Cross-linguistic differences and similarities in image descriptions

Cross-linguistic differences and similarities in image descriptions

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

Desmond Elliott
University of Edinburgh
d.elliott@ed.ac.uk

Piek Vossen
Vrije Universiteit Amsterdam
piek.vossen@vu.nl

Abstract

Automatic image description systems are commonly trained and evaluated on large image description datasets. Recently, researchers have started to collect such datasets for languages other than English. An unexplored question is how different these datasets are from English and, if there are any differences, what causes them to differ. This paper provides a cross-linguistic comparison of Dutch, English, and German image descriptions. We find that these descriptions are similar in many respects, but the familiarity of crowd workers with the subjects of the images has a noticeable influence on description specificity.

1 Introduction

Vision and language researchers have started to collect image description corpora for languages other than English, e.g. Chinese (Li et al., 2016), German (Elliott et al., 2016; Hitschler et al., 2016; Rajendran et al., 2016), Japanese (Miyazaki and Shimizu, 2016; Yoshikawa et al., 2017), French (Rajendran et al., 2016), and Turkish (Unal et al., 2016). The main aim of those efforts is to develop image description systems for non-English languages and to explore the related problems of cross-lingual image description (Elliott et al., 2015; Miyazaki and Shimizu, 2016) and machine translation in a visual context (Specia et al., 2016; Hitschler et al., 2016). We view these new corpora as sociological data that is in itself worth studying. Our research stems from the following question: *To what extent do speakers of different languages differ in their descriptions of the same images?* Considering this question, we developed a (non-exhaustive) list of factors that may influence the descriptions provided by crowd workers. Understanding the effect of these factors will enable us to improve the data collection process, and help us appreciate the challenges of natural language generation in a visual context:

1. **Task design effects**: There are many possible approaches to collecting descriptions of images. Previous research has re-used the Flickr8K (Hodosh et al., 2013) template and methodology. Baltaretu and Castro Ferreira (2016) showed that task design may influence the form of crowd-sourced descriptions.
2. **(Perceived) audience**: Speakers adapt the style of their messages to their audience (Bell, 1984). Knowing how the descriptions will be used may affect the style or quality of the corpora.
3. **Individual/Demographic factors**: Individual features, like the demographics or personal preferences of the workers, may explain part of the variation in the descriptions.
4. **Differences in (background) knowledge**: Workers can only provide as much information as they know. Besides educational factors, the background knowledge of a person can be influenced by where they currently live and where they have previously lived.
5. **Language differences**: Languages differ in how they package information, which may be reflected in the descriptions. This is close to, but separate from *linguistic relativity* (see e.g. (Deutscher, 2010; McWhorter, 2014)).

6. **Cultural differences**: Culture may influence the descriptions on a group level by affecting the social perspective of a population.

This paper focuses on the last three factors in a cross-linguistic corpus study of Dutch, German, and English image descriptions. Our work is a starting point for understanding the differences in descriptions between languages. The focus on the last three factors is a consequence of our corpus study: the first two factors require manipulating the experimental set-up, and the third factor requires data about the crowd workers that is not known (and should ideally also be controlled). As we will see in Sections 4 and 5, we *can* make claims about the last three factors based on the workers' language and geolocations.

We believe that studying differences between languages shows us which phenomena are robust across languages and thus important to consider when implementing and deploying models. Also, differences between languages can inform us about the feasibility of approaches to image description in different languages by translating existing English data (Li et al., 2016; Yoshikawa et al., 2017).

Our analysis combines quantitative and qualitative studies of a trilingual corpus of described images. We use the Flickr30K (Young et al., 2014) for English, Multi30K for German (Elliott et al., 2016), and a new corpus of Dutch descriptions (Section 3). We build on earlier work that studies the semantic and pragmatic properties of English descriptions (van Miltenburg, 2016; van Miltenburg et al., 2016). Those works study ethnicity marking, negation marking, and unwarranted inferences about the roles of people. The main finding of our analysis is that all of these properties are stable across Dutch, US English, and German (Section 4). We also show how differences in background knowledge can affect description specificity (Section 5). We make the Dutch corpus available online and we also release software to explore image description corpora with the descriptions in different languages side-by-side to encourage future work with different language families.[1]

2 Related work

We review work on the theory about the image description process, and work on automatic image description in other languages.

Describing an image. Erwin Panofsky's (Panofsky, 1939) hierarchy of meaning was originally intended as a guide for interpreting works of art. It has since been applied by Shatford (1986) and Jaimes and Chang (1999) in the context of indexing and searching for images in libraries. The hierarchy consists of three levels that build on each other.

1. **Pre-iconography**: giving a factual description of the contents of an image, and an expressional indication of the mood it conveys.
2. **Iconography**: giving a more *specific* description, informed by knowledge of the cultural context in which the image is situated.
3. **Iconology**: interpreting the image, establishing its cultural and intellectual significance.

This hierarchy is useful to think about for descriptions of images (Hodosh et al., 2013). As Panofsky (1939) notes, these levels require more knowledge as we move up the hierarchy. If we apply this hierarchy to the image description domain, we can say that image description corpora typically cover the first two levels. An important factor in the 'quality' of a description is the amount of *cultural* or *background* knowledge that informs the description. We will explore the influence of this factor in Section 5.

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., 2016; Miyazaki and Shimizu, 2016). Thus far, any differences have only been anecdotally described.

Li et al. (2016) 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,

[1] See: https://github.com/cltl/DutchDescriptions

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.

In this paper we collect a new dataset of Dutch image descriptions, but our work differs from previous work in two ways: (i) we aim to provide a more systematic overview of the differences between descriptions in three languages, and therefore (ii) we do not empirically evaluate system performance in reproducing the descriptions.

3 Collecting Dutch descriptions

We used Crowdflower to annotate 2,014 images from the validation and test splits of the Flickr30K corpus (Young et al., 2014) with five Dutch descriptions.

Following other work, our goal is to create a parallel 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 the supplementary materials.

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 (1), with a literal translation and original English description (verified using Google Translate).

(1) Response generated with Google Translate.

a. *Een paar kussen	(Description)
'A couple of kisses'	(Translation)
A couple kisses	(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."

4 Characterizing English, German, and Dutch image descriptions

We now examine the descriptions between languages in more detail, focusing on the *validation* subset of the Multi30K dataset (1,014 images, with 5,070 descriptions per language).

4.1 General statistics

Table 1 shows the mean sentence length (in tokens and words) for the three languages. The English descriptions are the longest, followed by the Dutch and

	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 1: Mean sentence length across languages.

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 confirmed by the number of word types: German has 31% more types than English (5709 versus 4355). Dutch has 19% more (5193).

Definiteness. The five most frequent bigrams that start a description (showing the typical subjects of the images) are given in Table 2. 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.

4.2 Replicating previous findings for negation and ethnicity marking

Previous work has studied the use of negation and ethnicity marking in English image description datasets (van Miltenburg et al., 2016; van Miltenburg, 2016) We now attempt to replicate these findings in the Dutch and German data.

Negations. van Miltenburg et al. (2016) 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 (2) and (3), respectively.

(2) 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(!)'

(3) 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. (2016) 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. 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 linguistic bias (Beukeboom, 2014).

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

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

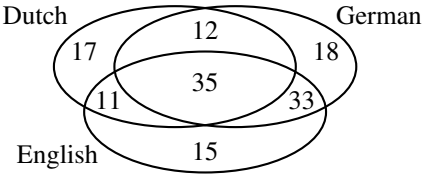

Figure 1: Venn diagram of ethnicity markers by Dutch, English, and German workers. Counts correspond to images.

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 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. However, we do *not* claim that this is evidence that people living in Germany and the U.S.A. are *more racist* than people living in the Netherlands. Rather than trying to interpret the meaning of this difference, we ask a different question: 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 default, as it is the dominant ethnicity in all three countries.[2] The marker *white* is only used to be consistent in the use of modifiers within 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 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.

[2] 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).

Figure 2: Image 4634063005. The older woman in the picture was often seen as the grandmother.

5 Familiarity and cultural differences

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 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 §6.2.

5.1 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 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:

(4) 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

Figure 3: Image 6408975653. This picture was taken at the Tuileries Garden in Paris, and shows the Luxor Obelisk and the Arc de Triomphe.

Figure 4: Image 4727348655. This picture shows a man wearing a Denver Broncos hat and jersey.

(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 5.

(5) 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.

The Denver Broncos. Figure 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

Figure 5: Image 4897113571. This picture shows the back of a street organ in the Netherlands.

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*. 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.

5.2 Objects

Familiarity also plays a role in labeling objects. Consider Figure 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 6).

(6) a. A **yellow truck** is standing on a busy street in front of the Swarovski store.
b. A **strange looking wood trailer** is parked in a street in front of stores.
c. An **unusual looking vehicle** parked in front of some stores.

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.

5.3 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.

6 Discussion

6.1 Description specificity

In Section 5 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. 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).

6.2 Limitations of translation approaches

One approach to image description in multiple languages is to use a translation system. For example, Li et al. (2016) 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. (2016) show that the former strategy achieves the best result, and argue that it is a promising approach because it requires no extra manual annotation.

Our observations in Section 5 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.

6.3 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 differing 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.

7 Conclusion

We studied a trilingually aligned corpus of described images to learn about how crowd workers of different languages described the same images. The main finding was that earlier observations about negation marking and ethnicity marking by English workers also hold for Dutch and German. Dutch and German workers also use negations in their image descriptions, showing that this is a robust phenomenon. Dutch and German workers also make unwarranted inferences about the images, this shows that crowd workers regularly include extra-visual information in their descriptions. In addition, Dutch and German workers also disproportionately mark non-white people in their descriptions, showing that image description corpora carry biases that we need to take into account when working with this data.

We also explored the role of familiarity in image description. We found images in our corpus that were easily described by workers of one language, but unidentifiable to the workers of another language. This has consequences for image description models trained on automatically translated training data: some images will not be properly described for the target audience. But the problem is more general. The success of image description systems trained on datasets of described images is limited by the knowledge of the annotators, regardless of the language. While the available data is useful for us to learn and discuss what human-like descriptions should look like, it can only take us so far. Full coverage systems that could tailor their descriptions to particular audiences are still out of reach.

We hope this work provides a starting point for conducting cross-linguistic comparisons of image descriptions. Future work includes replicating our analyses across more diverse families of languages, modifying the task design to contrast the results with our findings, and using our inspection tool to explore other linguistic phenomena. We are also interested in scaling up our analyses to larger corpora, which will require the development of automated comparison methods. We believe that these steps will bring us closer to an initial understanding of the diversity in image descriptions across different languages and social groups.

Acknowledgements

This research is funded through the NWO Spinoza prize, awarded to PV. DE is supported by an Amazon Research Award. We thank three anonymous reviewers for their questions and comments.

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, September 5-8.

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.

Allan Bell. 1984. Language style as audience design. *Language in Society*, 13(2):145—204.

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*, volume 31, pages 313–330. Psychology Press.

Centraal Bureau voor de Statistiek. 2016. Bevolking; generatie, geslacht, leeftijd en herkomstgroepering, 1 januari. Part of the CBS database, last modified 15 September 2016, September.

Paul Christophersen. 1939. *The articles: A study of their theory and use in English*. Copenhagen: Munksgaard.

Guy Deutscher. 2010. *Through the language glass: How Words Colour Your World*. New York: Metropolitan Books.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multilingual image description with neural sequence models. *CoRR*.

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.

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.

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*.

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.

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, March.

Alejandro Jaimes and Shih-Fu Chang. 1999. Conceptual framework for indexing visual information at multiple levels. In *Electronic Imaging*, pages 2–15. International Society for Optics and Photonics.

Roman Jakobson. 1972. Verbal communication. *Scientific American*, 227:72–80.

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.

Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016. Adding chinese captions to images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 271–275.

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.

Yo Matsumoto. 1995. The conversational condition on horn scales. *Linguistics and philosophy*, 18(1):21–60.

John H McWhorter. 2014. *The language hoax: Why the world looks the same in any language*. Oxford University Press, USA.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2939.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1780–1790.

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.

Erwin Panofsky. 1939. *Studies in Iconology*. Oxford University Press.

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*.

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.

Sara Shatford. 1986. Analyzing the subject of a picture: a theoretical approach. *Cataloging & classification quarterly*, 6(3):39–62.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine

translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation*, pages 543–553.

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.

Statistisches Bundesamt. 2013. Zensus 2011: 80,2 millionen einwohner lebten am 9. mai 2011 in deutschland. Press release, Nr. 188, May.

M. E. Unal, B. Citamak, S. Yagcioglu, A. Erdem, E. Erdem, N. Ikizler Cinbis, and R. 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)*.

Emiel van Miltenburg, Roser Morante, and Desmond Elliott. 2016. Pragmatic factors in image description: the case of negations. In *Proceedings of the 5^{th} Workshop on Vision and Language at ACL '16*.

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

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823*.

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.