A LARGE-SCALE MULTILINGUAL STUDY OF VISUAL CONSTRAINTS ON LINGUISTIC SELECTION OF DESCRIPTIONS

[1] The Hebrew University of Jerusalem

[2] University of Melbourne

Uri Berger[12], Lea Frermann[2], Gabriel Stanovsky[1], Omri Abend[1]

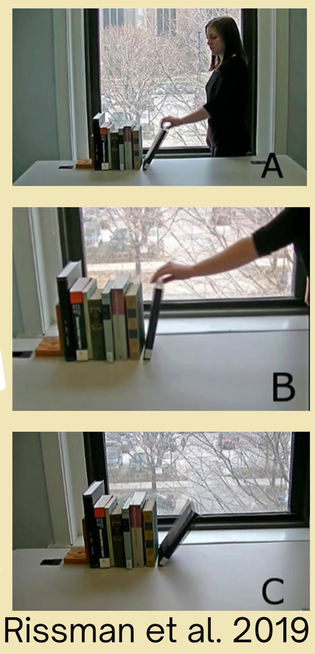
A large-scale study of how the content and style of images affect different linguistic properties (e.g., transitivity) of its descriptions, in 4 different languages.
We show effects of the visual modality on linguistic properties, mainly on the use of numerals and passive voice.

BACKGROUND & MOTIVATION

- We know **visual content** affects linguistic **semantic choiches**
- But: does it affect **structural choices**?

- Cognitive studies show that **visual content affect linguistic properties** of descriptions
  - E.g., cropping affects transitivity
- But:
  - Study a **single language**
  - Study a **single linguistic property**
  - **Small scale** (a few dozen participants)
  - Use **controlled** visual conditions

- Large datasets of **image-descriptions** are available
- But images are not labelled by **visual condition**



Rissman et al. 2019

METHODOLOGY

- Use existing multilingual **image-caption** datasets
- Use **semantic annotation** as **visual condition**
- Study **cross-lingual** properties of the **same image**
- Train **visual classifiers** to predict linguistic properties

DATA COLLECTION

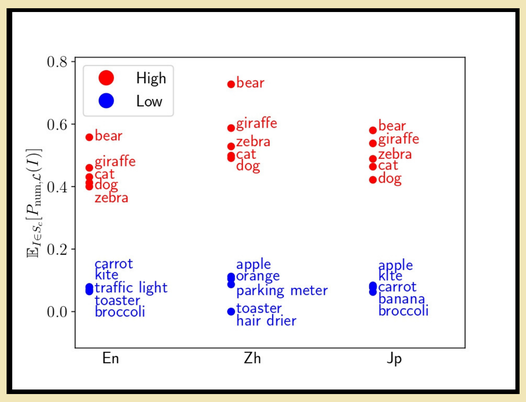
- Languages:
  - English
  - German
  - Chinse
  - Japanese
- Properties:
  - Numerals
  - Negation
  - Verb root
  - Transitivty
  - Passive
- Datasets:
  - 9 datasets
  - Multilingual
  - 604K images
  - 3M captions

- Properties annnotated **automtaically**
- Validated by **native speakers**

CORPUS ANALYSIS - NUMERALS

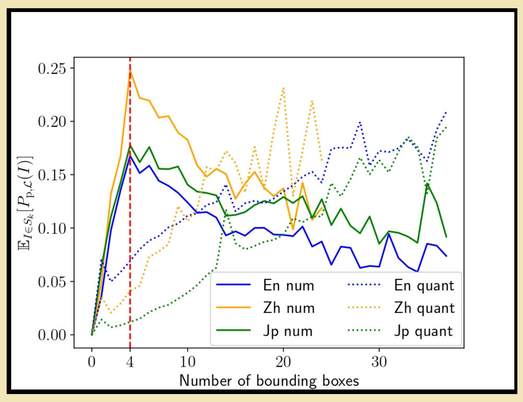
By object class

- Computed **mean usage of numerals** across **object classes**
- **Wild animal** classes' mean was high **across all languages**



By number of objects

- Computed **mean usage of numerals** across **numbers of objects** in the image
- In all languages, usage **increases to 4** and then decreases



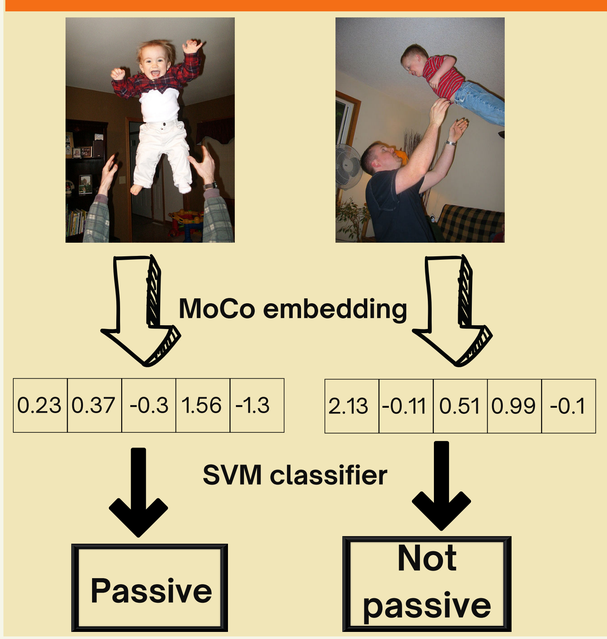
By role and pose

- Top images: same **role and pose,** annotators **used numerals**
- Bottom: annotators **didn't use numerals**

CORPUS ANALYSIS - PASSIVE VOICE

Images described using passive voice **in all languages**:
The passive agent is **centered** (by the **pose** of the camera or the **borders** of the image)

PREDICTING PROPERTIES FROM IMAGES



Language	Numerals	Passive	Negation	Transitivity	Verb root
English	**68.3**	66.8	62.5	64.6	58.8
German	**69.5**	58.5	51.5	62.0	57.8
Chinese	**80.6**	70.9	55.4	65.8	67.3
Japanese	**67.4**	-	-	-	-
Multilingual	**76.4**	66.2	62.6	64.7	63.1

CONCLUSION

- Above chance performance in **predicting linguistic properties** from **images**
- **Non-linguistic semantic context** (type, number, pose of objects) affects **linguistic properties**
- Intermediate step in **Image captioning:** Structure planning was shown to **improve generalization**

SCAN ME

REFERENCES

Rissman, Lilia, Amanda Woodward, and Susan Goldin-Meadow. "Occluding the face diminishes the conceptual accessibility of an animate agent." Language, cognition and neuroscience 34.3 (2019): 273-288.

CONTACT: