systematic cultural bias.[2]

Indeed, we did find that non-white people were often marked with adjectives such as *black, dark-skinned, Asian, Chinese*. In Dutch and German, white people were only marked to indicate a contrast between them and someone of a different ethnicity in the same image. The English data contains five exceptions to this rule, where white individuals were marked without any people of another ethnicity being present in the image. We do have to note, however, that there are other ways to *indirectly* mark someone as white, e.g. using adjectives like *blonde* or *brunette*.

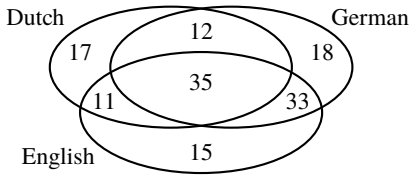


Figure 3.1 Venn diagram of ethnicity markers by Dutch, English, and German workers. Counts correspond to images.

Figure 3.1 shows a Venn-diagram of the use of race/ethnicity markers in Dutch, English, and German. We observe that English and German workers use these markers slightly more often than Dutch workers, but our sample size is not large enough to find any significant differences. Instead, we are interested to know what these groups have in common: what drives people to mention racial or ethnic features?

There are several reasons why people may mark race/ethnicity in their descriptions. One common theme is that annotators mark images where the people are dressed in traditional outfits. Examples include traditional dancers from South-East Asia, and Scotsmen wearing kilts. These items of clothing are *meant to* signal being part of a group, and the annotators picked up on this.

The distribution of the labels may be explained in terms of markedness (Jakobson, 1972) and reporting bias (Misra et al., 2016). In this explanation, white is seen as the unmarked

[1] We define 'co-occurrence between languages' as 'having at least one description for each language that shows the relevant phenomenon.'

[2] This bias is the same as what Beukeboom (2014) calls 'linguistic bias'. We followed this convention in the previous chapter, but feel that 'cultural' is more appropriate here, as it reflects the (apparent) shared bias between Western, majority-White cultures.

default, as it is the dominant ethnicity in all three countries.[3] The marker *white* is only used to be consistent in the use of modifiers within the same sentence. This reasoning also explains the observation by Miyazaki and Shimizu (2016) that Japanese crowd workers often used the labels *foreign* and *overseas* for the MS COCO images.

A final reason for crowd workers to mention ethnicity and skin color may be that the images are visually less interesting, but the description task still demands that the workers provide a description. Workers are thus pressured to find *something* worth mentioning about the image, because too general descriptions might get their work rejected. This is a general task effect that may have implications beyond racial/ethnic marking.

Speculation. van Miltenburg (2016) also found that that annotators often go beyond the content of the images in their descriptions, making *unwarranted inferences* about the pictures. If we find that Dutch and German crowd workers also make such inferences, we conclude that image descriptions in all three languages are *interpretations* of the images that may not necessarily be true.

We observed unwarranted inferences throughout the Dutch and German data, especially about women with infants, who were often seen as the mother. Figure 3.2 shows an image where both Dutch, English, and German workers suggested the woman is the *grandmother*. In the most extreme case, two KLM stewards in pantsuits were described by a German worker as well-dressed *Lesben* ('lesbians'). It would be undesirable for a model to associate all unseen images of air stewards with lesbians. We expect that having multiple descriptions alleviates this type of extreme example, but there is an open question about how to deal with more common types of speculation.

Figure 3.2 Picture showing an older woman and a young girl in a kitchen. The older woman in the picture was often seen as the grandmother. Picture by Ben Hoyt on Flickr.com (all rights reserved).

3.5.4 Familiarity

As the speakers of Dutch, English, and German have different backgrounds, some images may be more familiar to one group than to the others. Familiarity enables speakers to be more specific (but doesn't necessarily cause them to *be* more specific). We will look at three kinds of examples (selected after inspecting the full validation set), where differences in familiarity

[3] The US population is 75% white, according to the 2010 census (Humes et al., 2011). The Dutch and German census bureaus do not monitor ethnicity, and instead report that 77% of the Dutch population is Dutch/Frisian (Centraal Bureau voor de Statistiek, 2016) and 80% of the German population is German (Statistisches Bundesamt, 2013).

lead to differences in the description of named entities, objects, and sports. These examples are illustrative of a larger issue, namely that descriptions in one language may not be adequate for speakers of another language (even if they were perfectly translated). We discuss this issue in §3.7.1.

Named entities

The Dutch, English, and German descriptions differ in their use of place and entity names. We study two cases: one image that is more likely to be familiar to European workers (German and Dutch), and one that is more likely to be familiar to US workers (English).

The Tuileries Garden. Figure 3.3 shows a scene from the Tuileries Garden in Paris, a popular tourist attraction. It may be more likely for a European crowd worker to have visited this location than for an American crowd worker. Three Dutch people indeed included references to the actual location in their description. One mentioned the Arc de Triomphe in the background, one said that this picture is from a square in Paris, and the most specific description (correctly) identified the location:

(25) Een man zit aan de vijver van het Tuilleries park in Parijs.
'A man is sitting by the pond of the Tuileries park in Paris.'

Neither the German nor the American workers identified the location or the monuments by name (though one American worker thought this picture was taken at the Washington Monument). Instead of mentioning the location, the English and German workers describe the scene in more general terms. Two examples are given in Example 26.

(26) a. A person in a white sweatshirt is sitting in a chair near a pond and monument.
b. A man in a white hoodie relaxes in a chair by a fountain.

These examples reveal a common strategy to handle unfamiliarity: focus on something else you *do* know. This undermines the idea that crowd-sourced descriptions tell us what is relevant about the picture.

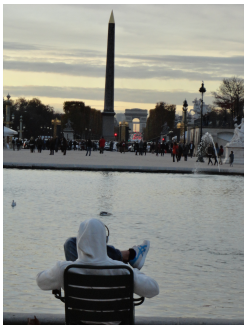


Figure 3.3 This picture was taken at the Tuileries Garden in Paris, and shows the Luxor Obelisk and the Arc de Triomphe. Image credit: eltpics (CC BY-NC) on Flickr.com.

Figure 3.4 This picture shows a man wearing a Denver Broncos hat and jersey. Picture taken by Bradley Gordon (CC BY) on Flickr.com.

The Denver Broncos. Figure 3.4 shows a man wearing a Denver Broncos hat and jersey. The Denver Broncos are an American Football team, which is not so well-known in Europe.

Two American crowd workers but neither the Dutch nor the German workers identified the Broncos jersey. Three out of five American workers also described the activity in the image as *tailgating*, a typical North-American phenomenon where people gather to enjoy an informal (often barbecue) meal on the parking lot outside a sports stadium. As this concept is not so prevalent in Dutch or German culture, there is no Dutch or German word, idiom, or collocation to describe tailgating. Such 'untranslatable' concepts are called *lexical gaps* (Lehrer, 1970). The presence of this gap means that the Dutch and German workers can only concretely describe the image without being able to relate the depicted event to any more abstract concept.

Objects

Familiarity also plays a role in labeling objects. Consider Figure 3.5, which shows (the backside of) a street organ in a shopping street in the Netherlands. All Dutch workers, as well as two German workers identified this object as a street organ, whereas the English workers are only able to provide very general descriptions (Example 27).

Figure 3.5 Picture showing the back of a street organ in the Netherlands. Original by user *rgarciasuarez74* on Flickr.com. License unknown.

(27) a. A trailer hitch is holding a **large contraption**.
b. A **yellow truck** is standing on a busy street in front of the Swarovski store.
c. A **strange looking wood trailer** is parked in a street in front of stores.
d. An **unusual looking vehicle** parked in front of some stores.
e. A **trailer** drives down a red brick road.

This example illustrates two strategies the crowd may use to provide descriptions for unfamiliar objects: (1) signal the unfamiliarity of the object using adjectives like *strange* and *unusual looking*. This is similar to the finding by Miyazaki and Shimizu (2016) that the Japanese crowd made frequent use of terms like *foreign* and *overseas* for the Western images from MS COCO. (2) use a more general cover term, like *vehicle*. Such terms may have a higher *visual dispersion* (Kiela et al., 2014), but they provide a safe back-off strategy.[4,5]

[4] Visual dispersion denotes the amount of differences between images corresponding to a particular term. Concrete, more specific terms tend to have a lower dispersion than abstract, more general terms. For example, the term *vehicle* corresponds to a much more diverse set of objects than the term *car*.

[5] See Blum and Levenston 1978 for a further discussion of strategies to avoid particular words or concepts.

Sports

We found that unfamiliarity with different kinds of sports leads to the misclassification of those sports. We focus on three sports: American Football, Rugby, and Soccer. Looking at images for these sports, we compared how the three different groups referred to them. We found that the German and Dutch groups patterned together, deviating from the American crowd workers.

As expected, the Dutch and German workers make the most mistakes categorizing American Football. For all seven pictures of American Football, there is at least one Dutch annotator who thinks it's a game of Rugby. For six of those, at least one German annotator made the same mistake. By contrast, workers from the US made more mistakes identifying rugby images. For all three pictures of Rugby, there is at least one American calling it Soccer or Football. For one of those images, a German annotator thought it was American Football. All Soccer images were universally recognized as Soccer.

3.5.5 Takeaway

The main takeaway for this section is that the observations from the previous chapter seem to hold cross-linguistically, or at least for Dutch, English, and German. For all three languages, we have found that:

1. Participants use negations in their descriptions. This shows that participants in all three languages reason about the images and whether they conform with their expectations.
2. Participants use racial and ethnic markers in their descriptions. In all three languages, White is the default, and is not mentioned unless there are any special circumstances.
3. Participants speculate about the images. This shows that in all three languages, participants actively interpret the images, and use their world knowledge to supplement the information that can be gleaned from the images.

In the introduction (§1.8), I have outlined different kinds of arguments that one may use based on corpus evidence. In the previous chapter, I have mainly employed the *existence* argument: we may observe these three different phenomena in English image description data. If we want a full solution to the image description task, then any complete system should be able to account for these phenomena. The current chapter has added *cross-linguistic evidence*: participants in different languages also use these linguistic mechanisms to talk about images. This shows that it is not just a quirk of English that people use negations in their descriptions, for example. Rather, speakers of all three languages find negations useful to describe the images in the Flickr30K dataset.

The introduction also lists *systematicity* as an argument for the importance of linguistic phenomena to be captured by any model of image description. Upon further consideration, there may be two kinds of systematicity: within images and across images. The street organ example (Figure 3.5) shows the former kind of systematicity, where all Dutch participants describing this image make reference to the street organ, and all use the same word (*draaiorgel*), even though that word is extremely rare in the corpus (it is only used for this image). If the image description experiment were to be repeated, I expect that all Dutch participants would again show the same behavior. The latter kind of systematicity (across images) refers to the same behavior being shown for multiple, similar images. For example, if we were to repeat the image description experiment with multiple different images of street organs, and we would again observe that US crowd-workers do not recognize these instruments, while Dutch

crowd-workers all recognize them and consistently use the same word to refer to them. Finally, we have seen one example of a systematic trait in Section 3.5.2, with the use of the indefinite article for entities (usually people) that the crowd-workers have not seen before.

The final kind of argument listed in the introduction is based on *frequency*: if a particular linguistic phenomenon occurs often in the data, then it may also be more important for an image description model to capture that phenomenon. Thus, it seems useful to know how often the phenomena discussed in this chapter occur. I leave the investigation of this issue for future research. Finally, the Dutch street organ example is made all the more salient by the fact that *draaiorgel* is a rare word. Nevertheless, all the Dutch crowd-workers used it to refer to the same entity. This strengthens the argument that knowledge of culturally relevant artifacts is an essential part of the ability to describe images.

3.6 Variation[6]

The previous chapter looked at variation in the use of entity labels. We will now turn to look at variation at the sentence level. For this, we will use the concept of *image specificity*, proposed by Jas and Parikh (2015). Whenever you ask multiple people to describe the same image, you rarely get the same description. The authors show that this variation is not consistent: some images elicit more variation than others. In their terminology: some images are *specific*, resulting in little variation in the descriptions, while others are more *ambiguous*. Jas and Parikh operationalize this idea by proposing a measure of image specificity, that computes the average similarity between the descriptions for each image. If the average is high, the image is said to be specific, and if the average is low, the image is said to be ambiguous.

Jas and Parikh (2015) show that their image specificity metric correlates well with human specificity ratings collected for the images from the image memorability dataset (Isola et al., 2011). With a Spearman's ρ of 0.69, their measure is close to human performance (0.72). To show that specificity is really a property of the image, Jas and Parikh (2015) carry out two experiments:

1. Replicating an image description task: if we ask another group of people to provide descriptions for the same set of images, do we then see the same amount of variation for each image? In their experiment, Jas and Parikh obtained a fairly strong correlation of 0.54 between groups, meaning that the variation did not just arise by chance.
2. A regression analysis: can we predict variation between the image descriptions on the basis of an image? Jas and Parikh reveal that image specificity can indeed be predicted from different properties of an image, such as the presence of people and large objects, the absence of generic buildings or blue skies, and the importance of objects that are visible. (Importance is calculated based on the number of mentions for certain objects in a set of image descriptions.)

But if image specificity is indeed a property of the image, we should also be able to correlate image specificity scores across different languages. We will test this hypothesis for Dutch, English, and German using existing datasets.

[6]This section is based on the research originally reported in van Miltenburg et al. 2018a.

3.6.1 The image specificity metric

Jas and Parikh compute image specificity by taking the average similarity between all descriptions of the same image. The similarity between pairs of sentences is determined using WordNet Fellbaum (1998):

1. For each word in the first sentence, compute the maximum path similarity between all possible synsets of that word and all possible synsets of all words in the second sentence. This is an alignment strategy to find the best matches between both sentences.
2. Repeat the process in the opposite direction: for each word in the second sentence, compute the maximum path similarity with the words in the first sentence.
3. Compute the average path similarity, weighted by the importance of each word (determined using TF-IDF on the entire description corpus under consideration).

Using this method, Jas and Parikh (2015) get a correlation of 0.69 with human specificity ratings, close to the inter-annotator correlation of 0.72. Their conclusion is that this is a reliable measure to estimate image specificity. One problem with this measure is that it requires a lexical resource (WordNet) that is not available for every language.[7] Since we want to run the evaluation corpus on the Dutch descriptions, and because the original implementation is relatively slow and difficult to modify, we re-implemented Jas and Parikh (2015)'s image specificity measure. Our reimplementation also achieves a correlation of 0.69 with the human ratings, and 0.99 with the original implementation.[8] Having validated our reimplementation, we replaced WordNet similarity with cosine similarity, using the GoogleNews word vectors (Mikolov et al., 2013a). With this modification, we achieve a correlation with human ratings of 0.71, and a correlation of 0.87 with the original implementation. We also ran the same measure using the FastText embeddings (Bojanowski et al., 2017), achieving a correlation of 0.69 with the human ratings and 0.86 with the original implementation. This means that the metric performs on par with Jas and Parikh's original measure, but captures slightly different information about the image descriptions.

3.6.2 Correlating image specificity between different languages

We used the embedding-based specificity metric to compare image descriptions in 3 different languages, using off-the-shelf embeddings (listed in Table 3.4). We compare English (ENG) descriptions from the Flickr30K dataset (Young et al., 2014) with German (DEU) and Dutch (NLD) descriptions for the same dataset (Elliott et al., 2016; van Miltenburg et al., 2017).

Table 3.5 presents the correlations between the scores. Our results show a striking difference between scores computed using word2vec embeddings and those computed using FastText embeddings. This difference seems to be due to poor performance of the German model, as the correlations between the Dutch and English scores are reasonably similar between word2vec and FastText (0.36 versus 0.40). The reason for this may be that the word2vec model has limited coverage, while the FastText model uses subword information to compute vectors for tokens that are out-of-vocabulary. This is especially important for languages like German, which uses more compounding and has a richer morphology than English. However, this

[7] Even though wordnets exist for Dutch (Postma et al., 2016b) and German (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010), we did not use them because they have lower coverage, and we do not need to worry about lemmatization.

[8] We also found that the WordNet lookup is the main bottleneck, and we can significantly speed up the algorithm by caching the word-to-word similarities. We used the built-in `@lru_cache` decorator in Python 3, storing a million input-output pairs.

Language	Type	Source
Dutch	word2vec	Mandera et al. 2017
English	word2vec	Mikolov et al. 2013a
German	word2vec	Müller 2015
All	FastText	Bojanowski et al. 2017

Table 3.4 Word embeddings used to compute the image specificity metric.

Comparison	Split	word2vec	FastText
NLD, DEU	val	0.23	**0.47**
NLD, ENG	val	0.36	**0.40**
DEU, ENG	val	0.18	**0.41**
DEU, ENG	train	0.16	**0.39**

Table 3.5 Spearman correlation between automated image specificity scores in different languages, using two sets of word embeddings.

does not explain why the correlations for Dutch (also a morphologically richer language than English) are relatively constant.

We observe that the scores based on the FastText embeddings have correlations between 0.39 and 0.47. This means that, to some extent, image specificity is indeed language-independent. In other words, the data suggests that some images just elicit more varied responses than others, and it does not matter whether you speak Dutch, English, or German. However, the explained variance is at most 22% ($R^2 = 0.47^2 = 0.2209$), so 78% of the variance is still unaccounted for. There are two ways to interpret this result, either based on the metric or on the data.

Metric. One could argue that a correlation between 0.39 and 0.47 is already impressive, given that the image specificity metric does not take compositionality into account; it just checks the similarity between the words used in the different descriptions, and ignores how the words are combined. With a metric that better captures the meaning of different sentences, we would achieve more accurate image specificity scores, which reduces noise and might give us a better correlation between the different languages.

Data. A different way to interpret the results is that this is as good as it gets, with the data that we currently have. Some images have a really clear subject, and will have very similar descriptions. But otherwise, the image description process is random and only somewhat constrained by the contents of each image. This may be caused by the way the descriptions were collected. The open-ended image description task is virtually unconstrained and lets participants do whatever they like. With a more targeted image description task, we may see an overall rise in image specificity (participants provide more similar descriptions) and a higher correlation between different languages (the effect of the image is stronger, because the noise resulting from the task is reduced). Alternatively, we might also see that, regardless of the task, people will still produce very different descriptions because they bring different experience and background knowledge to bear on the image description task.

Finally, we should note that the reliability of the image specificity metric could improve if we would have more descriptions per image. Vedantam et al. (2015) show that we could collect up to 50 descriptions per image and still find novel information about the image.

3.7 Conclusion

This chapter provided an overview of differences and similarities between image descriptions across different languages. I have shown that the phenomena observed in Chapter 2 (stereotyping, bias, using negations) occur in Dutch and German as well. Furthermore, we have seen that differences in cultural background may influence the descriptions that people produce. Of course, this is not specific to speakers of different languages; differences in background

knowledge between speakers of the same language should similarly influence the descriptions that those speakers may produce.

3.7.1 Implications for image description systems

Image specificity

The image specificity results emphasize the fact that image descriptions are very diverse, although it is not clear what causes this diversity. But also given the other results in this chapter, we may safely say that we cannot predict on the basis of an image alone how diverse the descriptions will be. Thus, one interesting avenue of research seems to be to explicitly model additional sources of variation. Some precedent for this already exists in the work of Wang et al. (2016), who present an image description model that learns multiple image description distributions simultaneously. Their model is able to produce multiple descriptions per image.

Description specificity

In Section 3.5.4 we observed that annotators differ in the specificity of their descriptions due to their familiarity with the depicted scenes or objects. One challenge for image description systems is to find the right level of specificity for their descriptions, despite this variation. Of course, what is 'the right level' also depends on the context in which the description should be produced. But if a system can identify the exact category of an object, it is probably more useful to produce e.g. *street organ* rather than *unusual looking vehicle*.

Besides familiarity, there are also other factors influencing label specificity. For example, cultures may have differences in their *basic level*; i.e. how specific speakers are generally expected to be (Rosch et al., 1976; Matsumoto, 1995). For this reason, *dog* is a more appropriate label than *affenpinscher* in most situations, even though the latter is more specific. Ideally, image description systems should recognize when to use a more general term, and when to go more into detail (Ordonez et al., 2015).

Limitations of translation approaches

One approach to image description in multiple languages is to use a translation system. For example, Li et al. (2016b) compare two strategies: *early* versus *late* translation. Using early translation, image descriptions are translated to the target language before training an image description system on the translated descriptions. Using late translation, an image description system is trained on the original data, and the output is translated. Li et al. (2016b) show that the former strategy achieves the best result, and argue that it is a promising approach because it requires no extra manual annotation. Following Li et al., others have used the translation strategy for Japanese, Turkish, and Italian (Yoshikawa et al., 2017; Samet et al., 2017; Masotti et al., 2017).

Our observations in Section 3.5.4 show that there are limits to what a translation-based approach can achieve. While translation provides a strong baseline, it can only capture those phenomena that are familiar to the crowd providing the descriptions. The street organ example shows that there exists a 'knowledge gap' between Dutch and English. Dutch users would certainly not be satisfied with street organs being labeled as *unusual looking vehicles*. If the translation-based approach is to be successful, future research should find out how to bridge such gaps.

So far we have not looked at frequency of culture specific items that would require native speakers to describe. Hence it is not clear how much the descriptions would be affected if we were only to use translations. Luckily, Frank et al. (2018) have carried out some experiments in this direction. They set up a rating task comparing original German descriptions with descriptions that were translated from English to German. Participants were asked to rate how well each of these descriptions described an image, using a seven-point Likert scale. The authors found that the original German descriptions were rated significantly better than the the translations from English. However, the effect size for this difference is very small, and the median difference between the two is negligible. Frank et al. (2018) conclude that, for practical purposes, the difference is so small that in this domain (the Flickr30K images) and with this combination of languages (German and English) translation is a good strategy. But, for languages pairs that are very different, or for domains where familiarity plays a bigger role, we still need to go beyond translation.

3.7.2 Limitations of this study

Our focus on Germanic languages from the Western world does not allow us to make general statements about how people describe images. A comparison with taxonomically and culturally different languages might help us uncover important factors that we have missed in this study. A surprising example comes from Baltaretu et al. (2016), who discuss how writing direction (left-to-right versus right-to-left) affects the way people process and recall visual scenes. This may have implications for the way that images are described by (or should be described for) speakers of languages that differ in this regard.

Finally, there are limits to what a corpus study can show. The phenomena described here are presented with post-hoc explanations. Plausible as these explanations may be, they are still hypotheses. We think these hypotheses are useful guides in thinking about image description, but they still remain to be validated experimentally.

3.7.3 Next chapter

Chapter 4 explores the image description process in more detail, using a real-time dataset of spoken image descriptions and eye-tracking data. I will provide examples that show evidence of human prediction and reasoning during the description process.

Chapter 4

Image description as a dynamic process

4.1 Introduction

Language production is a dynamic process. When people talk, they do not blurt out sentences one by one, as indivisible chunks. Rather, they build up their utterances over time, constantly monitoring what they have just said and what effect their utterances may have on their interlocutor. This chapter shows that image descriptions are no different. In the previous two chapters, we have used existing image description datasets for our research, such as Flickr30K and MS COCO (Young et al., 2014; Lin et al., 2014). However, existing datasets can only provide limited insight into the way humans produce image descriptions, because they only contain the *result* of that process, and do not tell us anything about *how the descriptions came about*. This kind of real-time information can be very insightful for developing image description systems, which is why we decided to collect a new dataset of spoken image descriptions, paired with eye-tracking data.

4.1.1 Contents of this chapter

This chapter introduces DIDEC, a corpus of spoken Dutch image descriptions with eye-tracking data. We explain how the corpus was created (§ 4.3), and provide general statistics about the resource, along with a short discussion of the annotated corrections, providing insight in the description process (§ 4.4). We also present an initial study, where we show that the eye-tracking data for the image description task is more coherent than the free-viewing data (§ 4.5). Section 4.6 offers suggestions for future research. Our corpus is freely available, along with an exploration interface, and all the materials that were used to create the dataset.[1]

4.1.2 Publications

This chapter is based on the following publication:

Emiel van Miltenburg, Ákos Kádar, Ruud Koolen, and Emiel Krahmer. 2018a. DIDEC: The Dutch Image Description and Eye-tracking Corpus. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*. Resource available at https://didec.uvt.nl

4.2 The Dutch Image Description and Eye-tracking Corpus

One important part of the human image description process is *visual attention*, i.e. which parts of the image people look at when they are asked to describe an image. Coco and Keller (2012) show that there are similarities between sequences of fixated objects in scan patterns and the sequences of words in the sentences that were produced about the images. This idea has been carried over to automatic image description systems in the form of *attention-based models*. Xu et al. (2015) show that one can improve the performance of an image description model

[1] Our resource is available at: http://didec.uvt.nl

by adding an attention module that learns to attend to salient parts of an image as it produces a description. Their model produces attention maps at every time step when it produces the next word. Lu et al. (2017a) improve this approach by having the model learn when visual information is or is not relevant to produce a particular word or phrase. We will discuss Xu et al.'s (2015) model in Chapter 6.

To better understand the role of visual attention in image description, we need a real-time dataset that shows us where the participants are looking as they are producing the descriptions. This chapter presents such a dataset: the Dutch Image Description and Eye-tracking Corpus (DIDEC). DIDEC contains 307 images from MS COCO that are both in SALICON (Jiang et al., 2015) and the Visual Genome dataset (Krishna et al., 2017). SALICON is a growing collection of mouse-tracking data, which is used to generate *attention maps*: heatmaps that show which parts of an image are salient and attract attention. The Visual Genome is a knowledge base that combines metadata from different sources about the images it contains. Thus, future researchers can use information from all these different sources in their analysis of our data.

Raw	Een hele kudde schapen ⟨uh⟩ met een man ⟨corr⟩ met een herder erachter en een pakezel.
Translation	A whole herd of sheep ⟨uh⟩ with a man ⟨corr⟩ with a shepherd behind them and a mule.
Normalized	Een hele kudde schapen met een herder erachter en een pakezel
Translation	A whole herd of sheep with a shepherd behind them and a mule.

Figure 4.1 Example item from DIDEC, with the annotated raw transcription, and the intended description. Left: image from MS COCO (originally by Jacinta Lluch Valero, CC BY-SA 2.0), Right: image overlaid with an eye-tracking heatmap. Glosses were only added for presentation in this chapter.

Each image in DIDEC is provided with spoken descriptions and real-time eye-tracking data. There are between 14 and 16 spoken descriptions per image. Each of these descriptions was manually transcribed and annotated. We provide the audio with two kinds of transcriptions (an example is given in Figure 4.1):

1. Raw descriptions, annotated with markers for repetitions, corrections, and (filled) pauses.
2. Normalized descriptions, without repetitions, and with the corrections suggested by the speaker.

Having these two kinds of descriptions enables us to develop a better understanding of the language production process, for example showing exactly where participants experience increased cognitive effort. The normalized descriptions facilitate comparison with written descriptions and improve searchability of the corpus. We also provide two kinds of eye-tracking data:

1. Free viewing: eye-tracking data collected without any concurrent task.

2. Description viewing: eye-tracking data collected simultaneously with the spoken descriptions.

These two sets of eye-tracking data allow us to study the influence of the description task on visual attention. Earlier studies have shown that different tasks may cause different patterns of visual attention (Buswell, 1935; Yarbus, 1967; Coco and Keller, 2014). Our eye-tracking data is complementary to the mouse-tracking data in SALICON, which can only be used to study *bottom-up* attention (driven by the image), and not *top-down* attention (driven by a specific task, such as image description; see our discussion in Section 4.5). Furthermore, because we collected *spoken* image descriptions, the descriptions are aligned with the eye-tracking data in the description viewing task. This is useful when studying phenomena like self-correction (Section 4.4.2).

4.3 Procedure

We carried out an eye tracking experiment consisting of two separate sub-experiments, which represented two tasks: (1) a free viewing task, during which participants looked at images while we tracked their eye movements, and (2) a task in which participants were asked to produce spoken descriptions of the images, while again their eye movements were recorded. There were different participants for the two sub-experiments, so no image was viewed twice by the same participant.

Data and Materials. Our image stimuli came from MS COCO (Lin et al., 2014), which contains over 160K images with 5 English descriptions each. We selected 307 images matching the following criteria: they should be in landscape orientation, and be part of both the SALICON and the Visual Genome dataset (Krishna et al., 2017). The latter was done for maximum compatibility with other projects.

In order to avoid lengthy experiments, we made three subsets of images, which we refer to as lists in the corpus: one list of 103 images, and two lists of 102 images. In both tasks, participants saw only one list of images. Participants were randomly assigned to one of the lists, with each between 14 and 16 participants. To avoid order effects, we made two versions of each list, which reflect the two fixed random orders in which the images were shown. We registered eye movements with an SMI RED 250 device, operated by the IviewX and the ExperimentCenter software packages.[2] We recorded the image descriptions using a headset microphone.

Free viewing versus Production viewing. In the free viewing task, subjects viewed images for three seconds while their eye movements were recorded. In the image description task, participants also viewed images, but this time they were also asked to produce a description of the current image (while their eye movements were again tracked). The instructions for this task were translated from the original MS COCO instructions. Participants could take as much time as needed for every trial to provide a proper description. In both tasks, every trial started with a cross in the middle of the screen, which had to be fixated for one second in order to launch the appearance of the image. All images in our study both occurred in the free viewing task and in the image description task, but always with different participants. This way, each image viewed by the participants was new to them, preventing any possible familiarity effects.

[2]The eye tracker had a sampling rate of 250 Hz. The stimulus materials were displayed on a 22 inch P2210 Dell monitor, with the resolution set to 1680 x 1050 pixels. The images were resized to 1267 x 950 pixels (without changing the aspect ratio), surrounded by grey borders. These borders were required because eye-tracking measurements outside the calibration area (i.e., in the most peripheral areas of the screen) are not reliable. The viewing distance was 70 cm.

Avoiding any confounding from familiarity effects also means we are forced to carry out a between-subjects analysis to study the effect of the task on the viewing patterns for the same image.

Transcription and annotation. After exporting the recordings for each trial, we automatically transcribed the descriptions using the built-in Dictation function from macOS Sierra 10.12.6.[3] The transcriptions were manually corrected by a native speaker of Dutch. To estimate the actual quality of the automatic transcriptions, we computed the word error rate (WER) for the automatic transcriptions, as compared to the corrected transcriptions.[4] This resulted in a WER of 37%.

We refer to the transcribed descriptions in the corpus as *literal descriptions*. In addition, the annotator marked repetitions, corrections, and (filled) pauses (*um, uh*, or silence longer than 2 seconds) by the speaker. We will later use these meta-linguistic annotations to gain more insight into the image description process. Finally, our annotator provided the *normalized descriptions*, without filled pauses or repetitions and with the repairs taken into account.

Participants. Our participants were 112 Dutch students who earned course credits for their participation: 54 students performed the free viewing task, while 58 students completed the image description task. We could not use the data of 19 participants (6 in the free viewing task; 13 in the image description task), since eye movements for these people were not recorded succesfully, or only partially. This was mainly due to the length of the experiments, and to the fact that speaking could distort the eyetracking signal. We tried to prevent this issue by calibrating participants' eyes to the eyetracker twice: once before the start, and once halfway. The final data set consists of data for 48 participants (34 women) in the free-viewing condition, with a mean age of 22 years and 3 months; and data for 45 participants (35 women) in the image description condition, with a mean age of 22 years and 6 months.[5]

Our experiment followed standard ethical procedures. After entering the lab, participants were seated in a soundproof booth, and read and signed the consent form. This form contained a general description of the experimental task, an indication of the duration of the experiment, contact information, and information about data storage. Participants needed to give explicit permission to make available their audio recordings and eye movement data for research purposes; otherwise, they would not participate. Also, participants were allowed to quit the experiment at any stage and still earn credits.

4.4 General results: the DIDEC corpus

In the description condition, 45 participants produced 4604 descriptions (59,248 tokens), leading to an average of 15 descriptions per image (min 14, max 16). The average description length for the normalized descriptions is 12.87 tokens (Median: 12, SD: 6.45). By comparison, the written English descriptions in MS COCO are shorter (average: 10.78 tokens) and have a lower variance in description length (SD: 2.65).[6] We checked to see if the difference in length

[3] This required us to emulate a microphone using SoundFlower 2.0b2, to use Audacity 2.1.0 to play the recordings and direct the output through the emulated microphone to the Dictation tool.

[4] We used the evaluation script from: https://github.com/belambert/asr-evaluation

[5] For 3 participants in the description task, and 4 participants in the free viewing task only a small subset of the eye-tracking data is missing (14 trials in total for the description task, and 7 trials in the free viewing task). We decided to keep these participants and treat the trials as missing data.

[6] We only counted the description lengths for the 307 images that are also in DIDEC. Since DIDEC lacks periods at the end of the descriptions, we also stripped them from the MS COCO descriptions. We used the SpaCy tokenizer to obtain the tokens.

is due to any differences between Dutch and English, using the Flickr30K validation set (data from Van Miltenburg et al., 2017). We found that the English descriptions are in fact *longer* than the Dutch ones (with a mean of 12.77 tokens for English (SD: 5.67) versus 10.47 tokens for Dutch (SD: 4.45)). These findings are in line with earlier findings from Drieman (1962a) and others that spoken descriptions are typically longer than written ones. We discuss the differences between spoken and written language in more detail in the next chapter.

We found a high degree of variation in description length across different participants. The difference between the lowest and highest median description length is 16.5 tokens (Lowest: 8, Highest: 24.5, Mean: 12.30, SD: 4.15). We also checked whether sentence length decreases with length of experiment, by correlating sentence length with the order in which the images were presented. We found a Spearman correlation of 0.06, suggesting that order had no effect on description length. Following this, we looked at the variation in description length between images. We found that the difference between the lowest and highest median description length is 15 tokens (Lowest: 6, Highest: 21, Mean: 11.75, SD: 2.46). We conclude that there is a greater variability between participants than between images.

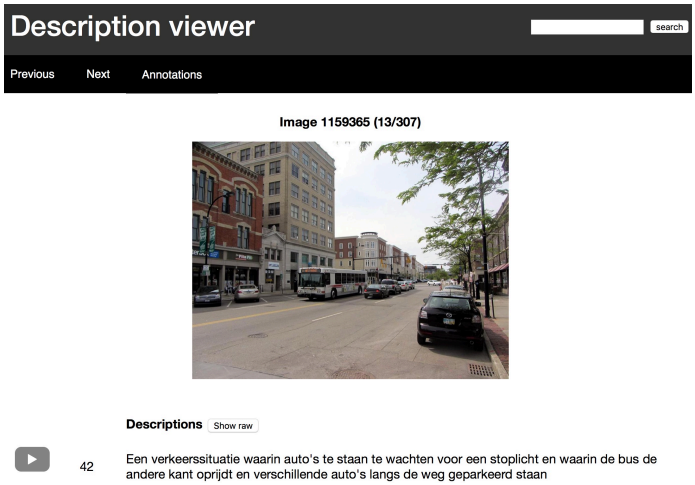


Figure 4.2 The description viewer provides a browser-based interface to the corpus. Users can browse through the images, search for specific words or annotations, and listen to the spoken descriptions. (Displayed image by David Wilson, CC BY 2.0)

4.4.1 Viewer tool

We made a description viewer application that allows users to browse through the images, read the annotated descriptions, and listen to the spoken descriptions. Users can also search the descriptions for particular annotations, or for the occurrence of particular words. The description viewer will then return a selection of the images where at least one of the descriptions contains that particular word or annotation. See Figure 4.2 for an impression of the interface, and see the appendix (§A.6) for details. The viewer tool can be downloaded along with our data from the corpus website.

4.4.2 Exploring the annotations in the dataset: descriptions with corrections

Recall that we also annotated basic meta-linguistic information to the raw descriptions, such as pauses, repetitions, and corrections. Table 4.1 shows the number of times each label was annotated. We chose to add these labels because they may inform us about the image description process. For example, one might expect participants to use more filled pauses and repetitions if the image is more complex or unclear (cf. Gatt et al., 2017). Repetitions, in this case, would signal initial uncertainty about the interpretation of the image, followed up by a confirmation that their initial interpretation was correct.

Tag	Meaning	Count
<uh>	Filled pause	1277
<corr>	Correction	693
<rep>	Repetition	139
<pause>	Pause	123
<?>	Inaudible	23

Table 4.1 Annotation counts.

Let us now look at some examples of corrections in the image description data. This will give us some idea of why people tend to make corrections in their descriptions, and what this process looks like. One of the first studies on this topic is provided by Levelt (1983), who discusses a corpus of 959 repairs that were spontaneously made by Dutch speakers after they were asked to describe visual patterns. The difference between DIDEC and Levelt's corpus is that the latter consists of abstract stimuli while DIDEC uses pictures of real-life situations. Levelt used his data to study monitoring (roughly: critically observing one's own speech production, as the production takes place), the use of editing terms (e.g. *uh, sorry, no, I mean* ...), and how people actually carry out repairs. Studies like these informed Levelt's (1989) seminal model of speech production. We will only look at four examples from our dataset, but we hope to show that these kinds of examples warrant further consideration. Looking through the data, many corrections are due to mispronunciations, as in (28).

(28) Een hele grote prie ⟨corr⟩ pizza met drie jongens
A very large pri ⟨corr⟩ pizza with three boys

This particular mispronunciation is a so-called *anticipation error*, one of the most frequent kinds of speech errors (Fromkin, 1971). As the speaker is saying *pizza*, she is already preparing to say *three*, and accidentally inserts the *r* in the onset of *pizza*. Besides mispronunciations, there are also more complex cases. Figure 4.1 already provided an interesting example, repeated for convenience in (29):

(29) Een hele kudde schapen ⟨uh⟩ met een man ⟨corr⟩ met een herder erachter en een pakezel.
A whole herd of sheep ⟨uh⟩ with a man ⟨corr⟩ with a shepherd behind them and a mule.

What is interesting about this example is that the original expression *with a man* was already correct. The correction *man* → *shepherd* was made to be more specific, so as to produce a more informative description. A possible reason why the speaker did not immediately say 'shepherd' instead of 'man' is that the former is a *(social) role* (Masolo et al., 2004). We cannot determine that the man is a shepherd based on his visual appearance alone, but rather we label

him as a shepherd on the basis of the context of him interacting with a herd of sheep. After making this inference, the original label is replaced.

Figure 4.3 Eye-tracking data for example (30). Numbers indicate the following: 1. Start of experiment, 2. Speech onset, 3. Speaker realizes her mistake: the group hasn't ordered yet, 4. Start of corrected description, 5. End of description. (Original image by Malcolm Manners, CC BY 2.0)

The example in (30) shows a correction after making an incorrect prediction about the situation in Figure 4.3. Initially the speaker thinks the group is already eating, but actually they haven't ordered yet.

(30) Gezelschap die aan het eten is of <corr> die in een restaurant zit en iets willen gaan bestellen.
Group of people that is eating or <corr> that is sitting in a restaurant and is about to order.

What is interesting here is that we can actually see the correction reflected in the eye-tracking data. Figure 4.3 shows the attention map along with the scanning pattern corresponding to the eye movements. (Underline colors in the example correspond to the colors in the figure.) The participant starts by scanning the situation and looking at the people at the table. During this time, she starts speaking, but then she realizes her mistake upon seeing the menu on the table. She then updates her beliefs about the situation and corrects her utterance. This is a good example of *predictive coding* (see e.g. Clark, 2013).

Finally, (31) provides an example of a participant who rephrases her description when she realizes that her description is ambiguous; Dutch *knuffel* could both mean 'hug/cuddle' and 'cuddly toy' while *knuffeldier* only means 'cuddly animal.' (We ignore the first correction here, but note that it is similar to the shepherd example.)

(31) Een vrouw die een meisje <corr> klein meisje een knuffel geeft **<corr>** knuffeldier.
A woman giving a girl <corr> little girl a cuddle **<corr>** cuddly animal.

The remainder of this chapter discusses task dependence in eye-tracking data: do we find any differences between the *free viewing* and the *production viewing* data?

4.5 Task-dependence in eye tracking

A potential issue in studying visual attention is that eye-tracking data may differ across tasks. In one of the first ever eye-tracking studies, Buswell (1935) shows that we can observe differences in eye-tracking behavior between people who are freely looking at an image, versus when they

are asked to look for particular objects in the same image. Yarbus (1967) presents a study in which participants are asked to carry out seven different visual tasks, and shows that we can observe differences in eye-movement patterns between each of the different tasks. He argues that "eye movements reflect the human thought process." Finally, Coco and Keller (2014) show that it is possible to train a classifier to distinguish eye-tracking data for three different tasks: object naming, scene description, and visual search. This suggests that in order to model different tasks, one should also collect different sets of eye-tracking data.

Bottom-up versus top-down attention. The literature on visual attention modeling identifies two kinds of salience. On the one hand, there is bottom-up, task-independent visual salience, which is typically image-driven. On the other hand, there is top-down, task-dependent salience, where attention is driven by the task that people may have in viewing the image (Borji and Itti, 2013; Itti and Koch, 2000). Visual attention models are usually designed to predict general, task-free salience (Bylinskii et al., 2016). This prediction task is exactly what the SALICON dataset was developed for.

Free viewing versus description viewing. DIDEC was developed with this top-down versus bottom-up distinction in mind, so that we could compare different modes of viewing the images. The free-viewing task corresponds to bottom-up attention; because there are no explicit instructions of where to look at or what to do, participants only have the image to guide their attention. As such, they are drawn towards the most salient parts of the image. The description viewing task corresponds to top-down attention; because our participants are asked to describe the images, their attention is also guided by what they think might be the most conceptually important parts of the images.

Analysis. To what extent do people differ in their visual attention between the two tasks? We decided to test this by comparing the attention maps computed on the basis of the eye-tracking data for both tasks. For each image, for each participant, we used their fixations to generate an attention map. Then, for each image, we computed the within-task and between-task average pairwise similarities between the attention maps.[7] By looking at the difference between the within-task similarity and the between-task similarity, we can see if there is consistently more agreement within each task than between the tasks.

	Compared to attention maps from the other task, attention maps from the same task are…		
Task	More similar	Equally similar	Less similar
Description viewing	300	0	7
Free viewing	116	0	191

Table 4.2 Results for the comparison between Free viewing and description viewing.

Results. Table 4.2 shows the results. We find that, on average, attention maps from the image description task tend to be more similar to each other than to the attention maps from the free viewing task. But when we look at the attention maps from the free viewing task, we see that they are only more similar to each other 38% of the time (116 out of 307). In 62% of the cases, the between-task similarity is higher than the within-task similarity for the free viewing data. Figure 4.4 shows the distribution of the scores. We conclude that the image

[7] We use existing code to analyze this data: https://github.com/NUS-VIP/salicon-evaluation/
The pairwise similarity between attention maps (`CC_score`) is computed using the Pearson correlation.

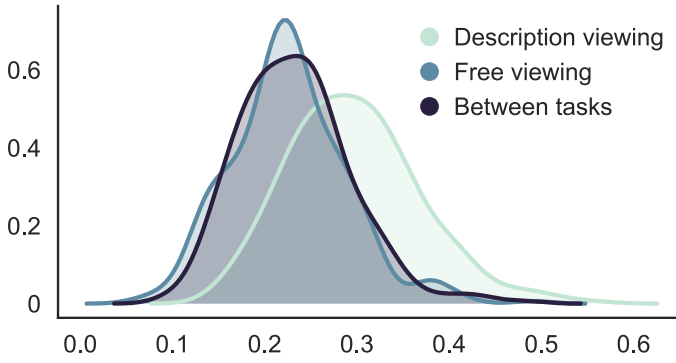


Figure 4.4 Distribution (Kernel Density Estimation) of the similarity scores within and between tasks.

description task reduces noise in the collection of eye-tracking data, and produces a more coherent set of attention maps.

4.6 Discussion and future research

We collected a corpus of Dutch image descriptions and eye-tracking data for 307 images, and provided an initial analysis of the self-corrections made by the participants. We have also presented two studies that show some uses of our data, but we believe many more analyses are possible. For reasons of space, we have not discussed the effect of modifying the modality of the image description task from written to spoken language, even though we know that modifying the prompt may have an effect on the response (e.g. Baltaretu and Castro Ferreira 2016). In a the next chapter, we compare spoken and written image descriptions in both Dutch and English. We still plan to semi-automatically annotate Speech Onset Times (SOT) using Praat (Boersma and Weenink, 2017), and to manually correct the output. We define SOT as the start of the utterance, including filled pauses (but excluding coughs and sighs). This is a measure of response time for each image, which is a proxy for the difficulty of producing a description, that could be correlated with e.g. image complexity (cf. Gatt et al., 2017).

Finally, the development of multilingual image description datasets (like Multi30K), has opened up new avenues of research, such as multimodal machine translation (Elliott et al., 2016, 2017). To the best of our knowledge, a dataset like DIDEC does not exist yet for any other language. We hope that our corpus may serve as an example, inspiring the development of parallel eye-tracking and image description datasets in other languages. This multilingual aspect is important because speakers of different languages may also display differences in familiarity with the contents of an image or, if their language uses a different writing directionality, different eye-tracking behavior (van Miltenburg et al., 2017; Baltaretu et al., 2016). We made all code and data used to build the corpus available on the corpus website, so as to encourage everyone to further study image description as a dynamic process.

4.7 Conclusion

This chapter presented image description as a dynamic process, using spoken descriptions to gain insight into the steps that are involved in formulating a description. We found evidence from self-corrections that people generate descriptions as they are interpreting the image (rather

than at the end of the interpretation process); whenever they make a wrong prediction about the contents of the image, they self-correct to make their descriptions congruent with the contents of the image. This provides further evidence that, in the image description task, people use world knowledge to reason about the contents of an image. Furthermore, from the shepherd example, it seems that people also self-correct towards more informative descriptions.

4.7.1 Implications for image description systems

What does this all mean for image description systems? Here we should distinguish two different goals that researchers in image description may have:

1. Building a cognitively plausible model of the human ability to describe images.
2. Building a useful tool to automatically generate image descriptions.

For researchers interested in the former, this chapter provides useful information regarding the timeline of the image description process, and how the two processes of image interpretation and image description overlap. However, most researchers are only interested in the latter. If you just want to have a black box that takes an image as input, and produces a description as output, then it does not matter *how* you get to a description, as long as the system works. What *does* matter is the content of the descriptions. And after three chapters (2-4) showing that people rely on world knowledge and past experience to produce more informative descriptions (e.g. noting that people in a restaurant *are about to order*), it seems clear that any system aiming to generate useful descriptions should also have some kind of knowledge component.

4.7.2 Next chapter

At this point, we have developed a deeper understanding of the canonical image description task, as it was introduced in Chapter 2. In that chapter, we have seen different pragmatic phenomena that occur in the two main image description datasets (Flickr30K and MS COCO). The subsequent Chapter 3 explored the influence of language on the resulting descriptions, and showed that the phenomena identified in Chapter 2 are not exclusive to English. The current chapter looked at the description process in more detail; how do people go about describing an image, in the canonical image description task? The next chapter looks at the different parameters of the task: how could we manipulate the task to get different kinds of descriptions? As an example, Chapter 5 manipulates the modality of the task, and tries to see whether we can find any differences between spoken and written descriptions.

Chapter 5

Task effects on image descriptions

5.1 Introduction

Chapter 2 described how the image descriptions in the Flickr30K and MS COCO datasets were collected through what I called *the canonical image description task*. Most current image description datasets have been collected using the same format. To some extent this is a good thing: using the same set-up (and the same images) means that the resulting descriptions can easily be compared. I could not have written Chapter 3 without having data from different languages, collected using the same set-up. But at the same time we should be aware that the canonical task is just one out of many possible formats that we could use to collect image descriptions. While the canonical format definitely served its purpose, yielding several useful corpora, we may ask ourselves: how does the format of the task affect the resulting descriptions?

5.1.1 Contents of this chapter

This chapter considers the ways in which the different configurations of the image description task may affect the resulting descriptions. We have already seen one example in Chapter 3, where I discussed how the language of the task may have an effect on the descriptions (through familiarity of the population with the contents of the images).

I will first discuss the canonical image description task, and the assumptions behind it (§5.2), after which I will look at the different variables that may influence the task (§5.3). Then we will turn to the main focus of this chapter: Sections 5.4–5.8 discuss the results of our preliminary study, looking for variables that differ between spoken and written descriptions. This study lays the foundations for future work, which should validate whether there is indeed a systematic difference between the two modalities.

5.1.2 Publications

This chapter is based on the following publications:

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*. Association for Computational Linguistics, Santiago de Compostela, Spain, pages 21–30

Emiel van Miltenburg, Ákos Kádar, Ruud Koolen, and Emiel Krahmer. 2018a. DIDEC: The Dutch Image Description and Eye-tracking Corpus. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*. Resource available at https://didec.uvt.nl

Emiel van Miltenburg, Ruud Koolen, and Emiel Krahmer. 2018b. Varying image description tasks: spoken versus written descriptions. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*

5.2 The image description task

Before discussing different factors affecting the outcome, let us first recapitulate the core properties of the canonical image description task, as used in the collection of the MS COCO and Flickr30K data (Lin et al., 2014; Young et al., 2014).

Materials. The materials for both Flickr30K and MS COCO are images that have been collected from Flickr.com, a social photo sharing platform. As shown in Chapter 3 of this thesis (Table 3.1), most other image description datasets have used the same images.

Participants. Participants are crowd-workers living in the relevant country or countries (as determined by the crowdsourcing platform using their IP-address). They are commonly asked to perform a short pre-test to determine whether they speak the target language.

Setting. Participants are presented with the guidelines for image description, along with some examples of 'good' and 'bad' descriptions. Following this, they are asked to provide image descriptions for a series of images, using a prompt that is similar to Figure 5.1. Participants are typically not told who or what the descriptions are for.

Figure 5.1 Prompt for the image description task, repeated from Chapter 2 of this thesis (Figure 2.5). Original picture taken by Luigi Cavasin (CC BY-NC-SA) on Flickr.com. Based on the example in (Rashtchian et al., 2010).

5.3 Factors influencing the image description task

Like any experiment or elicitation task, the image description task has many parameters that we could change, and that might affect the outcome (in other words: give us different kinds of descriptions). We can systematically analyze these parameters by looking at the different components of the speech situation. Biber (1988) usefully provides an overview of these components (reproduced in Figure 5.2). We will address each of these below.

1. **Participant roles and characteristics.** The canonical image description task only looks at the speaker/addressor, who is asked to formulate an image description. There is no physically present addressee (who the description is for) or audience (who may overhear the description), nor is it mentioned in the task who the descriptions are for. Furthermore, the authors of Flickr30K and MS COCO have not collected any demographic data from the participants of their crowdsourcing tasks. Thus, we know nothing about their personal or group characteristics (other than the fact that their IP-address is from the United States).

1. **Participant roles and characteristics**
1.1. Communicative roles of participants
  – Addressor(s)
  – Addressee(s)
  – Audience
1.2. Personal characteristics
  – Stable: personality, interests, beliefs, etc.
  – Temporary: mood, emotions, etc.
1.3. Group characteristics
  – Social class, ethnic group, gender, age, occupation, education, etc.

2. **Relations among participants**
2.1. Social role relations
  – relative social power, status, etc.
2.2. Personal relations
  – like, respect, etc.
2.3. Extent of shared knowledge
  – Cultural world knowledge
  – Specific personal knowledge
2.4. 'Plurality' of participants

3. **Setting**
3.1. Physical context
3.2. Temporal context
3.3. Superordinate activity type
3.4. Extent to which space and time are shared by participants

4. **Topic**

5. **Purpose**
5.1. Conventional goals
5.2. Personal goals

6. **Social evaluation**
6.1. Evaluation of the communicative event
  – Values shared by whole culture
  – Values held by sub-cultures or individuals
6.2. Speaker's attitudes toward content
  – Feelings, judgements, attitudinal 'stance'
  – Key: tone or manner of speech
  – Degree of commitment towards the content, epistemological 'stance'

7. **Relations of participants to the text**

8. **Channel**
8.1. Primary channel
  – Speech, writing, drums, signs, etc.
8.2. Number of sub-channels available

Figure 5.2 List of 'components of the speech situation', compiled by Douglas Biber. Based on Table 2.1 from Biber 1988, page 30. Biber notes that this taxonomy "draws heavily on Brown and Fraser (1979) and (Hymes, 1974, Chapter 2)."

2. **Relations among participants.** This category is not applicable, because the canonical image description task only looks at speakers, and ignores any other conversational agents.
3. **Setting.** The images for Flickr30K and MS COCO are described by crowd-workers from the comfort of their own computer or smartphone. Participants are not asked to imagine some other context, either.
4. **Topic.** The topic for each of the descriptions is the content of the relevant image, that speakers are asked to provide a description for.
5. **Purpose.** The canonical image description task does not provide any reason for the participants to provide their descriptions, so participants are left to infer on their own what the task is about. As a consequence, we might see variation in the descriptions arising from different interpretations of the task.
6. **Social evaluation.** Biber (1988) refers here to the standards that exist regarding different kinds of language use. In the canonical image description task, explicit standards can be found in the guidelines provided to the participants, with examples of 'good' and 'bad' descriptions. Furthermore, participants first have to take a pre-test to ensure that their spelling and grammar are up to the standards of the image description task. Not following these standards may result in their work being rejected by the authors of the task, which means that workers would not get paid, and would see their worker rating decrease on the Mechanical Turk platform. Beyond the

standards set by the authors of the task, there are also implicit standards: how people believe they are supposed to write.

7. **Relations of participants to the text.** Biber (1988) cites Chafe (1982) as one of the few researchers looking at the affordances of different kinds of messages. Writers can compose a text as quickly and as carefully as they want, while speakers have to produce their texts online, in real time. The same constraints hold for readers and hearers: readers can take as much time as they like, but hearers have to process speech in real time. For the canonical image description task, the participants writing the descriptions are allowed to take as much time as they want.

8. **Channel.** For the canonical image description task, descriptions have to be written, rather than spoken by the participants. Biber notes that writing only offers the addressor one sub-channel: the lexical/syntactic channel. That is: writers can only express themselves through careful combinations of words. By contrast, speakers can *also* convey meaning through prosody (stress and intonation) and paralinguistic means (e.g. gestures). These might help them better convey what an image is about.

This characterization of the canonical image description task raises questions about the influence of each of the different components on the resulting descriptions. In the remainder of this chapter, we will explore the impact of the primary channel: can we observe a difference between spoken and written descriptions?

5.4 Investigating the difference between spoken and written descriptions

One of the motivations behind automatic image description research is to support blind or visually impaired people (e.g. Gella and Mitchell, 2016), and indeed *apps* are starting to appear which describe visual content for blind users (e.g. *TapTapSee* or Microsoft's *Seeing AI*[1]). These apps are commonly used together with screen readers, which convert on-screen text to speech. Given this presentation through speech, it is worth asking: should we not also collect *spoken* rather than *written* training data? That might give us more natural-sounding descriptions. But a big downside of collecting spoken training data is that it also requires a costly transcription procedure (unless we go for an *end-to-end* approach, see Chrupała et al., 2017). An alternative is to try to understand the differences between written and spoken image descriptions. Once we know those differences, and we know what kind of descriptions users prefer, we may be able to direct image description systems to produce more human-like descriptions, similar to the way we can modify the style of the descriptions, for example with positive/negative sentiment (Mathews et al., 2016), or humorous descriptions (Gan et al., 2017).

This chapter presents an exploratory study of the differences between spoken and written image descriptions, for two languages: English and Dutch. We provide an overview of the variables that have been found to differ between spoken and written language, and see whether these differences also hold between English spoken and written image descriptions. Following this, we repeat the same experiment for Dutch. Our main findings are that spoken descriptions (1) tend to be longer than written descriptions, (2) contain more adverbs than written descriptions, (3) contain more pseudo-quantifiers and allness terms (DeVito, 1966), and (4) tend to reflect the certainty of the speaker's beliefs more-so than written descriptions. Our work paves the way for a future controlled replication study, and follow-up studies to assess what kind of descriptions users prefer. All of our code and data is available online.[2]

[1] TapTapSee: https://taptapseeapp.com/; Seeing AI: https://www.microsoft.com/en-us/seeing-ai/

[2] Our code and data is available at https://github.com/cltl/Spoken-versus-Written

5.5 Technical background: Manipulating the image description task

Recently, researchers have started to manipulate the image description task to obtain a better understanding of how this influences the resulting descriptions. This section presents a brief list of variables that have been considered in the literature.

Language. The most common modification is the *language* in which the task is carried out (e.g. Elliott et al., 2016; Li et al., 2016; Miyazaki and Shimizu, 2016). This is typically done to be able to train an image description system in a different language, but Van Miltenburg et al. (2017) use this manipulation to show that speakers of different languages may also provide different descriptions. For example, speakers of American English described sports fans barbecuing on a parking lot as *tailgating*, a concept unknown to Dutch and German speakers.

Style. Another possible manipulation is the requested *style* of the descriptions. Gan et al. (2017) asked crowd workers to provide 'humorous' and 'romantic' descriptions, but found that it is impossible to control the quality of the resulting descriptions. So they, like Mathews et al. (2016), further changed the description task to a *description editing* task.

Content. Gella and Mitchell (2016) emphasize the importance of *emotional or descriptive content* and *humor* in the image, and explicitly ask for these to be annotated. This makes the elicited descriptions useful for training an assistive image description system which can provide descriptions for blind people.

Task demands. Baltaretu and Castro Ferreira (2016) present variations on an *object description* task (the *ReferIt* task, by Kazemzadeh et al. (2014)). The authors show that asking participants to work very fast, or produce thorough or creative descriptions, results in very different kinds of descriptions.[3]

While the studies listed above cover a wide range of variables, there are many more possibilities that are still unexplored. Van Miltenburg et al. (2017) provide a (non-exhaustive) list of other factors that may influence the image description process. This chapter aims to identify the role of the channel through which descriptions are communicated.

5.6 Theoretical background: Spoken versus written language

The differences between spoken and written language have been thoroughly studied in the linguistics literature since the 1960s. Extensive overviews are provided by Akinnaso (1982), Chafe and Danielewicz (1987), Chafe and Tannen (1987), Biber (1988), and Miller and Fernandez-Vest (2006). Why should we study differences between spoken and written image descriptions, when so many linguists before us have studied differences between spoken and written language? Because spoken and written language are not monoliths. Biber (1988) notes that there is often as much variation within each modality, as there is between the two modalities. Biber attributes this variation to situational, functional, and processing considerations (p. 24-25). So while there may be general tendencies for particular linguistic phenomena to occur more in written than in spoken language (or vice versa), the only way to know for sure how

[3] Relatedly, Schwartz et al. (2017) show for a different task (the ROC story cloze task, Mostafazadeh et al. 2016), that variations in the elicitation task (writing either a coherent or an incoherent ending to a story) may cause participants to provide qualitatively different text responses. They note that this creates a confound in NLP evaluation tasks, where the 'right' and 'wrong' answers are elicited through different writing tasks. Indeed, Poliak et al. (2018) have shown that 'hypothesis-only' baselines (without access to the context) perform above chance on many different Natural Language Inference datasets. These results provide an additional argument that researchers in Natural Language Processing should take elicitation tasks more seriously and treat them as experiments like those in Linguistics and Psychology.

speech differs from writing in a particular domain is to investigate that particular domain. For image description, the seminal study by Drieman 1962a; 1962b is of particular interest to us. Drieman asked eight participants to describe two realistic paintings (one by Renoir and one by Weissenbruch), providing either spoken or written descriptions. He found that written texts (1) are shorter; but (2) have have longer words (fewer words of one syllable, more words of more than one syllable); (3) have more attributive adjectives[4]; and (4) a more varied vocabulary. The drawback of this study is its limited size. Moreover, it is unclear if Drieman's conclusions extend to one-sentence image descriptions like those in MS COCO and Flickr30K. This is what we intend to study.

Following Drieman's study, researchers have proposed many other variables that seem to correlate with the speech/writing distinction. After surveying the literature on spoken versus written language, Biber (1988) presents an extensive list of linguistic features. The features used in this Chapter are based on Biber's list, see Section 5.7.3 for an overview. Noted in almost all surveys is the *ephemeral nature* of speech; whereas writing samples can be edited and reworded, speech cannot be edited the same way. Hence, spoken language also contains false starts, speech errors, and subsequent repairs. But despite those flaws, we must not think of spoken language as somehow *inferior* to written language. Halliday (1989) notes that the two are simply different media that serve different functions, which may require different forms of language. It is our task, as language users, to pick the right form (and medium) for the right job. If we find significant differences between spoken and written language, we should ask ourselves: now that we know about these differences in the way people describe images, which form is the most suitable for an image description system?

5.7 Data and methods for analyzing image descriptions

We present an analysis for both Dutch and English image descriptions. For each language, we took existing sets of spoken and written image descriptions, and automatically computed their differences in terms of the literature discussed above. The rationale here is that, even if these corpora are not perfectly comparable, they do provide an indication of the extent to which spoken and written image descriptions may differ. If we find structural differences between spoken and written image descriptions, it may be worth it to explore these differences further in a more controlled environment. If we fail to find any differences, we should conclude that there is no evidence for the effect of modality on the image description task. But, as we will see later, there do seem to be structural differences between spoken and written descriptions in both Dutch and English.

5.7.1 English data

For the written sample, we use the Flickr30K and the MS COCO datasets. Both were collected through Mechanical Turk, and have 5 written descriptions per image. We only use the training splits from both datasets, so that we remain ignorant of the properties of the validation and test splits. Figure 5.3 provides the instructions for both datasets. One of the main differences between the two is that the MS COCO instructions explicitly forbid the use of *there is* at the start of a sentence, which leads to the use of different syntactic constructions. Otherwise the instructions are very similar.

[4] In English, this means that the adjective is used in the prenominal position (*the **good** book*) rather than postnominal (*the book is **good***). The same holds for Dutch.

MS COCO instructions

1. Describe all the important parts of the scene.
2. Do not start the sentences with "There is."
3. Do not describe unimportant details.
4. Do not describe things that might have happened in the future or past.
5. Do not describe what a person might say.
6. Do not give people proper names.
7. The sentences should contain at least 8 words.

Flickr30K instructions

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.

Figure 5.3 Instructions for the written English data. MS COCO instructions are from Chen et al. (2015). Flickr30K instructions are from the appendix of Hodosh et al. (2013), edited for brevity.

For the spoken sample, we use the Places Audio Caption Corpus, Part 1 (Harwath et al., 2016; Harwath and Glass, 2017), which contains about 230,000 spoken descriptions for a selection of images that were equally sampled from the 205 scene categories in the Places205 dataset Zhou et al. (2014). The spoken descriptions were collected through Mechanical Turk using the Spoke framework (Saylor, 2015). These were then automatically transcribed by Harwath et al. (2016) using the Google Speech API. Because the transcriptions were not manually corrected, they have a word error rate of about 20%. It is unclear how participants were instructed to describe the image. The authors only mention that the descriptions are free-form, and that they should describe the salient objects in the scene.[5]

Image selection. The images from Flickr30K, MS COCO, and Places205 were all collected from online sources. Flickr30K and MS COCO exclusively use images from Flickr,[6] while Places also contains images found through general image search engines (Google and Bing). The main difference between the datasets is in the kind of images that are included. For the Flickr30K dataset, the authors downloaded images from six different user groups on the Flickr website.[7] The MS COCO authors compiled a list of 91 object categories, and searched for different object+object combinations of different categories on Flickr. They also selected 60 scene categories from the SUN database (Xiao et al., 2010), and searched for different object+scene combinations to diversify their data. Finally, the Places205 dataset is built by querying different search engines for adjective+scene combinations. The 205 scenes come

[5] We contacted the authors for more information about the crowd-sourcing task, but have not received any response.

[6] A social image sharing platform, see: www.flickr.com.

[7] These user groups are: *strangers!; Wild-Child (Kids in Action); Dogs in Action (Read the Rules); Outdoor Activities; Action Photography; Flickr-Social (two or more people in the photo).* See Hodosh et al. 2013 for the full methodology.

from the SUN database, and the adjectives come from a manually curated list. Examples are: *messy, spare, sunny*.

Comparability. To what extent can we compare the descriptions in these datasets? Ideally, we would have one set of images that is provided with both spoken and written descriptions. But if the tasks are similar enough, and we compare the image descriptions on a large scale, we may still be able to confirm general tendencies of spoken versus written data, e.g. that spoken descriptions tend to be longer than written ones (Drieman, 1962a), or use more self-reference terms (DeVito, 1966). What we *cannot* do, is compare how often particular properties or kinds of entities are mentioned, because the distribution of those properties or entities might be dramatically different. Generally speaking, using different sets of images also means that we can never exclude the possibility that the underlying cause of the differences between the descriptions lies with the images rather than the modality. However, as the sets of images become more similar, chances of the images being a major source of the differences between the written and spoken descriptions become smaller. So how big *are* the differences between existing datasets?

	Flickr30K		MS COCO		Places	
#	Word	Count	Word	Count	Word	Count
1	man	42595	man	48847	picture	36020
2	woman	22197	people	25723	people	26094
3	people	17338	woman	22992	building	25735
4	shirt	14341	table	21104	trees	22449
5	girl	9656	street	20527	water	20324
6	men	9499	person	16857	man	18609
7	boy	9399	top	14755	front	16584
8	dog	9093	field	14597	background	15484
9	street	8012	group	14450	side	15254
10	group	7852	tennis	13411	room	12985

Table 5.1 Top-10 most frequent nouns for all three datasets. Flickr30K and MS COCO are fairly similar (they have a larger overlap), but Places differs from the other two.

To answer this question, we tagged the descriptions in all three datasets using the SpaCy part-of-speech tagger.[8] Table 5.1 shows the top-10 most frequent nouns in all three datasets. These correspond to the most frequent entities. We observe that while the Flickr30K and MS COCO datasets are fairly similar (sharing 5 words in their top-10), the Places dataset stands out from the other two (sharing only 2 words). So the only spoken English descriptions that are available, describe images that are fairly different from the other datasets. Luckily we have more comparable data for Dutch.

5.7.2 Dutch Data

For the written sample, we use the data collected by Van Miltenburg et al. (2017). The authors crowdsourced Dutch descriptions for the Flickr30K validation and test sets (1014 + 1000 images, with 5 descriptions per image). The annotation task was translated from the Flickr30K and Multi30K templates (Elliott et al., 2016), to stay as close to these datasets as possible. We

[8] We use version 2.0.4. See: http://spacy.io/

only use the validation split for our comparison, so that we remain ignorant of the properties of the test set.

For our spoken sample, we use data from our Dutch Image Description and Eye-tracking Corpus (DIDEC, van Miltenburg et al. 2018a; also discussed in Chapter 4 of this thesis). 45 Dutch students participated in a lab experiment where they were asked to describe a series of images, while we also measured their eye movements. We used the 307 images from MS COCO that both appear in SALICON and the Visual Genome dataset (Jiang et al., 2015; Krishna et al., 2017). We transcribed and annotated the recorded descriptions, so that we ended up with three layers: (1) a raw layer; (2) an annotated layer indicating (filled) pauses, corrections and repetitions; and (3) a normalized layer, with the 'intended' description. In total, we collected 14-16 descriptions per image, resulting in a grand total of 4604 descriptions for the entire dataset. This study uses the normalized descriptions, so that our metrics are unaffected by corrections and repetitions.

5.7.3 Preprocessing, metrics, and hypotheses

We tokenize, tag, and parse the descriptions using SpaCy. Then, we compute the following metrics:

1. **Average token length** Drieman (1962a) and others have found that the tokens in spoken language are shorter than those in written language. We measure token length in terms of syllables (following e.g. Drieman 1962a) and characters (following e.g. Biber 1988), using Hunspell to obtain the syllables.[9]
2. **Average description length** Drieman (1962a) and others have shown that spoken language has a higher sentence length than written language. We measure description length in tokens and syllables.
3. **Mean-segmental type-token ratio** (MSTTR) corresponds to the average number of types per 1000 tokens (Johnson, 1944). It is used as a measure of lexical variation. Because it is computed for a fixed number of tokens, it is unaffected by corpus size or sentence length. Drieman (1962a) shows that written language is more diverse than spoken language. One issue is that the Places Audio Caption Corpus has only one description per image, versus five descriptions per image for MS COCO and Flickr30K. This means that for every description in Flickr30K or MS COCO, there are four very similar descriptions, which makes these corpora less diverse overall. For a fair comparison, we treat Flickr30K and MS COCO as collections of five similar corpora, compute MSTTR for each of these, and report the average.
4. **Attributive adjectives** Drieman shows that spoken language contains fewer attributive adjectives than written language. We use SpaCy's tagger and parser to determine if an adjective is attributive or not. We consider a token to be an attributive adjective if its part-of-speech tag is ADJ, and it has an *amod* dependency relation with a head that is either tagged as NOUN or PROPN. In other words: if it's an adjective modifying a noun.
5. **Adverbs** We count all tokens with the ADV part-of-speech tag. The literature shows mixed results for the use of adverbs: Harrell (1957) studied children's production of stories, and found *fewer* adverbs in spoken than in written language, while Chafe and Danielewicz (1987) show that adverbs are used *more* in conversation and letters, and less in lectures and academic writing. They explain this pattern by arguing that the key variable is not *modality* but *involvement*. Whenever people are more involved with their audience or their environment, they also tend

[9] Hunspell is the spell checker from LibreOffice, which has a powerful hyphenation function. See: https://hunspell.github.io for more details. We use the Pyphen library (https:/github.com/Kozea/Pyphen) as an interface.

to use more locative or temporal adverbials. And whenever they are more *detached* (talking about more abstract ideas), they tend to use fewer adverbs.

6. **Prepositions** Chafe and Danielewicz (1987) show that prepositions are used more in (academic) written language. We count all tokens with the ADP part-of-speech tag.

The metrics below are computed by matching the tokenized descriptions with different sets of words.

7. **Consciousness-of-projection terms** DeVito (1966) defines these as: "words which indicate that the observed is in part a function of the observer." He shows that these words are more frequently used in speech than in writing. Since DeVito does not provide a list of the terms used in his work, we compiled our own list containing the following words: *apparently, appear, appears, certainly, clearly, definitely, likely, may, maybe, might, obviously, perhaps, possibly, presumably, probably, seem, seemed, seemingly, seems, surely*. The consciousness-of-projection terms contain Biber's 1988 set of *possibility modals* and *seem and appear*.

8. **Self-reference terms** DeVito (1966) also shows that self-reference terms (first-person pronouns and phrases like *the author*) are used more in spoken than in written language. We only use *I, me, my* as self-reference terms, since phrases like *the author* are not relevant in this domain.

9. **Positive allness terms** DeVito (1966) shows that spoken language contains more 'allness terms' than written language. For DeVito, these include both positive (*all, every, always*) and negative (*none, never*) terms. Following more recent work, which also focuses explicitly on negations (Biber et al., 1999), we decided to distinguish between the two. As *positive* allness terms, we use the words *all, each* and *every*.

10. **Negations** (Biber et al., 1999, Chapter 3) show that spoken language contains more negations than written language. For the *negative* allness terms, we focus on explicit, non-affixal negations: *n't, neither, never, no, nobody, none, nor, not, nothing, nowhere*. (Using the terminology from Tottie (1980).)

11. **Pseudo-quantifiers** While DeVito (1966) did not find any significant differences in the use of exact numerals, between spoken and written language, he did find such differences in the usage of terms like *many*, that are "loosely indicative of amount or size." We use the following terms: *few, lots, many, much, plenty, some* and *a lot*.

Feature	Terms
Consciousness-of-projection	*Lijkt, lijken, waarschijnlijk, misschien, duidelijk, mogelijk, zeker*
Self-reference	*Ik, me, mij*
Positive allness	*Alle, elke, iedere, iedereen*
Negations	*Geen, niet, niemand, nergens, noch, nooit, niets*
Pseudo-quantifiers	*Veel, vele, weinig, enkele, een paar, een hoop, grote hoeveelheid, kleine hoeveelheid*

Table 5.2 Dutch terms that were used for each feature.

Table 5.2 shows the Dutch terms used for each feature. For all features except average token length, average description length, and MSTTR, we report the average number of occurrences per description, and per 1000 tokens. We also compute the Propositional Idea Density (PID) for the spoken and written descriptions. PID corresponds to the average number of propositional ideas per word in a text (Turner and Greene, 1977). According Turner and Greene's annotation

scheme, sentence (32a) breaks down into the five ideas expressed in (32b).[10] Because the nine words in (32a) express five ideas, the PID for this sentence is $5/9 = 0.56$.

(32) a. The old gray mare has a very large nose
b. HAS(MARE, NOSE), OLD(MARE), GRAY(MARE), LARGE(NOSE), VERY(LARGE)(NOSE)

We expect that written language has a higher PID than spoken language. In other words: that spoken language uses more words to convey the same amount of information. This hypothesis is based on the idea that written language is edited or condensed to convey as much information as possible. For example, Chafe and Danielewicz (1987) show that nominalizations (e.g. *categorization, development*) occur more often in written language. They argue that the spoken alternatives for nominalizations are often much longer: several clauses instead of one. Another example comes from Ravid and Berman (2006), who show that written narratives contains relatively more *propositional* content ("events, descriptions, and interpretations") and less *ancillary* content ("nonnovel, nonreferential, or nonnarrative"). Spoken narratives are said contain more ancillary content for communicative purposes. We use existing tools to measure idea density. For English, we use the Computerized Propositional Idea Density Rater Brown et al. (2008).[11] For Dutch, we use the tool developed by Marckx (2017).

Because this is an exploratory study, we will only report descriptive statistics. These allow us to formulate hypotheses about the differences between spoken versus written image descriptions. We can test these hypotheses in a future study with spoken and written descriptions for the same images, collected in the same controlled setting.

5.8 Results

This section presents an overview of the different metrics for the Dutch and the English data. We first present the English results, followed by the Dutch results, and end with a summary of our main findings.

5.8.1 English results

Name	Descriptions	Tokens	Types	MSTTR
MS COCO	414,113	4,348,698	23,450	0.32
Flickr30K	145,000	1,787,693	17,784	0.38
Places	229,388	4,765,891	31,800	0.34

Table 5.3 General metrics for the three datasets: number of descriptions, tokens, and types, along with the mean-segmental type-token ratio.

Tables 5.3 and 5.4 show the results for the English descriptions. We immediately see that, in line with the literature, spoken image descriptions are almost twice as long as their written counterparts. With almost half the number of descriptions of MS COCO, the Places dataset has significantly more tokens. Based on the literature, we might also expect spoken descriptions to use shorter words than written descriptions. This is indeed the case when we

[10] This example was taken from (Brown et al., 2008).

[11] We use CPIDR version 3.2.3738.41169 on OS X 10.13.2, using Wine version 1.8-rc4.

	TokLen		DescLen		Attributives		Adverbs		Prepositions	
Name	Syll	Char	Syll	Tok	Desc	‰	Desc	‰	Desc	‰
MS COCO	1.29	4.01	13.53	10.50	0.64	60.97	0.16	14.99	1.75	166.37
Flickr30K	1.30	4.11	16.05	12.33	0.97	78.81	0.15	12.20	1.91	154.78
Places	1.26	4.08	26.27	20.78	1.39	66.77	0.97	46.67	3.06	147.14

	Consciousness		Self-reference		Allness		Negations		PseudoQuant	
Name	Desc	‰	Desc	‰	Desc	‰	Desc	‰	Desc	‰
MS COCO	0.00	0.22	0.00	0.09	0.02	1.56	0.00	0.42	0.06	6.01
Flickr30K	0.01	0.63	0.00	0.09	0.02	1.30	0.00	0.35	0.04	2.88
Places	0.08	4.07	0.05	2.49	0.07	3.22	0.06	2.88	0.24	11.66

Table 5.4 Results for our analysis of MS COCO, Flickr30K (both written), and the Places Audio Caption Corpus (spoken). For the top table, columns correspond to: average token length (in syllables and in characters), average description length (in syllables and in tokens), features 4-6 (per description and per 1000 tokens). The bottom table shows features 7-11 (per description and per 1000 tokens).

look at syllable length, but when we look at the number of characters, tokens in the MS COCO dataset have a shorter average length. We conclude there is no clear difference in token length between spoken and written image descriptions.

MSTTR. We next look at the richness of the vocabulary used by the crowd workers. Following Drieman's work, we expected that written descriptions would have a higher type-token ratio than spoken descriptions. This expectation is not borne out by the data. The MSTTR score for the Places data falls between the scores for MS COCO and Flickr30K. A possible explanation for this result is that spoken language is typically produced without any preparation, which leads speakers to 'fall back' on a more basic vocabulary. But with the Places dataset, participants could think of a description before they pressed the 'record' button, alleviating cognitive constraints on language production.

Adjectives and prepositions. For the remaining features, we report the average number of occurrences per description, as well as per 1000 tokens. Based on Drieman's work, we thought that attributive adjectives might occur more in written descriptions, but when we look at Table 5.4, we find a mixed result: spoken descriptions contain *more* attributive adjectives per description, but *fewer* attributive adjectives per 1000 tokens than the written descriptions in the Flickr30K dataset. This is possible because the spoken descriptions are longer than the written ones. We conclude that there is no clear difference between written and spoken descriptions in the use of attributive adjectives. We draw the same conclusion for the use of prepositions.

Adverbs and other features. We observe that spoken descriptions contain more adverbs than written ones; three times more adverb tokens than MS COCO, and almost four times more than Flickr30K. The same holds for consciousness-of-projection terms, self-reference terms, positive allness terms, negations, and pseudo-quantifiers: all these kinds of terms are used more often in spoken than in written image descriptions.

Propositional idea density. Figure 5.4 shows the distribution of Propositional Idea Density scores for each of the three datasets, visualized using Kernel Density Estimation. We computed the PID scores over 100 samples of 1000 descriptions for each dataset. We observe that the

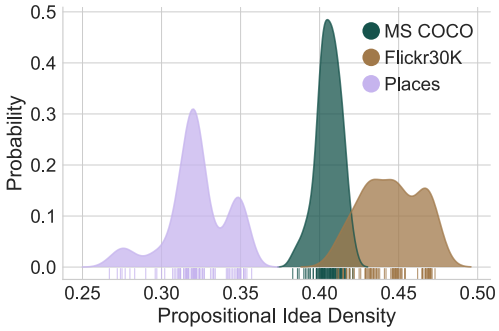


Figure 5.4 Distribution of the Propositional Idea Density scores for each of the three English datasets, computed over $3 \cdot 100$ sets of 1000 descriptions. The lines on the x-axis show individual scores.

spoken descriptions have a lower PID than both written datasets, confirming the hypothesis that spoken descriptions use more words to convey the same amount of propositional information. Of course, the extra-propositional information may be useful as well, e.g. to convey pragmatic messages. Future research should look into whether users prefer the spoken or the written variant.

5.8.2 Dutch results

Tables 5.6 and 5.5 show the results for the Dutch descriptions. As with the English descriptions, we observe that the spoken descriptions are longer than their written counterparts, albeit to a lesser extent. Whereas the English spoken descriptions were almost twice as long as the written descriptions, the Dutch spoken descriptions are only two tokens longer on average.

Name	Descriptions	Tokens	Types	MSTTR
Written	5,070	52,548	5,141	0.39
Spoken	4,604	57,805	4,179	0.37

Table 5.5 General statistics for the Dutch corpora.

	TokLen		DescLen		Attributives		Adverbs		Prepositions	
Name	Syll	Char	Syll	Tok	Desc	‰	Desc	‰	Desc	‰
Written	1.47	4.6	15.22	10.36	0.52	50.37	0.22	21.56	1.91	184.03
Spoken	1.49	4.58	18.7	12.56	0.5	39.51	0.67	52.76	1.83	144.69

	Consciousness		Self-reference		Allness		Negations		PseudoQuant	
Name	Desc	‰	Desc	‰	Desc	‰	Desc	‰	Desc	‰
Written	0.01	0.84	0.00	0.04	0	0.04	0.00	0.21	0.02	1.69
Spoken	0.03	2.22	0.02	1.53	0	0.33	0.01	0.79	0.06	4.78

Table 5.6 Results for our analysis of the Dutch spoken and written descriptions.

Token length and MSTTR. We do not find any major differences in terms of token length or mean-segmental type-token ratios. The spoken descriptions *are* slightly less diverse, but not by a large margin. Unlike the English spoken data, the participants for the Dutch spoken data did not have any time to prepare, since the experiment immediately started recording as the picture was presented. We hypothesize that the differences that Drieman found might have been due to the length of the spoken and written samples, and that with a description spanning multiple sentences, speakers are perhaps more likely to repeat themselves, leading to less diversity in their descriptions.[12]

Adjectives and prepositions. In contrast to the English descriptions, we do observe a difference in the use of attributive adjectives between spoken and written descriptions. Written description contain slightly more attributive adjectives per description (even though written descriptions are shorter on average), and significantly more attributive adjectives per 1000 tokens. We also find that written descriptions contain more prepositions than spoken descriptions. These findings are in line with Drieman's original results.

Adverbs and other features. We find that spoken descriptions contain more than twice as many adverbs than written descriptions, mirroring the results for English. And, just like in English, we find that spoken descriptions also contain more negations, pseudo-quantifiers, and consciousness-of-projection, self-reference, and allness terms.

Propositional idea density. We also computed the Propositional Idea Density for both written and spoken descriptions, but we found little difference between the two: 0.44 for written descriptions versus 0.46 for their spoken counterparts. This is a far cry from the highly contrastive results we found for English. We conclude that there is no clear difference for Dutch spoken and written descriptions, though we should note that Marckx (2017) translated the rules to compute propositional idea density from English to Dutch. It may be the case that the Dutch PID rater overlooked linguistic constructions for communicating propositional ideas that only exist in Dutch.

5.8.3 Summary of our findings

Looking at the results for both Dutch and English, we have found that: (1) Spoken descriptions are likely to be longer than written descriptions and, in English, seem to have a lower propositional information density than written descriptions. (2) Spoken descriptions contain more adverbs than written descriptions. (3) Spoken descriptions contain more pseudo-quantifiers and allness terms. (4) Speakers have a bigger tendency to "show themselves" in their descriptions than writers, who are less *involved* (in the sense of Chafe and Danielewicz 1987). We can see this in the use of more consciousness-of-projection and self-reference terms. Akinnaso (1982) calls this *egocentric language*, indicating "that the observed is in part a function of the observer" (p. 102). It has been shown that negations in image descriptions often reflect the author's expectations about the image they are describing van Miltenburg et al. (2016a).

Some of the 'negative' findings (where, unlike earlier work, we find no difference between spoken and written language) may be explained in functionalist terms. E.g. token length may not be a function of spoken versus written language, but rather of *topic* or *register*; abstract or formal language tends to use longer words than concrete or informal language. Another explanation comes from the fact that Drieman (1962a) used *paintings* as a stimuli, which also

[12] We did use normalized rather than raw spoken descriptions in our analysis, but the entire corpus of spoken Dutch descriptions contains only 139 repetitions/false starts, which is unlikely to have a strong effect over 57K+ tokens.

come with a particular vocabulary, whereas MS COCO, Flickr30K, and the Places Audio Caption corpus use real-life photographs, which do not elicit the same kind of expert language.

5.9 Future research

We performed an exploratory study to find differences between spoken and written image descriptions in both Dutch and English. We found four main differences, summarized in the previous section. Where should we go from here? We offer two directions to consider.

5.9.1 Controlled replication.

As Akinnaso (1982) notes, Drieman's study carefully controlled for (1) the topic of the descriptions; (2) the circumstances in which participants were asked to provide the descriptions; and (3) participants' background and level of linguistic knowledge. Changing any of these factors between the written and spoken condition makes the resulting data less comparable. Because we used existing datasets, we were not able to control for these. Although we believe that our main findings *should* hold up, the only way to know for sure is to carry out a follow-up study. The benefit of this exploratory study is that we have compiled a freely available set of tools to analyze spoken versus written language, and we have narrowed down the potential differences between spoken and written descriptions to four main differences. We can now also begin to study how potential users feel about these differences.

5.9.2 What do users want?

Having found differences between spoken and written language, we should now ask ourselves: what kind of descriptions would users of image description technology prefer? Research on this topic goes back to user studies of ALT-text on the internet. For example, Petrie et al. (2005) asked a group of blind people about the type of content they would like to be described. They found that there is no single answer to this question, because descriptions are context dependent. But generally speaking, blind users like to know about objects, buildings, and people; activities; the use of color; the purpose of the image; the emotion and atmosphere; and the location where the picture was taken. Gella and Mitchell (2016) asked a panel of visually impaired users about automatic image captioning, and also found that users want to hear about humor and emotional content (besides concrete, literal content). While these studies are important for our understanding of the needs of blind users, they only focus on *what* should be described, and not so much on *how* images should be described, which is still an open question. Possibly the most interesting feature to explore in the context of this chapter is the use of subjective language. There is already some evidence that visually impaired users of image description technology appreciate expressions of (un)certainty (Zhao et al., 2017b). Furthermore, the datasets discussed in this chapter all use pictures from Flickr, or unspecified images from the web. But Gella and Mitchell found that blind users would also like to have image description technology for personal, news, and social media images. It is unclear how these should be described, and whether these kinds of images would elicit similar differences between spoken and written descriptions.

5.10 Conclusion

This chapter discussed the effect of the image description task format on the elicited descriptions. Prior to this chapter, we have already seen the influence of language/culture on the elicited descriptions (Chapter 3), and this chapter discussed the effect of modality *within* a particular language (either English or Dutch). As discussed in the summary (Section 5.8.3), there seem to be systematic differences between spoken and written image descriptions, but these differences still need to be confirmed in a controlled follow-up study.

5.10.1 Implications for image description systems

Since this is the last chapter of Part I of this thesis, we will not just reflect on the implications of this chapter alone, but also take stock of the implications of Chapters 2–5 taken together. So far we have discussed the human image description process from different angles: *what* do image descriptions look like? (Chapters 2 and 3), *how* do they come about? (Chapter 4), and *why* do they look the way they do? (this chapter).

Chapter 2 looked at general linguistic properties of image descriptions in the Flickr30K and MS COCO datasets. We have seen that these descriptions are very diverse, with different crowd-workers focusing on different aspects of the images they were asked to describe. Moreover, the descriptions also show how crowd-workers use their world knowledge to interpret and contextualize the images.

Chapter 3 showed that the different pragmatic phenomena observed in Chapter 2 can also be observed for Dutch and German image descriptions that were collected through a very similar image description task. At the same time, we found that differences in background knowledge may lead workers to provide different kinds of descriptions.

Chapter 4 looked at image description from a real-time perspective. We collected the *spoken* image descriptions that also form the basis for this chapter. Looking at those descriptions, we found that people seem to interpret images as they are describing them. During this process, speakers actively predict what the image is likely to be about, and correct themselves if those predictions turn out to be wrong.

With these three chapters, we have characterized human image description as a dynamic process in which people use their background knowledge to interpret and contextualize an image. We have also seen, in three different languages, that different speakers may provide different descriptions for the same image. These differences may partly be explained by differences in background knowledge, but there are likely more factors to be involved. Either way, the fact that there are so many ways to describe an individual image raises the question: what is the right way to describe an image? Can there even be such a thing as *the* right description? What is 'the right description' differs from situation to situation. That is where the current chapter comes in.

This chapter presented an exploration of the different factors that may influence the kinds of descriptions that people will provide for a given image. We have started with an overview of these, but have focused on modality because of its implications for assistive image description systems that generate spoken descriptions through text-to-speech. If human-spoken image descriptions are systematically different from human-written ones, then we may also want to investigate whether users find the spoken variety more natural. Given the findings in this chapter, we believe that it may be interesting to investigate whether people prefer more subjective descriptions, that reflect the speaker's (uncertain) interpretation of the image (e.g. 'It looks like a dog' or 'It is probably a dog' versus the more objective 'It's a dog'). More

generally, we believe that the factors identified by Biber (1988) (discussed in Section 5.3 of this chapter) may be a useful guide in exploring the context-dependence of image description. As noted in Footnote 3, researchers in NLP should be careful to control for these variables in their elicitation tasks, so as to obtain more reliable datasets.

5.10.2 Next part

The next part of this thesis looks at automatic image description systems. We will first provide a general introduction of these systems, and how they work. Following this, we will look at their performance and how they currently compare to humans performing the same task. We will show that current systems are still lagging behind: they are still prone to making errors that no human would make (Chapter 6), and the generated descriptions show a lack of diversity (Chapter 7).

Part II

Machines and images

Chapter 6

Automatic image description: a first impression

6.1 Introduction

The field of automatic image description lies at the intersection of Natural Language Processing and Computer Vision (Bernardi et al., 2016). It is closely related to the field of Natural Language Generation (NLG, see Gatt and Krahmer 2018 for an overview). Researchers in this area aim to produce systems which can automatically generate understandable text, either on the basis of data (e.g. stock market figures, patient data, images), or on the basis of another text (e.g. for the purpose of summarization or text simplification). Generally speaking, there are two approaches to build NLG systems: (1) writing rules and templates that specify what the generated text should look like; and (2) training a system to learn the correct behavior from example data. We will ignore planning-based approaches (see the discussion in Gatt and Krahmer 2018), because we are not aware of any planning-based image description systems. Although early work in automatic image description used a rule-based approach (e.g. Kulkarni et al. 2011; Mitchell et al. 2012; Elliott and Keller 2013), most recent work takes a more data-driven approach (e.g. Vinyals et al. 2015; Xu et al. 2015; Wu et al. 2017a; Dai et al. 2017). Hence, in this chapter, we will focus on the latter.

6.1.1 Goal of this chapter

This chapter serves as an introduction to the second part of this thesis, and presents an overview of the components used in current image description systems. Besides introducing important terms and concepts from the literature, we will also provide an error analysis of a data-driven automatic image description system (Xu et al., 2015). This will give us an indication of the quality of state-of-the-art image description technology.

6.1.2 Structure

This chapter consists of two parts. Following this introduction, we first cover the basics of neural networks, and then discuss the typical components of neural image description systems (CNNs in §6.3 and RNNs in §6.4). Finally, we will describe Generative Adversarial Networks (§6.5). Our aim here is to give the reader a basic understanding of how neural architectures work, without diving into their formal definitions. We conclude this part with a section on possible future improvements (§6.6).

The second part of this chapter (§6.7-6.11) focuses on error analysis, a practice used to understand the strengths and weaknesses of individual systems. We present a taxonomy of errors made by Xu et al.'s (2015) system, and annotate those errors to get a sense of their distribution. Through this annotation effort, we show where there is still room for improvement for this system (and, by extension, systems with a similar architecture).

6.1.3 Sources

This chapter draws from several overviews of the field, in particular:

- Bernardi et al.'s (2016) survey of the automatic image description literature.
- Gatt and Krahmer's (2018) survey of the Natural Language Generation literature.
- Goldberg's (2017) overview of neural network methods for natural language processing.
- Goodfellow et al.'s (2016) book on deep learning.

The second half of this chapter is based on the following work:

Emiel van Miltenburg and Desmond Elliott. 2017. Room for improvement in automatic image description: an error analysis. *arXiv preprint arXiv:1704.04198*

6.2 Neural networks

Neural networks are machine learning models that are loosely inspired by the human brain. They consist of artificial *neurons*, that are usually connected to each other in *layers* (groups of neurons). Figure 6.1 shows a schematic of an 'actual' neuron. It consists of a cell body that receives input through its dendrites. If the signal is strong enough to surpass a threshold value, the cell fires a signal through the axon to the axon terminals, which pass the signal through to other cells. Figure 6.2 shows an artificial neuron. The input nodes ($x_{1\ldots n}$) are attached to the neuron through weighted connections. The neuron takes the sum of the inputs multiplied by their weights, and transforms the result through some predefined function: $f(\sum_1^n x_i \cdot w_i)$. When fed with example ⟨input, output⟩ pairs, neural networks are programmed to learn a mapping between the input and the output, by modifying the *weights* on their connections by back-propagation (Rumelhart et al., 1986). This supervised learning process is referred to as *training*.

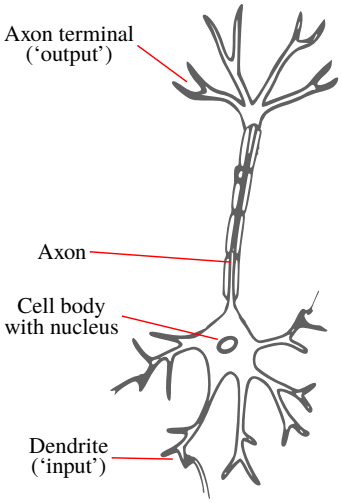


Figure 6.1 A neuron. Original image by edgato on Openclipart.org (public domain).

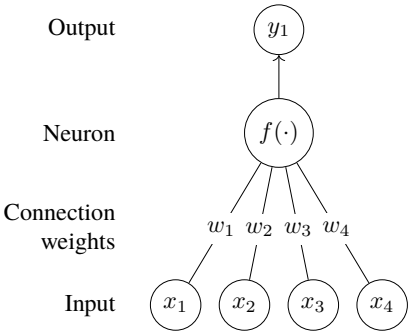


Figure 6.2 Artificial neuron.

To illustrate the training process, let us suppose that the neuron just computes the identity function $f(x) = x$. Further suppose that the task of the network is to predict whether a number is even or odd, and that numbers are fed to the neuron in binary form. E.g. 1 is represented as `[0,0,0,1]`, four is represented as `[0,1,0,0]`, and nine is represented as

`[1,0,0,1]`.[1] For *odd*, the network should output 1, and for *even*, the network should output 0. Finally, assume that weights are initialized as random floats. With an initialization of $w_{1,2,3,4} = [0.1, 0.4, 0.3, 0.9]$, an input of `[1,0,0,1]` would yield 1.0 (the sum of all inputs multiplied by their weights). Accidental success! But for other numbers, this initialization would not give the right result. For example, with an input of 6 (`[0,1,1,0]`) the neuron would yield 0.7. During training, the weights that contributed to the error are adjusted to generate a better result in the future. Eventually, the weights for this example should end up as $[0, 0, 0, 1.0]$, because only the final digit of the binary number provides relevant information about whether the number is odd or even.

A neural network is simply a collection of neurons that are all connected together as a directed acyclic graph. Figure 6.3 shows an example of such a network, with the neurons organized in layers. The input layer feeds into a hidden layer, which is connected to another hidden layer, which feeds into the output layer. The hidden layers are called 'hidden' because they are not as directly accessible in the same way as the input or the output. Having multiple of these layers is useful, because they allow for more complex transformations of the input data, which in turn allows us to solve more complex problems. An intriguing property of neural networks is that there is no symbolic representation of 'what is learned.' Knowledge of how to solve the problem is stored holistically as connection weights.

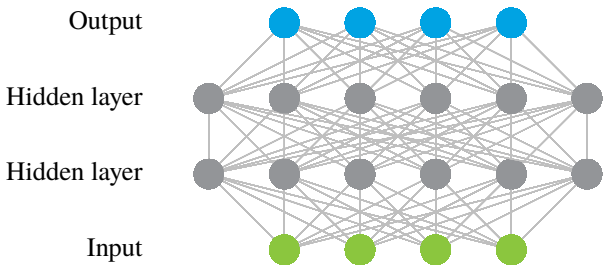


Figure 6.3 A neural network with two hidden layers.

6.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs, LeCun et al. 1998) are commonly used for Computer Vision tasks, such as optical character recognition (OCR) and image labeling (Russakovsky et al., 2015). Rather than taking a one-dimensional vector as their input (like most neural networks), CNNs operate over two-dimensional grids or matrices. This is useful for image processing, because digital images can be represented as matrices with pixel values. Figure 6.4 shows an example, using a picture of the number four.

Let's say we want to build a system that automatically recognizes handwritten numbers, by correlating different visual features with the desired output. The image on the left of Figure 6.4 shows three relevant features (according to human intuition), highlighted in blue:

[1] The position of the digits in a binary number correspond to powers of 2, starting from 2^0 in the rightmost position. To obtain the value of a binary number, multiply the value in each position with the corresponding power of two, and sum the results. Hence $[1, 0, 1, 1] = (1 * 2^3) + (0 * 2^2) + (1 * 2^1) + (1 * 2^0) = 11$. Because $2^{1 \ldots n}$ are all even, the rightmost position in a binary number (corresponding to $2^0 = 1$) determines whether the number is odd or even. So the problem reduces to: 'is the last digit 1 (odd) or 0 (even)?'

Figure 6.4 Left: highlighted parts of the image that are relevant for classifying the image as the number four. Middle: Illustration of a filter sliding (convolving) over an image. Right: highlighted parts of the image that are relevant for classifying the image as the number zero. Original images (drawings of the numbers four and zero) by Jon Phillips, from OpenClipart.org.

a diagonal line, two crossing lines, and the lower end of an upright line. Whenever we can identify all those parts in a picture of a number, we can be fairly sure that the number should be four. With enough training examples, the CNN learns that this combination of features is strongly correlated with the number four. A zero, on the other hand, would have more curved features and no crossing lines. Some relevant features (again according to human intuition) are indicated in the image on the right in Figure 6.4. Presence of these features heightens the probability of the image being an example of the number zero, and lowers the probability of the image being an example of the number four. An attractive property of CNNs is that we do not need to specify any of these features. Rather, the network learns to find relevant patterns by itself. For more details on how this works for digit recognition, see LeCun et al. 1998.

Convolutional Neural Networks consist of multiple layers, where each layer learns more abstract patterns than the previous one (combining lower-level features). Convolutional layers recognize images by sliding filters over an image, and computing a function between the filter and the image at every step. This sliding is shown in Figure 6.4 by the image in the middle. After having computed the matches for all filters at all locations, we have a new grid of values that gets sent to the next layer. Following the convolutional layers, CNNs usually end with a fully connected layer (or a set of fully connected layers) that serves to make predictions about the input. With modern CNNs, the network architecture can become quite complex, as illustrated in Figure 6.5, which shows the CNN known as GoogLeNet (Szegedy et al., 2015). This network is over 30 layers deep, and uses multiple convolutional modules (with different filter sizes) per layer (all but five of the blue boxes in Figure 6.5).[2]

GoogLeNet won the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) in 2014 (Russakovsky et al., 2015). The goal of this challenge is for systems to correctly classify 1000 different types of objects, in a large collection of images. CNNs have become the standard approach to this task, since Krizhevsky et al.'s (2012) AlexNet system won the 2012 challenge with a 10% lower error-rate than the first runner-up (which used a set of hand-crafted features).

The success of AlexNet led other researchers to explore why CNNs are so successful, and what kind of features are learned by Convolutional Neural Networks. Zeiler and Fergus (2014) and Yosinski et al. (2015) visualize what different layers in image-labeling CNNs respond to. They show that the feature maps from the lower layers correspond to low-level features (corners and edges), while feature maps in higher layers correspond more closely to the different classes that the CNN is trained to recognize (dogs' faces, birds' legs). Soon after, researchers realized that CNNs trained for the ILSVRC could also be more broadly applied. For example, Donahue et al. (2014) trained a CNN on the 2012 ImageNet data, and then showed that the features

[2]The other blue boxes correspond to fully connected layers preceding the yellow object classification layers. There are two FC-layers for both of the 'intermediate' classification layers, and one FC-layer for the final classification.

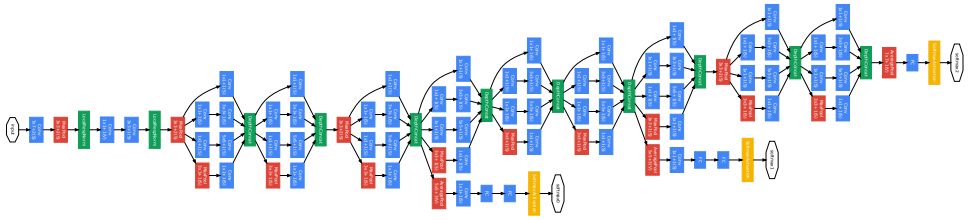


Figure 6.5 The full convolutional neural network from Szegedy et al. (2015), also known as GoogLeNet (in honor of Yann LeCun's work on CNNs). Image copied from Szegedy et al. 2015. All but five blue boxes show convolutional layers. The network makes predictions about objects depicted in the image at three different stages. This is indicated with the white octagonal boxes.

learned by the CNN were also useful to classify images as being *indoors* or *outdoors*. This is an example of *transfer learning*, where knowledge about one task can be carried over to another.[3] Around the same time, different researchers also found that image features extracted using a CNN could also serve as an image representation, which could be used to produce image descriptions (Kiros et al., 2014; Mao et al., 2015; Vinyals et al., 2015). We will specifically look at the system by Vinyals et al. (2015) in the next section.

6.4 Recurrent Neural Networks

Many neural network architectures, such as the Multilayer Perceptron, have a fixed input size. This makes it difficult to work with text data, because sentences can be arbitrarily long; there is no upper bound to how long a sentence can be. Recurrent Neural Networks (RNNs, Elman 1990) are designed to handle (text) sequences of arbitrary length. In recent years, RNNs have become one of the workhorses of Natural Language Processing. This section provides a general introduction to RNNs and how they are used. For a more extensive overview, see Lipton et al. 2015; Goodfellow et al. 2016; Goldberg 2017.

We will assume that text sequences are represented as lists of tokenized words (even though one might also choose to represent text as a sequence of characters). We can feed text into an RNN by providing the tokens one-by-one, in separate *time steps*. Alternatively, RNNs can also produce sequences of text, generating sentences word-by-word.

6.4.1 Model architecture

The basic RNN architecture is illustrated in Figure 6.6. It consists of an input X (provided at time step t), an RNN unit, and an output H (the Hypothesis at time step t). The RNN unit is connected to itself, which means that at every time step, it sends some information from its hidden state to itself as input for the next time step. Instead of representing RNN using this recursive loop, we can also present them *unrolled* as in Figure 6.7. This presentation shows the entire sequence of time steps.

[3] See Kornblith et al. 2018 for a recent discussion of transfer learning using ImageNet models.

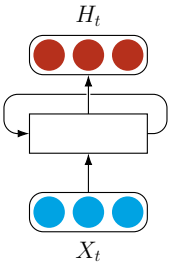


Figure 6.6 Recurrent Neural Network.

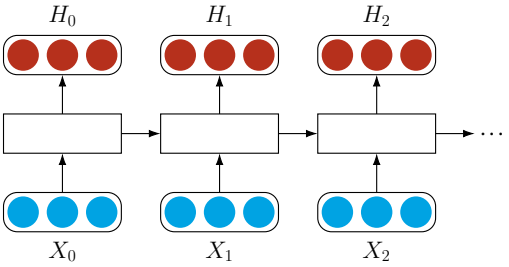


Figure 6.7 Unrolled Recurrent Neural Network.

6.4.2 Uses of RNNs

A basic use of RNNs is to apply them for *sequence labeling* tasks, such as Part-of-Speech tagging and Named Entity Recognition, where there is a one-to-one mapping between the input and the output. Table 6.1 shows an example sentence with its tokens associated with part-of-speech tags and entity labels. For every input token X_t, the RNN can use the preceding tokens $X_{0...t-1}$ to decide upon the right tag or label for X_t. The predicted tag or label is the one that has the highest probability, given the current input and the sequential data observed so far.

Tokens:	Keith	Richards	performed	in	Arnhem	.
Tags:	PROPN	PROPN	VERB	ADP	PROPN	PUNCT
Labels:	PERSON	PERSON	-	-	LOCATION	-

Table 6.1 Table showing a mapping from the input (the tokenized sentence *Keith Richards performed in Amsterdam*) to possible outputs: either part-of-speech tags or entity labels. These kinds of mappings could be learned by an RNN model. (PROPN stands for *proper noun*, ADP for *adposition*, and PUNCT for *punctuation*.)

6.4.3 Different kinds of RNNs

Although basic RNNs work well for many sequence modeling problems, different researchers have proposed extensions or modifications to improve their performance.

Bidirectional RNNs. As Figure 6.7 shows, standard RNNs only operate in one direction: either from left to right, or from right to left. But whatever direction we go in, the RNN cannot use the next tokens ($X_{(t+1)...n}$) predict a label for the current token, even though that additional context could be very useful. Bidirectional RNNs (Schuster and Paliwal, 1997) solve this problem by having two RNNs operate over the input sequence: one that goes from left to right, another that goes from right to left. An additional layer uses the information extracted by both RNNs at the same time steps to make predictions about the input.

Gated RNNs. The problem with basic RNNs, as they were originally conceived, is that they struggle with longer dependencies; with long sequences, it is difficult for the network to 'remember' information from the beginning of the sequence, all the way up to the end (Bengio et al., 1994). This lead to the introduction of *gated* RNNs: recurrent neural networks where

the RNN modules have a memory component, with *gates* that determine whether to keep remembering, or to forget particular information. Gated RNNs learn by themselves (using the training data) how to control those gates. The most common gated RNNs are Long Short-Term Memory networks (LSTMs, Hochreiter and Schmidhuber 1997) and Gated Recurrent Units (GRUs, Cho et al. 2014).

6.4.4 Encoding and decoding sentences

RNNs are also used to produce representations of sequential data, that can be fed to other machine learning components, such as classifiers. An RNN-classifier can be used to predict properties of a sequence, e.g. whether a sequence of words forms a grammatical Dutch sentence, or whether the sequence carries positive or negative sentiment. Figure 6.8 provides an illustration. In this scenario, we can say that the RNN is used as an *encoder*.

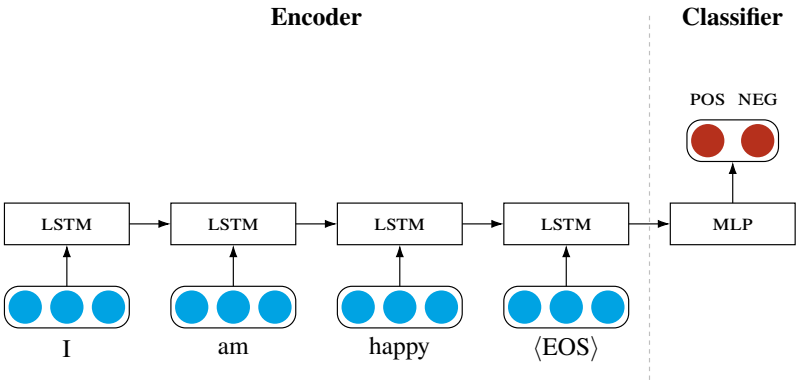


Figure 6.8 Recurrent Neural Network used to classify the polarity of a sentence as either positive or negative. ⟨EOS⟩ is a token that signals the end of the sentence. This example uses a multilayer perceptron (MLP, a neural network with at least three layers: an input layer, one or more hidden layers, and an output layer) to classify polarity, based on the output of the LSTM.

The reverse is also possible. RNN-*decoders* take vector representations as their input and produce sequences as their output. Those vector representations can also be generated by another RNN. This technique is often used for *sequence-to-sequence* (or *seq2seq*) problems such as Machine Translation. The idea (proposed by Cho et al. 2014 and Sutskever et al. 2014, illustrated in Figure 6.9) is that the message from one language is projected into a shared semantic space between the encoder and the decoder, and the decoder uses that representation to reproduce the message in another language.

Rather than decoding a message from one language into another, Vinyals et al. (2015) propose to use an LSTM-decoder to produce image descriptions based on vector representations of images. They use a pre-trained convolutional network model to compute feature vectors for the images in the Flickr30K and MS COCO image description datasets (Young et al., 2014; Lin et al., 2014), and train an LSTM to produce descriptions for those images, based on the extracted features. Figure 6.10 provides an illustration. Note that the image is only provided at the start of the generation process, rather than at every time step (as in Mao et al. 2015, for example).

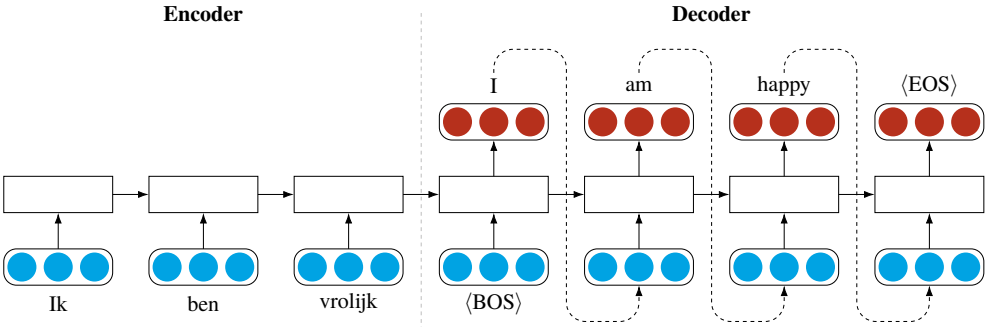


Figure 6.9 RNN used to translate a sentence. The RNN on the left is used to encode the Dutch source sentence, while the RNN on the right is used to decode the hidden representation into English. Note that the decoder uses the predicted words from each previous time step (H_{t-1}) to predict the next word. ⟨BOS⟩ is a token that signals the start of the sentence to be decoded.

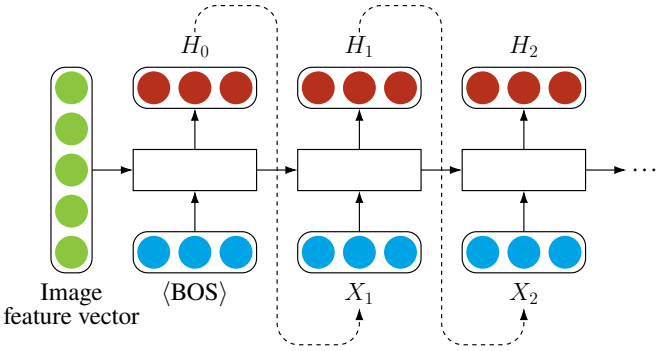


Figure 6.10 RNN initialized with an image feature vector, and a beginning-of-sentence (BOS) token. The dashed lines indicate that the hypothesis H_{t-1} may be used at inference time as the input for the next time step.

6.4.5 Attention mechanisms

Regular conditioned RNNs can only look at the image as a whole, because the visual feature vector does not contain any spatial information. The standard approach of using the penultimate layer of an image labeling CNN means that the feature vector only contains information about *what* is in the image, not *where* it is. This limits the kind of descriptions that these models can generate. After all: it is hard to talk about what you cannot see. To tackle this issue, Xu et al. (2015) present an image description system with an *attention module*, illustrated in Figure 6.11.[4]

The idea behind the attention module is that the system should learn to identify salient (visually important) parts of an image, and attend to those regions while describing them. To achieve this, the attention module receives two inputs: (1) a set of *feature maps*, corresponding

[4] Attention modules have also been used to improve machine translation systems (e.g. Bahdanau et al. 2015). The idea is that, for every word the system produces in the target language, it should have evidence from the source language. Attention-based machine translation models explicitly learn where to look for this evidence in the source sentence. For a more elaborate discussion, see (Olah and Carter, 2016).

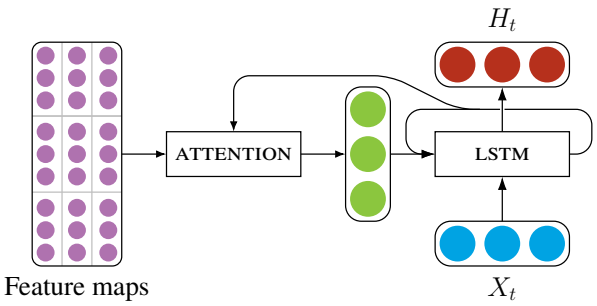


Figure 6.11 RNN with an attention module. The model is first provided with a set of feature maps, corresponding to different regions of the image. At every time step, the attention module learns to identify relevant parts of the image for the system to describe. As its input, the module takes takes the feature maps and the RNN's hidden state from the previous time step. The attention model produces a single feature map for the RNN to use in generating the next word.

to different parts of the image; and (2) a feature vector from the LSTM module, containing information about the previous time steps. The attention module is trained to produce a feature vector with visual information that is relevant for the current time step. (At the first time step, t_0, the attention module returns the average over all feature maps.) The LSTM uses this information, along with its regular inputs, to produce the next word and to provide feedback to the attention layer, so that it can select relevant regions to attend to in the next time step.

Ideas like the incorporation of an attention module are continually refined, by researchers trying to use their intuitions about the problem and hard-coding those ideas to constrain the learning task. One example comes from Lu et al. (2017b), who argue that not all words in a description depend on visual information. Some are inherently non-visual (e.g. *the* or *of*), while others can easily be predicted from the preceding words (e.g. *person talking on a cell* ... –answer: *phone*). Therefore, Lu et al. (2017b) propose to add another module to the architecture: a *visual sentinel*. The system now also has to learn whether or not to look at the image, but at least it isn't forced to look at (and use) the image anymore if it is not relevant.

6.5 Generative Adversarial Networks

Generative Adversarial Networks (GANs) were proposed by Goodfellow et al. (2014) for the problem of learning generative models of data. The idea is to train two neural networks that compete with each other: one tries to generate realistic images, while the other tries to discriminate between real and artificially generated images. This drives the generator to produce images that fall into the same distribution as the training images. GANs have been highly successful at generating realistic images and videos, and this success has led others to propose adversarial training for other applications as well. Recently, Dai et al. (2017) and Shetty et al. (2017) have proposed different GAN-based image description systems. What makes GANs successful at producing more diverse descriptions is the presence of an additional *objective*: not only do they have to be good at predicting the next word at every time step, but they also have to make sure that the *entire* description is human-like as well.

6.6 Takeaway

The previous sections (§6.2-6.5) discussed the building blocks for data-driven image description systems (based on neural networks) that have become standard for automatic image description. Understanding how current systems work also helps to see (at an abstract level) how they could be improved in future research:

1. Improve the visual component. If we can extract better features from the image, or information about the image, then the descriptions may become more reliable and accurate. One way to approach this problem is to design models that perform better on the ImageNet Visual Recognition Challenge, and then use their internal representation of the image (rather than representations from existing feature extractors). See Kornblith et al. (2018) for a discussion of this idea.

2. Improve the generation component. If we change how to act on the visual information, we may be able to produce higher quality descriptions. Possible changes are:

   - Change the the kind of feedback the system receives while training. GAN-based models are an example of this. These models don't necessarily perform better than other models (as measured by BLEU, Meteor, and other automated metrics), but they do generate more diverse descriptions (Dai et al., 2017; Shetty et al., 2017, see also Chapter 7 of this thesis).
   - Add other sources of information for the system to use in the description process. This idea has been explored in the context of Visual Question Answering (Antol et al., 2015), see e.g. (Wu et al., 2017a).

At the same time, the limitations of current data-driven image description systems are also clear: they don't do much more than correlate image features with sequences of words. Judea Pearl, in an interview (Hartnett, 2018) about his recent book (Pearl and Mackenzie, 2018) calls this *curve fitting*. He argues that, for real intelligent behavior, we need systems to reason about the world. This has traditionally been the domain of more formal, rule-based systems (which tend to be restricted to a small domain, because rule-writing is very labor-intensive). It is unclear where the field is going, but these are fertile grounds for the development of hybrid systems that enjoy the best of both worlds.

6.7 Evaluation

We only know how good or bad a system is once we have evaluated it. The question of how to evaluate Natural Language Processing systems has a surprisingly short history; just thirty years ago, system evaluation was considered a controversial topic (Paroubek et al., 2007). Nowadays, no NLP engineering paper is published without some form of evaluation, and automatic image description is no exception.

6.7.1 Evaluation of automatic image descriptions

As Bernardi et al. (2016) note in their survey of the image description literature, automatic image description systems have been evaluated in two ways: either through human judgments or through automated metrics. We briefly discuss each of these below.

Human judgments

Early work on image description was evaluated with text-based similarity measures *and* a human judgment study (Bernardi et al., 2016). This type of judgment study involves asking humans to rate whether the descriptions accurately describe the image, are grammatically correct, are relevant for the image, are human-like, *inter-alia*, using a Likert-scale survey. The main criticisms of human judgment studies is they are expensive to perform and difficult to replicate without access to the same subject pool and control samples (e.g. Papineni et al. 2002; Hodosh and Hockenmaier 2016). Nevertheless, these studies are the clearest indication of overall performance differences between models.

Automatic evaluation

Recent advances in automatic image description have mostly been evaluated with text-based similarity metrics. These metrics compare automatically generated descriptions (the *hypotheses*) for a set of images with the (human-generated) *reference descriptions* associated with those images. Jurafsky and Martin (2009) note that the intuition behind these metrics "derives from Miller and Beebe-Center (1958), who pointed out that a good MT output is one that is very similar to a human translation." Examples are:

BLEU (Papineni et al., 2002) computes the amount of n-gram overlap between the hypothesis and the reference descriptions, using a modified n-gram precision metric. In other words, BLEU asks: to what extent can we find the same n-grams from the hypothesis in the reference descriptions?

ROUGE (Lin, 2004) computes the extent to which the hypothesis overlaps with the references, using a recall-based approach. In other words, ROUGE asks: how much of the information in the references is also captured by the hypothesis?

TER (Snover et al., 2006) computes the minimum amount of edits needed to transform the hypothesis into the closest reference.

Meteor (Banerjee and Lavie, 2005; Denkowski and Lavie, 2014) is similar to BLEU and ROUGE but adds the ability to match synonyms and paraphrases, using WordNet and a paraphrase table.

Metrics like these make it easy for researchers to benchmark the effect of their modeling decisions in terms of overall quality, but they are not informative about the strengths and weaknesses of a proposed model. This is especially true for n-gram based metrics, such as BLEU, which measure grammatical fluency and not semantic adequacy (Reiter and Belz, 2009).[5] Elliott and Keller (2014) show that BLEU, Meteor, ROUGE and TER have at best a moderate correlation with human ratings of image description quality. More recently, different researchers have proposed other metrics for image description evaluation:

CIDEr (Vedantam et al., 2015) is similar to existing metrics that compare a hypothesis with a set of reference descriptions, except that it gives a higher weight to words that are more informative (as computed using the TF-IDF score for each word).

[5] We may also note the current trend to highlight the shortcomings of BLEU in particular, e.g. with Reiter's (2018) structured review of the validity of BLEU, and Sulem et al.'s (2018) analysis showing that BLEU is not suitable for text simplification evaluation.

SPICE (Anderson et al., 2016) uses the reference descriptions to build a 'scene graph', which can be represented as a set of propositions about the picture. Automatically generated descriptions are also parsed into a scene graph, and the SPICE metric measures the extent to which the two scene graphs overlap (expressed as an F1-score, combining precision and recall over the propositions).

Word Mover's Distance (WMD; Kusner et al. 2015) was originally developed to measure document similarity. Kilickaya et al. (2017) modified this metric for the evaluation of image descriptions. Rather than directly working with the tokens in the hypothesis and references, the WMD metric uses word embeddings to compute the distance between the hypothesis and each individual reference description.

Kilickaya et al. (2017) show similar results to Elliott and Keller's (2014) study. For all metrics listed above (except for TER), they found that these metrics have at best a moderate Spearman correlation (between 0.44 and 0.64) with human judgments.[6] Furthermore, the authors find that the different metrics seem to capture different aspects of description quality. In particular, they note that Meteor, SPICE, and WMD seem to complement each other; after combining these metrics, the authors obtain a Spearman correlation of 0.66 with human judgments. Having that said, there is still room for improvement of these metrics. We will discuss some possibilities in Section 6.13.1.

6.8 Error analysis

Error analysis is the process of identifying the mistakes that a system makes, and ordering those mistakes into coherent subgroups. This categorization reveals the distribution of the different kinds of errors, so that we know (if we used a representative sample) which errors occur most often, and which occur less frequently. The remainder of this chapter presents a coarse- and fine-grained analysis of the descriptions generated by a state-of-the-art attention-based model (Xu et al., 2015), trained on the Flickr30K dataset (Young et al., 2014). The goal is to assess the qualities of a state-of-the-art model to illustrate the recent progress in this area and the challenges that lie ahead.

6.8.1 Coarse-grained analysis

Our coarse-grained analysis quantifies whether the descriptions are accurate or inaccurate. (We define accurate to mean that the description is congruent with the image, without it necessarily being the "best" or most complete description.) This is similar to the human judgment studies discussed earlier. Our coarse-grained analysis is a binarized version of the correctness scale from Mitchell et al. (2012). Figure 6.12 provides some examples of image descriptions with differing amounts of errors. Each of these would be classified as INACCURATE.

6.8.2 Fine-grained analysis

Our fine-grained analysis takes the image descriptions that have been classed as INACCURATE, and further classifies them in terms of our taxonomy of errors, presented in Section 6.9. This gives us an indication of the distribution of errors that the system produces.

[6] Their Table 3 shows correlations with human judgments on the Flickr8k dataset from Elliott and Keller 2014. The same table also shows that all metrics have weaker Spearman correlations with data from Aditya et al. (2015) (between 0.39 and 0.44), but these numbers conflate *correctness* and *thoroughness*.

One error

A woman in a **red** shirt is standing in front of a building

Two errors

A man in a **yellow** helmet rides a **bike** in the air

Three errors

A blond **woman** in a **white** shirt is **blowing** her teeth

Four errors

A little **boy** in a **white shirt** playing **soccer**

Figure 6.12 Examples of images with 1–4 errors. The annotated errors are marked in boldface. Original images by: Feggy Art (CC BY-NC-ND 2.0), el Reino (All rights reserved), Edbury Enegren (CC BY-NC-SA), and Neil Smith (CC BY-NC-SA), all through Flickr.com.

Our work is most closely related to earlier work by Hodosh and Hockenmaier (2016), who propose an evaluation of image description systems using binary forced-choice tasks, where systems have to choose the best description for a given image. For each image, the system can choose between the original description or a manipulated description. By controlling the manipulations, the authors are able to check for weaknesses in image description systems. Their error categories (i.e. the different kinds of manipulations) partially overlap with ours, though we provide a more fine-grained typology.

6.9 Error categories

We developed a non-exhaustive categorisation of errors by inspecting the descriptions generated by an attention-based image description model (Xu et al., 2015). We trained the model on the Flickr30K dataset (Young et al., 2014), with 300-dimensional word embeddings, a 1000-dimensional GRU hidden layer (Cho et al., 2014), and '$\text{CONV}_{5,4}$' image features from the VGG-19 convolutional neural network (Simonyan and Zisserman, 2015). We generated 1,014 descriptions with a beam width of five hypotheses, recording a Meteor score of 17.4 on the

Flickr30K test set. All our code and data is available online.[7]

In total, we identified 20 common types of errors, which we grouped into four main categories: PEOPLE, SUBJECT, OBJECT, and GENERAL. We developed annotation guidelines with examples for each type of error. The error categories and types of errors are described below. For the full guidelines, see Appendix D.

People Image description models often make mistakes that are specific to the description of people. Types of errors in this category are AGE (e.g. *woman* instead of *girl*), GENDER (*man* instead of *woman*), TYPE OF CLOTHING (*shirt* instead of *jacket*), and COLOR OF CLOTHING (*red shirt* instead of *blue shirt*).

Subject Mistakes relating to the subject of the description. This category contains the following types of errors: WRONG when the wrong entity in the image is chosen as the subject, SIMILAR when the model mis-identifies the subject for something visually similar (e.g. *guitar* instead of *violin*), NON-EXISTENT when nothing close to the mentioned entity is present in the image, and EXTRA SUBJECT when an additional (nonexistent) entity is described along with the correct entity.

Object Similar to **Subject**.

General Mistakes that are not specific to people. Error types in this category are: STANCE for posture-related mistakes, ACTIVITY for wrongly identified activities, POSITION for mistakes in spatial relations within the image, NUMBER for counting errors (too few/many entities mentioned), SCENE/EVENT/LOCATION for mis-identifications of the scene, event, or location, COLOR for non-clothing entities that are mistakenly attributed with a color, OTHER for any unforeseen mistakes, and GENERALLY UNRELATED for descriptions that do not seem to have any relation with the image. In these cases, it is impossible for annotators to assign any error category to the description. E.g. if the first image in Figure 6.12 were to be described as *A dog runs through the snow*.

6.10 Annotation tasks

We define two error annotation tasks: The **coarse-grained annotation** task is a binary categorization problem, where an annotator determines for every description whether it is accurate. The **fine-grained annotation** task is a multiclass categorization problem, given the error types presented in the previous section. Each inaccurate description is annotated with one or more error types. We can think of this task as a means to assess the *semantic edit distance* between a generated description and the closest accurate alternative.

In total, one annotator categorized all 1,014 generated descriptions into the coarse-grained groups: accurate and inaccurate descriptions. The same annotator then performed the fine-grained annotation. We validated the annotation scheme by double-annotating a random selection of 100 descriptions (10% of the data, annotating both coarse and fine-grained) to determine whether the annotation guidelines provide a reliable basis for annotating the errors.

6.10.1 Results for the coarse-grained task

In the coarse-grained annotation task, 812 out of 1014 descriptions (80%) were judged to be inaccurate. We achieved a good inter-annotator agreement of Cohen's κ=0.67, with an accuracy of 91%. The discrepancy between these numbers is explained by the label distribution: the INACCURATE category is so dominant that any disagreement yields a high penalty in κ. Out

[7] See: https://github.com/evanmiltenburg/ErrorAnalysis

of the 100 double-annotated descriptions, the first and second annotator judged 86 and 81 descriptions to be inaccurate, with agreement on 79 descriptions.

6.10.2 Evaluating the fine-grained annotations

We found 1,265 errors in 812 descriptions, which is an average of 1.56 errors / description. Tables 6.2 and 6.3 show the number of errors per image, and the distribution of error types across the dataset. Surprisingly, the most common error category is GENERALLY UNRELATED (264 times). Errors from the GENERAL and PEOPLE categories are much more frequent than the other two. Taken together, the SUBJECT category is least common. Our intuition is that this is because mistakes in decoding the subject from the language model affect the entire sentence; the choice of subject influences the probability of all subsequent words, leading to a generally unrelated sentence.

Number of errors	1	2	3	4
Frequency	486	221	83	22

Table 6.2 The distribution of error annotations. Top: the number of errors for a single description. Bottom: how many descriptions have exactly that many errors.

Type	Count
generally unrelated	264
color of clothing	195
activity	168
type of clothing	104
gender	98
scene/event/location	91
number	61

Type	Count
non-existent object	47
age	40
stance	38
position	37
extra subject	34
similar-object	31
other	20

Type	Count
color	14
non-existent subject	11
wrong-object	7
similar-subject	3
extra object	1
wrong-subject	1

Table 6.3 Number of times each error was annotated in our fine-grained analysis.

The fine-grained annotation task is inherently ambiguous because inaccurate descriptions might be corrected in many different ways. The first image in Figure 6.12 illustrates this ambiguity. The generated description for this image is given in Example (33a). This description could either be corrected to (33b) or (33c), depending on whether one assumes the mistake is in the color or the type of clothing.

(33) a. A woman in a **red shirt** is standing in front of a building
b. A woman in a **black shirt** is standing ...
c. A woman in a **red skirt** is standing ...

Subjectivity and ambiguity are inherent to the task of image description; describing an image in one simple sentence means that you have to make a choice about what to include in your description. But this subjectivity also means that it is difficult to provide a proper intrinsic evaluation for the annotation task: different choices about how to describe an image may be equally valid. To quantify the extent of this issue, we treat the double annotation for the fine-grained task as a retrieval problem, i.e. how many error types are also found by the

Type	BLEU	Δ	Meteor	Δ
Baseline	17.8	—	17.2	—
Color of clothing	18.8	1.0	17.5	0.3
Activity	18.5	0.7	17.7	0.5
Type of clothing	18.1	0.3	17.4	0.2
Gender	18.6	0.8	17.6	0.4
Scene/event/location	18.0	0.2	17.4	0.2

Table 6.4 Error categories and the BLEU-4 and Meteor scores after correcting the errors. Δ indicates improvement in the scores between the modified descriptions and the original descriptions.

second annotator? For the fine-grained annotation task, we ended up double-annotating 79 descriptions that both annotators agreed contained at least one inaccuracy. For these cases, we achieved a precision of 0.54, with a recall of 0.55. Based on this observation, we decided to carry out an *extrinsic* evaluation: how useful are the fine-grained annotations for guiding future research on model development? We discuss this evaluation below.

6.11 Correcting the errors

Now we have observed the frequency of each type of error, we can ask: what is the effect of addressing these errors on the automatic evaluation metrics? We selected the five most common error types (excluding GENERALLY UNRELATED), and manually corrected each error *without* looking at the reference descriptions. If a description is annotated with multiple errors, we only correct the relevant error. We tried to be conservative in our corrections; e.g. for COLOR OF CLOTHING errors, if the system wrote e.g. *white shirt* instead of *checkered/leopard print/...shirt*, we left the description untouched, rather than insert the pattern. For the ACTIVITY errors, we tried to change as little as possible but editing the activity often also entails changing the object as well. For example, a sentence that read *A man in a suit is holding a sign.* was changed to *A man in a suit is talking.* because the man wasn't holding anything and leaving out the object would produce an ungrammatical sentence. If a change would entail completely re-ordering the sentence, we leave the generated description untouched.

Table 6.4 presents the BLEU and Meteor scores for the validation set before and after correction. For example, after only correcting the colors of clothing, we find a one-point improvement for the BLEU score with respect to the original model.

We did not investigate whether these effects are cumulative, i.e. what happens if we correct *all* errors. Presumably, they are cumulative, but this task is not suitable for such an investigation because the corrections need to be restrictions in order for the improvement estimation to be accurate. If we allowed annotators to correct all the errors in a sentence, we would be giving them *carte blanche* to rewrite everything, turning the analysis into an evaluation of human performance.

6.12 Takeaway

Sections 6.8–6.11 provided an extensive error analysis for image descriptions generated by a state-of-the-art attention-based model. The main contributions of this analysis are:

1. Providing a taxonomy of common errors in automatically generated image descriptions.

2. Quantifying the weaknesses of the model. We posit that any model with a similar architecture will have similar weaknesses.
3. Quantifying the possible improvement of this model if those weaknesses are addressed.

We focused on the nature of the inaccurate descriptions, and looked at different errors that these contain. But what about the accurate descriptions? The descriptions that *are* accurate, are also much more general than the human descriptions, which usually include small, but salient details. We propose the following rule: if the majority of the human descriptions comments on an aspect of the image that is not addressed by a generated description, then that aspect could be improved. This idea is operationalized in the next chapter, where we propose the *local recall metric* (§7.4.2).

We see two other perspectives to build on the observations from this error analysis.

Automated error analysis: As noted earlier, Hodosh and Hockenmaier (2016) carried out a study in which they evaluate image description models using binary forced-choice tasks, where models have to choose which description best describes a particular image. The choices are carefully manipulated, so that each task evaluates the model's performance in one area (e.g. recognizing scenes). Our taxonomy of errors could be used to extend the range of available tasks, for example with a task to evaluate the use of color terms;

Extending existing models: Table 6.4 provides an indication of how much a model could improve by incorporating a dedicated module to detect color, actions, type of clothing, gender, and scenes. We expect that our work will encourage researchers in vision & language to investigate this possibility. More generally, we hope that our taxonomy of error types will help others to go beyond similarity-based metrics, and to look at their model's output through a qualitative lens.

Our results cast doubt on some of the findings from Anderson et al. (2016). Their paper, proposing the SPICE metric, argues that we can use SPICE to evaluate model performance in more detail than 'global' metrics like BLEU and Meteor. Specifically, the authors claim that SPICE is useful to evaluate whether models are able to identify relevant objects, relations between objects, and whether objects have particular attributes. For attributes, the authors identify three subcategories: COLOR, COUNT, and SIZE. Referring to their Table 2 (comparing different system outputs with human performance), Anderson et al. (2016) argue that the models from Fang et al. (2015) and Vinyals et al. (2015) "outperform the human baseline in their use of object color attributes." This is a surprising result, given our findings with Xu et al.'s (2015) attention-based model (which has been shown to outperform both Vinyals et al.'s (2015) and Fang et al.'s (2015) model). Humans are not likely to make the same mistakes as in Figure 6.12 (e.g. saying *white shirt* instead of *black shirt*), and we found many errors like this. Furthermore, we have to ask ourselves what it means to 'outperform the human baseline' on the SPICE metric. Participants of the image description task and image description systems are asked to do two different things. Humans receive very few examples of 'proper' descriptions, and produce texts about the images that capture the main contents of those images, based on their individual ideas of what a description should look like. Systems receive a large amount of training data, and are asked to produce descriptions that are similar to what they have seen before. Thus, their task is to generate 'average' descriptions that are close to what human participants have produced before. Because automated metrics evaluate descriptions based on human reference data, they are biased towards the image description systems, whose task is closer to the how they are evaluated. We conclude that 'outperforming the human baseline' in terms of the SPICE metric may not be a good indicator of actual performance, and that

human-level performance has not been achieved yet.[8]

6.13 Conclusion

This chapter presented an introduction to automatic image description, focusing on data-driven models. The first part of this chapter showed the building blocks that form the core of these models, while the second part of this chapter showed the shortcomings of one particular instantiation. It is clear that current approaches to automatic image description still leave plenty of room for improvement. If many of the generated descriptions seem generally unrelated to the images, and a recent model makes a substantial amount of errors in identifying color of clothing, then we are still a long way from the kind of reasoning about the images that we see in human descriptions (as shown in the first part of this dissertation).

6.13.1 Implications for image description research

Even though image description systems try to find the descriptions with the highest probability, given the input image, the error analysis shows that the generated descriptions are not necessarily faithful to the images themselves. This raises the question: how could we force image description models to remain faithful? (Aside from using better image representations, to reduce the noise in the input.)

We have already seen one approach in Chapter 2, when we were discussing ways to tackle bias in image description (§2.11). Burns et al.'s (2018) Equalizer model forces itself to use correctly gendered terms, and if the model finds no evidence of gender in the image, it uses a gender-neutral term (e.g. *person, snowboarder*). Another proposal was recently made in the *Shortcomings in Vision & Language* workshop, where researchers in Vision & Language discussed weaknesses of current systems combining Computer Vision and Natural Language Processing. Madhyastha et al. (2018) note that current image description evaluation metrics do not take the images into account. Rather, they compute the similarity between the generated description and a set of reference descriptions. Optimizing for these metrics will not improve the accuracy of the generated descriptions. Instead, Madhyastha et al. suggest to use metrics that take image content into account as well. They propose to use pre-trained object detectors, and to compare generated image descriptions with the set of detected objects. While this is still a crude metric (for example, it does not take actions into account), it does show us a way forward to make descriptions more closely match the images they are supposed to be describing.

6.13.2 Next chapter

Having looked at the *accuracy* of automatically generated image descriptions, the next chapter will also look at the *diversity* of automatic image descriptions. We will compare the output of

[8] However, we should acknowledge the difference between *models* and *architectures*. Xu et al. (2015) have shown that they were able to train a well-performing model using their attention-based architecture. Although the model that we analyzed has the same architecture as Xu et al.'s (2015) model, it *is* a different model. And although it is *plausible* that Vinyals et al.'s (2015) and Fang et al.'s (2015) models would make similar mistakes as our model, it is not yet certain that they do. If we want to conclusively show that both Vinyals et al.'s (2015) and Fang et al.'s (2015) model still do not perform at human level, with regard to object color attributes, we should carry out another error analysis with the model outputs as they were evaluated by Anderson et al. (2016).

nine different systems with human reference data. As we will see, automatic image descriptions use a much smaller vocabulary, and are much less diverse than their human-generated counterparts.

Chapter 7

Measuring diversity

7.1 Introduction

Automatic image description is a challenging task because natural language and the visual world both exhibit a wide range of variation (Bernardi et al., 2016). Computational image description models are trained to generalize over datasets of images with multiple human descriptions, but much of the variation present in these descriptions is lost in a trained model. Dai et al. (2017) note that the descriptions generated by recurrent neural networks using a maximum-likelihood objective are "overly rigid and lacking in variability." This rigidity and lack of variability in the output of state-of-the-art models is unfortunate because human descriptions are the exact opposite of this: Devlin et al. (2015) found that humans typically produce unique descriptions, i.e. only 4.8% of the human-described evaluation data in the MS COCO dataset (Lin et al., 2014) also occur in the training data. In sum: human-generated image descriptions are much more diverse than automatically generated descriptions. The first step in addressing this issue is to find ways of measuring and analyzing the difference in diversity between human- and machine-generated output. Once we are able to measure this difference, then we can look for ways to reduce it. This chapter provides an overview of different ways to measure the diversity of automatic image descriptions, and compares the performance of 9 recent image description systems with human reference data for the MS COCO dataset.

7.1.1 Contents of this chapter

This chapter starts with a background section (§7.2), where we will discuss different definitions of diversity, and some of the metrics that are currently used to assess the diversity of automatically generated text. We observe that there is a lack of consensus in this area, which means that it is hard to compare the results from different systems, because they tend to use different metrics.

In the next section, we present six different metrics to assess the diversity of automatically generated English image descriptions, and compare them using nine state-of-the-art image description systems (Section 7.3). Besides covering existing metrics, like TTR and average sentence length, we also propose two word recall metrics that provide more information about the output vocabulary (Section 7.4).

We also investigate the compositional capacity of the different systems, by examining how many different compound nouns and prepositional phrases they can produce. We use these metrics to analyze how image description systems differ from human descriptions (Section 7.5). It is not our goal to evaluate the quality of the descriptions, though future research may find that more diverse descriptions are also more attractive for human readers (Section 7.6.2).

The main finding of our analysis is that recent GAN-based systems (Dai et al., 2017; Shetty et al., 2017), designed to produce more human-like image descriptions, do indeed produce more diverse output than the other MLE-based systems, but this increased diversity still mostly comes from the head of the vocabulary (i.e. the most frequent words in the training set). In

order to support future analyses, we release a toolkit to assess the output of any system and to compare the results with existing approaches.[1]

7.1.2 Publications

This chapter was edited from the following publication:

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*

7.2 Background

The lack of variability in machine-generated text is not limited to automatic image description; it is a general problem in natural language generation. Simply put: automatically generated text quickly becomes boring or repetitive. Recent efforts to address this problem include using maximum mutual information as an objective function, rather than the likelihood of the output, to improve the variablity of a neural conversation model Li et al. (2016a). Castro Ferreira et al. (2016) focused on the deterministic nature of NLG systems, in the sense that they repeatedly use the same referential forms to refer to the same entity in longer stretches of text. They addressed this problem by explicitly training their model to mimic human variability for referring expression generation.

In the image description literature, there have been two recent approaches to generating diverse outputs: (i) learn different description distributions simultaneously to generate multiple different descriptions for the same image (Wang et al., 2016); and (ii) augmenting a model with an additional (conditional) Generative Adversarial Network objective (Goodfellow et al., 2014; Mirza and Osindero, 2014, GAN) to generate more natural and diverse descriptions. In this setting, the caption generator tries to fool a discriminator that is trying to distinguish human image descriptions from machine-generated ones (Dai et al., 2017; Shetty et al., 2017). From these papers, two definitions of diversity emerge:

Local diversity: The ability to generate many different descriptions for the same image.

Global diversity: The ability to use (many different combinations of) many different words.

The former is *local* because it can be evaluated for individual images. The latter is *global*, because it is a property at the corpus level. This chapter focuses on global diversity, which means that we will study whether systems are able to produce as many different words and phrases as humans do in their descriptions of images. We know that word frequencies follow a *Zipfian* (or *power law*) distribution (Zipf, 1949; Van Heuven et al., 2014; Corral et al., 2015), which means that a small subset of the vocabulary accounts for the largest part of the data. Natural language processing systems trained on corpus data are sensitive to this, and tend to overfit on the head of the distribution (e.g. Postma et al. 2016a). We will show that this also holds for the output of image description systems: all systems considered in this chapter mainly use the top 20% most frequent words.

In this chapter, we consider the following question: **How can we measure the diversity of the output generated by an image description model?** There is currently a lack of consensus about how to measure the diversity of model output but the metrics used thus-far fall into four broad areas:

[1] Toolkit: https://github.com/evanmiltenburg/MeasureDiversity

(i) Modified[2] type-token ratio: the number of distinct unigrams or bigrams, divided by the total number of generated words (Li et al., 2016a; Shetty et al., 2017).
(ii) mBLEU: compute the average BLEU score (Papineni et al., 2002) between each description and the other descriptions generated for the same image. This metric can only be used to evaluate models that produce multiple descriptions per image (Wang et al., 2016; Shetty et al., 2017).
(iii) Model-internal: a Generative Adversarial evaluator network that judges whether descriptions are more natural-sounding and semantically relevant than human descriptions (Dai et al., 2017); and
(iv) vocabulary size and the proportion of uniquely generated sentences (Shetty et al., 2017).

In addition to this lack of consensus about which metrics should be used to measure diversity, it is not known how state-of-the-art systems differ in terms of output diversity because it has not been standard practice to report this type of performance statistic. For this reason, we present an overview of six different diversity metrics, and compare their results for nine different image description systems. We hope that these results can serve as a reference point for other researchers interested in generating more diverse image descriptions.

7.3 Existing metrics

This section discusses six general metrics to measure output diversity at the word level, along with a method to visually inspect the differences between systems. All of these methods require tokenized image descriptions – we use SpaCy 2.0.4 for this purpose and lowercase all of the tokens.[3] The validation data is different from the system output, in that it consists of 5 reference descriptions per image, while the systems only produce one description per image. Hence, for the validation data, we compute each score 5 times – once per reference description – and report the average.

1. The **average sentence length (ASL)** corresponds to the mean number of tokens per sentence.
2. The **standard deviation of the sentence length (SDSL)** is a measure of how much systems vary in their description lengths.
3. The **number of types** measures the number of unique word types in the output vocabulary.
4. The **mean segmented type-token ratio ($\mathbf{TTR}_1$)** is the average number of types per 1000 tokens (Johnson, 1944). It is not affected by sentence length because it is computed for a fixed number of tokens. It is more difficult to artificially increase than the number of types because it is an average.
5. The **bigram TTR ($\mathbf{TTR}_2$)** is the average number of bigram types per 1000 bigram tokens. This is based on Li et al.'s (2016a) diversity metric (looking at bigram diversity), and the MSTTR metric (using a fixed size, averaging over multiple samples) so that it is not biased by description length.
6. The **percentage novel descriptions (%Novel)** refers to the generated descriptions that do not occur in the training data. Note that there may be duplicates among the novel descriptions.

[2] This is similar to the type-token ratio (TTR; number of types divided by number of tokens), except that it is customary to compute TTR over a fixed number of tokens, as TTR decreases with corpus size (Youmans, 1990).

[3] See https://spacy.io for more information about SpaCy.

7.3.1 Systems

For any analysis of output diversity, it is essential to have the generated descriptions. Unfortunately, this data is generally not available for most published systems. We contacted the authors of papers that appeared in relevant conferences and journals between 2016–2017[4], and received nine responses with descriptions generated for the MS COCO validation set. All these systems are listed in Table 7.1. With the exception of the two GAN-based systems (Dai et al., 2017; Shetty et al., 2017), the other systems are based on a conditioned recurrent neural network, trained using a Maximum Likelihood (MLE) objective.

7.3.2 Results

Table 7.1 presents the results for the metrics discussed above. We discuss each of them in turn.

Average sentence length. We observe that all models produce shorter sentences than humans, on average, perhaps also conveying less information. It also means that the BLEU *brevity penalty* (Papineni et al., 2002) and Meteor *length penalty* (Denkowski and Lavie, 2014) are affecting the metric scores. However, *producing shorter sentences* does not necessarily mean *producing worse descriptions*.

Standard deviation of sentence length. We observe that the GAN-based systems vary more than most other systems, but the systems by Liu et al. (2017) and Vinyals et al. (2017) have more variation than other MLE-based systems. Humans vary much more than any model in the length of their descriptions.

Number of types. The model by Liu et al. (2017) produces the fewest distinct word types (598), which severely limits the output diversity of the system. The two GAN-based models produce the most distinct word types: 1,922 and 2,611. This is still much lower than the human type count, which averages at 9,200. The total number of types in the validation set is much higher, at 17,557.

TTR$_{\{1,2\}}$ We find that the GAN-based models again outperform the rest. But in terms of variation, there is still much room for improvement before they reach human parity.

Percentage novel descriptions. We find that the model by Vinyals et al. (2017) outperforms the rest (90.5% novel), with the GAN-based systems following close behind at 87.7% and 80.5% novel. The remainder of the systems reproduce a sentence from the training data approximately 50% of the time.

We visualize the differences between the systems using a **type-token curve (TTC)**, which shows how the number of types develops as one reads more output tokens (Youmans, 1990). This curve was originally proposed to compare different texts, which means that sentence order is fixed. With automatic image description, we do not have this constraint. Rather than taking a single sample, and reading the image descriptions in a single order, we randomized the order of the descriptions ten times, and computed the average TTC for the validation data for each system. Figure 7.1 shows the type-token curves for the validation data and all systems. We observe that the TTC for the human reference data develops much more rapidly than that of the systems. Moreover, we can clearly see how the two GAN-based systems stand out from the others in producing more diverse output.

We now inspect how strongly the different existing metrics correlate with each other. Figure 7.2 shows the correlation matrix between the different general metrics for measuring

[4] We surveyed AAAI, ACL, BMVC, COLING, CVPR, EACL, EMNLP, ICCV, ICLR, ICPR, IJCAI, NAACL, and NIPS.

Type	System	BLEU	Meteor	ASL	SDSL	Types	TTR_1	TTR_2	%Novel	Cov	Loc_5
MLE	Liu et al. 2017	32.3	25.8	10.3	1.32	598	0.17	0.38	50.1	0.05	0.70
	Mun et al. 2017	32.6	25.7	9.4	1.12	1009	0.16	0.38	50.0	0.08	0.78
	Shetty et al. 2016	31.9	25.2	9.0	1.03	1112	0.15	0.34	43.0	0.08	0.74
	Tavakoli et al. 2017	28.7	23.5	9.2	1.03	917	0.15	0.33	38.8	0.07	0.66
	Vinyals et al. 2017	32.1	25.7	10.1	1.28	953	0.21	0.43	90.5	0.07	0.69
	Wu et al. 2016	31.0	25.0	9.1	1.03	849	0.14	0.32	44.5	0.06	0.72
	Zhou et al. 2017	30.0	24.8	9.3	1.20	1334	0.22	0.51	60.1	0.10	0.80
GAN	Dai et al. 2017	20.7	22.4	9.8	1.63	1922	0.23	0.55	87.7	0.15	0.76
	Shetty et al. 2017	–	23.6	9.4	1.31	2611	0.24	0.54	80.5	0.20	0.71
	Validation data	–	–	11.3	2.61	9200	0.32	0.72	95.3	–	–

Table 7.1 System results: BLEU and Meteor scores; average sentence length; standard deviation of sentence length; mean-segmented type-token ratio (TTR); bigram TTR; percentage novel descriptions; coverage; and local recall with importance class 5. BLEU/Meteor scores are originally reported values, except for Dai et al. (2017) and Vinyals et al. (2017), which we computed on the validation set.

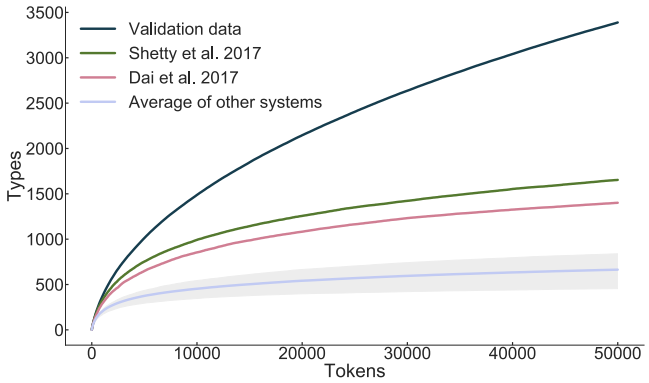


Figure 7.1 Type-token curves for nine systems. The validation data grows much faster than any of the systems and the GAN-based systems clearly outperform the other systems (shaded, with a line plotting the average performance).

diversity. We observe that TTR_1 and TTR_2 are almost perfectly correlated. We conclude from this that a single type-token ratio measure is enough to capture differences between systems in their use of different types. The number of novel descriptions is strongly correlated with the type-token ratio. An intuitive explanation for this is that whenever a model produces more varied output, it is also more likely to produce novel output. In this light, it is interesting to observe the lower correlation between the number of types and the percentage of novel sentences. An explanation for this may be that producing more different types in total does not necessarily mean more diverse output. A system has to *consistently* produce more different types to have an impact.

	ASL	SDSL	Types	TTR1	TTR2	Novel
ASL	1.00	0.83	0.10	0.57	0.52	0.70
SDSL		1.00	0.33	0.85	0.80	0.77
Types			1.00	0.68	0.78	0.43
TTR1				1.00	0.95	0.82
TTR2					1.00	0.83
Novel						1.00

Figure 7.2 Absolute Spearman correlation between the different diversity metrics, computed over the results for the 9 different systems

7.4 Image description as word recall

We argue that image description can be simplified to a word recall problem, where the goal is simply to produce a bag of words that should overlap with the reference data. By ignoring sentence structure, we can focus on the richness of the vocabulary, and study system performance for different classes of words. We distinguish between global recall, looking at the corpus as a whole, and local recall, looking at the corpus image-by-image. We also introduce ranking measures based on these concepts.

7.4.1 Global recall

We formally define the *global recall* metrics using Equations 7.1–7.3. The sets TRAIN, EVAL, GEN correspond to the words that are in the training set, evaluation set, or those generated by the model. Any word type that is both in TRAIN and EVAL is *learnable* from the training data (Eq. 7.1).[5] Recalled words are those that are both learnable and generated by the model (Eq. 7.2). We quantify the success of a system as the percentage of learnable words it can recall, i.e. coverage (Eq. 7.3). Since the set of learnable word types is a subset of the word types in EVAL (this follows from (Eq. 7.1)), systems that are trained on the training data alone cannot recall *all* word types in EVAL. We define this limit in (Eq. 7.4). Intuitively, a model that has a higher coverage (Eq. 7.3) can recall more types from the learnable set (Eq. 7.1), therefore the model is producing a more globally diverse output.

$$\text{Learnable} = \text{TRAIN} \cap \text{EVAL} \tag{7.1}$$

$$\text{Recalled} = \text{GEN} \cap \text{Learnable} \tag{7.2}$$

$$\text{Coverage} = \frac{|\text{Recalled}|}{|\text{Learnable}|} \tag{7.3}$$

$$\text{Limit} = \frac{|\text{Learnable}|}{|\text{Eval}|} \tag{7.4}$$

Using the coverage metric to evaluate the nine systems, we find that the GAN-based systems of Shetty et al. (2017) and Dai et al. (2017) once again achieve the highest scores, achieving 15-20% coverage. This still leaves much room for improvement. We further explore coverage for 10 different subsets of the learnable word types, ranging from the 10% most to the 10% least frequent types in the validation data (based on the counts in the validation set).

Figure 7.3 shows the results. We see that the two GAN-based systems achieve almost 90% coverage of the most frequent types, but this score quickly degrades. Other systems only achieve about 60% coverage of the head, and degrade even more quickly than the GAN-based systems. Furthermore, we observe that the GAN-based systems only achieve better coverage than the other systems on the head of the distribution. Dai et al.'s system is only better for the 0 – 20% most frequent terms (part A), and Shetty et al.'s (2017) system still shows higher coverage than the others up to the 60% mark (part B), but there is no difference for the rest of the lexicon (part C). We emphasize that, for global recall, a system only has to use a type *once* for it to be counted. The Limit for the MS COCO validation set is 0.75. This means that the other 25% (4356 words) in the validation set cannot be learned on the basis of the training set.

7.4.2 Local recall

Local recall considers each image in the evaluation data as a separate word recall problem. We define the local target set as the union of the descriptions (sets of words, D) for an image I_i (Eq. 7.5). The goal is to recall the content words that are important to describe the image. We used SpaCy 2.0.4 to tag the descriptions and we only use adjectives, verbs, nouns, and adverbs as content words for the analysis.

Recalled words are those that are generated for a specific image I_i and occur in the local target set (Eq. 7.6). We define the importance of a word w for an image I in terms of the

[5] We ignore zero-shot learning approaches that could learn to describe images using words outside the training data.

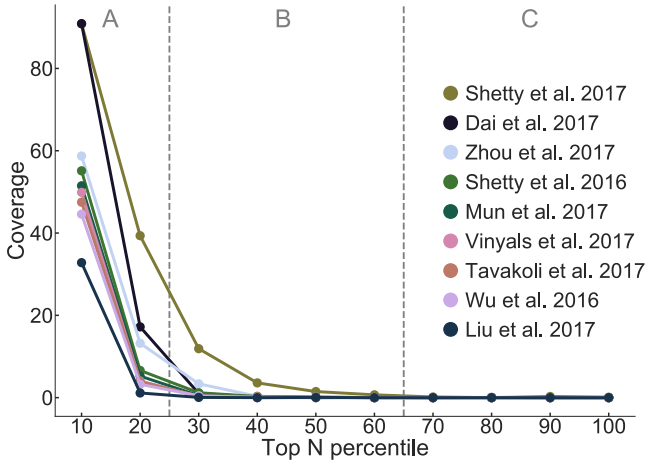


Figure 7.3 Coverage (equation 7.3) for different subsets of the learnable words. Recall for all systems is best for the top 10% most frequent words, but immediately drops for the next 10% of most frequent words.

number of descriptions D that the word w occurs in (Eq. 7.7), resulting in a value between 1 and N (here N=5, as there are 5 descriptions per image). We use the importance metric to measure how well a system recalls the essential (with a score of 5) or the majority (3 or higher) words.

$$\text{Local}_i = \bigcup_{D_j \in I_i} \{w : w \in D_j\} \tag{7.5}$$

$$\text{Recalled}_i = \text{Gen}_i \cap \text{Local}_i \tag{7.6}$$

$$\text{Importance}(w, I) = |\{D : w \in D \wedge D \in I\}| \tag{7.7}$$

$$\text{Local recall score}_k = \frac{1}{|\text{Val}|} \sum_{I \in \text{Val}} \frac{|\{w : w \in \text{Recalled}_i \wedge \text{Importance}(w, I_i) = k\}|}{|\{w : w \in \text{Local}_i \wedge \text{Importance}(w, I_i) = k\}|} \tag{7.8}$$

The *local recall score* for words of k importance is computed by dividing the total number of recalled words with an importance of k by the total number of words with an importance of k (Eq. 7.8).

Figure 7.4 shows the scores for all 9 systems. All models achieve local recall scores between 65% and 80% for types that are mentioned in all five references. This time, the GAN-based models do not outperform the rest, although they still have recalls around 75%. Although local recall is not strictly about diversity in output vocabulary, it does test each system's ability to use the right words at the right time (even if those words are rare).

Figure 7.5 shows the correlations between coverage (Eq. 7.3) and the local recall metric with the existing measures of diversity that were discussed earlier. We find that coverage and the number of types are perfectly correlated. Future work may find that these two measures do not always correlate perfectly, since coverage is based on the word types in the evaluation set. If future systems start producing more word types that are not in the evaluation set, we would see a divergence between coverage and number of types. Local recall (Loc_5 in the table), does not correlate as strongly with the other metrics.

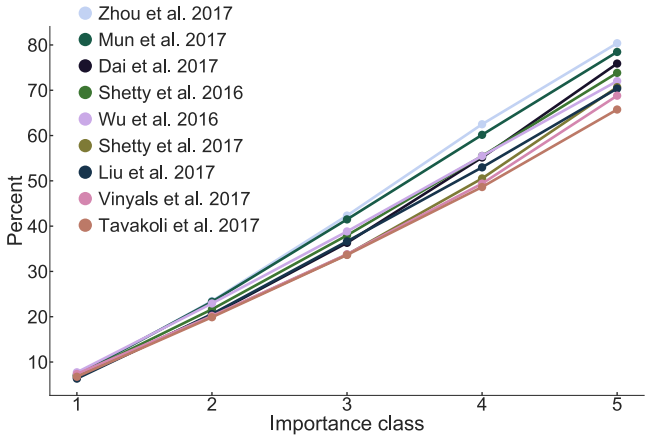


Figure 7.4 Local recall scores for all systems for each word importance class. Systems have low recall for words that occur only once in the reference descriptions, but their recall grows to 65-80% when all five references mention the same word.

	ASL	SDSL	Types	TTR1	TTR2	Novel	Cov	Loc5
Cov	0.10	0.33	1.00	0.68	0.78	0.43	1.00	0.50
Loc5	0.17	0.02	0.50	0.15	0.37	0.10	0.50	1.00

Figure 7.5 Spearman correlations between our coverage and local recall metric and the existing metrics.

7.4.3 Global ranking of omitted words

Instead of using local and global recall to produce scores summarizing model performance, we can use these metrics to construct a ranking of words typically produced by a model, or that a model typically fails to produce. We refer to ranking on the basis of global recall as *global ranking*. The most straightforward way to use global ranking is to construct a frequency table for all words in the evaluation set that are not recalled by a model. This gives us a list of the *most common omissions* for that model. Table 7.2a presents the 15 most frequent words that *all* systems failed to produce. The first ranking is based on the frequency in the training set; the second ranking on the basis of the validation set frequency. The advantage of the former is that we see which words are omitted even though there is sufficient evidence. The advantage of the latter is that we see which words are omitted, even though there are sufficient contexts in which those words could have been used.

Two types that immediately stand out are *'s* and *n't*. One possible reason that both these types were never produced by any system is that they are (cognitively) complex. The possessive *'s* indicates abstract relations between animate entities and objects that vary from scene to scene, making it difficult to learn how to use this type on the basis of visual information alone. The use of negations like *n't* typically requires the speaker to reason about whether or not an image conforms with their expectations (van Miltenburg et al., 2016a). Another difficult case is *thrown*, which refers to a throwing action taking place before the picture was taken. Completing the top-3 in both rankings are *elderly* (253 occurrences in the training data, 140 in the validation data) and *toast* (237 and 124). These are less complex than the examples mentioned above, and could be determined on the basis of visual information alone. Further research is needed to determine why these words could not be produced by any system.

Train	Validation
's	's
elderly	elderly
toast	toast
we	we
whole	thrown
laughing	whole
displays	ham
meadow	located
located	driver
ham	mat
nicely	n't
n't	heading
almost	displays
more	amongst
picking	simple

(a) Global ranking

Absolute	Relative	Relative_{10}
man	pillow	door
dog	kitten	paper
woman	flag	van
people	turkey	pink
cat	milk	head
umbrella	ice	doll
dogs	chips	hair
sign	rainbow	pool
pizza	potatoes	fork
ball	map	tray
cake	eggs	carrot
bear	cream	girls
bed	butter	apple
table	strawberries	women
elephant	pregnant	rice

(b) Local ranking

Table 7.2 Global and local rankings of omitted words. These rankings show the most frequent words that *are not* produced at all (Global ranking), or that are most commonly omitted by the 9 image description systems (Local ranking).

7.4.4 Local ranking of omitted words

We refer to rankings produced on the basis of local recall as *local ranking*. With local ranking, we can look at the words that models failed to produce most often. We will only look at the words with importance class $k = 5$. Table 7.2b presents three local rankings:

1. An *absolute* ranking, where we look at the aggregate number of times each word was missed by the models under investigation.

2. A *relative* ranking, where we look at the rate at which each word was missed (Eq. 7.9). In the case of a tie, the most frequent word 'wins', so that words with the largest impact on model performance are ranked higher.

$$\text{MissRatio}(w) = \frac{\text{missed}(w)}{\text{missed}(w) + \text{recalled}(w)} \tag{7.9}$$

3. A relative ranking with an *occurrence threshold*, where each word with importance class $k = 5$ has to occur at least $n = 10$ times for each system. This eliminates words from the ranking that occur only a few times, but that are missed by all systems (MissRatio(w) = 1).

All three rankings provide a starting point to explore system performance. For example, in the first ranking, we observe that some of the most common terms in the MS COCO dataset overall (*man* and *woman*) are often missed by image description systems, when all annotators *do* use those terms. Since these words are ranked high, they have a big impact on the quality of the descriptions. A natural next step (to be addressed in future research) would be to look at example descriptions where systems fail to produce *man* or *woman* and identify potential causes of this behavior (e.g. an inability to determine people's gender using only visual information).

7.5 Compound nouns and prepositional phrases

Beyond the word level, we can look at how words are combined to form new phrases (i.e. *compositionality*; Szabó 2017). We detect compound nouns using a part-of-speech tagger (SpaCy 2.0.4), assuming that any sequence of nouns is a nominal compound. We also compute the *compound ratio*: the average number of compounds per description. Figure 7.6 and Table 7.3 (next page) show the results.

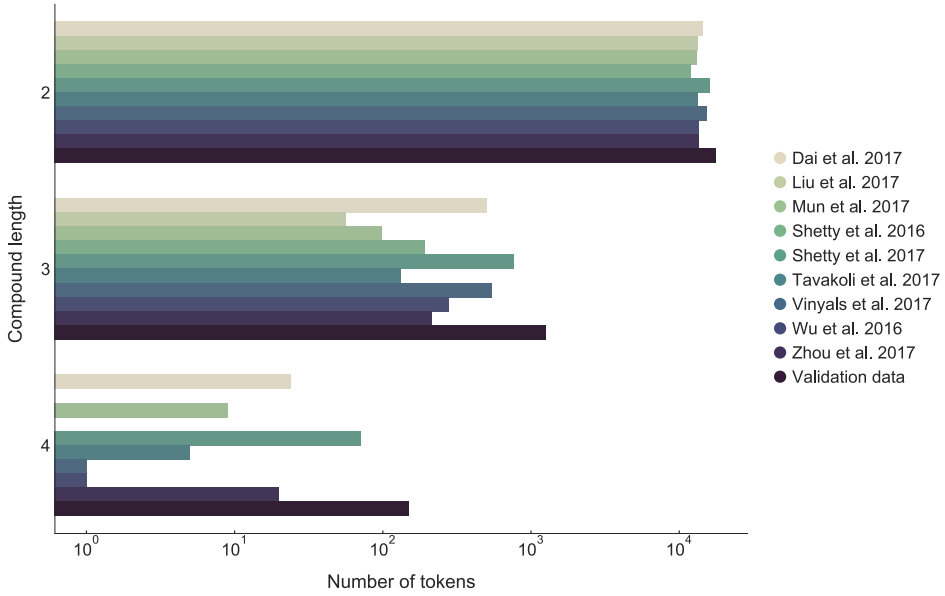


Figure 7.6 Histograms showing the number of tokens with compound length 2, 3, and 4. The validation data and the two GAN-based systems Dai et al. (2017); Shetty et al. (2017) clearly have more compound tokens than the other systems.

We observe that the human reference data has a larger number of compound nouns, resulting in a higher compound ratio. When we separate the compounds by length, we see that humans produce most compounds in any category, and the GAN-based systems (Dai et al., 2017; Shetty et al., 2017) produce more compounds of length 3 and 4 than the other systems. The system by Vinyals et al. (2017) also stands out in this regard. Finally, we see that the GAN-based systems produce more compound types of length 2 than any other system, but there is still a big gap between the GAN-based systems and human performance.

We detect prepositional phrases (PPs), such as *in the kitchen*, using SpaCy's part-of-speech tagger and dependency parser. First, we identify each preposition in the description (e.g. *in, with, on*). Then we inspect the subtree headed by those prepositions. For each of those subtrees, we count their depth in terms of PP embeddings, e.g. *on top of a pan on a table* (34) has a depth of 3.

(34) [**on** top [**of** a pan [**on** a table]]]

We also compute the *preposition ratio*, which is the average number of prepositions per description. Table 7.3 and Figure 7.7 show the results. We do not see a big difference between the validation data and the systems. The only difference is that humans produce more types of

	Compound stats		PP stats	
	Ratio	Types-2	Ratio	Types-1
Liu et al. 2017	0.33	122	1.86	1145
Mun et al. 2017	0.33	300	1.74	2423
Shetty et al. 2016	0.30	319	1.65	2426
Tavakoli et al. 2017	0.33	259	1.72	1888
Vinyals et al. 2017	0.39	275	1.74	1678
Wu et al. 2016	0.34	237	1.69	1732
Zhou et al. 2017	0.34	472	1.71	3451
Dai et al. 2017	0.37	2576	1.78	11709
Shetty et al. 2017	0.42	1446	1.58	8439
Validation data	0.47	6089	1.74	22237

Table 7.3 Statistics for nominal compounds and prepositional phrases. Compound ratio corresponds to the number of compounds per description. Types-2 refers to the number of compound types of length 2. Preposition ratio corresponds to the number of prepositional phrases per description. Types-1 refers to the number of PP types of depth 1.

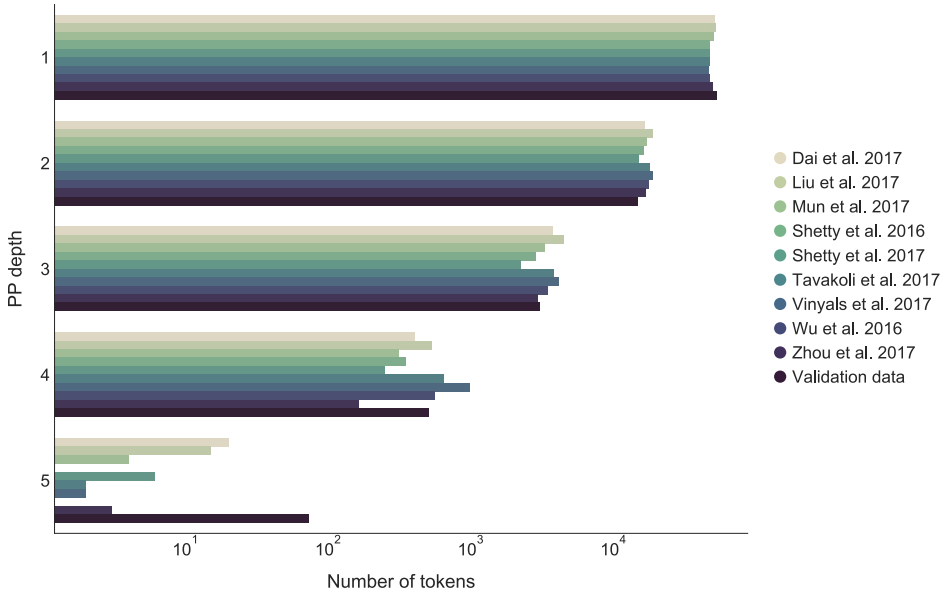


Figure 7.7 Histogram showing the number of tokens with PP-depth 1–5, for all 9 systems and the MS COCO validation data. We do not observe a clear difference between GAN-based systems and other systems in terms of PP depth.

PPs with depth 1: twice as many as the System by Dai et al. (2017). We conclude that image description systems still have much to gain in terms of compositionality. For further discussion of this topic, also see recent work by Lake and Baroni (2017).

7.6 Discussion and Future Research

7.6.1 Other metrics

In addition to the metrics proposed in this chapter, there are other options that could be explored in future work. Follow-up studies could look at metrics based on the **frequency distribution** of words in the training and validation data. We already mentioned Shetty et al.'s (2017) use of frequency ratios in the introduction. Their approach could be extended (perhaps also using log-likelihood; Rayson and Garside 2000) to produce a ranking of words that are over- or underused by a particular system. Overused words could be further analyzed by computing a '**local precision**' metric, measuring how often a generated word is also used in at least one reference description. Ferraro et al. (2015a) present other metrics in their survey of datasets for vision and language research, including:

Yngve and Frazier measurements of syntactic complexity (Yngve, 1960; Frazier, 1985). Ferraro et al. (2015a) found that the MS COCO and Flickr30K datasets have the most complex sentences, compared to other vision & language datasets. It is still an open question whether machine-generated descriptions are of equal complexity and, if not, what are the differences.

Abstract-to-concrete ratio The authors also compare the proportion of abstract words that each corpus contains. They count abstract words by using a list of abstract words compiled in earlier work. In the literature, there are two definitions of abstractness and concreteness. Concrete words are either said to be (1) more closely tied to perception, or (2) more specific (Spreen and Schulz, 1966; Theijssen et al., 2011). It is unclear which is meant by Ferraro et al., but it would be interesting to see whether machine-generated descriptions are more closely tied to perception than human descriptions, who also speculate about the context of the images (van Miltenburg, 2016).

Part-of-speech distribution Ferraro et al. (2015a) compared the distribution of *nouns, verbs, adjectives*, and *other* parts of speech. Our work on detecting prepositional phrases and compound nouns (Section 7.5) suggests that differences in the distribution of parts of speech between human- and machine-generated descriptions could be an interesting avenue to explore.

Besides the measures discussed above, it would also be interesting to study some types of linguistic phenomena in more detail. For example, van Miltenburg et al. (2016a) provide a thorough overview of the uses of *negations* in human-generated image descriptions. Even though this is a low-frequent (or long-tail) phenomenon, studying a subset of the image descriptions informs us about the human image description process, and the *cognitive requirements* to produce a description containing a negation. It remains to be seen whether image description systems could produce similar descriptions.

7.6.2 Limitations and human validation

Earlier work has shown that automated evaluation metrics do not correlate well with human judgments (Elliott and Keller, 2014; Kilickaya et al., 2017). For this reason, we should not blindly trust evaluation metrics in their assessment of system performance. Still, this chapter only includes automatic, intrinsic metrics. This is by design: we want to gain insight into the descriptions, not to evaluate their quality.

While you cannot evaluate a system using only automated metrics, they do tell us something about how a system behaves. Future researchers could try to improve the diversity metrics while maintaining or improving the quality of the descriptions (ideally measured by human judgments). At that point, we should determine if more diverse descriptions (as measured by the metrics covered in this chapter) are perceived by humans as more interesting to read. One issue is that it is unclear how human judgments could be used to rate the diversity of the generated descriptions, because diversity is a *global* property of the data. In other words: you cannot judge the diversity of a single description, because that is not what diversity is about. You can only judge the diversity of a larger collection of descriptions. One way to do this might be to generate descriptions for sets of very similar images, and have participants rate the diversity of different batches of descriptions.

Finally, it is important to note that diversity is closely tied to description specificity. Descriptions can be made more diverse by, for example, replacing the commonly used word *man* with something more specific, e.g. *baker, business man, student, gardener*. The benefit of this (next to the increased diversity/interestingness) is that it makes the descriptions more informative. But it also comes with a risk: more specific descriptions have a higher risk of being wrong, and making an image description system produce more specific descriptions might also lead to more biased output (cf. Chapters 2 and 3). Moreover, the system would run the risk of being *overly specific* (see the discussion in Chapter 3, especially §3.7.1).

7.7 Conclusion

We explored several metrics to analyze the richness of computer-generated image descriptions, most of which focus on diversity at the word level. In our analysis of the output of nine state-of-the-art systems, we found that there are clear differences between human and system output: humans produce more word types; more different types when averaged over multiple 1000-token samples; more compound nouns per description; more long compound nouns; and more compound noun types than image description systems. Not all of these observations hold for prepositional phrases: humans *don't* produce more prepositional phrases per description, and neither do they produce more embedded prepositional phrases, however, they *do* produce a larger number of different prepositional phrases than the systems. At the sentence level, we found that humans produce longer descriptions, vary more in their description length, and produce more novel descriptions. We also found that GAN-based systems produce more diverse descriptions than MLE-based systems. However, we caution that the GAN-based systems are the only ones in our evaluation that are designed with diversity in mind. Further research is needed to find out what kind of approach is best for producing diverse descriptions.

We also proposed to frame image description as a word recall task to further explore the differences highlighted above. *Global recall* looks at the types from all the validation data that are learnable from the training data. *Local recall* measures whether systems are able to produce content words that are mentioned in n reference descriptions for a single image. These metrics show that there is plenty of room for improvement, both in terms of vocabulary size,

as well as using the right words at the right time. One way to approach this challenge is by *ranking* terms that are often missed by a system, and looking for ways to learn when to use these words.

We provide all the code and data to to apply the metrics discussed in this chapter and compare systems. We encourage readers to use this overview to start exploring the output of their own image description systems, but note that the metrics covered here are just the tip of the iceberg. As more researchers focus on producing more diverse descriptions, we will hopefully also develop a better understanding of what makes a description human-like. Formalizing these notions enables us to measure our progress towards richer and more diverse descriptions.

Chapter 8

Final conclusion

This thesis set out to study the extent to which automatic image description systems are able to generate human-like descriptions. This question was split into three separate research questions:

1. How can we characterize human image descriptions? Specifically, what does the image description process look like, what do people choose to describe, to what extent do they differ in how they describe the same images, and how objective are their descriptions?
2. How can we characterize automatic image descriptions? Specifically, what does the image description process look like, how accurate are the automatically generated descriptions, and are they as diverse as human-generated descriptions?
3. Should we even want to mimic humans in all respects? Specifically, are all examples in current image description datasets suitable to be generated by automatic image description systems? If not, what kinds of examples should we avoid?

The first question aims to understand what human image descriptions look like, so as to see what kind of descriptions current data-driven systems aspire to produce. The second question aims to understand where stand in the development of systems producing human-like descriptions. The third, over-arching question, is meant to reflect on the differences between humans and machines. Is it wise to copy all human image description behavior?

8.1 What have we learned?

This thesis is split up into two parts. The first part focused on image description from a human perspective, where we looked at how humans describe images, and what the implications are of this for automatic image description. The second part of this thesis looked at image description from a machine perspective, assessing the state of current automatic image description systems. This section provides a summary of what we have learned from these two parts, followed by a reflection on (un)desirable image description behavior.

8.1.1 Image description from a human perspective

In the first part of this thesis, we have seen that there are three main properties of human image descriptions that have implications for automatic image description systems: (1) they are subjective, (2) they require reasoning, (3) they are task-dependent. We will now discuss these properties in turn.

Human descriptions are subjective

The canonical image description task is not deterministic; when you present the same image to five different participants, chances are that you will end up with five different descriptions. Assuming that this variation is not completely random, we have to conclude that the image descriptions are subjective. In other words, they depend on the participants' interpretation of

the task itself, their interpretation of the images, and their personal thoughts, feelings, and associations with the images. Chapters 2 and and 3 have shown several ways in which human image descriptions for the same image differ from each other:

1. They may present the same facts from a different perspective.
2. They may mention (or omit) different parts of the same image.
3. They may make reference to the same objects at different levels of granularity. That is: they may be more or less specific in the terms (and modifiers) that they use. The specificity of a description may depend on the background knowledge of the speaker and of the perceived background knowledge of the hearer.
4. They may rely on different interpretations of the same image. Actions in particular are underspecified in still images, because by definition photographs (presented in isolation) do not show any movement. For example, the difference between throwing and catching a ball may not be apparent from a picture of someone throwing a ball with two hands.
5. They may rely on different inferences based on the content of the image and the knowledge and beliefs of the participants.

This variation is not necessarily a bad thing. We are still in the early stages of image description research, and the diversity found in current image description datasets allows us to reflect on the question of what image descriptions should look like in the first place.

Human descriptions require reasoning and world knowledge

The subjectivity of the descriptions already hints at the idea that image description requires reasoning and world knowledge. After all: the descriptions depend on how different participants *interpret* the task and the images presented to them. We have seen further evidence that participants actively reason about the images in Chapters 2, 3, and 4.

Chapter 2 presented our basic findings for the English descriptions in the Flickr30K and MS COCO corpora. We found that crowd-workers often go beyond the contents of the images, and add their own inferences to their descriptions (e.g. about the goals, activities, ethnicity, or occupation of the people in the images). These unwarranted inferences are unexpected, because participants were instructed to not make any unfounded assumptions. Furthermore, the use of negations and adjectives also shows how crowd-workers compare the images to their past experiences and mark aspects of the images that are unusual or that deviate from the norm. The use of negations also shows that participants are reasoning about what is happening outside the frame, and about what happened before and after the picture was taken.

Chapter 3 showed that our findings also hold for other languages, and provided additional evidence from the comparison of Dutch, English, and German descriptions that differences in world knowledge affect the specificity of the descriptions. For example, American crowd-workers were unable to identify a traditional Dutch street organ, whereas every Dutch crowd-worker used the same term (*draaiorgel*) to refer to the instrument.

Chapter 4 further supported our claim that image description requires reasoning, by providing real-time evidence of participants reasoning about the images. By eliciting spoken image descriptions together with eye-tracking data, we were able to see what participants were looking at as they were describing the images. Participants seem to be actively predicting what the images were about as they are describing the images. Furthermore, their self-corrections indicate that they reason about (1) the appropriate level of specificity for their descriptions, and (2) whether their descriptions might be ambiguous.

Having shown that the descriptions in current image description datasets are the result of a higher-level reasoning process (rather than a one-to-one mapping of visual features to text), it seems clear that if we want automatic image description systems to be able to produce human-like descriptions, then they should also be able to perform this kind of reasoning.

Human descriptions are task-dependent

Chapter 5 considered the effect of the format of the image description task on the resulting descriptions. The chapter argues that the canonical image description task has just one out of many possible formats, and provides an overview of all the different parameters that one might manipulate to influence the outcome. Focusing on spoken versus written descriptions, we have found that speakers are more likely to 'show themselves' in their descriptions than writers. For example, they seem to use more *consciousness-of-projection* terms, indicating how certain they are about their observations. Future research should investigate whether users appreciate the spoken style more (or less) than the written style.

Given that differences in the image description task lead to different descriptions, we may ask ourselves whether the canonical format actually provides the best set-up for the task. This is important, because with the use of image description corpora for training automatic image description systems, we are implicitly telling models that this is what image descriptions should look like.

Towards an understanding of the human image description process

In his *Tractatus*, the philosopher Ludwig Wittgenstein noted that, though incorrect, his propositions were useful to gain a deeper understanding of the relation between language and reality. After gaining this newfound understanding, we can abandon the propositions and move on. Or in Wittgenstein's words:

> **6.54** - My propositions serve as elucidations in the following way: anyone who understands me eventually recognizes them as nonsensical, when he has used them –as steps– to climb beyond them. (He must, so to speak, throw away the ladder after he has climbed up it.)
> He must transcend these propositions, and then he will see the world aright.
>
> (Wittgenstein, 1921/1961)

This idea has come to be known as *Wittgenstein's Ladder* (although others have used this metaphor before him, see Gakis 2010). Datasets such as Flickr30K and MS COCO are similar: they are useful for us to gain a better understanding of how people describe images, but, having reached this level of understanding, it is clear that we need more controlled data. For example, it would be useful to specify the goal of the task, so that participants know how their descriptions will be used. This would enable them to adjust their descriptions accordingly, which would reduce variation in the descriptions. We have discussed other factors influencing the descriptions in Section 5.3. In a more controlled experiment, we could start to systematically manipulate these factors to see how they influence the image description process. Furthermore, it would be useful to retain participant IDs, so that it is possible to study individual variation in image description.[1]

If we want to make the goal of the image description task more explicit, then more work is also needed to explore different applications of image description technology. We will

[1] This data is available for the Dutch image description data collected for this thesis, as well as the German part of the Multi30K corpus (Elliott et al., 2016). It may also be available for other image description corpora, but this metadata is typically not listed with the publication of the data.

discuss this in more depth in Section 8.2, but for now it is important to recognize that different applications may also have different requirements regarding the form and content of the descriptions. This in turn means that we would need different image description corpora to study how images should be described for a particular task, in a particular domain. We may then find that the cognitive requirements for producing suitable image descriptions may differ between tasks and domains.

8.1.2 Image description from a machine perspective

The second part of this thesis focused on image description from a machine perspective. We have identified three main properties of current approaches. They (1) are inherently limited, (2) produce flawed descriptions, and (3) produce generic descriptions. We will now discuss these properties in turn.

Current approaches are inherently limited

Chapter 6 presented an overview of current image description technology, introducing different kinds of neural networks. But given that their goal is to produce human-like image descriptions, we have to ask ourselves: are they up to the task? As I have argued above in Section 8.1.1, it is clear that human image descriptions are subjective (depending on the participants' interpretation of the task itself, their interpretation of the images, and their personal thoughts, feelings, and associations with the images), require world knowledge, and are highly contextual. However, looking at the general architectures that are used for automatic image description systems, it is clear that they assume a simple one-to-one mapping from images to text. There are (typically) no components that use external resources to reason about the images. Thus there is a clear contrast between what humans do, and what automatic image description systems are designed to do. As noted in the introduction of this thesis (§1.6), there are two possible ways to resolve this issue: either we should (1) build more advanced image description systems, or we should (2) change the (currently implicit) goal of trying to match human descriptions as closely as possible, and formulate a more restrictive standard for what image descriptions should look like.

Current approaches produce flawed descriptions

Following an overview of the general architecture of automatic image description systems, Chapter 6 provided an error analysis for one specific model: Xu et al.'s (2015) attention-based architecture, trained for the Flickr30K dataset. Error analyses for image descriptions are subjective by nature, because classifying the type of error means that the annotator has to reason about what the model is supposed to say. Nevertheless, error analysis is useful to get a general sense of a model's strengths and weaknesses.

Our results indicate that about 80 percent of the generated descriptions contains at least one error. Most of the errors fall in the GENERALLY UNRELATED category, which means that the description does not seem to have any relation to the image. After this category, most errors fall into the categories COLOR OF CLOTHING (e.g. *green shirt* instead of *red shirt*), ACTIVITY (*walking* instead of *running*), TYPE OF CLOTHING (*shirt* instead of *coat*), and GENDER (*man* instead of *woman*). Furthermore, many of the errors made by the system are unlikely to be made by humans. For example, Figure 6.12 in Chapter 6 shows a little girl in a pink dress holding a large ball in her hands. This image is described by the system as *A little boy in a white shirt playing soccer*. Descriptions like these show us that we are still far away from human-level

automatic image descriptions. Despite the fact that we only looked at the performance of one model, we expect that other image description systems with similar architectures will also make errors like these. The distribution of errors will probably differ, but there is no fundamental reason to expect that another model will not produce any mistakes regarding color of clothing, for example. What is needed, is some way to ensure the *visual fidelity* of the descriptions (cf. Madhyastha et al. 2018), so that the automatically generated descriptions will not only be similar to the human-generated descriptions, but also correspond to the contents of the image.

Current approaches produce generic descriptions

Having looked at the content of the automatically generated descriptions, Chapter 7 examined the diversity of the output of 9 different automatic image description systems. We asked to what extent these systems display the same amount of variation as the human-generated descriptions, and whether these systems were able to use particular labels that all human annotators agreed on. In both of these areas, we found that there seems to be much room for improvement. Automatic image description systems tend to only use a small portion of the vocabulary that is available from the training data. Furthermore, if human annotators all agree that a particular term should be used in the description of an image, systems only use that term in 80% of the cases. Finally, image description systems seem to lag behind humans in terms of compositionality; they use fewer kinds of compound nouns, and fewer kinds of prepositional phrases. This may indicate that automatic image description systems are less expressive than humans. At the same time, we shouldn't necessarily take humans as the standard to aspire to. In some cases, it might actually be beneficial for a system to produce relatively predictable descriptions, with only a limited vocabulary. More research is needed to establish when to use a more diverse vocabulary, and when generic descriptions would suffice. Either way, Chapter 7 provides a first step towards a better operationalization of diversity in image descriptions.

8.1.3 How human-like should automatic image descriptions be?

The third sub-question is difficult to answer, because it is not clear what it means for a description to be human-like. As noted above, the descriptions in the Flickr30K and MS COCO datasets are very diverse, and different annotators have different ideas about what an image description should contain. For the sake of simplicity, let us say that a system is fully human-like if it is able to produce any of the different kinds of descriptions that we see in existing image description datasets. Based on the above, there are three answers to the third sub-question:

Computability. Some kinds of descriptions are easier to produce than others. For example, a description like 'A man in a red shirt is walking down the street' is relatively straightforward, compared to descriptions containing negations, or interpretations of how people might be feeling in a particular situation. The latter require much more reasoning and background knowledge (e.g. about how different experiences may affect someone's mood). It may not be feasible for current systems to produce these more advanced kinds of descriptions. We will also discuss this in Section 8.3.

Systematicity and predictability. The wide range of variation displayed by human image descriptions also makes the descriptions themselves unpredictable. This thesis posits that the amount of variation may (at least in part) be due to the fact that participants of the image description task differ in their understanding of what the task is about. With a clearer problem

definition (stating what the image descriptions should be used for), we might be able to establish some standards of what a proper image description should look like. (See §8.2 below for further discussion.) Following these standards, we should see less variation in the descriptions, which should also make the output of image description systems more predictable (and easier to evaluate). This predictability may help users understand when a system generates a particular kind of description, which also helps them make inferences about what likely *isn't* in the image (because otherwise the system would have told them; cf. Grice 1975).

Truthfulness and fairness. For an image description system to be usable, it should provide reliable descriptions, that treat all subjects fairly and without prejudice. We have seen in Chapters 2 and 3 that participants of the image description task don't restrict themselves to the contents of the images, but often speculate about what is happening in the image, what caused the events in the image, and what is likely to happen. In their speculations, people often resort to stereotypes. Furthermore, people display biases in the way that they mark people and situations that are different from what they perceive to be the default. It would not be advisable for systems to display the same behavior, because of the potential for this behavior to be harmful or offensive (next to the fact that speculations and generalizations are simply not always true).

8.2 Application: supporting blind and visually impaired people

What could automatic image description systems be used for? This section will discuss one of the most important applications for automatic image description technology: supporting blind and visually impaired users in their interaction with the world around them. As I have argued in this thesis, current image description datasets display an overwhelming amount of variation, with many different ways to describe the same image. This section argues that we need to talk to potential users of image description technology to understand what is the best way to describe any particular image (§8.2.1), provides an overview of existing research using image description technology to help blind and visually impaired users (§8.2.2), and discusses possible next steps (§8.2.3).

8.2.1 Developing sign-language gloves: A cautionary tale

In developing any application, it is important to keep the end users in mind, and to try and understand their needs. One of the best examples of what *not* to do is the development of sign-language gloves. In an article titled *Why sign-language gloves don't help deaf people*, Michael Erard (2017) describes how different groups of researchers developed high-tech gloves for deaf people to wear, so that their gestures could automatically be translated into spoken English.[2] The main problem with these kinds of gloves is that they misconstrue the problem. Many sign-language gloves only focus on what the hands do (e.g. finger-spelling). But sign-language also uses arm-gestures, facial expressions, and lip movements, which are not captured by the gloves. Thus, the gloves cannot possibly translate the entire message. Furthermore, there is no way for hearing people to respond, so the conversation remains one-way traffic. The moral of the story is that, in developing assistive technology, we should always involve the potential users themselves. Ideally, they should be consulted from the beginning, so that the research does not start out on the wrong foot, and our solutions are actually useful in practice.

[2] Also see the the open letter by Forshay et al. (2016).

8.2.2 Existing research on supporting the blind

The automatic image description literature regularly refers to the potential of this technology to help blind or visually impaired people,[3] but we are still in the early stages of establishing what these people actually want or need, in terms of image descriptions. Existing research can be categorized as follows:[4]

Alt-text

Petrie et al. (2005) provide an overview of existing guidelines for *alt-text*: 'alternative text' to be displayed instead of images for visually impaired users browsing the web using a screen reader. The authors also describe the results of a series of interviews with visually impaired users, asking them how images on the web should be described. Petrie et al.'s (2005) conclusion is that descriptions are very context-dependent, but the following elements should usually be included:

1. Objects, buildings, and people in the image.
2. Activities taking place in the image.
3. The use of color.
4. The purpose of the image.
5. Emotion and atmosphere.
6. The location of the depicted events or activities.

Since this study predates most of today's social media outlets, or at least their widespread use,[5] it does not tell us which properties of images are important in the context of social media. Furthermore, these guidelines are also not informative about life outside the web; how should real-life situations be described?[6]

Automatic image description

Gella and Mitchell (2016) contrast the capabilities of automatic image description systems with the needs or blind or visually impaired people. They note that current automatic image description systems mostly focus on objects, attributes, and actions. Talking to blind or visually impaired people, however, Gella and Mitchell found that users would also like to have a description of the emotion and atmosphere, and whether the image is humorous or not (which perhaps coincides with what Petrie et al. call the purpose of the image). Furthermore, they would like to see descriptions for different types of domains: personal, news, and social media images.

Studies about automatic image description for social media images have been carried out by MacLeod et al. (2017); Zhao et al. (2017b), and Wu et al. (2017b). MacLeod et al. (2017) carried out a user study with automatically generated descriptions for images from Twitter. They provided blind or visually impaired people with actual tweets, that were enriched with

[3] For example: Mao et al. 2015; Elliott et al. 2016; Lu et al. 2017a; Yao et al. 2017; Yoshikawa et al. 2017.

[4] I will ignore related areas, such as object detection, depth estimation, (micro-)navigation, text extraction and text summarization. See Weiss et al. 2018 for a short survey.

[5] Facebook was introduced to college students in 2005, and Twitter was launched in 2006; see Boyd and Ellison 2007 for a timeline.

[6] This question also gives rise to an ethical dilemma: determining what to describe also means that you are effectively withholding information about other parts of the image. How should we handle this responsibility? We leave this question for future research.

automatically generated descriptions. Their first experiment was a think-aloud study, where users were asked to describe their experiences with the automatically generated descriptions. The authors note that users generally trusted the descriptions (without double-checking the information), despite the fact that they were often wrong. Moreover, in cases where the descriptions did not line up with the content of the Tweet, the users tried to provide explanations for why the Tweet-caption combination could still be coherent, rather than dismissing the captions for being implausible. MacLeod et al. (2017) note that this bears some risk for users of automatic image description software, because they may wrongly act upon misleading descriptions. Thus it is important to clearly communicate the accuracy of automatically generated image descriptions to the users. In a follow-up experiment, the authors looked at different ways to communicate (un)certainty about descriptions that are (in)congruent with the images they are associated with. They found that negatively framed descriptions encourage users to remain skeptical about the descriptions in situations where the system is uncertain. Examples of negative frames are: *I have absolutely no idea but my best guess is . . . ; I am not completely sure, but I think it's . . .* . This works better than positive framing (e.g. *I'm only sort of confident, but. . . ; I'm pretty sure it's . . .*), where users are more likely to accept the descriptions as valid.

Zhao et al. (2017a) interviewed 12 visually impaired participants to understand their experiences with photo sharing on Facebook. The authors developed an automatic image description system to aid visually impaired users of the mobile Facebook application. Afterwards, they evaluated this application using a seven-day diary study with six visually impaired users. Based on the 12 interviews, the authors identified three aspects that users would like to know before uploading an image to Facebook:

1. Key visual elements: main landmarks and objects depicted in the image.
2. People: the identities and relative location of the people in the image.
3. Photo quality: technical (focus, lighting), composition (e.g. no people cut off), and subject behavior (e.g. smiling, no eyes closed).

The diary study indicated that users found the application helpful, but they were unsure about the reliability of the descriptions. Having used the application they also had further requests to improve the descriptions. They should provide information about:

4. The kind and color of different objects, especially for common objects like flowers. For example, 'flowers' could be specified to 'yellow tulips'.
5. Non-salient items, especially those that may help distinguish multiple similar images.[7]
6. The luminance and the *level* of blurriness (some blurriness may be acceptable).

Wu et al. (2017b) present another user evaluation for Facebook's Automatic Alt Text (AAT) functionality. Their participants noted two further improvements that they would like to see:

7. The ability to extract and recognize text.
8. More detailed descriptions of people, "including their identity, age, gender, clothing, action, and emotional state."

Finally, Zhao et al. (2017b) also found that their participants were *re-appropriating* the app to organize their photo collections. This also shows that there is room for the development of

[7] The idea to automatically produce pragmatically informative descriptions that distinguish an image from similar images, is explored by Andreas and Klein (2016) and Cohn-Gordon et al. (2018).

personal photo organization applications, which may have different requirements than social media image descriptions.

Visual Question Answering and the VizWiz grand challenge

Following initial work on Visual Question Answering (Antol et al., 2015; Goyal et al., 2017), where computers are asked to answer different questions about a set of images, Gurari et al. (2018) presented the VizWiz grand challenge. The VizWiz dataset consists of 31,000 questions from blind people, about pictures they took themselves. This dataset represents a real-life application (VizWiz; Bigham et al. 2010), which blind people use to answer everyday questions, such as: *what type of soup is this?* or *what temperature is the oven set to?* Because the pictures are taken by blind users (who cannot see the screen), the images are often of low quality, and the questions are spoken rather than written. The VizWiz grand challenge consists of two subtasks: 1. predicting the answer to a visual question; and 2. predicting the *answerability* of a visual question.

The VizWiz grand challenge is a great addition to the existing multimodal Natural Language Processing and Computer Vision tasks, because it confronts us with the noise and uncertainty of real-life data. Moreover, the dataset itself is a very rich source of information about the domains that blind and visually impaired people are interested in. For example, we may use the subjects of the questions and images to understand what kind of information should be highlighted in automatic image descriptions.

8.2.3 Future research supporting blind and visually impaired people

Summarizing the above, there is a growing list of aspects that are *generally* important for automatic image description systems describe. But it is still unclear:

1. Which of those aspects are relevant to mention, given a particular image and context.
2. How specific the description of those aspects should be.
3. What is the best way to phrase the descriptions.

The image description literature has generally avoided these issues by delegating them to the crowd-workers annotating the images. A technical solution is still far on the horizon, because formulating a suitable description, mentioning the relevant aspects of an image, at the right level of specificity is still too difficult for current technology. (The next section discusses *third-wave* approaches that should be able to provide satisfying descriptions to users.) An alternative would be to take a Q&A-style approach (similar to Visual Dialog; Das et al. 2017), where the system would generate a 'basic description' and the user can ask for specific details. The basic description would then serve as a starting point for the conversation. Whatever approach we end up taking, we should always keep the end users in mind. By involving them in the process, we can establish clear guidelines to develop image description solutions that actually address the needs of blind and visually impaired people. These guidelines in turn allow us to develop evaluation metrics that show our progress in generating suitable descriptions.

8.3 Automatic image description in the context of Artificial Intelligence

The work in this thesis can be seen as part of the more general area of Artificial Intelligence. This section aims to present a short overview of the recent progress in this field.

8.3.1 Three waves of AI

In a recent DARPA[8] video, Launchbury (2017) describes the development of Artificial Intelligence (AI) as coming in three waves:

1. **Handcrafted knowledge**: this first wave of development involves experts translating knowledge from a particular domain into formal rules for computers to follow. This works very well for narrow domains, where the computer can take a set of basic facts and reason through their implications. The downside, according to Launchbury is that rule-based systems are less suited to learn from experience, to abstract away from specific problems and apply their knowledge in a different domain. Furthermore, they are not able to perceive the outside world and see what's going on. In Launchbury's (2017) words, they "stumble when it comes to the natural world."

2. **Statistical learning**: this second wave of development focuses on the ability to extract knowledge from data. AI systems in this second wave are much better at perceiving the world and learning from data to adapt to new situations. At the same time, these systems are limited in terms of logical reasoning. Launchbury (2017) summarizes the strengths and weaknesses of second-wave AI systems by saying that they have "nuanced capabilities to classify data and to predict the consequences of data, but they don't really have any ability to understand the context in which they're taking place and they have minimal capability to reason." Hence, DARPA is foreseeing a third wave:

3. **Contextual adaptation**: Launchbury (2017) describes this future wave as one where "the systems themselves over time will build underlying explanatory models that allow them to characterize real-world phenomena." An important feature of these systems is the ability to properly *explain* their decisions.[9] Furthermore, third-wave systems should be able to learn from only a handful of examples, rather than the thousands of training examples required for current statistical learning systems.

The automatic image description systems demonstrated in this thesis are clearly part of the second wave of AI; current systems mostly aim to generate 'the most probable description' given an image, without developing an explanatory model that could tell us *why* an image should be described in a particular way. Current systems are also unable to adapt to the context in which they are providing their descriptions. Chapters 6 and 7, then, are an exploration of second wave systems and the limits of this kind of technology.

8.3.2 Requirements

How do we move from second to third-wave AI? In a recent paper, Lake et al. (2017) present an overview of the requirements for "building machines that learn and think like people." They broadly categorize these requirements into three sets of ingredients:

1. **"Start-up software"** This first set of ingredients corresponds to cognitive capabilities that children have from an early age:

[8] DARPA is the Defence Advanced Research Projects Agency, which funds scientific research in the United States.

[9] The ability to explain decisions is not just nice to have. Not only do explanations signal a deeper understanding of the problem that the system is built to solve, they also satisfy public demands for transparency in automated decision making systems. See Goodman and Flaxman 2017; Selbst and Powles 2017 for a discussion of the European 'right to an explanation' and Lipton 2016 for a discussion of what it means for a machine learning model to be interpretable.

Intuitive physics: infants have a basic understanding of how the physical world works, and they know, for example, which kinds of movements are possible and impossible. They can use (and improve) this understanding with every new task they learn.

Intuitive psychology: infants can attribute mental states (goals, beliefs, desires, intentions, knowledge) to other people, which helps them reason about other people's behavior. In turn, this helps them infer other properties about the world (e.g. which objects are good and which are bad).

2. **Learning** Lake et al. (2017) note that they "view learning as a form of model building, or explaining observed data through the construction of *causal* models of the world" (emphasis in original). These models of the world include the intuitive notions of physics and psychology that infants start out with, and that gradually improves as they learn. The authors argue that compositionality and learning-to-learn are essential ingredients to make rapid model learning possible.

Compositionality is the key to understand complex scenes or objects. Rather than treat each complex scene or object as completely new, we can begin to understand those scenes or objects by decomposing them into their primitive parts. This makes the reasoning process more efficient, and it improves generalization, because each encounter with a complex scene or object informs us about the properties of its more primitive parts (and vice versa), which we can use in the next situation where we encounter those parts again.

Causality means knowing or reasoning about how different situations come to be; providing an explanation. Lake et al. (2017) argue that people also understand scenes like the ones in the Flickr30K and MS COCO dataset by building causal models. Specifically: "human-level scene understanding involves composing a story that explains the perceptual observation, drawing upon and integrating the ingredients of intuitive physics, intuitive psychology, and compositionality." In other words, understanding a scene requires us to identify the individual components and to be aware of what they might contribute to the scene (compositionality), it requires us to reason about the way that the objects in the scene are held together (intuitive physics), and it requires us to think about the goals and intentions of the people in the scene (intuitive psychology), in order to construct a coherent story about what is going on. Lake et al. (2017) note that causality might also help us understand the role of unfamiliar objects in a scene.

The errors that we have seen in Chapter 6 of this thesis are either foundational errors (the visual features being flat-out misleading), or they could be the result of missing one or more of the ingredients listed so far. Lake et al. (2017) note that image description systems often seem to get the key objects correct, but are unable to relate these objects to each other (and thus they do not build the right causal model –if they build causal models at all).

Learning-to-learn refers to the idea that previous learning experiences can make it easier to learn new tasks (Harlow, 1949). Lake et al. (2017) note that this is similar to *transfer learning, multi-task learning* or *representation learning* in the field of Machine Learning. The authors note that, while these concepts are already used, there is still room for improvement, because humans are still much more efficient at leveraging their past experiences to learn to perform new tasks. One way to improve learning-to-learn skills is to focus on the ingredients listed earlier.

3. **Speed** Rich and complex causal models that humans develop about the world (such as the ones that Lake et al. (2017) propose), are typically slow, as they may require multiple

reasoning steps to get to the answer. Lake et al. (2017) observe that this contrasts with speed of perception and thought. Somehow, the authors note, humans successfully combine rich models with efficient inference. Though Lake et al. (2017) do not make the connection, this is reminiscent of Kahneman's (2011) theory of *Thinking Fast and Slow*. He argues that we have two modes of thought, which he refers to as System 1 and System 2. System 1-thinking is fast, instinctive, and emotional, while System 2-thinking is slower, deliberative, and logical. We may also interpret one of the main examples from Chapter 4 of this thesis in these terms. Figure 4.3 showed a picture of a restaurant with people sitting around a table. The participant describing this image immediately inferred from the setting that the group of people was eating. But as the participant continued to described the image, they found that the group wasn't in fact eating anything yet; they were still looking at their menus. In this example, the quick interpretation of the image would be a good example of System-1 thinking, which was later corrected as the participant collected more information and had more time to think about the image.

The requirements laid out by Lake et al. (2017) are based on findings from a wide range of disciplines. And these are just the *cognitive* requirements. If we want artificially intelligent systems to have any role in society, we also need to think about the ethical implications of developing such systems (e.g. Hovy and Spruit 2016; Friedman et al. 2013; IEEE 2018). In short: it is impossible to study AI in isolation. Developing AI means talking to many different groups of researchers. For this conversation to be successful, it is important to make our research as accessible as possible. The next paragraph highlights ways of doing so.

8.3.3 A way forward: more interaction with related fields

Epstein et al. (2018) discuss the rise of Artificial Intelligence as a field, and note that there are strong incentives to develop new systems that improve upon the state-of-the-art performance on particular tasks, but there is less emphasis on the study of those systems themselves. This gives rise to a *knowledge gap* in AI: development of AI systems moves faster than our understanding of them.[10] What we need, Epstein et al. argue, is a centralized platform where researchers can upload their systems, and others can easily test them, without the need to install anything on their own computer. This would allow social scientists (and I would add: linguists and philosophers) to test the biases and competence of AI systems without requiring any technological knowledge.

Another solution to address the AI knowledge gap is to create shared events with researchers from other fields. One such example is the *Workshop on Building Linguistically Generalizable Natural Language Processing Systems*, which aims to bring together linguists and NLP researchers (Ettinger et al., 2017). The first edition of the workshop also featured a *build it, break it*-challenge, where there are two kinds of participants: the builders and the breakers. The former aim to build NLP systems that are robust to linguistic variation, while the latter aim to construct difficult test cases that might trip up the systems. The first edition saw four breaker teams submit test cases for the builders to evaluate their systems on. Those test cases focused on (morpho)syntactic, semantic, and pragmatic phenomena, as well as the ability to use world

[10] This echoes the sentiment from Ali Rahimi and Ben Recht's controversial acceptance speech for the NIPS 2017 'test of time' award (Rahimi and Recht, 2017). They argued that AI and machine learning are still at a pre-scientific stage, similar to alchemy before it developed into physics and chemistry. We have some idea of what works and what doesn't, but we don't know *why*. Jordan (2018) uses the metaphor of civil engineering: "While the building blocks have begun to emerge, the principles for putting these blocks together have not yet emerged, and so the blocks are currently put together in ad-hoc ways."

knowledge to reason about the examples. These test cases help us to better understand model performance in terms of phenomena that are well-studied in linguistics. In a way, the *build it, break it*-challenge is a real-life version of the platform that Epstein et al. (2018) propose. But the organization of shared workshops has the additional benefit of engaging with each other in person. At the same time, a permanent platform where researchers can continuously interact with existing systems allows for more experimentation.

In short: publishing papers and open-sourcing code and data is not enough. We need to think more about the accessibility of our work, and whether it is also feasible for non-technical researchers to study the fruits of AI and NLP research. Opening up the field to 'outsiders' may help us deepen our understanding of what AI and NLP are capable of.

8.4 Future research

There are different ways to make a contribution to NLP. David Marr (1982) posited that a *full* description of any cognitive system[11] requires an explanation at three levels:[12]

1. The computational level: what task is the system solving?
2. The algorithmic level: how does it actually solve the task?
3. The implementational level: how is this algorithm physically realized?

Most work in NLP seems to focus on the algorithmic level: assuming a well-defined task, can we find a better solution to that task? This thesis has mostly on the computational level, trying to give a better characterization of the task of automatic image description, through analyzing existing image description data. One of the main problems (or perhaps even the main problem) with image description research right now is that the task is not well-defined. What we need is a combination of:

1. **User studies** asking potential users of image description systems what the descriptions should look like. These studies should identify different classes of properties that image descriptions should have. We have already seen some of these kinds of studies in our discussion of image description for blind and visually impaired people (§8.2), but user studies shouldn't be limited to this target group only. Others may also benefit from image description applications, e.g. users of voice assistants like Siri, Google Home, or Alexa.

2. **Metric development** where researchers determine for a given feature how to measure whether a particular system is able to competently produce descriptions with that feature. For example, whether a system is able to use negations in its image descriptions. Having more fine-grained test sets with targeted evaluation metrics hopefully allows for a 'divide-and-conquer' situation where different groups work towards solving different sub-problems of automatic image description.

3. **Feasibility studies** where we look at which features are feasible for an image description system to produce. These studies could either target a single feature, or see to what extent a particular is able to competently produce a wider array of different features. In these kinds of contexts, it is often proposed to develop a 'summary score' to see how well systems are doing overall. I would argue against this idea, because it is not clear what such summary statistics

[11] A cognitive system could be defined here as 'any information processing system,' which could equally apply to both humans and machines trying to produce a description for a given image.

[12] This has come to be known in Cognitive Science as the *Tri-Level Hypothesis* (Dawson, 1998).

mean. Is a system with an overall score of 0.8 better than one with an overall score of 0.75? That depends on how important you think the individual features are that make up the overall score, and this importance may differ from situation to situation.

As noted in the previous section, we cannot do this alone. It takes a concerted effort of researchers in NLP, human-computer-interaction, and linguistics to bring us towards a future where computers can finally handle all the pragmatic factors in automatic image description.

Bibliography

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292* .

Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. 2017. Physiognomy's new clothes. Medium. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.

F Niyi Akinnaso. 1982. On the differences between spoken and written language. *Language and speech* 25(2):97–125.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*. Springer, pages 382–398.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1173–1182. https://doi.org/10.18653/v1/D16-1125.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. pages 2425–2433.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '98, pages 86–90. https://doi.org/10.3115/980845.980860.

Adriana Baltaretu and Thiago Castro Ferreira. 2016. Task demands and individual variation in referring expressions. In *Proceedings of the 9th International Natural Language Generation conference*. Association for Computational Linguistics, Edinburgh, UK, pages 89–93.

Adriana Baltaretu, Emiel J Krahmer, Carel van Wijk, and Alfons Maes. 2016. Talking about relations: Factors influencing the production of relational descriptions. *Frontiers in psychology* 7.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 65–72. http://www.aclweb.org/anthology/W/W05/W05-0909.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Pa-*

pers. Association for Computational Linguistics, Belgium, Brussels, pages 304–323. http://www.aclweb.org/anthology/W18-6402.

Roland Barthes. 1957. *Mythologies*. New York: Hill and Wang. Translated by Annette Lavers, 1972.

Roland Barthes. 1961. The photographic message. In Susan Sontag, editor, *A Barthes Reader*, 1994. New York: Hill and Wang, pages 194–210.

Roland Barthes. 1978. Rhetoric of the image. In *Image-music-text*, Farrar, Straus and Giroux. Translated by Stephen Heath.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *European conference on computer vision (ICCV)*. Springer, pages 404–417.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5(2):157–166.

Anton Benz and Katja Jasinskaja. 2017. Questions under discussion: From sentence to discourse. *Discourse Processes* 54(3):177–186. https://doi.org/10.1080/0163853X.2017.1316038.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research* 55:409–442.

Camiel J. Beukeboom. 2014. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. In J. Laszlo, J. Forgas, and O. Vincze, editors, *Social cognition and communication*, Psychology Press, volume 31, pages 313–330. Author's pdf: http://dare.ubvu.vu.nl/handle/1871/47698.

Camiel J Beukeboom, Catrin Finkenauer, and Daniël HJ Wigboldus. 2010. The negation bias: when negations signal stereotypic expectancies. *Journal of personality and social psychology* 99(6):978.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, pages 333–342.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008. http://stacks.iop.org/1742-5468/2008/i=10/a=P10008.

Shoshana Blum and Eddie A Levenston. 1978. Universals of lexical simplification. *Language learning* 28(2):399–415.

Paul Boersma and David Weenink. 2017. Praat: doing phonetics by computer [computer program]. Version 6.0.35, downloaded from http://www.praat.org/.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146. https://transacl.org/ojs/index.php/tacl/article/view/999.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*. pages 4349–4357.

Ali Borji and Laurent Itti. 2013. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(1):185–207.

Danah M. Boyd and Nicole B Ellison. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication* 13(1):210–230.

Cati Brown, Tony Snodgrass, Susan J Kemper, Ruth Herman, and Michael A Covington. 2008. Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior research methods* 40(2):540–545.

Penelope Brown and Colin Fraser. 1979. Speech as a marker of situation. In Klaus R. Scherer and Howard Giles, editors, *Social markers in speech*, Cambridge: Cambridge University Press, pages 33–62.

Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. *arXiv preprint arXiv:1803.09797* .

Guy Thomas Buswell. 1935. How people look at pictures: a study of the psychology and perception in art. .

Zora Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2016. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605, 2016* .

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, pages 568–577.

Centraal Bureau voor de Statistiek. 2016. Bevolking; generatie, geslacht, leeftijd en herkomstgroepering, 1 januari. Part of the CBS database, last modified 15 September 2016. http://statline.cbs.nl/.

Wallace Chafe. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen, editor, *Spoken and written language: exploring orality and literacy*, Norwood, N.J.: Ablex., pages 35–54.

Wallace Chafe and Jane Danielewicz. 1987. Properties of spoken and written language. In R. Horowitz and F.J. Samuels, editors, *Comprehending oral and written language*, New York: Academic Press.

Wallace Chafe and Deborah Tannen. 1987. The relation between written and spoken language. *Annual Review of Anthropology* 16(1):383–407.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting*

of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '09, pages 602–610. http://dl.acm.org/citation.cfm?id=1690219.1690231.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR* abs/1504.00325. http://arxiv.org/abs/1504.00325.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1724–1734. http://www.aclweb.org/anthology/D14-1179.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Hans Dam Christensen. 2017. Rethinking image indexing? *Journal of the Association for Information Science and Technology* 68(7):1782–1785. https://doi.org/10.1002/asi.23812.

Paul Christophersen. 1939. *The articles: A study of their theory and use in English*. Copenhagen: Munksgaard.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 613–622.

Andy Clark. 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences* 36(3):181–204.

Eve V Clark. 1997. Conceptual perspective and lexical choice in acquisition. *Cognition* 64(1):1 – 37. https://doi.org/10.1016/S0010-0277(97)00010-3.

Moreno I Coco and Frank Keller. 2012. Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science* 36(7):1204–1223.

Moreno I Coco and Frank Keller. 2014. Classification of visual and linguistic tasks using eye-movement features. *Journal of vision* 14(3):11–11.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, pages 439–443. https://doi.org/10.18653/v1/N18-2070.

Álvaro Corral, Gemma Boleda, and Ramon Ferrer-i Cancho. 2015. Zipf's law for word frequencies: word forms versus lemmas in long texts. *PloS one* 10(7):e0129031.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the 2017 International Conference on Computer Vision*. Venice, Italy, pages 2970–2979.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Michael R.W. Dawson. 1998. *Understanding Cognitive Science*. Wiley.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, pages 248–255.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Joseph A. DeVito. 1966. Psychogrammatical factors in oral and written discourse by skilled communicators. *Speech Monographs* 33(1):73–76.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, pages 100–105.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*. pages 647–655.

Gerard HJ Drieman. 1962a. Differences between written and spoken language: An exploratory study, I. quantitative approach. *Acta Psychologica* 20:36–57.

Gerard HJ Drieman. 1962b. Differences between written and spoken language: An exploratory study, II. qualitative approach. *Acta Psychologica* 20:78–100.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark, pages 215–233. http://www.aclweb.org/anthology/W17-4718.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR* abs/1510.04709. http://arxiv.org/abs/1510.04709.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*. Association for Computational Linguistics, Berlin, Germany, pages 70–74. http://anthology.aclweb.org/W16-3210.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1292–1302. http://www.aclweb.org/anthology/D13-1128.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 452–457.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science* 14(2):179–211.

Peter GB Enser. 1995. Progress in documentation pictorial information retrieval. *Journal of documentation* 51(2):126–170.

Ziv Epstein, Blakeley H Payne, Judy Hanwen Shen, Abhimanyu Dubey, Bjarke Felbo, Matthew Groh, Nick Obradovich, Manuel Cebrian, and Iyad Rahwan. 2018. Closing the ai knowledge gap. *arXiv preprint arXiv:1803.07233* .

Michael Erard. 2017. Why sign-language gloves don't help deaf people. *The Atlantic* https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable nlp systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1–10. https://doi.org/10.18653/v1/W17-5401.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 1473–1482.

Ethan Fast, William McGrath, Pranav Rajpurkar, and Michael S. Bernstein. 2016. Augur: Mining human behaviors from fiction to power interactive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, CHI '16, pages 237–247. https://doi.org/10.1145/2858036.2858528.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. 2015a. A survey of current datasets for vision and language research. In *EMNLP*. Lisbon, Portugal, pages 207–213.

Francis Ferraro, Nasrin Mostafazadeh, Lucy Vanderwende, Jacob Devlin, Michel Galley, Margaret Mitchell, et al. 2015b. A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 201–213.

Charles J Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences* 280(1):20–32.

Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter Van Attveldt. 2018. Studying Muslim Stereotyping through Microportrait Extraction. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan.

Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, pages 411–412.

Lance Forshay, Kristi Winter, and Emily M. Bender. 2016. Open letter to UW on "SignAloud" project. Open letter. http://depts.washington.edu/asluw/SignAloud-openletter.pdf.

Karën Fort, Gilles Adda, and K Bretonnel Cohen. 2011. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics* 37(2):413–420.

Stella Frank, Desmond Elliott, and Lucia Specia. 2018. Assessing multilingual multimodal

image description: Studies of native speaker preferences and translator choices. *Natural Language Engineering* 24(3):393–413. https://doi.org/10.1017/S1351324918000074.

Lyn Frazier. 1985. Syntactic complexity. In D. R. Dowty, L. Karttunen, and A. M. Zwicky, editors, *Natural language parsing: Psychological, computational, and theoretical perspectives*, Cambridge University Press, Cambridge, pages 129–189.

Batya Friedman, Peter H Kahn Jr, Alan Borning, and Alina Huldtgren. 2013. Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory*, Springer, pages 55–95.

Victoria A Fromkin. 1971. The non-anomalous nature of anomalous utterances. *Language* pages 27–52.

Ruka Funaki and Hideki Nakayama. 2015. Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 585–590. https://doi.org/10.18653/v1/D15-1070.

Dimitris Gakis. 2010. Throwing away the ladder before climbing it. In Elisabeth Nemeth, Richard Heinrich, and Wolfram Pichler, editors, *Papers of the 33rd International Wittgenstein Symposium*. Kirchberg am Wechsel: ALWS, pages 98–100. http://wittgensteinrepository.org/agora-alws/article/view/2891/3506.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61:65–170.

Albert Gatt, Emiel Krahmer, Kees van Deemter, and Roger P.G. van Gompel. 2017. Reference production as search: The impact of domain size on the production of distinguishing descriptions. *Cognitive Science* 41:1457–1492.

Albert Gatt, Marc Tanti, Adrian Muscat, Patrizia Paggio, Reuben A Farrugia, Claudia Borg, Kenneth P Camilleri, Mike Rosner, and Lonneke van der Plas. 2018. Face2text: Collecting an annotated image description corpus for the generation of rich face descriptions. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC'18)*.

Spandana Gella and Margaret Mitchell. 2016. Residual multiple instance learning for visually impaired image descriptions. In *11th Women in Machine Learning Workshop*.

James J. Gibson. 1977. The theory of affordances. In R. E. Shaw and J. Bransford, editors, *Perceiving, Acting, and Knowing*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies* 10(1):1–309.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. Montreal, Canada, pages 2672–2680.

Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38(3):50–57.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74(6):1464.

Herbert Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics*, New York: Academic Press, volume 3, pages 41–58.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage*. volume 5, page 10.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *arXiv preprint arXiv:1802.08218* .

Michael Alexander Kirkwood Halliday. 1989. *Spoken and written language*. Language Education. Oxford University Press.

Birgit Hamp and Helmut Feldweg. 1997. Germanet – a lexical-semantic net for german. In *Proceedings of the ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Jonathon S Hare, Paul H Lewis, Peter GB Enser, and Christine J Sandom. 2006. Mind the gap: Another look at the problem of the semantic gap in image retrieval. In *Multimedia Content Analysis, Management, and Retrieval 2006*. International Society for Optics and Photonics, volume 6073, page 607309.

Harry F Harlow. 1949. The formation of learning sets. *Psychological review* 56(1):51.

Lester E Harrell. 1957. *A comparison of the development of oral and written language in school-age children*, volume 22 of *Monographs of the Society for Research in Child Development*. Wiley.

Kevin Hartnett. 2018. To build truly intelligent machines, teach them cause and effect. *Quanta Magazine* https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/.

David Harwath and James Glass. 2017. Learning word-like units from joint audio-visual analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 506–517.

David Harwath, Antonio Torralba, and James Glass. 2016. Unsupervised learning of spoken language with visual context. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, Curran Associates, Inc., pages 1858–1866.

Martin Haspelmath. 2006. Against markedness (and what to replace it with). *Journal of linguistics* 42(1):25–70.

Irene Heim. 1982. *The semantics of definite and indefinite noun phrases*. Ph.D. thesis, University of Massachusetts. New edition typeset in 2011 by Anders J. Schoubye and Ephraim Glick.

Verena Henrich and Erhard Hinrichs. 2010. Gernedit - the germanet editing tool. In *Proceedings of the ACL 2010 System Demonstrations*. Association for Computational Linguistics, Uppsala, Sweden, pages 19–24. http://www.aclweb.org/anthology/P10-4004.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Micah Hodosh and Julia Hockenmaier. 2016. Focused evaluation for image description with binary forced-choice tasks. In *Workshop on Vision and Language, Annual Meeting of the Association for Computational Linguistics*. volume 3.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47:853–899.

Laurence Horn. 1984. Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. *Meaning, form, and use in context: Linguistic applications* 11:42.

Laurence R. Horn. 1972. *On the Semantic Properties of Logical Operators in English*. Ph.D. thesis, UCLA, Los Angeles.

Laurence R. Horn. 1989. *A natural history of negation*. CSLI Publications.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 591–598. http://anthology.aclweb.org/P16-2096.

Karen R. Humes, Nicholas A. Jones, and Roberto R. Ramirez. 2011. Overview of race and hispanic origin: 2010. Published by the United States Census Bureau. https://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf.

Dell H. Hymes. 1974. *Foundations in Sociolinguistics*. Philadelphia: University of Pennsylvania Press.

IEEE. 2018. Ethically aligned design. First draft, published by the The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. Available through: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf. Retrieved March 2018.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* .

Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pages 145–152.

Laurent Itti and Christof Koch. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* 40(10):1489 – 1506.

Alejandro Jaimes and Shih-Fu Chang. 1999. Conceptual framework for indexing visual information at multiple levels. In *Internet Imaging*. International Society for Optics and Photonics, volume 3964, pages 2–16.

Roman Jakobson. 1972. Verbal communication. *Scientific American* 227:72–80.

Mainak Jas and Devi Parikh. 2015. Image specificity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 2727–2736.

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 1072–1080.

Wendell Johnson. 1944. I. a program of research. *Psychological Monographs* 56(2):1.

Michael Jordan. 2018. Artificial intelligence — the revolution hasn't happened yet. *Medium* https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7.

Dan Jurafsky and James H Martin. 2009. *Speech and language processing*. Pearson Education, second edition edition.

D. Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux. https://books.google.nl/books?id=ZuKTvERuPG8C.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 787–798. http://www.aclweb.org/anthology/D14-1086.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. pages 835–841.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 199–209.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* .

Simon Kornblith, Jonathon Shlens, and Quoc V Le. 2018. Do better imagenet models transfer better? *arXiv preprint arXiv:1805.08974* .

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision* 123(1):32–73. https://doi.org/10.1007/s11263-016-0981-7.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. pages 1097–1105.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *CVPR 2011*. pages 1601–1608. https://doi.org/10.1109/CVPR.2011.5995466.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*. pages 957–966.

Brenden M Lake and Marco Baroni. 2017. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint arXiv:1711.00350* .

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences* 40.

John Launchbury. 2017. A darpa perspective on artificial intelligence. Technical report, Defense Advanced Research Projects Agency (DARPA). Published on YouTube by DARPAtv. https://www.youtube.com/watch?v=-O01G3tSYpU.

Sara Shatford Layne. 1994. Some issues in the indexing of images. *Journal of the American Society for Information Science* 45(8):583–588. https://doi.org/10.1002/(SICI)1097-4571(199409)45:8<583::AID-ASI13>3.0.CO;2-N.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Daniel J. Lee. 2016. Racial bias and the validity of the implicit association test. Working paper 53, United Nations University World Institute for Development Economics Research (UNI-WIDER).

Geoffrey Leech. 1983. *Principles of pragmatics*. London and New York: Longman.

Adrienne Lehrer. 1970. Notes on lexical gaps. *Journal of Linguistics* 6(2):257–261. http://www.jstor.org/stable/4175082.

Willem J.M. Levelt. 1983. Monitoring and self-repair in speech. *Cognition* 14(1):41 – 104. https://doi.org/https://doi.org/10.1016/0010-0277(83)90026-4.

Willem JM Levelt. 1989. *Speaking: From intention to articulation*. MIT press.

Willem JM Levelt. 1999. Producing spoken language: A blueprint of the speaker. In *The neurocognition of language*, Oxford University Press, pages 83–122.

Stephen C Levinson. 1983. *Pragmatics*. Cambridge textbooks in linguistics. Cambridge University Press.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL:HLT*. ACL, San Diego, California, pages 110–119.

Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016b. Adding chinese captions to images. In *Proceedings of the 2016 ACM International Conference on Multimedia Retrieval*. ACM, pages 271–275.

Xirong Li, Xiaoxu Wang, Chaoxi Xu, Weiyu Lan, Qijie Wei, Gang Yang, and Jieping Xu. 2018. COCO-CN for cross-lingual image tagging, captioning and retrieval. *CoRR* abs/1805.08661. http://arxiv.org/abs/1805.08661.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Association for Computational Linguistics, Barcelona, Spain, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, pages 740–755.

Zachary C. Lipton. 2016. The mythos of model interpretability. In *ICML 2016 Workshop on Human Interpretability in Machine Learning (WHI 2016)*.

Zachary C Lipton, John Berkowitz, and Charles Elkan. 2015. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019* .

Chang Liu, Fuchun Sun, Changhu Wang, Feng Wang, and Alan Yuille. 2017. Mat: A multimodal attentive translator for image captioning. In *IJCAI*. pages 4033–4039.

Alessandro Lopopolo and Emiel van Miltenburg. 2015. Sound-based distributional models. In *Proceedings of the 11th International Conference on Computational Semantics*. Association for Computational Linguistics, London, UK, pages 70–75.

David G Lowe. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. IEEE, volume 2, pages 1150–1157.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017a. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017b. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. volume 6.

Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, pages 5988–5999.

Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2018. Estimating visual fidelity in image captions. Extended abstract, presented at the Workshop on Shortcomings in Vision and Language (SiVL), collocated with ECCV 2018.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language* 92:57–78.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proceedings of ICLR*. https://arxiv.org/abs/1412.6632.

Silke Marckx. 2017. *Propositional Idea Density in Patients with Alzheimer's Disease: An Exploratory Study*. Master's thesis, Universiteit Antwerpen.

Karen Markey. 1983. Computer-assisted construction of a thematic catalog of primary and secondary subject matter. *Visual Resources* 3(1):16–49.

David Marr. 1982. *Vision: A computational approach*. San Francisco, Freeman & Co.

Jacob M Marszalek, Carolyn Barber, Julie Kohlhart, and B Holmes Cooper. 2011. Sample size in psychological research over the past 30 years. *Perceptual and motor skills* 112(2):331–348.

Claudio Masolo, Laure Vieu, Emanuele Bottazzi, Carola Catenacci, Roberta Ferrario, Aldo Gangemi, and Nicola Guarino. 2004. Social roles and their descriptions. In *Proceedings*

of the Ninth International Conference on Principles of Knowledge Representation and Reasoning. AAAI Press, KR'04, pages 267–277.

Caterina Masotti, Danilo Croce, and Roberto Basili. 2017. Deep learning for automatic image captioning in poor training conditions. In Giorgio Satta Roberto Basili, Malvina Nissim, editor, *Proceedings of the Forth Italian Conference on Computational Linguistics (CLiC-it 2017)*. CEUR-WS. http://ceur-ws.org/Vol-2006/paper030.pdf.

Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *AAAI*. pages 3574–3580.

Yo Matsumoto. 1995. The conversational condition on horn scales. *Linguistics and philosophy* 18(1):21–60.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 746–751. http://www.aclweb.org/anthology/N13-1090.

George A Miller and J.G. Beebe-Center. 1958. Some psychological methods for evaluating the quality of translations. *Mechanical translation* 3:73–80.

Jim Miller and M. M. Jocelyne Fernandez-Vest. 2006. Spoken and written language. In Giuliano Bernini and Marcia L. Schwartz, editors, *Pragmatic organization of discourse in the languages of Europe*, Berlin ; New York : Mouton de Gruyter, Empirical approaches to language typology. EUROTYP.

Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv:1411.1784* .

Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pages 2930–2939. https://doi.org/10.1109/CVPR.2016.320.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daume III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 747–756. http://www.aclweb.org/anthology/E12-1076.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1780–1790. http://www.aclweb.org/anthology/P16-1168.

Roser Morante, Anthony Liekens, and Walter Daelemans. 2008. Learning the scope of negation in biomedical texts. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, pages 715–724. http://www.aclweb.org/anthology/D08-1075.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation

for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 839–849. http://www.aclweb.org/anthology/N16-1098.

Alison Mountz. 2009. The other. In Carolyn Gallaher, Carl T Dahlman, Mary Gilmartin, Alison Mountz, and Peter Shirlow, editors, *Key concepts in political geography*, Sage.

Andreas Müller. 2015. German word embeddings. Available from GitHub at: http://devmount.github.io/GermanWordEmbeddings/.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2017. Text-guided attention model for image captioning. In *AAAI Conference on Artificial Intelligence*.

Charles Kay Ogden and Ivor Armstrong Richards. 1923. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*, volume 29. K. Paul, Trench, Trubner & Company, Limited.

Chris Olah and Shan Carter. 2016. Attention and augmented recurrent neural networks. *Distill* https://doi.org/10.23915/distill.00001.

Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42(3):145–175.

Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2015. Predicting entry-level categories. *International Journal of Computer Vision* pages 1–15. https://doi.org/10.1007/s11263-015-0815-z.

Susanne Ørnager. 1997. Image retrieval: Theoretical analysis an empirical user studies on accessing information in images. In *Asis' 97: Proceedings of the 60th Asis Annual Meeting, Washington, Dc, November 1-6, 1997*. Information Today, pages 202–211.

Susanne Ørnager and Haakon Lund. 2018. Images in social media: Categorization and organization of images and their collections. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 10(1):i–101. https://doi.org/10.2200/S00821ED1V01Y201712ICR062.

Erwin Panofsky. 1939. *Studies in Iconology: Humanist Themes in the Art of the Renaissance*. Oxford University Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Patrick Paroubek, Stéphane Chaudiron, and Lynette Hirschman. 2007. Principles of evaluation in natural language processing. *Traitement Automatique des Langues* 48(1):7–31.

J. Pearl and D. Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Marco Pennacchiotti, Diego De Cao, Roberto Basili, Danilo Croce, and Michael Roth. 2008. Automatic induction of framenet lexical units. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 457–465.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Helen Petrie, Chandra Harrison, and Sundeep Dev. 2005. Describing images on the web: a survey of current practice and prospects for the future. *Proceedings of Human Computer Interaction International (HCII)* 71.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*. pages 2641–2649.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, New Orleans, Louisiana, pages 180–191. http://www.aclweb.org/anthology/S18-2023.

Ronald Poppe. 2010. A survey on vision-based human action recognition. *Image and Vision Computing* 28(6):976 – 990.

Marten Postma, Ruben Izquierdo, Eneko Agirre, German Rigau, and Piek Vossen. 2016a. Addressing the mfs bias in wsd systems. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016b. Open dutch wordnet. In *Proceedings of the Eighth Global Wordnet Conference*. Bucharest, Romania.

Alexander J Quinn and Benjamin B Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, pages 1403–1412.

Ali Rahimi and Ben Recht. 2017. Reflections on random kitchen sinks. Talk given at the occasion of the NIPS 2017 'test of time' award. Video: https://www.youtube.com/watch?v=Qi1Yry33TQE. Text: https://web.archive.org/web/20180818034101/http://www.argmin.net/2017/12/05/kitchen-sinks/.

Janarthanan Rajendran, Mitesh M Khapra, Sarath Chandar, and Balaraman Ravindran. 2016. Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, pages 139–147.

Dorit Ravid and Ruth A. Berman. 2006. Information density in the development of spoken and written narratives in english and hebrew. *Discourse Processes* 41(2):117–149.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora - Volume 9*. ACL, Stroudsburg, PA, USA, pages 1–6.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics* 44(–):393–401. https://doi.org/10.1162/coli_a_00322.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics* 35(4):529–558.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering* 3(01):57–87.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics* pages 91–136.

Suzanne Romaine. 2001. A corpus-based view of gender in british and american english. *Gender Across Languages: The linguistic representation of women and men* 1:153–175.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology* 8(3):382–439.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature* 323(6088):533.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Disability Studies* 20:33–53.

N. Samet, S. Hiçsönmez, P. Duygulu, and E. Akbaş. 2017. Could we create a training set for image captioning using automatic translation? In *2017 25th Signal Processing and Communications Applications Conference (SIU)*. pages 1–4. https://doi.org/10.1109/SIU.2017.7960638.

Patricia Saylor. 2015. *Spoke: A framework for building speech-enabled websites*. Master's thesis, Massachusetts Institute of Technology.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. L. Erlbaum Associates.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. pages 15–25.

Denise Sekaquaptewa, Penelope Espinoza, Mischa Thompson, Patrick Vargas, and William von Hippel. 2003. Stereotypic explanatory bias: Implicit stereotyping as a predictor of discrimination. *Journal of Experimental Social Psychology* 39(1):75–82.

Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law* 7(4):233–242. https://doi.org/10.1093/idpl/ipx022.

Sara Shatford. 1986. Analyzing the subject of a picture: a theoretical approach. *Cataloging & classification quarterly* 6(3):39–62.

Rakshith Shetty, Hamed R.-Tavakoli, and Jorma Laaksonen. 2016. Exploiting scene context for image captioning. In *Proceedings of the 2016 ACM Workshop on Vision and Language Integration Meets Multimedia Fusion*. ACM, New York, NY, USA, iV&L-MM '16, pages 1–8.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *ICCV*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence* 22(12):1349–1380.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*.

Amanda Song, Linjie Li, Chad Atalla, and Garrison Cottrell. 2017. Learning to see people like people: Predicting the social perception of faces. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Karen Spärck-Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1):11–21. https://doi.org/10.1108/eb026526.

Otfried Spreen and Rudolph W Schulz. 1966. Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior* 5(5):459–468.

Rachele Sprugnoli, Giovanni Moretti, Luisa Bentivogli, and Diego Giuliani. 2016. Creating a ground truth multilingual dataset of news and talk show transcriptions through crowdsourcing. *Language Resources and Evaluation* pages 1–35.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. *Social communication* pages 163–187.

Statistisches Bundesamt. 2013. Zensus 2011: 80,2 millionen einwohner lebten am 9. mai 2011 in deutschland. Press release, Nr. 188. https://www.destatis.de/.

B. Stewart. 2010. Getting the picture: An exploratory study of current indexing practices in providing subject access to historic photographs. *The Canadian Journal of Information and Library Science* 34:297.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995* Accepted for publication as a short paper at EMNLP 2018.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 3104–3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

Zoltán Gendler Szabó. 2017. Compositionality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University. Summer 2017 edition.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hamed R Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying attention to descriptions generated by image captioning models. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pages 2506–2515.

DL Theijssen, H van Halteren, LWJ Boves, and NHJ Oostdijk. 2011. On the difficulty of making concreteness concrete. *Computational Linguistics in the Netherlands Journal* 1:61–77.

Alexander Todorov, Peter Mende-Siedlecki, and Ron Dotsch. 2013. Social judgments from faces. *Current opinion in neurobiology* 23(3):373–380.

Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review* 113(4):766.

Gunnel Tottie. 1980. Affixal and non-affixal negation in English: Two systems in (almost) complementary distribution. *Studia linguistica* 34(2):101–123.

Althea Turner and Edith Greene. 1977. *The construction and use of a propositional text base*. Institute for the Study of Intellectual Behavior, University of Colorado Boulder.

Mesut Erhan Unal, Begum Citamak, Semih Yagcioglu, Aykut Erdem, Erkut Erdem, Nazli Ikizler Cinbis, and Ruket Cakici. 2016. Tasviret: Görüntülerden otomatik türkçe açıklama oluşturma İçin bir denektaçı veri kümesi (tasviret: A benchmark dataset for automatic turkish description generation from images). In *IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU 2016)*.

Roberto Valenti, Nicu Sebe, and Theo Gevers. 2012. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing* 21(2):802–815.

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: A new and improved word frequency database for british english. *The Quarterly Journal of Experimental Psychology* 67(6):1176–1190.

Emiel van Miltenburg. 2016. Stereotyping and bias in the flickr30k dataset. In Jens Edlund, Dirk Heylen, and Patrizia Paggio, editors, *Proceedings of Multimodal Corpora: Computer vision and language processing (MMC 2016)*. pages 1–4.

Emiel van Miltenburg. 2017. Pragmatic descriptions of perceptual stimuli. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, pages 1–10.

Emiel van Miltenburg and Desmond Elliott. 2017. Room for improvement in automatic image description: an error analysis. *arXiv preprint arXiv:1704.04198* .

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*. Association for Computational Linguistics, Santiago de Compostela, Spain, pages 21–30.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Talking about other people: an endless range of possibilities. In *Proceedings of the 11th International Conference on*

Natural Language Generation. Association for Computational Linguistics, pages 415–420. http://aclweb.org/anthology/W18-6550.

Emiel van Miltenburg, Ákos Kádar, Ruud Koolen, and Emiel Krahmer. 2018a. DIDEC: The Dutch Image Description and Eye-tracking Corpus. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*. Resource available at https://didec.uvt.nl.

Emiel van Miltenburg, Ruud Koolen, and Emiel Krahmer. 2018b. Varying image description tasks: spoken versus written descriptions. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Emiel van Miltenburg, Roser Morante, and Desmond Elliott. 2016a. Pragmatic factors in image description: The case of negations. In *Proceedings of the 5th Workshop on Vision and Language*. Association for Computational Linguistics, Berlin, Germany, pages 54–59.

Emiel van Miltenburg, Benjamin Timmermans, and Lora Aroyo. 2016b. The vu sound corpus: Adding more fine-grained annotations to the freesound database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Portorož, Slovenia.

Chantal van Son, Emiel van Miltenburg, and Roser Morante. 2016. Building a dictionary of affixal negations. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 49–56. http://aclweb.org/anthology/W16-5007.

Bob van Tiel. 2014. *Quantity matters: Implicatures, typicality and truth*. Ph.D. thesis, Radboud Universiteit Nijmegen.

Bob Van Tiel, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33(1):137–175. https://doi.org/10.1093/jos/ffu017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 3156–3164.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(4):652–663.

Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How blind people interact with visual content on social networking services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, pages 1584–1595.

Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, and Yueting Zhuang. 2016. Diverse image captioning via grouptalk. In *IJCAI*. AAAI Press, pages 2957–2964.

Martin Weiss, Margaux Luck, Roger Girgis, Chris Pal, and Joseph Paul Cohen. 2018. A survey of mobile computing for the visually impaired. *CoRR* abs/1811.10120. http://arxiv.org/abs/1811.10120.

Ludwig Wittgenstein. 1921/1961. *Tractatus Logico-Philosophicus*. Routledge & Kegan Paul. Translated by David Pears and Brian McGuinness. Available online through http://people.umass.edu/klement/tlp/.

Jennifer Wortman Vaughan. 2018. Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research* 18(193):1–46. http://jmlr.org/papers/v18/17-234.html.

Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. 2017a. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017b. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *CSCW*. pages 1180–1192.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, pages 3485–3492.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. pages 2048–2057.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *The IEEE International Conference on Computer Vision (ICCV)*.

Alfred L Yarbus. 1967. *Eye movements and vision*. Springer.

Victor H Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American philosophical society* 104(5):444–466.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823* .

Jason Yosinski, Jeff Clune, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. In *ICML Workshop on Deep Learning*. https://arxiv.org/abs/1506.06579.

Gilbert Youmans. 1990. Measuring lexical style and competence: The type-token vocabulary curve. *Style* pages 584–599.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67–78.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, pages 818–833.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017a. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2979–2989. https://www.aclweb.org/anthology/D17-1323.

Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2017b. The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. *Proceedings of the ACM on Human-Computer Interaction* 1(CSCW):121.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 487–495.

George Kingsley Zipf. 1949. *Human behaviour and the principle of least effort: an introduction to human ecology.*. Cambridge, MA: Addison-Wesley.

Appendix A

Annotation and inspection tools

A.1 Introduction

How do you search or annotate a corpus of image descriptions? Ideally, we should have a program that displays the images and their descriptions together on the screen. For an annotation tool, it would also be good to have some kind of form, to be able to add or edit information about an image and its descriptions. Since there are few (if any) programs that provide this functionality, I developed several different tools to do this.

In my experience, one of the easiest way to build inspection or annotation tools is to create a small web application. This way, the interface can be created using HTML templates, and it can be viewed in any modern browser (eliminating the need to develop a separate graphical user interface). As a back-end, I usually rely on Flask, a Python module to build small web-apps.[1] These apps can either be hosted locally (with no need for an external server) or online (using a remote host). This also means that it is easy to convert annotation tools into corpus demonstration tools.

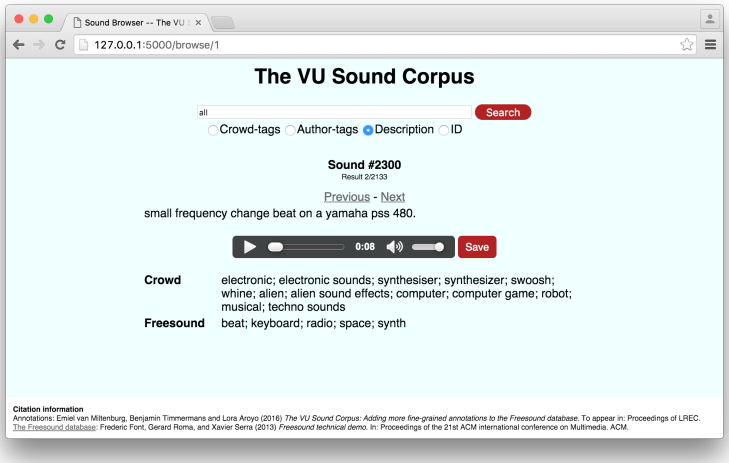


Figure A.1 Screenshot of the browsing tool for the VU Sound Corpus.

A.2 Exploring the VU sound corpus

I developed my first inspection tool to enable others to easily search the VU Sound Corpus (van Miltenburg et al., 2016b), and inspect our data. Figure A.1 shows a screenshot of this tool, which can be downloaded through: https://github.com/evanmiltenburg/SoundBrowser.

[1] See http://flask.pocoo.org

Users can either browse the sounds one-by-one, or search for any of the original tags (provided by the authors of the sounds), crowd-annotated tags, the descriptions, and the sound identifiers. The interface combines information from different sources (audio files and metadata), and allows for quick inspection of the data.

A.3 Annotating image descriptions

The next annotation tool was developed to annotate stereotyping behavior in the Flickr30K corpus (van Miltenburg, 2016). Figure A.2 shows a screenshot of this tool, which can be downloaded through: https://github.com/evanmiltenburg/Flickr30k-Image-Viewer.

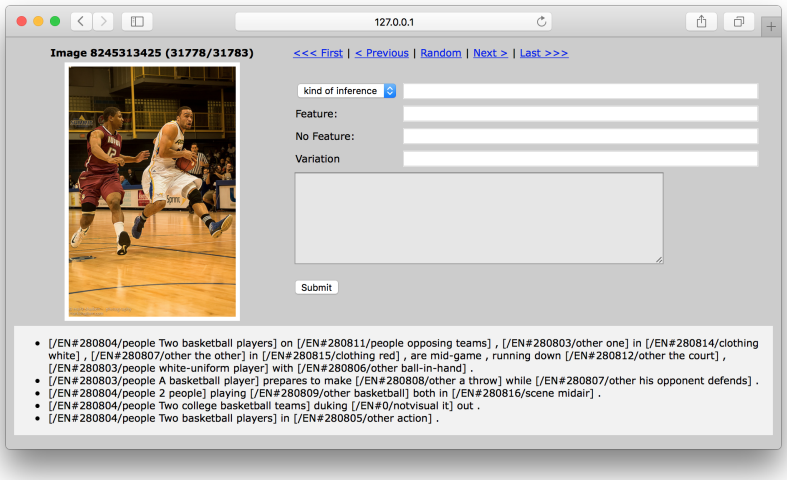


Figure A.2 Screenshot of the annotation tool for the Flickr30K images.

This tool includes a form that is intended to take notes about the images. The form includes a drop-down menu with different kinds of unwarranted inferences, and several text fields to make additional annotations. The color scheme is set in different shades of gray, which is less straining on the eyes than having a white background.

A.4 Annotating negations

We developed another annotation tool to annotate uses of negations in the Flickr30K corpus (van Miltenburg et al., 2016a). Figure A.3 shows a screenshot of this tool, which can be downloaded through: https://github.com/evanmiltenburg/annotating-negations

This tool presents a different workflow, in which users can categorize multiple sentences at the same time. If an annotator recognizes a particular pattern in the data, they can use the search box to find all sentences matching that pattern. They can then select a category for those sentences and annotate the sentences with the same category all at once (after going through the results and deselecting sentences which, on closer inspection, shouldn't fall under the same category). The annotation task is split up into several smaller tasks, one for each different negation term.

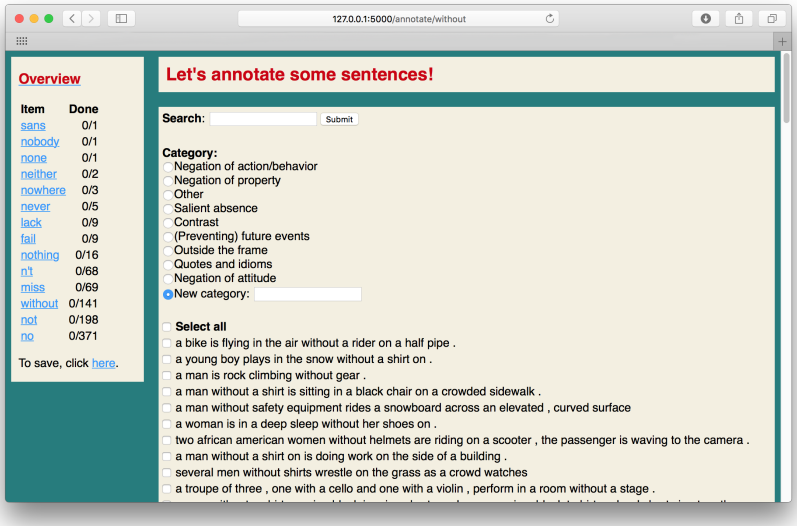


Figure A.3 Screenshot of the annotation tool to categorize different uses of negations.

A.5 Comparing image descriptions across languages

For our next inspection tool, we had to present data from three diffent corpora on a screen, namely: Dutch, German, and English image descriptions (van Miltenburg et al., 2017). Figure A.4 shows a screenshot of this tool, which can be downloaded through: https://github.com/cltl/DutchDescriptions

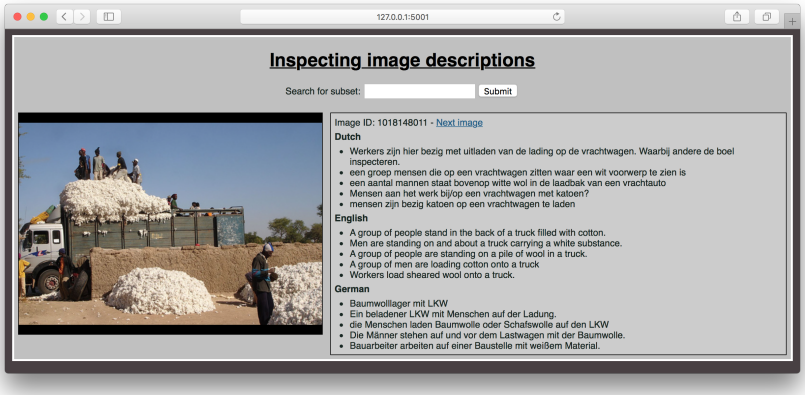


Figure A.4 Screenshot of the inspection tool to compare image descriptions in different languages.

The search box allows users to look up any word or phrase in any of the languages. They will then be taken to a results page, where they can browse through all the images matching the query in any of the descriptions.

A.6 Inspecting spoken image descriptions

For the Dutch Image Description and Eye-tracking Dataset (DIDEC; van Miltenburg et al. 2018a), we developed an inspection tool to browse the spoken image descriptions. Figure A.5 shows a screenshot of this tool, which can be downloaded through: https://didec.uvt.nl/pages/interfaces.html.

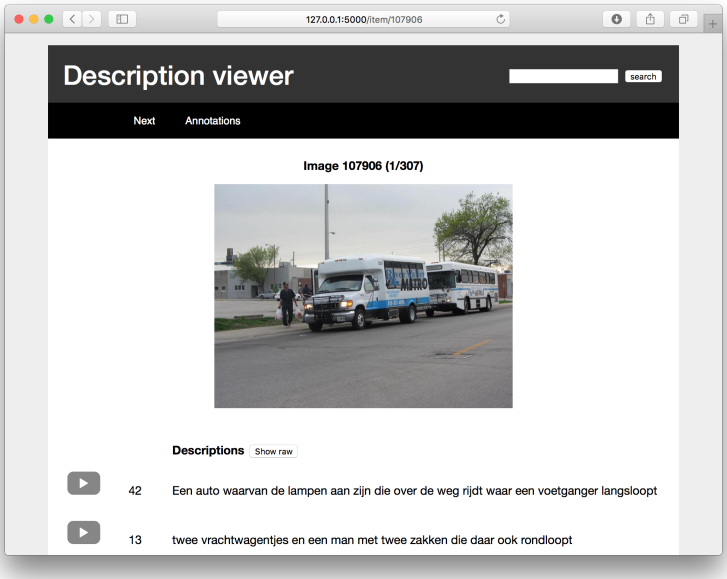


Figure A.5 Screenshot of the inspection tool for the spoken Dutch descriptions.

Users can search for descriptions containing particular words and phrases, and browse through the results. They can also choose to only look at descriptions containing corrections, repetitions, pauses, or filled pauses. Finally, users can toggle between showing and hiding the raw transcriptions for each recording.

Appendix B

Instructions for collecting Dutch image descriptions

B.1 About this appendix

This appendix contains the instructions for the Dutch crowdsourcing task from Chapter 3, translated from Hodosh et al. (2013). Accordingly, the rest of this appendix is in Dutch.

B.2 Prompt

Beschrijf de afbeelding in één volledige, maar eenvoudige zin.

B.3 Richtlijnen

Beschrijf elk van de volgende vijf afbeeldingen met één Nederlandse zin.

- Geef een accurate beschrijving van de activiteiten, mensen, dieren, en objecten die je ziet in de afbeelding.
- Elke beschrijving moet bestaan uit één zin, die maximaal 100 karakters bevat.
- Probeer kort en bondig te zijn.
- Let erop dat de spelling en grammatica van de zinnen in orde is.
- Wij accepteren jouw resultaten als je een goede beschrijving geeft voor alle vijf de afbeeldingen, en als alles ingevuld is.
- Alleen moedertaalsprekers van het Nederlands kunnen meedoen. Gebruikers van Google Translate worden afgewezen.

B.4 Voorbeelden van goede en slechte beschrijvingen.

1. De hond draagt een rode sombrero.
   Heel goed: beide hoofdobjecten worden kort en bondig beschreven.
2. Een witte hond met een rode hoed.
   Acceptabel: een onvolledige zin (met alleen het onderwerp) is acceptabel.
3. De witte hond draagt een roze halsband.
   Acceptabel: de hond wordt beschreven, maar de hoed wordt genegeerd.
4. De rode hoed is versierd met gouden pailletten.
   Slecht: de hond wordt genegeerd.
5. De hond is boos omdat hij honger heeft.
   Slecht: dit is speculatief.
6. Een hond/De hond.
   Zeer slecht: deze beschrijving zou kunnen slaan op elke beschrijving van elke hond. De beschrijving is niet specifiek genoeg.

Appendix C

Instructions for the DIDEC experiments

C.1 Introduction

This appendix provides the instructions and consent forms for the experiments reported as:

Emiel van Miltenburg, Ákos Kádar, Ruud Koolen, and Emiel Krahmer. 2018a. DIDEC: The Dutch Image Description and Eye-tracking Corpus. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*. Resource available at https://didec.uvt.nl

C.2 Instructions

This section presents the instructions for the free viewing task and the description viewing task. Since both experiments were carried out in Dutch, the instructions are in Dutch as well. The instructions for the production viewing task were translated from Hodosh et al. (2013).

C.2.1 Free viewing

In dit laatste deel van het experiment ga je foto's bekijken. Je krijgt zometeen achtereenvolgens zo'n 100 foto's te zien. Iedere foto is in beeld voor drie seconden. Je enige taak is om iedere foto zo nauwkeurig mogelijk te bekijken.

Het is belangrijk dat je de foto's serieus bekijkt: aan het einde van het experiment krijg je een geheugentaak waarin je wordt gevraagd of je bepaalde foto's eerder hebt gezien.

De procedure om naar de volgende foto te gaan is hetzelfde als in het vorige deel van het experiment: er verschijnt steeds eerst een kruis in beeld. Als je vervolgens gedurende 1 seconde naar dat kruis kijkt, verschijnt de foto in beeld, en kan je deze gedurende drie seconden gaan bekijken.

Zodra je je zometeen richt tot het computerscherm, gaan we eerst weer de eyetracker afstemmen op je ogen. Als we dat gedaan hebben, kan je drie keer oefenen met de taak. Daarna begint het eigenlijke experiment.

N.B. Probeer tijdens het experiment zo stil mogelijk te blijven zitten!

C.2.2 Description viewing

In dit tweede deel van het experiment verzamelen we gesproken beschrijvingen van foto's als geheel. Je krijgt zometeen achtereenvolgens ongeveer 100 foto's te zien. Je taak is om iedere foto nauwkeurig te bekijken, en vervolgens te beschrijven wat je ziet. Je kunt simpelweg benoemen wat je opvalt: situaties, gebeurtenissen, maar ook andere dingen die te zien zijn, zoals mensen, dieren of objecten. Het is de bedoeling dat je iedere foto in één zin beschrijft.

Om je een beeld te geven van het soort beschrijvingen dat we verwachten, zie je hieronder twee foto's met een mogelijke beschrijving:

Voorbeeldfoto 1

Mogelijke beschrijving

"De man met de knuppel maakt zich klaar om te gaan slaan terwijl de scheidsrechter toekijkt."

Voorbeeldfoto 2

Mogelijke beschrijving

"Een paard loopt voor een wagen met daarop een grote hoeveelheid hooi en twee mensen."

Bij het geven van de beschrijvingen willen we je vragen om je te houden aan de volgende richtlijnen:

1. Start de beschrijving niet met "Er is…" of "Ik zie…"
2. Beschrijf geen onbelangrijke details.
3. Beschrijf wat je ziet op de foto, dus geen gebeurtenissen die mogelijk hebben plaatsgevonden in verleden of toekomst.
4. Beschrijf niet wat een persoon zou kunnen zeggen.
5. Geef geen namen aan mensen.
6. Voor iedere foto moet je beschrijving minimaal 8 woorden bevatten.

Het is belangrijk dat je de foto's serieus bekijkt: als je klaar bent met het beschrijven van de foto's, krijg je een geheugentaak waarin je wordt gevraagd of je bepaalde foto's eerder hebt gezien. Het experiment start met twee foto's waarmee je de taak kunt oefenen.

N.B. Probeer tijdens het experiment zo stil mogelijk te blijven zitten!

C.3 Consent forms

This section provides the consent forms (in Dutch) for both tasks.

C.3.1 Free viewing: Informatie & Consentverklaring

Titel: Afbeeldingen bekijken
Doel en procedure onderzoek: In dit onderzoek ga je simpelweg afbeeldingen bekijken, waarbij iedere afbeelding drie seconden in beeld is. Het experiment vindt plaats in een geluidsdichte cabine. Je neemt plaats achter de computer. Met een instructie word je voorbereid op de taak die je uit gaat voeren. Na het lezen van de instructie mag je vragen stellen als

je iets niet begrijpt. Je wordt –waar mogelijk– verzocht om tijdens het experiment alleen te communiceren over de taak die je uitvoert. Tijdens het experiment worden je oogbewegingen geregistreerd. Er worden geen video-opnames gemaakt.
Duur onderzoek: het onderzoek duurt ongeveer 20 minuten en je kunt er 0,5 proefpersoonuur mee verdienen.
Privacy en vertrouwelijkheid: Alle data die worden verzameld zullen hoogst vertrouwelijk behandeld worden. Je privacy wordt gewaarborgd. Je naam zal in geen enkel geval verbonden worden aan de resultaten. De data worden tenminste 5 jaar bewaard. Dit is in lijn met de voorgeschreven termijn uit de Nederlandse Gedragscode Wetenschapsbeoefening. Jouw identiteit als proefpersoon is op geen enkele manier te achterhalen.
Vrijwillige deelname: Je loopt geen enkel risico als je aan dit experiment deelneemt, en je deelname heeft dan ook geen negatieve lichamelijke of geestelijke gevolgen. Je kunt geen goede of foute dingen doen. Je bent evengoed niet verplicht om aan dit onderzoek deel te nemen. Op het moment dat je besluit om deel te gaan nemen, kun je op elk moment je deelname aan het onderzoek opzeggen zonder dat dit gevolgen heeft. Je bent niet verplicht om vragen te beantwoorden die je niet wilt beantwoorden, en mag te allen tijde de ruimte verlaten en het experiment afbreken.
Contact: Mocht je na afloop van dit onderzoek nog vragen hebben, dan kun je contact opnemen met de onderzoeksleider, dr. Ruud Koolen. Dit kan direct na afloop van het experiment, maar ook in een later stadium (per telefoon: **********, per e-mail: **********, of in persoon: kamer **********). Voor meer informatie over de richtlijnen waaraan onderzoeken dienen te voldoen, kan je kijken naar het proefpersonenreglement en de ethische richtlijnen onder Course Information van de Proefpersonenpool op Blackboard.
Expliciete toestemming voor het registreren van je stem en oogbewegingen: Hierbij geef ik toestemming…
…om mijn geregistreerde oogbewegingen te gebruiken voor onderzoeksdoeleinden□ Ja □ Nee

Ik heb de gelegenheid gehad deze Informatie & Consentverklaring te lezen en het onderzoek is aan mij uitgelegd. Ik heb de mogelijkheid gehad om vragen te stellen over het onderzoek en mijn vragen zijn beantwoord. Ik ben bereid om te participeren in het onderzoek 'Afbeeldingen bekijken en beschrijven'.

_____ Naam proefpersoon
_____ Handtekening proefpersoon
_____ Man/vrouw
_____ Handtekening proefleider
_____ Datum
_____ Leeftijd
_____ Datum

C.3.2 Description viewing: Informatie & Consentverklaring

Titel: Afbeeldingen bekijken en beschrijven.
Doel en procedure onderzoek: In dit onderzoek ga je afbeeldingen bekijken, en beschrijven wat er te zien is. Het experiment vindt plaats in een geluidsdichte cabine. Je neemt plaats achter de computer. Ieder deel van het experiment verloopt in twee fases: een oefenfase, en de fase waarin je het betreffende deel van het experiment daadwerkelijk doorloopt. In de oefenfase word je iedere keer voorbereid op de taak die je uit gaat voeren. Tijdens en na deze fase mag

je vragen stellen als je iets niet begrijpt. Je wordt –waar mogelijk– verzocht om tijdens het experiment alleen te communiceren over de taak die je uitvoert. Tijdens het experiment worden je oogbewegingen geregistreerd, en worden er geluidsopnames gemaakt van jou als spreker. Er worden geen video-opnames gemaakt.

Duur onderzoek: Het onderzoek duurt ongeveer 60 minuten en je kunt er 1 proefpersoonuur mee verdienen.

Privacy en vertrouwelijkheid: Alle data die worden verzameld – waaronder de geluidsopnames van de spreker – zullen hoogst vertrouwelijk behandeld worden. Je privacy wordt gewaarborgd. Je naam zal in geen enkel geval verbonden worden aan de resultaten. De geluidsopnames worden anoniem opgeslagen (je naam wordt niet vermeld in de bestandsnaam), en tenminste 5 jaar bewaard. Dit is in lijn met de voorgeschreven termijn uit de Nederlandse Gedragscode Wetenschapsbeoefening. Na afloop van het experiment worden de opnames uitgeschreven. Deze uitgeschreven spraak wordt eveneens anoniem opgeslagen, op een dusdanige manier dat jouw identiteit als proefpersoon op geen enkele manier is te achterhalen.

Vrijwillige deelname: Je loopt geen enkel risico als je aan dit experiment deelneemt, en je deelname heeft dan ook geen negatieve lichamelijke of geestelijke gevolgen. Je kunt geen goede of foute dingen doen of zeggen. Je bent evengoed niet verplicht om aan dit onderzoek deel te nemen. Op het moment dat je besluit om deel te gaan nemen, kun je op elk moment je deelname aan het onderzoek opzeggen zonder dat dit gevolgen heeft. Je bent niet verplicht om vragen te beantwoorden die je niet wilt beantwoorden, en mag te allen tijde de ruimte verlaten en het experiment afbreken.

Contact: Mocht je na afloop van dit onderzoek nog vragen hebben, dan kun je contact opnemen met de onderzoeksleider, dr. Ruud Koolen. Dit kan direct na afloop van het experiment, maar ook in een later stadium (per telefoon: **********, per e-mail: **********, of in persoon: kamer **********). Voor meer informatie over de richtlijnen waaraan onderzoeken dienen te voldoen, kan je kijken naar het proefpersonenreglement en de ethische richtlijnen onder Course Information van de Proefpersonenpool op Blackboard.

Expliciete toestemming voor het registreren van je stem en oogbewegingen: Hierbij geef ik toestemming…

…om mijn audio-opnames te gebruiken voor onderzoeksdoeleinden ☐ Ja ☐ Nee

…om mijn geregistreerde oogbewegingen te gebruiken voor onderzoeksdoeleinden ☐ Ja ☐ Nee

Ik heb de gelegenheid gehad deze Informatie & Consentverklaring te lezen en het onderzoek is aan mij uitgelegd. Ik heb de mogelijkheid gehad om vragen te stellen over het onderzoek en mijn vragen zijn beantwoord. Ik ben bereid om te participeren in het onderzoek ‘Afbeeldingen bekijken en beschrijven’.

_____ Naam proefpersoon
_____ Handtekening proefpersoon
_____ Man/vrouw
_____ Handtekening proefleider
_____ Datum
_____ Leeftijd
_____ Datum

Appendix D

Guidelines for error analysis

D.1 Introduction

This document provides guidelines for the annotation of automatically generated image descriptions. Our goal is to assess the semantic competence of image description models. In other words: are the descriptions at least 'technically' correct? This is a low bar, as we ignore fluency and usefulness, which are also desirable properties for an NLG system. We define two tasks:

1. **A binary decision task**, where annotators judge whether or not a description is congruent with an image.

2. **A categorization task**, where annotators select error categories that apply for incongruent descriptions.

These tasks are strongly related: if a description is incongruent, it should fall into one of the error categories, and vice versa. Hence, annotators for either task need to be familiar with our taxonomy of errors.

People	Subject	Object	General	General
Age	Wrong	Wrong	Stance	Scene/event/location
Gender	Similar	Similar	Activity	Other
Type of clothing	Inexistent	Inexistent	Position	Color
Color of clothing	Extra subject	Extra object	Number	Generally unrelated

Table D.1 Error categories for incongruent image descriptions. The organization of these categories corresponds to the organization of the categories in the annotation environment.

D.2 Error categories

All our error categories are provided in Table D.1. There are four main categories: People, Subject, Object, and General. I tried to strike a balance between specificity and amount of categories. No doubt some of these could be further subcategorized, but more categories means the annotation task might become overwhelming.

D.2.1 Short description

Here's a short description of each category, and each of the subcategories. The next subsection provides examples for each of these.

People Image description models often make mistakes that are specific to the description of people. Subcategories are AGE (e.g. *woman* instead of *girl*), GENDER (*man* instead of *woman*),

TYPE OF CLOTHING (*shirt* instead of *jacket*), and COLOR OF CLOTHING (*red shirt* instead of *blue shirt*).

Subject Mistakes relating to the subject of the description. We use the following subcategories: WRONG when the wrong entity in the image is chosen as the subject, SIMILAR when the image description system mis-identifies the subject for something visually similar (e.g. *guitar* instead of *violin*), INEXISTENT when nothing close to the mentioned entity is present in the image, and EXTRA SUBJECT/OBJECT when an additional (nonexistent) entity is mentioned besides the correct entity.

Object See **subject**.

General Mistakes that are not specific to people. The subcategories are as follows: STANCE for posture-related mistakes, ACTIVITY for wrongly identified activities, POSITION for mistakes in spatial relations within the image, NUMBER for any counting errors (too few/many entities mentioned), SCENE/EVENT/LOCATION for misidentifications of the scene, event, or location, COLOR for non-clothing entities that are mistakenly said to have a particular color, OTHER for any unforeseen mistakes, and GENERALLY UNRELATED for generally unrelated descriptions, that are beyond repair. This is usually the case when more than 2–3 error (sub)categories are applicable.

D.2.2 Examples

A **man** is climbing a rock
Category: Age

A **girl** playing soccer
Category: Gender

A girl in a yellow **shirt** is standing on the beach
Category: Type of clothing

A man in a **blue** shirt and blue jeans is working on a ladder
Category: Color of clothing

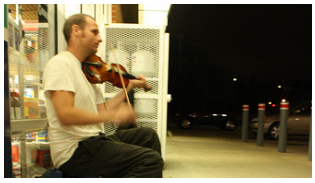
A **boy** jumps over a hurdle
Category: Wrong subject

A woman in a blue shirt is standing in front of a blue car
Category: Inexistent subject

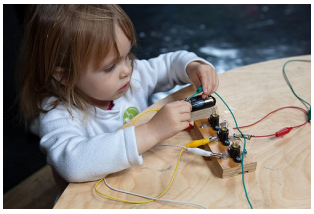
Two police officers are posing for a picture
Category: Similar subject, number

A man in a white shirt **and a man in a white shirt** are preparing food
Category: Extra subject

A young boy is holding a **little girl**
Category: Wrong object



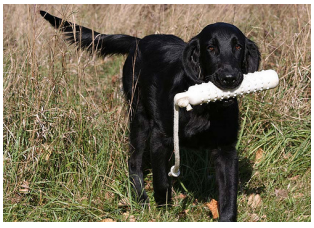
A man is playing a **guitar**
Category: Similar object



A young girl in a white shirt is playing with a **guitar**
Category: Inexistent object

A man with a tennis racket **and a tennis racket**
Category: Extra object

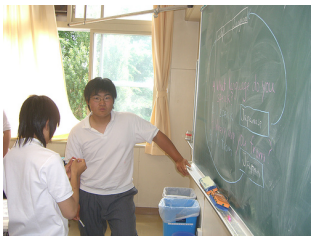
A man in a brown jacket is **standing** in front of a wall
Category: Stance



A black dog **runs through** the grass
Category: Activity

Two men are playing instruments
Category: Number

A little girl in a white dress is walking **in** the water
Category: Position



A man in a white shirt and a woman in a white shirt are standing **in a hallway**
Category: Scene/event/location

A black **and white** dog is playing in the snow
Category: Color

A group of people standing in the snow
Category: Generally unrelated

A group of people are standing in **a fire**
Category: Other

D.2.3 Important contrasts

While the categories are fairly straightforward, there are cases where it is easy to get confused between a pair of categories. Here are additional guidelines for difficult cases that I have encountered.

- STANCE versus ACTIVITY: Use the former when the difference is static, e.g. *standing* vs. *sitting*. Use the latter if the difference is dynamic, e.g. *standing* versus *walking*.
- SCENE/EVENT/LOCATION versus POSITION: Use the former when the surroundings are not correct. Use the latter when position within the surroundings is not correct.
- EXTRA SUBJECT/OBJECT versus NUMBER: Use the former when the subject/object is wrongfully extended with a conjunction (e.g. *and a woman in a white shirt*). Use the latter when there's a general mismatch in number (*a, one, two, three, a group of*).
- SIMILAR OBJECT versus POSITION: This conflict arises in cases where e.g. *... is sitting on a bench* is used instead of *... is sitting on a chair*. In all these cases, use *similar object*. (Even if there is an actual bench in the image.)

D.3 Task descriptions & instructions

Now that we have seen the different error categories, we can describe the two main tasks as follows:

Task 1: Congruency Judge whether the generated description is congruent (no error categories apply) or incongruent (at least one error category applies).

Task 2: Categorizing incongruent descriptions Annotate the 'semantic edit distance' between the generated description and the closest valid description that you can imagine. Tick all the error categories corresponding to the things you would have to change. If the generated description is unrelated to the image, or if you feel that there are too many changes necessary to get to a valid description, select GENERALLY UNRELATED.

The threshold for when a description is generally unrelated is undefined. In general, I feel like type/color of clothing don't really hurt the relation between description and image as much as e.g. having the wrong verb. So it all comes down to your intuition.

D.4 Evaluation: correcting the errors

This is a separate task that serves both as an evaluation of Task 2, and as an indication of system performance if all errors identified in Task 2 are addressed. The correction task works as follows.

1. Select an error type to correct. E.g. COLOR OF CLOTHING.
2. Go through all images annotated with this type, and correct *only* the relevant error.
3. When all relevant errors are corrected, we evaluate the results using BLEU/Meteor.

It is important for this task to be conservative in editing the descriptions. Try to change as little as possible. If a change would require restructuring the entire sentence, leave the description as it is. We'd rather underestimate than overestimate the improvement from fixing the errors. Otherwise we'd just be evaluating how good humans are at writing descriptions. So e.g. for colors, *only* change color terms into other color terms. For gender, only change *man* ↔ *woman* and *boy* ↔ *girl*, not *man* ↔ *girl*. That would be changing the age along with the gender.

Glossary

AlexNet A deep convolutional neural network model that won the 2012 ImageNet Large-Scale Visual Recognition Challenge, beating the competition by 10% (Krizhevsky et al., 2012). This achievement convinced many researchers to use convolutional neural network-based models for computer vision.

Annotation The process of providing data with additional information about its contents, usually by labeling or describing the data.

Attributive adjective Adjective that is used in the prenominal position (*the **good** book*) rather than postnominal (*the book is **good***).

Attention mechanism Part of sequence modeling neural networks that learns 'where to look' in the input data for every step in the generation process.

Bias tendency to describe one social group (e.g. black people) differently than another social group (e.g. white people), even though both groups are comparable, and there isn't a particular reason to treat the groups differently.

BLEU An n-gram based sentence similarity metric, commonly used to evaluate machine translation and image description systems.

Bounding box A set of coordinates (usually forming a rectangle) that enclose an object or entity in an image.

CIDEr Stands for Consensus-based Image Description Evaluation Vedantam et al. (2015). This n-gram-based metric compares the generated description with the reference descriptions, discounting popular words (using TF-IDF).

Clustering The process of ordering a collection of data points into groups. Examples of clustering algorithms are *k-nearest neighbour* (grouping data points into k clusters based on their proximity to each other) and the Louvain method.

Competence-Performance distinction Distinction drawn by Chomsky (1965) between language behavior (performance), and language as a cognitive system (competence). Chomsky argued that linguistics should focus on the latter, in analogy to physicists studying (idealized) models of reality rather than reality itself. The goal of linguistics, then, is to find a grammar model that is able to generate all and only possible sentences of a given language.

Computational linguistics The branch of linguistics that uses computational approaches to study and model natural language.

Consciousness-of-projection terms Words that indicate the certainty that an observer has about their interpretation of a particular situation. For example: *apparently, appear, appears, certainly, clearly, definitely*.

Convolutional Neural Network (CNN) A type of (deep) neural network that is specifically designed to take two-dimensional data (usually images) as input. CNNs are often used to extract *image features* that are useful for further computation.

Corpora Plural form of *Corpus*.

Corpus A (large) body of data. This work mainly uses corpora of annotated images.

Crowdsourcing Outsourcing small jobs to online *crowd workers*, through services like Mechanical Turk, Crowdflower, or Prolific. Often used for online surveys and annotation tasks.

Crowd workers People who carry out crowdsourcing tasks. Anyone can register an account with a crowdsourcing platform and do these jobs from their home.

Deep learning Machine learning with neural networks containing many hidden layers. The size of these models means that they have a large amount of connection weights, and the optimization of these weights requires large amounts of data.

Description specificity The level of specificity for a particular description. Descriptions with narrower terms (e.g. *labrador*) are more specific than those using broader terms (e.g. *animal*).

Diversity The amount of variation in a corpus. This thesis recognizes two subkinds: *local* and *global* diversity.

Downstream task A downstream task is a task that depends on systems or models trained for another, more basic task.

Error analysis The process of identifying the mistakes that a system makes, and ordering those mistakes into coherent subgroups. This categorization reveals the distribution of the different kinds of errors, so that we know (if we used a representative sample) which errors occur most often, and which occur less frequently.

Eye-tracking Measuring human participants' eye movements, as they are looking at a computer screen.

Feature extractor A system that produces meaningful representations for some input.

Feature vector A vector representing important features for some relevant input, that are useful for further computation.

Flickr8K Image description corpus, consisting of 8000 images, with 5 descriptions per image (Hodosh et al., 2013).

Flickr30K Image description corpus, consisting of 30,000 images, with 5 descriptions per image (Young et al., 2014). This corpus is also provided with entity annotations.

Free-viewing Watching different images without any objective.

Generative Adversarial Network (GAN) GANs are pairs of networks that are trained by having them compete against each other. The Generator network tries to produce realistic (or human-like) output, while the Discriminator network tries to distinguish between actual examples and generated examples. Researchers are usually interested in the former network.

Global diversity Variation computed over a whole corpus of image descriptions, rather than on an image-by-image basis.

Global recall The amount of different word types that are produced by an image description system, relative to the amount of different types produced by humans.

Iconography The second level of Panofsky's meaning hierarchy: giving a more specific description of the image, also using information about the historical and cultural context in which the image was produced.

Iconology The third level of Panofsky's meaning hierarchy: interpreting the image, establishing its cultural and intellectual significance.

Image features One or more feature vectors that represent the contents of an image.

Image specificity The amount of variation (in image descriptions) that is elicited by an image. Specific images lead to a lower diversity in the associated descriptions.

Linguistic bias A bias in language use that is visible through the distribution of terms used to describe entities in a particular category, as compared to entities from an *a priori* comparable category (e.g. black versus white people).

Local diversity The variation in image descriptions generated for one specific image.

Local recall The amount of different word types that are produced by an image description system for a specific image, relative to the amount of different types produced by humans for that same image. Words that are used by multiple annotators can be said to have a higher importance than words that are only used by a single annotator.

Long Short-Term Memory (LSTM) A type of recurrent neural network that is better at capturing longer dependencies within a sequence. Since natural language is full of such dependencies (e.g. verb agreement), LSTMs are a popular choice to model sentences (either as sequences of words or as sequences of characters).

Louvain method A network clustering method developed by Blondel et al. (2008).

METEOR Metric for Evaluation of Translation with Explicit ORdering. An n-gram based similarity metric that is used to evaluate automatic image descriptions (Banerjee and Lavie, 2005; Denkowski and Lavie, 2014). It is similar to BLEU and ROUGE but adds the ability to match synonyms and paraphrases, using WordNet and a paraphrase table.

Mean-segmental type-token ratio (MSTTR) The mean Type-Token Ratio (TTR), computed over multiple windows of a fixed number of tokens (typically 100 or 1000).

Multitask learning (MTL) A machine learning strategy to use training signals from multiple (related) tasks to make a model generalize better on a particular task, through the use of shared representations between tasks.

Multilayer perceptron (MLP) A neural network with at least three layers: an input layer, one or more hidden layers, and an output layer

MS COCO A large collection of images annotated with 5 descriptions per image, and labels for 80 object categories.

Natural language generation (NLG) A subfield of natural language processing that is concerned with the production of natural language.

Natural language processing (NLP) A branch of computer science and artificial intelligence that aims to build systems to process natural language.

Negation Expression signaling that something is not the case.

Neural Network A machine learning approach based on artificial neurons that are connected to each other (loosely inspired by the human brain). Information flow between the neurons is modulated by *weights* that determine how strongly the signal from one neuron should be transmitted to another. Neural networks can be *trained* by using example $\langle$*input, output*$\rangle$ pairs, and optimizing the weights in such a way that the result for a particular input is close to the expected output.

Of/About-distinction Distinction between what a picture is *Of* and what it is *About*. Shatford (1986) argues that the first two levels of Panofsky's meaning hierarchy consist of these two aspects. 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.

Panofsky's meaning hierarchy A distinction between three levels of understanding, originally developed by Erwin Panofsky (1939) in the context of renaissance paintings, but now more broadly applied. The three levels are: pre-iconography, iconography, and iconology.

Perspective-taking Deciding how to frame the description of a particular situation.

Pragmatics The study of how language is used, and how that use provides utterances with an additional layer of meaning.

Pragmatic gap The gap between what is visually identifiable in an image, and what people choose to report about the image. This is related to *content determination* in the classic Natural Language Generation pipeline Reiter and Dale (1997).

Pre-iconography The first level of Panofsky's meaning hierarchy: giving a low-level description of the contents of a picture (factual description), and the mood it conveys (expressional description).

Production-viewing Watching different images without the objective to produce image descriptions.

Propositional Idea Density (PID) the average number of propositional ideas per word in a text (Turner and Greene, 1977). It is believed that written language has a higher PID than spoken language; fewer words are used to express the same amount of ideas.

Pseudo-quantifier A word that is "loosely indicative of amount or size" (DeVito, 1966), such as: *few, lots, many, much, plenty, some* and *a lot*.

Recall Amount of items that are retrieved, relative to a set of relevant words that could have been retrieved.

Recurrent Neural Network (RNN) A type of neural network that not only produces an output vector, but also passes information to itself from one time step to the next. This makes RNNs useful to model sequential information.

ROUGE Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004). An evaluation metric that computes the extent to which the hypothesis overlaps with the references, using a recall-based approach. In other words, ROUGE asks: how much of the information in the references is also captured by the hypothesis?

Self-reference terms Words like *I, me*, and *my* that refer to the speaker.

Semantic gap the difference in the amounts of information that people and machines can extract from an image.

Shared task A competition for researchers to build a system to perform a particular task (developed by the organizers). All teams run their system on the same data, so that they can compare their results and see which techniques perform best on the task.

SPICE Semantic Propositional Image Caption Evaluation, an evaluation metric proposed by Anderson et al. (2016). The difference between SPICE and other metrics is that SPICE

converts the reference descriptions into a scene graph, and uses this graph (rather than textual similarity) for the evaluation.

Stereotype 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.

TER Translation Edit Rate. An metric for machine translation evaluation, proposed by Snover et al. (2006).

TF-IDF Term frequency–Inverse document frequency. This is a measure of term importance. Term frequency refers to the number of times a term occurs in a document. Inverse Document Frequency was proposed as a measure of informativeness by Karen Spärck-Jones (1972), who observed that terms that occur in *all* documents do not provide any distinguishing information. TF-IDF is used in the CIDEr metric to give more importance to informative words, in the evaluation of image descriptions.

Three waves of AI The idea that the development of Artificial Intelligence is coming in waves. The earliest wave was based on *rule-based systems*, followed by a wave of *statistical learning*, and we are now awaiting the third wave of *context-sensitive systems* that are able to explain their decisions.

Token An instance of a word or n-gram.

Tri-level hypothesis The hypothesis (by David Marr) that a *full* description of any cognitive system requires an explanation at three levels:

1. The computational level: what problem is the system solving?
2. The algorithmic level: how does it actually solve the problem?
3. The implementational level: how is this algorithm physically realized?

This hypothesis makes the assumption that *cognition is information processing*, a key assumption that stems from the *Cognitive Revolution* in the 1950s.

Type A word or n-gram. Types can be opposed to *tokens*, which are specific *instances* of words or n-grams.

Type-Token Ratio (TTR) The number of *Types*, divided by the number of *Tokens*. Compare with the Mean-Segmental Type-Token Ratio.

Unwarranted inference An inference that is plausible given the situation, but that is not justified by the facts at hand.

Vector A mathematical object that you can think of as a list of numerical values, that can be used to represent the meaning of words or the contents of an image in a high-dimensional vector space.

Vector space Formally, a collection of vectors. We can reason about the meaning of words in terms of word vectors that are embedded in a high-dimensional 'meaning space'. Reasoning takes place by performing mathematical operations using the vectors in this space. The most famous example is the analogy *man is to woman as king is to* ... (queen). Mikolov et al. (2013b) have shown that the vector spaces generated using the `word2vec` algorithm allow us to solve this analogy by computing: $\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}}$. The result of this operation is a vector that is close to the embedding of *queen*.

VGG An image classification model from the Visual Geometry Group at the University of Oxford. The model was published by Simonyan and Zisserman (2015), and has been used in

automatic image description for the internal representation that it builds up as it tries to classify an image. This representation can also be used as an input for image description systems, to condition the language model used to generate the descriptions.

Word Mover's Distance (WMD) Word Mover's Distance is a measure of document similarity, developed by Kusner et al. (2015). It was repurposed as an image description evaluation metric by Kilickaya et al. (2017).

WordNet A database that organizes lemmas through lexical relations between *synsets* (synonym sets). Examples of relations are hyponymy (DOG is a kind of ANIMAL) and antonymy (HOT is the opposite of COLD). The best-known wordnet is Princeton WordNet Fellbaum (1998), but there are also other wordnets, such as Open Dutch WordNet (Postma et al., 2016b) and GermaNet (Hamp and Feldweg, 1997).

Summary

This thesis aims to deepen our understanding of the gap in performance between humans and machines, for the task of (automatic) image description. In image recognition, this gap has been referred to as the *semantic gap* (Smeulders et al., 2000), but in image description there is also a *pragmatic* component because one has to decide which parts of the image are relevant to describe.

Chapter 1 provides the theoretical framework for this thesis, and discusses the main research question in terms of the gap between human and machine performance.

Chapter 2 presents an overview of the different properties of human-generated image descriptions. This overview is based on two different datasets of human-described images: Flickr30K (Young et al., 2014) and MS COCO (Lin et al., 2014). A key assumption behind these datasets is that the descriptions are objective, and don't contain any form of speculation. But looking at the descriptions, we find that they are very diverse (revealing the many different choices that speakers have to make when producing a description), and contain different kinds of stereotypes and biases. Thus, we have to conclude that image description data, at least in the Flickr30K and MS COCO datasets, is subjective. Chapter 2 also coins the term *unwarranted inference* for those descriptions that go beyond what can be derived from the image itself.

Chapter 3 looks at image descriptions in other languages. Specifically, this chapter looks at the differences and similarities between Dutch, English, and German descriptions. Chapter 3 also describes the collection of a dataset of written Dutch image descriptions for the Flickr30K validation and test sets. Looking at the data, Dutch and German image descriptions exhibit the same phenomena that were described in Chapter 2 (i.e. bias, unwarranted inferences). We can thus conclude that the image description task as it is currently construed seems to lead participants to provide subjective descriptions. Next to the observation that Dutch, English, and German all show signs of subjectivity, Chapter 3 also finds differences between the descriptions in the three languages: speakers from different countries (the Netherlands, Germany, and the United States of America) provide descriptions at different levels of specificity, depending on their familiarity with the scenes, locations, and objects depicted in the images. This shows that background knowledge plays an important role in human image description.

Chapter 4 looks at image description as a dynamic process. Rather than study the results of the image description task, this chapter uses an eye-tracking experiment to study image descriptions as they are generated. Chapter 4 describes the collection of DIDEC: the Dutch Image Description and Eye-tracking Corpus, and provides a preliminary analysis of the data. Evidence from speech errors shows that people produce image descriptions as they are interpreting the image. During this process, they make predictions about what the image is likely to be about (again, using their background knowledge). When the predictions are wrong, speakers self-correct to provide a correct description of the image. Finally, speakers may also make their descriptions more precise, so as to avoid any ambiguities for the hearer. These processes are hidden from us when we just look at the end product of the image description process, but they provide useful information about how people actually produce image descriptions. Furthermore, these findings again highlight the need for background knowledge in the image description process.

Chapter 5 provides a discussion of task effects on image descriptions. We know from the preceding chapters that the canonical image description task elicits a wide range of different descriptions, that are often subjective, and that may differ depending on the language of the task or the past experiences of the participants. This chapter presents an overview of all the factors that may influence the image description process (taken from Biber 1988), and focuses on the differences between spoken and written language. An exploratory study shows that spoken image descriptions may differ from written image descriptions: spoken descriptions are likely to be longer, contain more adverbs, pseudo-quantifiers, and allness terms, and speakers are more likely to "show themselves" in their descriptions (as evidenced by their use of self-reference terms and consciousness-of-projection terms).

With Chapters 2–5, this thesis provides a general impression of human image descriptions. In short: they are diverse, often subjective, and regularly show signs of (pragmatic) reasoning and participants' reliance on background knowledge. The second part of this thesis (chapters 6 and 7) provide an indication of the current performance of automatic image description systems.

Chapter 6 provides a general introduction to automatic image description systems, and how they are typically evaluated. Its main contribution is an in-depth error analysis of the output of a well-known image description system (Xu et al., 2015). Chapter 6 shows the difficulty of categorizing flawed image descriptions, because they are often ambiguous in the sense that they can be interpreted as being the result of different kinds of mistakes. Regardless of the exact nature of these errors, it is clear that the mistakes that the model makes are unlike the mistakes that any human would make.

Chapter 7 aims to characterize the diversity of the descriptions generated by humans and machines. The chapter provides an overview of the existing ways to measure diversity, and presents several additional metrics (both generally applicable metrics, as well as metrics that are specifically geared to image description). Chapter 7 shows that human-generated image descriptions are much more diverse than automatically generated image descriptions, and that GAN-based systems seem to produce more diverse descriptions than models trained without an adversarial objective. The takeaway from this chapter is twofold. First, it tells us that diversity is a multifaceted property that can and should be measured in different ways. Focusing on only one diversity metric means that you lose sight of other aspects of diversity. Second, this chapter shows us that there is still much room for improvement in the generation of more diverse (and thus more human-like) image descriptions.

Conclusion. As a whole, this thesis provides a more thorough characterization of the problem of how to generate image descriptions. Looking at the subjective nature of human-generated image descriptions, it seems clear that we probably shouldn't want machines to copy all human image description behavior. But then the question arises: what *should* automatically generated image descriptions look like? As with most scientific questions, the answer is: it depends. This thesis makes two suggestions: 1. Take the cognitive complexity of the descriptions into account. If you want to develop a system that produces descriptions that require high-level reasoning (for example, descriptions containing negations), then the system architecture should support this kind of reasoning. Alternatively, you could also choose to focus on easier descriptions. But whatever you choose, it should be a conscious choice. 2. Talk to the users (for example, blind or visually impaired people, or users of virtual assistants such as Siri or Alexa), identify their needs, and develop image description guidelines to match their needs. Future research should (continue to) try and understand users' needs, and be honest about what image description systems can and cannot do. This requires a deep understanding of the linguistic aspects of image description, for which this thesis provides a starting point.

Samenvatting in het Nederlands

Het doel van dit proefschrift is om beter te begrijpen hoe mensen en computers verschillen in hun vermogen om afbeeldingen te beschrijven. Het verschil tussen mensen en computers wordt in de literatuur over automatische beeldherkenning ook wel *the semantic gap* genoemd (Smeulders et al., 2000). Afhankelijk van hoe optimistisch of pessimistisch je bent over de kwaliteit van automatische beeldherkenning op dit moment, kun je dat vertalen als 'het semantische verschil' of 'de semantische kloof.' Bij automatische beeld*beschrijving* komt daar nog een uitdaging bij: naast het begrip van de afbeelding, wordt het systeem ook gevraagd om een keuze te maken over wat er relevant genoeg is om te beschrijven, en hoe dat dan beschreven moet worden. Tussen mens en computer zit er momenteel een flinke pragmatische kloof.

Hoofdstuk 1 geeft een algemene inleiding, en bespreekt de hoofdvraag uit dit onderzoek in termen van de kloof tussen mensen en machines. De rest van dit proefschrift bestudeert eerst (hoofdstuk 2–5) hoe mensen afbeeldingen beschrijven, en vervolgens (hoofdstukken 6 en 7) hoe computers dat doen.

Hoofdstuk 2 geeft een overzicht van de verschillende eigenschappen van door mensen gegenereerde beschrijvingen van afbeeldingen. Dit overzicht is gebaseerd op twee verschillende datasets van afbeeldingen die door mensen beschreven zijn: Flickr30K (Young et al., 2014) en MS COCO (Lin et al., 2014). Een belangrijke aanname achter deze datasets is dat de beschrijvingen objectief zijn en geen enkele vorm van speculatie bevatten. Maar als we naar de beschrijvingen kijken, zien we dat ze heel divers zijn (wat al laat zien dat er veel verschillende keuzes zijn die sprekers moeten maken bij het produceren van een beschrijving), en dat de beschrijvingen verschillende soorten stereotypen bevatten, en verschillende bevolkingsgroepen anders behandelen. Dat leidt ons tot de conclusie dat bestaande datasets met door mensen gegenereerde beschrijvingen (of in ieder geval Flickr30K en MS COCO) subjectief zijn. Hoofdstuk 2 introduceert ook de term *unwarranted inference* ('ongegronde gevolgtrekking') voor beschrijvingen die gebaseerd zijn op aannames over de afbeeldingen, in plaats van op de afbeeldingen zelf.

Hoofdstuk 3 gaat in op beschrijvingen in andere talen. Specifiek kijkt dit hoofdstuk naar de verschillen en overeenkomsten tussen Nederlandse, Engelse en Duitse beeldbeschrijvingen. Hoofdstuk 3 beschrijft ook de verzameling van een dataset met geschreven Nederlandse beeldbeschrijvingen voor de validatie- en testset van de Flickr30K-data. Als we naar deze data kijken, vertonen Nederlandse en Duitse beeldbeschrijvingen veel overeenkomsten met de Engelse beschrijvingen uit hoofdstuk 2; net als bij de Engelse data, bevatten de Nederlandse en Duitse beschrijvingen vaak speculaties, en zien we ongelijkheden in de manier waarop verschillende bevolkingsgroepen worden beschreven. Het lijkt er dus op, dat de standaard beeldomschrijvingstaak aanleiding geeft om subjectieve beschrijvingen te produceren. Naast de overeenkomsten tussen de Nederlandse, Engelse, en Duitse beschrijvingen, worden er in hoofdstuk 3 ook verschillen gevonden: sprekers van de verschillende talen lijken specifiekere beschrijvingen te geven voor scènes, locaties en objecten die hen bekend voorkomen. Dit laat zien dat achtergrondkennis een belangrijke rol speelt bij het beschrijven van afbeeldingen.

Hoofdstuk 4 beschouwt beeldbeschrijving als een dynamisch proces. In plaats van de resultaten van de beeldbeschrijvingstaak te bestuderen (zoals in hoofdstuk 2 en 3), wordt er in

dit hoofdstuk een eye-tracking-experiment gebruikt om de beschrijvingen van de afbeelding te bestuderen terwijl deze worden gegenereerd. Hoofdstuk 4 beschrijft de verzameling van DIDEC: *the Dutch Image Description and Eye-tracking Corpus* (een corpus van gesproken beschrijvingen, met opnames van de oogbewegingen van de participanten terwijl ze de afbeeldingen beschrijven). Uit de versprekingen die mensen maken tijdens het beschrijven van de afbeeldingen, kunnen we afleiden dat ze al beginnen te praten voordat ze de afbeeldingen volledig geinterpreteerd hebben. Tijdens het beschrijvingsproces maken ze voorspellingen over waar de afbeelding waarschijnlijk over gaat (op basis van hun achtergrondkennis). Als die voorspellingen verkeerd zijn, corrigeren sprekers zichzelf om tot een foutloze beschrijving te komen. Ten slotte kunnen sprekers ook hun beschrijvingen specifieker maken, om dubbelzinnigheden voor de toehoorder te voorkomen. Deze observaties blijven voor ons verborgen als we alleen naar het eindproduct van de beschrijvingstaak kijken, en laten daarmee de meerwaarde zien van het bestuderen van gesproken beschrijvingen: *real-time* data biedt nuttige informatie over hoe mensen daadwerkelijk beschrijvingen produceren. Bovendien benadrukken deze bevindingen opnieuw de behoefte aan achtergrondkennis in het beeldbeschrijvingsproces.

Hoofdstuk 5 geeft een overzicht van verschillende taakeffecten op beeldbeschrijvingen. We weten uit de voorgaande hoofdstukken dat de canonieke beeldbeschrijvingstaak zorgt voor een diverse verzameling van beschrijvingen, die vaak subjectief zijn, en daarnaast ook afhankelijk zijn van de taal of de eerdere ervaringen van de participanten. Dit hoofdstuk geeft een overzicht van alle factoren die van invloed kunnen zijn op het beeldbeschrijvingsproces (gebaseerd op eerder werk van Biber 1988), en richt zich op de verschillen tussen gesproken en geschreven taal. Een verkennend onderzoek toont aan dat gesproken beschrijvingen lijken te verschillen van geschreven beschrijvingen: gesproken beschrijvingen zijn vaak langer, bevatten meer bijwoorden, pseudo-kwantoren en universele kwantoren, en sprekers zullen zichzelf eerder "laten zien" in hun beschrijvingen (onder andere door te verwijzen naar zichzelf, of door aan te geven hoe zeker ze zijn van hun interpretatie).

Met hoofdstuk 2 – 5 geeft dit proefschrift een algemeen beeld van de manier waarop mensen afbeeldingen beschrijven: menselijke beschrijvingen zijn divers, te begrijpen als het resultaat van een (pragmatisch) redeneerproces, en zijn afhankelijk van de achtergrondkennis van de participanten. Het tweede deel van dit proefschrift (hoofdstuk 6 en 7) geeft een beeld van de huidige prestaties van automatische beeldbeschrijvingssystemen. Deze gegenereerde beschrijvingen zijn minder divers en de systemen maken vaak fouten. Tot op zekere hoogte zijn deze eigenschappen terug te voeren op de manier waarop de systemen ontworpen zijn.

Hoofdstuk 6 geeft een algemene inleiding in de techniek achter automatische beeldbeschrijvingssystemen en hoe deze doorgaans worden geëvalueerd. De belangrijkste bijdrage is een analyse van de fouten in de uitvoer van een bekend beeldbeschrijvingssysteem (Xu et al., 2015). Hoofdstuk 6 laat zien hoe lastig het is om gebrekkige beschrijvingen te categoriseren, omdat ze vaak dubbelzinnig zijn; ze kunnen worden geïnterpreteerd als het resultaat van verschillende soorten herkenningsfouten. Desalniettemin is het duidelijk dat de fouten die het systeem van Xu et al. (2015) maakt, verschillen van de fouten die een mens zou maken.

Hoofdstuk 7 heeft tot doel de diversiteit van beschrijvingen die door mensen en machines worden gegenereerd, te karakteriseren. Het hoofdstuk biedt een overzicht van de bestaande manieren om diversiteit te meten en presenteert verschillende nieuwe metrieken (zowel algemeen toepasbaar als metrieken die specifiek zijn toegespitst op beeldbeschrijving). Hoofdstuk 7 laat zien dat door mensen gegenereerde beeldbeschrijvingen veel gevarieerder zijn dan automatisch gegenereerde beeldbeschrijvingen, en dat GAN-gebaseerde systemen meer diverse beschrijvingen lijken te produceren dan modellen die zijn getraind zonder GAN. De boodschap van dit hoofdstuk is tweeledig. Ten eerste laat dit hoofdstuk zien dat diversiteit een veelzijdige

eigenschap is die op verschillende manieren kan en moet worden gemeten. Door je te concentreren op slechts één diversiteitsmetriek, verlies je andere belangrijke aspecten van diversiteit uit het oog. Ten tweede laat dit hoofdstuk ons zien dat er nog veel ruimte is voor verbetering bij het genereren van meer diverse beeldbeschrijvingen.

Samenvattend biedt dit proefschrift een grondiger karakterisering van de uitdaging om automatisch menselijke beschrijvingen van afbeeldingen te genereren. Kijkend naar de subjectiviteit van door mensen gegenereerde beeldbeschrijvingen, kunnen we concluderen dat het niet wenselijk is om al het menselijke gedrag te kopiëren. Maar dan rijst de vraag: hoe zouden automatisch gegenereerde beeldbeschrijvingen er dan uit moeten zien? Zoals bij de meeste wetenschappelijke vragen, is het antwoord: het hangt ervan af. Dit proefschrift doet twee suggesties: 1. Houd rekening met de cognitieve complexiteit van de beschrijvingen. Als je een systeem wil ontwikkelen dat cognitief veeleisende beschrijvingen produceert (bijvoorbeeld met negaties), moet de systeemarchitectuur dat ook ondersteunen. Je kunt er ook voor kiezen om je te concentreren op eenvoudigere beschrijvingen. Maar wat je ook kiest, het moet een bewuste keuze zijn. 2. Praat met de gebruikers (bijvoorbeeld blinden en slechtzienden, of mensen die een virtuele assistent zoals Siri of Alexa gebruiken), identificeer hun behoeften en ontwikkel richtlijnen die daarop aansluiten. Toekomstig onderzoek moet (blijven) proberen de behoeften van gebruikers te begrijpen en eerlijk zijn over wat beeldbeschrijvingssystemen wel of niet kunnen. Dit vereist begrip van de taalkundige aspecten van beeldbeschrijving, waarvoor dit proefschrift een uitgangspunt biedt.

SIKS dissertation series

This thesis is part of the SIKS dissertation series (2019-25). Earlier titles in this series are printed below.

2011 01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models
02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language
03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems
04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference
05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.
06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage
07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction
08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues
09 Tim de Jong (OU), Contextualised Mobile Media for Learning
10 Bart Bogaert (UvT), Cloud Content Contention
11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective
12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining
13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling
14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets
15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval
16 Maarten Schadd (UM), Selective Search in Games of Different Complexity
17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness
18 Mark Ponsen (UM), Strategic Decision-Making in complex games
19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles
20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach
21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems
22 Junte Zhang (UVA), System Evaluation of Archival Description and Access
23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media
24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics
26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns
28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure
29 Faisal Kamiran (TUE), Discrimination-aware Classification
30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions

31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science
33 Tom van der Weide (UU), Arguing to Motivate Decisions
34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training
36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach
37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference
38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization
39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games
40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development
41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control
42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution
43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge
44 Boris Reuderink (UT), Robust Brain-Computer Interfaces
45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection
46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression
48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

---

2012 01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda
02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories
04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications
05 Marijn Plomp (UU), Maturing Interorganisational Information Systems
06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks
07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories
09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms
10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment
11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems
13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
18 Eltjo Poort (VU), Improving Solution Architecting Practices

19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution
20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
26 Emile de Maat (UVA), Making Sense of Legal Text
27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
28 Nancy Pascall (UvT), Engendering Technology Empowering Women
29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval
30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications
35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41 Sebastian Kelle (OU), Game Design Patterns for Learning
42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43 Withdrawn
44 Anna Tordai (VU), On Combining Alignment Techniques
45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering
51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

---

2013 01 Viorel Milea (EUR), News Analytics for Financial Decision Support
02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics
04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
07 Giel van Lankveld (UvT), Quantifying Individual Player Differences

08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services
12 Marian Razavian (VU), Knowledge-driven Migration to Services
13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling
20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
22 Tom Claassen (RUN), Causal Discovery and Logic
23 Patricio de Alencar Silva (UvT), Value Activity Monitoring
24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning
27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance
28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience
29 Iwan de Kok (UT), Listening Heads
30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support
31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications
32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development
33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere
34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search
35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction
36 Than Lam Hoang (TUe), Pattern Mining in Data Streams
37 Dirk Börner (OUN), Ambient Learning Displays
38 Eelco den Heijer (VU), Autonomous Evolutionary Art
39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems
40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games
41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning
43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

---

2014 01 Nicola Barile (UU), Studies in Learning Monotone Models from Data
02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method
03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions
04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation
05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability

06 Damian Tamburri (VU), Supporting Networked Software Development
07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior
08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints
09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
10 Ivan Salvador Razo Zapata (VU), Service Value Networks
11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support
12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control
13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains
14 Yangyang Shi (TUD), Language Models With Meta-information
15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria
17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations
19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
24 Davide Ceolin (VU), Trusting Semi-structured Web Data
25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
26 Tim Baarslag (TUD), What to Bid and When to Stop
27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
28 Anna Chmielowiec (VU), Decentralized k-Clique Matching
29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
30 Peter de Cock (UvT), Anticipating Criminal Behaviour
31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data
33 Tesfa Tegegne (RUN), Service Discovery in eHealth
34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.
35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach
36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying
38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.
39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital
40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education
41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text
42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models
43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments
44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.

45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach
46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity
47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

2015 01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response
02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls
03 Twan van Laarhoven (RUN), Machine learning for network data
04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments
05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding
06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis
08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning
11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins
12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
13 Giuseppe Procaccianti (VU), Energy-Efficient Software
14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation
16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners
20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination
21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
22 Zhemin Zhu (UT), Co-occurrence Rate Networks
23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage
24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection
26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure
27 Sándor Héman (CWI), Updating compressed colomn stores
28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO
29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning
31 Yakup Koç (TUD), On the robustness of Power Grids
32 Jerome Gard (UL), Corporate Venture Management in SMEs
33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources
34 Victor de Graaf (UT), Gesocial Recommender Systems
35 Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VU), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UVA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UVA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (UvT), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

---

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UVA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TUE), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UVA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VU), Logics for causal inference under uncertainty
23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VU), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (UvT), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VU), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TUE), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TUE), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17 Jianpeng Zhang (TUE), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech
30 Wouter Beek (VU), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
05 Sebastiaan van Zelst (TUE), Process Mining with Streaming Data
06 Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
09 Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems
10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VU), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN), Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VU), Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dimitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing