DIDEC: The Dutch Image Description and Eye-tracking Corpus

Emiel van Miltenburg
Vrije Universiteit Amsterdam
Emiel.van.Miltenburg@vu.nl

Ákos Kádár
Tilburg University
A.Kadar@tilburguniversity.edu

Ruud Koolen
Tilburg University
R.M.F.Koolen@tilburguniversity.edu

Emiel Krahmer
Tilburg University
E.J.Krahmer@tilburguniversity.edu

Abstract

We present a corpus of spoken Dutch image descriptions, paired with two sets of eye-tracking data: *free viewing*, where participants look at images without any particular purpose, and *description viewing*, where we track eye movements while participants produce spoken descriptions of the images they are viewing. This paper describes the data collection procedure and the corpus itself, and provides an initial analysis of self-corrections in image descriptions. We also present two studies showing the potential of this data. Though these studies mainly serve as an example, we do find two interesting results: (1) the eye-tracking data for the description viewing task is more coherent than for the free-viewing task; (2) variation in image descriptions (also called *image specificity*; Jas and Parikh, 2015) is only moderately correlated across different languages. Our corpus can be used to gain a deeper understanding of the image description task, particularly how visual attention is correlated with the image description process.

Title and Abstract in Dutch

DIDEC: Een corpus van afbeeldingen met Nederlandstalige beschrijvingen en eye-tracking data

Wij presenteren DIDEC, een corpus van foto's met gesproken Nederlandse beschrijvingen en twee verschillende soorten *eye-tracking* data: ofwel verzameld tijdens het beschrijven van de afbeeldingen, ofwel verzameld terwijl de proefpersonen alleen maar keken naar de afbeeldingen (zonder ze te hoeven beschrijven). Dit artikel beschrijft de dataverzameling, alsook een eerste analyse van de zelf-correcties in de beschrijvingen. Daarnaast beschrijven we twee voorbeeldstudies om aan te geven wat er mogelijk is met DIDEC. Deze studies geven twee interessante resultaten: (1) de eye-tracking data verzameld tijdens het beschrijven van de afbeeldingen is eenduidiger dan de data verzameld tijdens het bekijken van de afbeeldingen; (2) de variatie in de beschrijvingen voor iedere afbeelding (ook wel *afbeeldingsspecificiteit* genoemd; Jas and Parikh, 2015) is slechts matig gecorreleerd tussen verschillende talen (Duits, Nederlands, Engels). Ons corpus kan gebruikt worden om beter te begrijpen hoe mensen afbeeldingen beschrijven, en in het bijzonder wat de rol is van visuele aandacht op het beschrijvingsproces.

1 Introduction

Automatic image description is a task at the intersection of Computer Vision (CV) and Natural Language Processing (NLP). The goal is for machines to automatically produce natural language descriptions for any image (Bernardi et al., 2016). The field of automatic image description saw an explosive growth in 2014 with the release of the Flickr30K and MS COCO datasets: two corpora of images collected from Flickr, with 5 crowd-sourced descriptions per image (Young et al., 2014; Lin et al., 2014). These resources enabled researchers to train end-to-end systems that automatically learn a mapping between images and text (Vinyals et al., 2015), but also to better understand how humans describe images (e.g.,

Proceedings of the 27th International Conference on Computational Linguistics, pages 3658–3669
Santa Fe, New Mexico, USA, August 20-26, 2018.

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 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 paper.

Van Miltenburg et al., 2016). 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 process information can be very insightful for developing image description systems, which is why we decided to collect a new dataset.

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 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. (2017) improve this approach by having the model learn when visual information is or is not relevant to produce a particular word or phrase.

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. We present 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.

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 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 get 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). This difference is further discussed in Section 4. 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 3.2).

Contributions. This paper introduces DIDEC, a corpus of spoken image descriptions with eye-tracking data. We explain how the corpus was created (§ 2), and provide general statistics about the resource, along with a short discussion of the annotated corrections, providing insight in the description process (§ 3). Then, we present two initial studies to show different possible uses of our dataset. The first study focuses on the effect of task on visual attention, where we show that the eye-tracking data for the image description task is more coherent than the free-viewing data (§ 4). The second study looks at *image specificity* (Jas and Parikh, 2015) across different languages, and whether it is possible to predict image specificity from eye-tracking data. We provide a more efficient, multilingual re-implementation of Jas and Parikh's measure, and show that image specificity is only moderately correlated across different languages, and cannot be predicted directly from attention map similarity (§ 5). Our corpus is freely available, along with an exploration interface, and all the materials that were used to create the dataset.[1]

2 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 200K 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

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

[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.

familiarity effects. 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 get an idea of 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-score 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.

3 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 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 (1962) and others that spoken descriptions are typically longer than written ones. We discuss the differences between spoken and written language in more detail in (van Miltenburg et al., 2018).

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

[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.

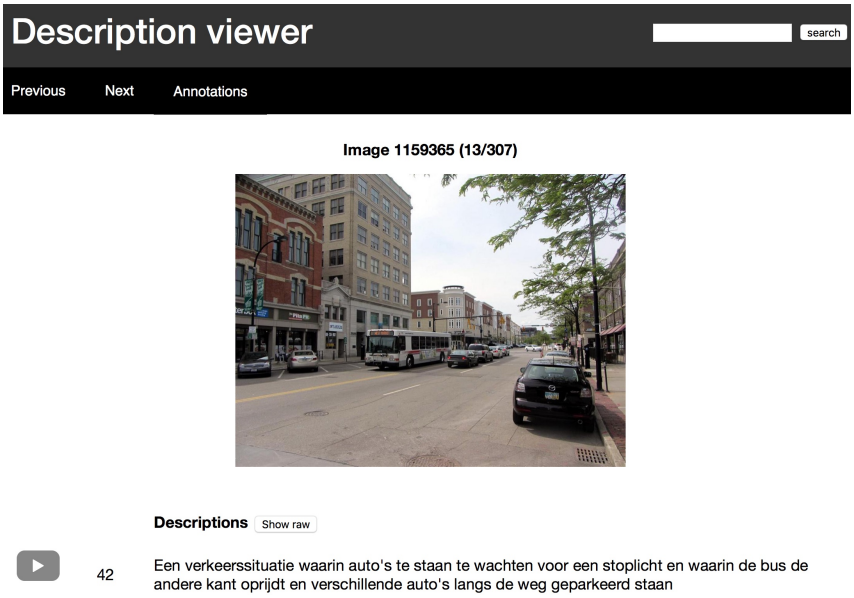

Figure 2: The description viewer provides a browser-based interface to the corpus. Users can browse through the images, search for specific words or annotations, and listen to the spoken descriptions. (Displayed image by David Wilson, CC BY 2.0)

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.

3.1 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 2 for an impression of the interface. The viewer tool can be downloaded along with our data from the corpus website.

3.2 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 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 1: Annotation counts.

We will 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), 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 seminal model of speech

Figure 3: Eye-tracking data for example (3). 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)

production (Levelt, 1989). For reasons of space, 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 (1).

(1) 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 1 already provided an interesting example, repeated for convenience in (2):

(2) 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.

The example in (3) shows a correction after making an incorrect prediction about the situation in Figure 3. Initially the speaker thinks the group is already eating, but actually they haven't ordered yet.

(3) 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 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, (4) 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.)

(4) 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 paper discusses two case studies – one comparing eye movements during free viewing versus image description, and one looking into the effects of image specificity on the description

	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 2: Results for the comparison between Free viewing and description viewing.

produced– both highlighting the potential usage of the DIDEC corpus.

4 Task-dependence in eye tracking: free viewing versus producing descriptions

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 our 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.

Results. Table 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 shows the distribution of the scores. We conclude that the image description task reduces noise in the collection of eye-tracking data, and produces a more

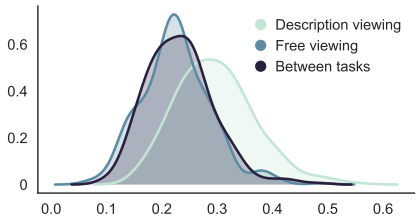

Figure 4: Distribution of the similarity scores within and between tasks.

[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.

coherent set of attention maps.

5 Image specificity

Whenever you ask multiple people to describe the same image, you rarely get the same description. Jas and Parikh (2015) 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*. Following up on this observation, Jas and Parikh (2015) propose an automated measure for image specificity in English, and show that it 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 first test this hypothesis for Dutch, English, and German using existing datasets, and then replicate the correlation between Dutch and English specificity scores using DIDEC. In a second step, we will see if we can directly correlate differences in eye-tracking behavior with the Dutch specificity scores. This experiment builds on the results from Coco and Keller (2012), who showed that scan patterns can be used to predict what someone will say about an image.

5.1 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. 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., 2013). 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

[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. (2013)
German	word2vec	Müller (2015)
All	FastText	Bojanowski et al. (2017)

Table 3: Word embeddings used to compute the image specificity metric.

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

Table 4: Spearman correlation between automated image specificity scores in different languages, using two sets of word embeddings.

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.

5.2 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).[9] 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). Note that the Dutch data comes from a different dataset than ours, which allows for comparison with both English *and* German.

Table 4 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. 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.

5.3 Replicating the results using DIDEC

Having found a fairly consistent Spearman correlation of about 0.4 between the different languages, we asked ourselves whether this result could be replicated. We pre-registered our hypothesis (yes it can) with the Open Science Foundation before collecting the image description data.[10] After the data collection procedure was finished, we computed the correlation between the Dutch and English image specificity scores for the 307 images in our dataset. We found a Spearman correlation of 0.23, which is lower than expected. One possible explanation for this is that the descriptions in MS COCO are shorter and have a lower variance in description length (as mentioned in §3), which reduces noise in the comparison. To see if this is indeed the case, we repeated the experiment with only 5 descriptions per image, always selecting the median 5 descriptions when sorted by length. With this approach, we find an even lower correlation of 0.20. We conclude that image content only has a limited influence on description diversity.

5.4 Predicting image specificity from eye-tracking data: a negative result

We now turn to look at whether the image specificity scores are correlated with our eye-tracking data. As noted above, this experiment builds on Coco and Keller's (2012) work, which shows that scan patterns can be used to predict what people will say about an image. Rather than taking Coco and Keller's more

[9] Even though wordnets exist for Dutch (Postma et al., 2016) 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.

[10] Our preregistration can be found at: `https://osf.io/6pc2t/register/565fb3678c5e4a66b5582f67` We deviated from our preregistration in two areas: (1) the experiment took more time than expected, so we split the experiment in two parts; (2) we had to discard more data than expected, so we collected data from more participants to compensate.

advanced approach (involving image segmentation), we use a more naive strategy: (1) We first compute the average pairwise similarity scores between the attention maps. If our participants agreed on where to look, this score will be higher than if our participants attended to different parts of the image. (2) We then correlate those scores with the image specificity scores we computed earlier.

We found no correlation between the eye-tracking data and the image specificity scores. Because Coco and Keller (2012) did find correlations between eye-tracking similarity and description similarity, we conclude that our naive approach is probably too weak to detect any effect. This shows us the limitations of using raw eye-tracking data, which may be too noisy to be compared directly for these kinds of fine-grained predictions. In cases like this, following Coco and Keller, it is better to combine our eye-tracking data with the object-level annotations in MS COCO.

6 Conclusion

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 possible 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 companion paper, we compare spoken and written image descriptions in both Dutch and English (van Miltenburg et al., 2018). 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; Elliott et al., 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.

7 Acknowledgments

Emiel van Miltenburg is supported via the 2013 NWO Spinoza grant awarded to Piek Vossen. Ákos Kádár is supported through an NWO Aspasia grant awarded to Afra Alishahi. We thank Rein Cozijn for his technical assistance in setting up the eye-tracking experiments, and Kim Tenfelde for her transcription and annotation efforts.

References

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*, pages 89–93, Edinburgh, UK, September 5-8. Association for Computational Linguistics.

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*, 7:103.

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.

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.

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.

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 Frdo Durand. 2016. What do different evaluation metrics tell us about saliency models?

Andy Clark. 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

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.

Gerard HJ Drieman. 1962. Differences between written and spoken language: An exploratory study, I. quantitative approach. *Acta Psychologica*, 20:36–57.

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*, pages 70–74, Berlin, Germany, August. Association for Computational Linguistics.

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*, Copenhagen, Denmark, September.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Victoria A Fromkin. 1971. The non-anomalous nature of anomalous utterances. *Language*, pages 27–52.

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.

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*.

Verena Henrich and Erhard Hinrichs. 2010. Gernedit - the germanet editing tool. In *Proceedings of the ACL 2010 System Demonstrations*, pages 19–24, Uppsala, Sweden, July. Association for Computational Linguistics.

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.

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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

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, May.

Willem J.M. Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41 – 104.

Willem JM Levelt. 1989. *Speaking: From intention to articulation*. MIT press.

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*, pages 740–755. Springer.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July.

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.

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*, KR'04, pages 267–277. AAAI Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Andreas Müller. 2015. German word embeddings. Available from GitHub at: `http://devmount.github.io/GermanWordEmbeddings/`.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open dutch wordnet. In *Proceedings of the Eighth Global Wordnet Conference*, Bucharest, Romania, January 27-30.

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

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*, pages 21–30, Santiago de Compostela, Spain, September. Association for Computational Linguistics.

Emiel van Miltenburg, Ruud Koolen, and Emiel Krahmer. 2018. Varying image description tasks: spoken versus written descriptions. Submitted.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

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 Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul. PMLR.

Alfred L Yarbus. 1967. *Eye movements and vision*. Springer.

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.