

# Cross-linguistic differences and similarities in image descriptions

Emiel van Miltenburg @evanmiltenburg

Desmond Elliott @delliott

Piek Vossen @piekvossen

# Image descriptions



A boy wearing blue and yellow walking on a cliff edge.

A boy in yellow shorts is standing on top of a cliff.

A young child is standing alone on some jagged rocks.

A little boy standing high in the air on a rock.

Child stands near edge of cliff.

# Why describe images?

- **Proxy for image understanding**
- **Understanding reference and pragmatics in language**
- Aiding visually impaired users
- Human-Computer interaction

# Image understanding

Range of tasks

- Image description
- Visual Question Answering
- Visual Dialog
- Visual Storytelling

Each tasks tests **different aspects** of image understanding!

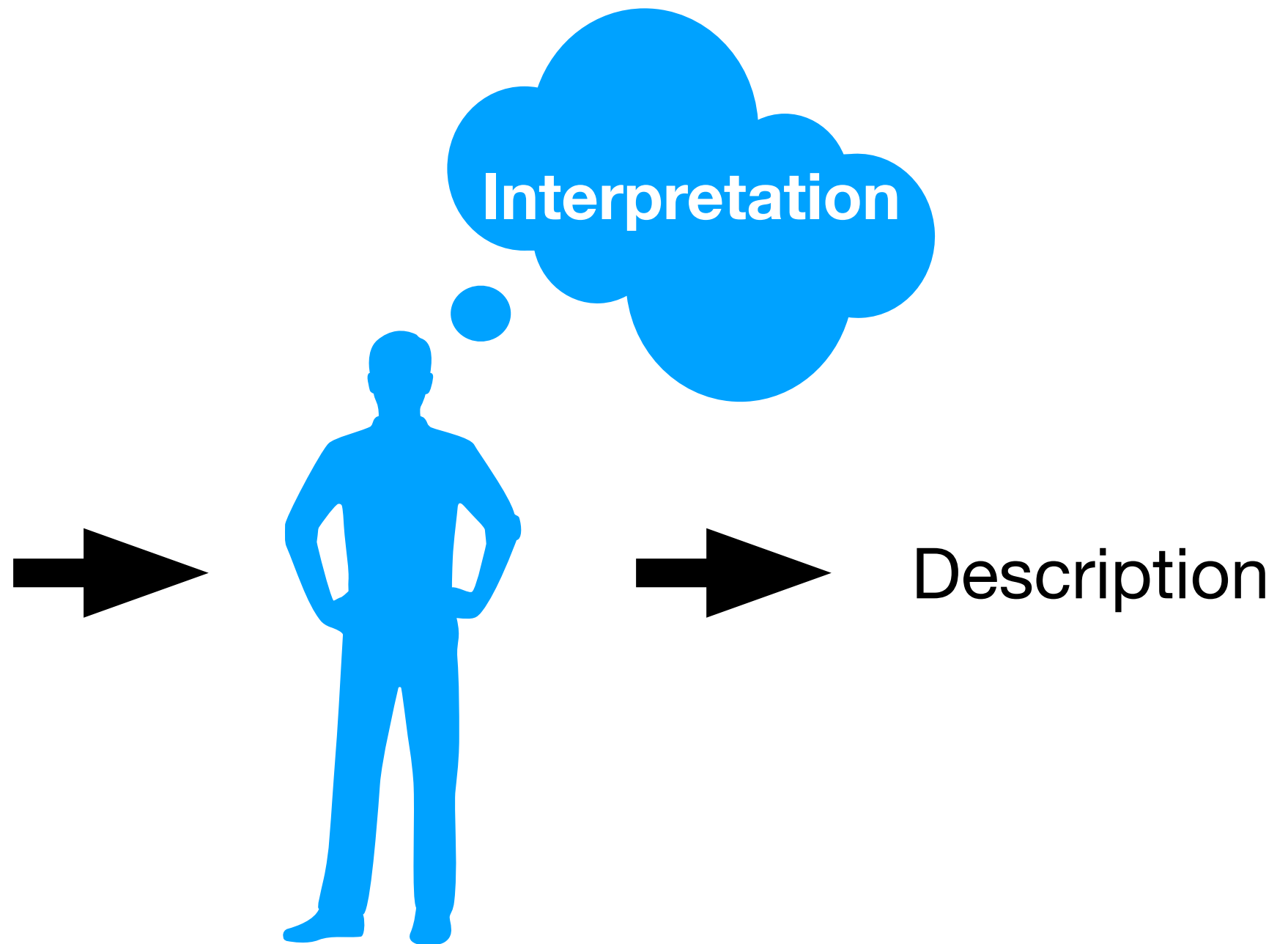
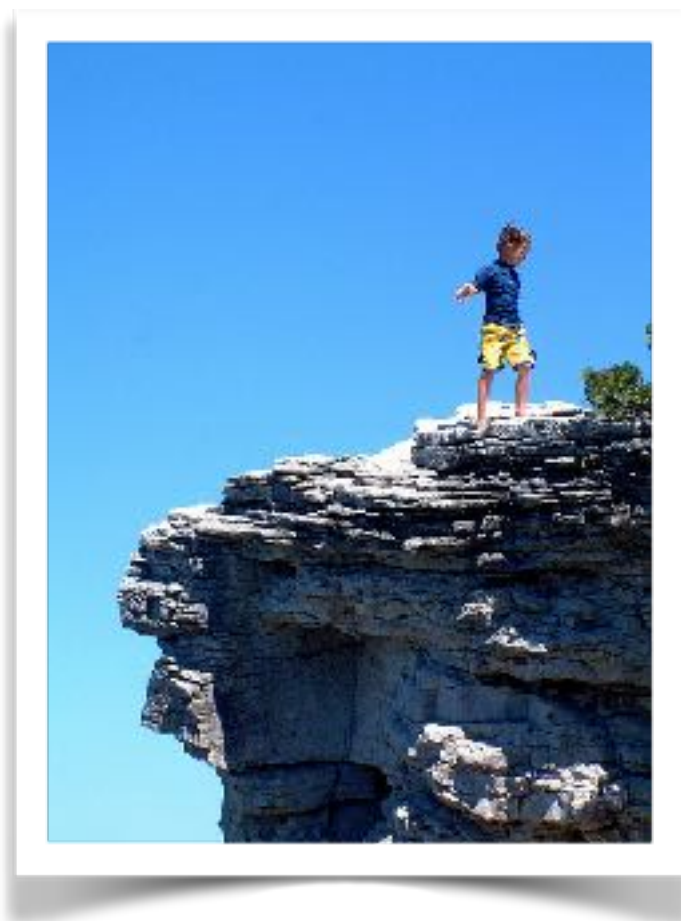
# Image understanding

Range of tasks

So what is image description all about?

**Variation & Perspective**

# A model



Differences in perspective lead to differences in the descriptions

# Sources of variation

This talk

✓ Background knowledge

✓ Cultural differences

✓ Language differences

✗ Task design

✗ Audience

✗ Demographic factors

# Sources of variation

This talk

✓ Background knowledge



**Corpus**

✓ Cultural differences



✓ Language differences



✗ Task design



**Experiment**

✗ Audience






✗ Demographic factors





# Data



-  Girl in red jumping for joy (Flickr30K)
-  Eine Frau freut sich und springt (Multi30K)
-  Een vrouw in een rode jas springt in de lucht (New!)

# Collecting data



Please describe the image in a short but complete sentence

# Dutch descriptions

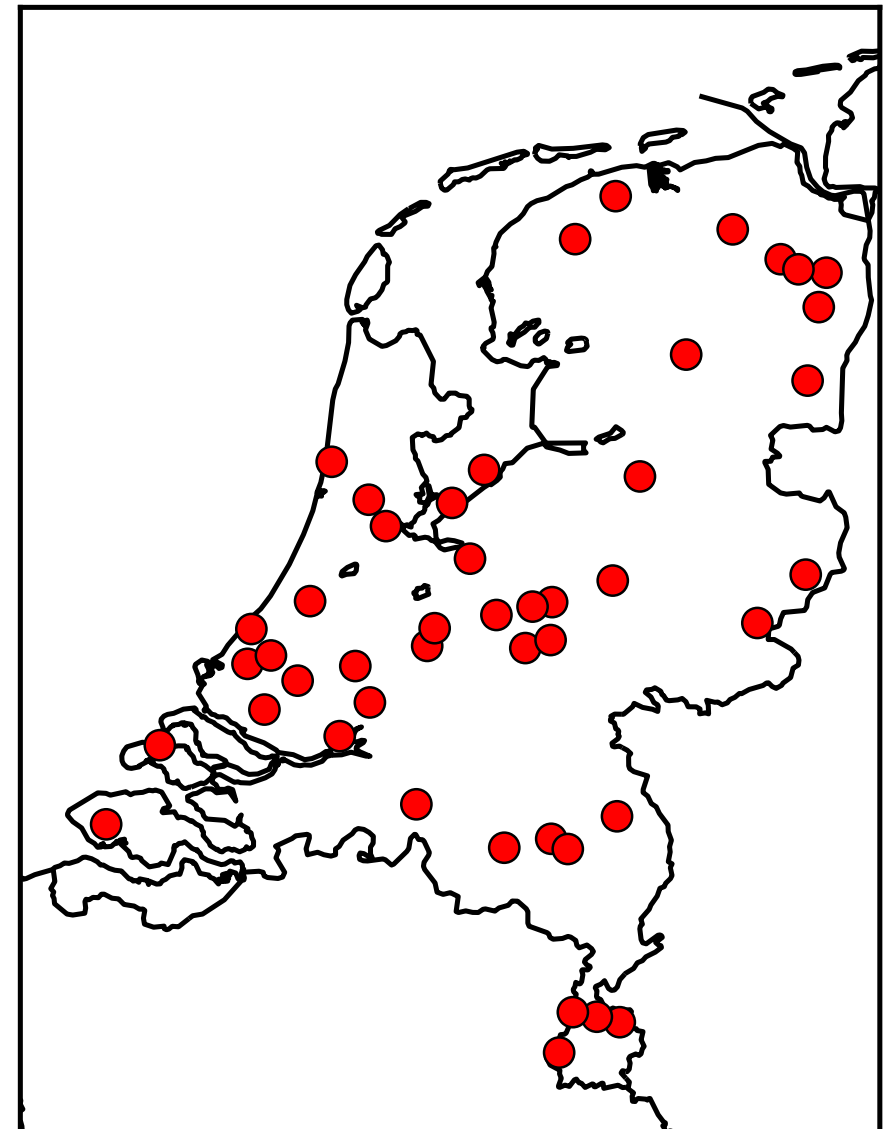
Collected through CrowdFlower

1,000 development images

1,014 test images

5 descriptions per image

4 months, due to **small crowd**



# Manual checking

People use Google Translate to cheat!

\*Een paar kussen

(Description)

`A couple of kisses'

(Gloss)

A couple kisses

(Original from Google Translate)

\*Mensen het kopen van vis

(Description)

`People the buying of fish'

(Gloss)

People buying fish

(Original from Google Translate)

# Our study

# Phenomena

This talk

- ✓ Negations
- ✓ Stereotype & Bias
- ✓ Familiarity
- ✗ Definiteness

# Methodology

Talk to me for a demo!

127.0.0.1:5001

## Inspecting image descriptions

Search for subset:




Image ID: 1018148011 - [Next Image](#)

**Dutch**

- Werkers zijn hier bezig met uitladen van de lading op de vrachtwagen. Waarbij andere de boel inspecteren.
- een groep mensen die op een vrachtwagen zitten waar een wit voorwerp te zien is
- een aantal mannen staat bovenop witte wol in de laadbak van een vrachtauto
- Mensen aan het werk bij/op een vrachtwagen met katoen?
- mensen zijn bezig katoen op een vrachtwagen te laden

**English**

- A group of people stand in the back of a truck filled with cotton.
- Men are standing on and about a truck carrying a white substance.
- A group of people are standing on a pile of wool in a truck.
- A group of men are loading cotton onto a truck
- Workers load sheared wool onto a truck.

**German**

- Baumwollager mit LKW
- Ein beladener LKW mit Menschen auf der Ladung.
- die Menschen laden Baumwolle oder Schafswolle auf den LKW
- Die Männer stehen auf und vor dem Lastwagen mit der Baumwolle.
- Bauarbeiter arbeiten auf einer Baustelle mit weißem Material.

<https://github.com/cltl/DutchDescriptions>



# Negations

Following Van Miltenburg et al. (2016)



🇩🇪 Eine Ansammlung von Menschen, sommerlich gekleidet schaut auf ein Ereignis, das **nicht** im Bild ist

🇺🇸 “A group of people, dressed in summer clothes, watches an event **not** shown in the image.”



# Negations

Following Van Miltenburg et al. (2016)

🇳🇱 Vrouw snijdt broodje zonder te kijken(!)

🇺🇸 “Woman slices bun without looking (!)”



# Results (1)

All languages use negations in their descriptions.

**11** Dutch, **27** English, **20** German

# Results (2)

Little overlap between languages in their use.

Variation: *shirtless, half naked, not wearing a shirt, without a shirt*

# Take-home message

Negations signal the need for **background knowledge**

Negations aren't necessary (other linguistic means available)

But, sign of a more general phenomenon:

**Reasoning about scenes** takes place in every language

# Stereotypes

Following Van Miltenburg (2016)



*A worker is being scolded by her boss in a stern lecture.  
A manager talks to an employee about job performance.  
A hot, blond girl getting criticized by her boss.  
Sonic employees talking about work.*

# Stereotypes

- Speculation, “unwarranted inferences”
- Workers going *beyond the contents of the image*
- Hard to detect! Inspect every image.



# “Mothers”





# Stereotyping results

- Stereotypes occur in all languages
- Hard to detect, **except** social roles like *mother*



# Stereotyping results

- Stereotypes occur in all languages
- Hard to detect, **except** social roles like *mother*
- Gray area: **stereotype** vs. **highly likely to be true**
- Example: “breastfeeding —> mother” almost always true.

# Take-home message

Human annotations are **subjective**. This:

- + Makes the descriptions more **specific**, but
- Creates additional **noise** for data-driven image description

# Bias

The language we use reflects the way we perceive the world.

People report things that stand out to them, that are Other.

(Beukeboom 2014; Misra et al. 2016)

# Ethnicity



Two young **African American** boys sitting at a desk in a classroom.

# Ethnicity



Two young  
boys sitting at a desk in a



**A Chinese woman** is eating  
some kind of dessert.



# Ethnicity



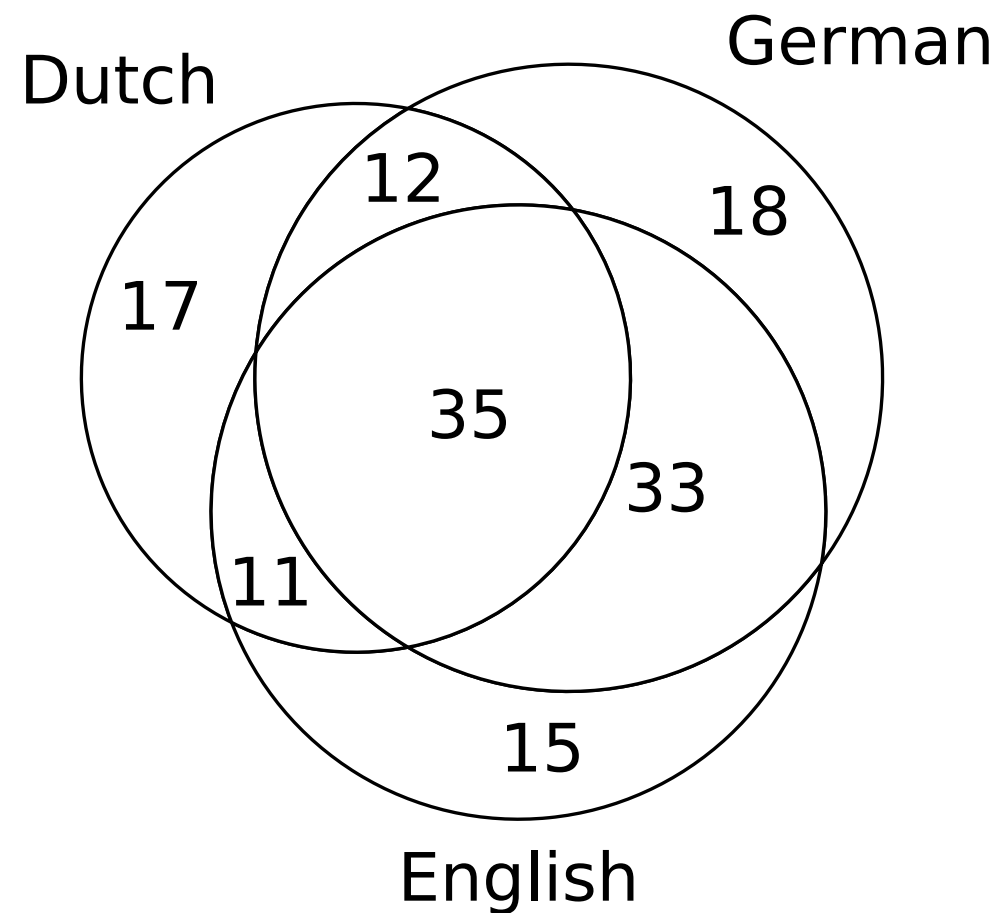
Two young  
boys sit



**A blond woman** in an army green hat and dress sitting next to a **Japanese woman** with sunglasses on her head and a t-shirt.

# Results (1)

Usage of nationality/ethnicity markers



# Results (2)

## Take-home message

- All languages use markers
- ‘White’ almost exclusively for contrast
- But this isn’t exclusively Western! (Miyazaki and Shimizu 2016; Li et al. 2016)

**Is this racist? necessary? useful?**



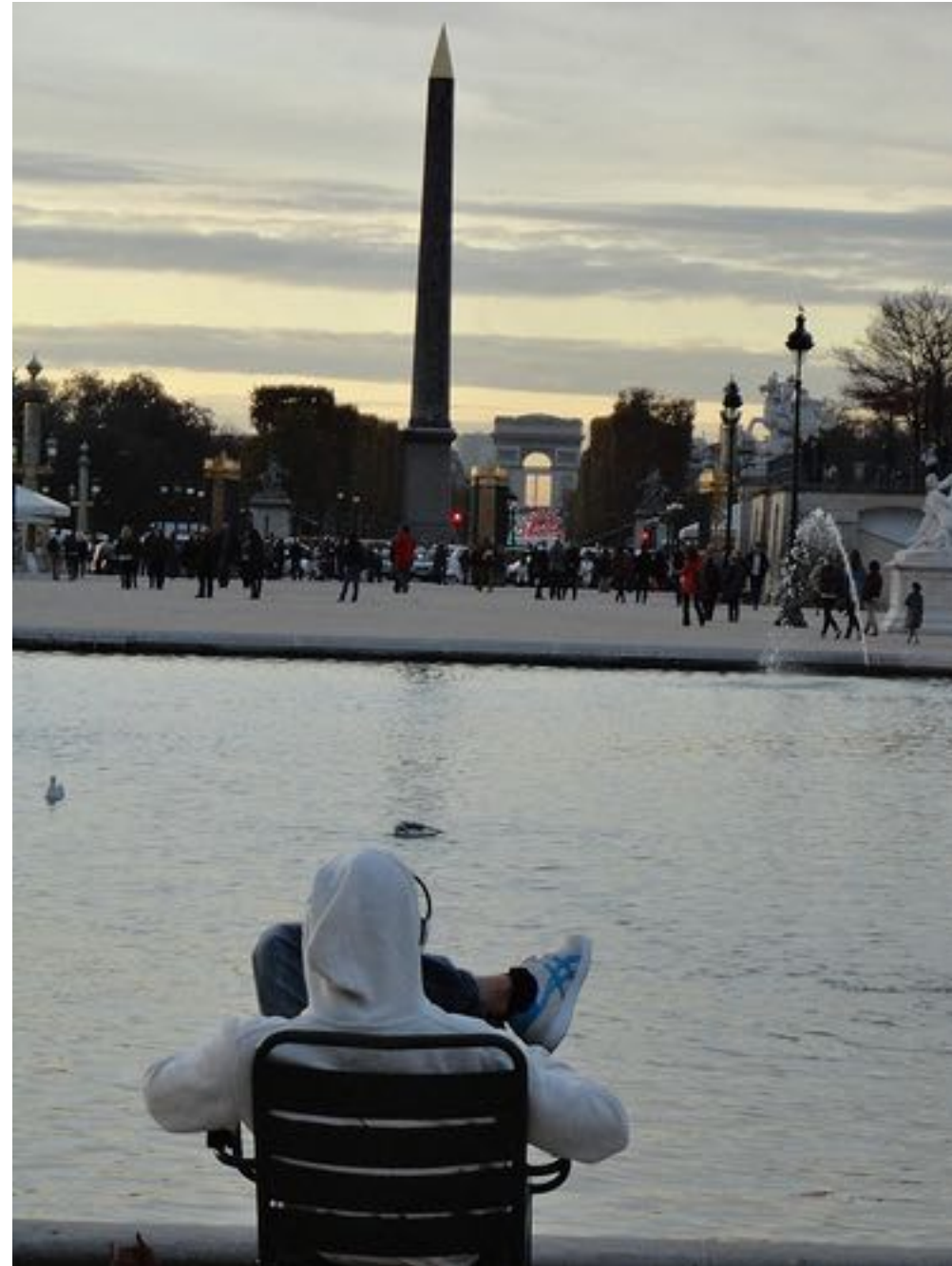
# Familiarity

People can only be as specific as their knowledge allows them to.

This leads to differences between languages/populations!

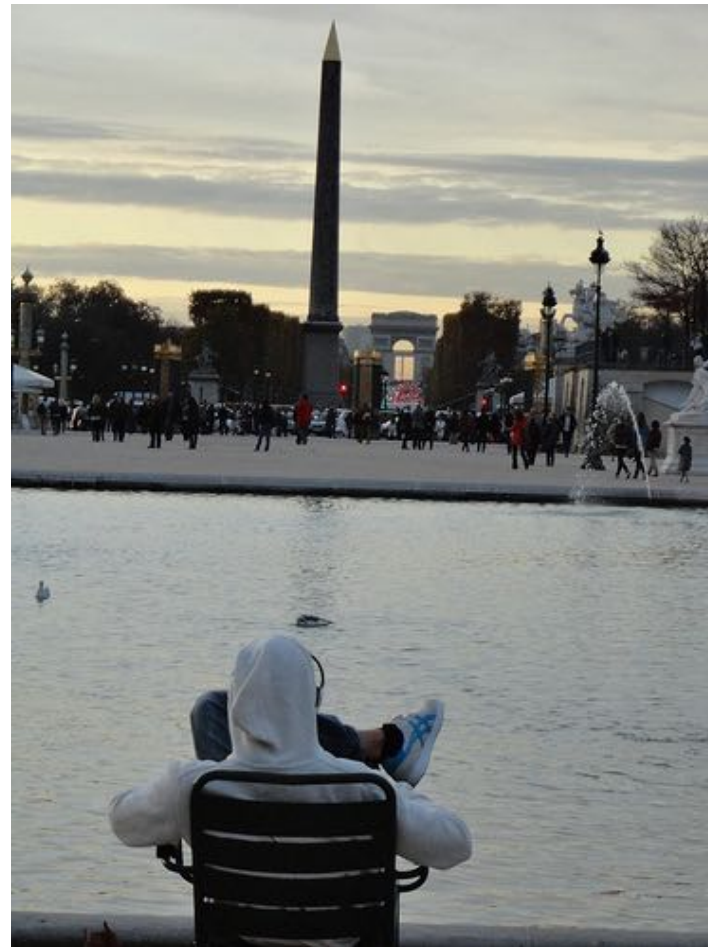
# Familiarity

Where is this?



# Familiarity

Jardin des Tuileries - Paris



🇳🇱 'A man sits by the pond in the **Tuileries park in Paris**'

🇺🇸 A person wearing a white hoodie gazes across the water at the **Washington Monument** during dusk.

🇩🇪 'A man is sitting in a chair, watching **the obelisk.**'

# Familiarity

What is going on?





# Familiarity

Tailgating outside a sports stadium



🇳🇱 ‘People **standing** in a parking lot with barbecues.’

🇺🇸 A man in a **Denver Broncos** jersey is **tailgating** with his friends.

🇩🇪 A man in a **sports jersey** is **standing** next to his friends in a parking lot.

# Familiarity

What is this?





# Familiarity

A street organ in Amsterdam (*Kalverstraat*)



🇳🇱 ‘A **street organ** in a shopping street with pedestrians.’

🇺🇸 A **strange looking wood trailer** is parked in a street in front of stores.

🇩🇪 Mixed responses. 2/5: music organ, 3/5: strange vehicle.

# Take-home message

- We cannot just translate NLG systems!
- Image description takes more than the ability to see.

How to tailor descriptions to an audience?

How can we integrate a knowledge component?



# Limitations of this study

- We've only looked at three Germanic languages.
- There's much more to explore, e.g. negation in Turkish:



“Two brown hunting dogs **unable to share** the black object they found in the grass.”

(data from the Tasviret corpus, Unal et al. 2016, translated using Google Translate)

# Conclusion

Image descriptions in the Flickr30K data:

- are **inherently subjective**
- involve **reasoning over situations**
- depend on **world knowledge**

Open question: what do we want descriptions to look like?



**Thank you!**



**Danke!**



**Bedankt!**

**GitHub** <https://github.com/cltl/DutchDescriptions>

**E-mail:** [emiel.van.miltenburg@vu.nl](mailto:emiel.van.miltenburg@vu.nl)

**Twitter:** @evanmiltenburg, @delliott, @piekvossen