

# Human Biases in Multilingual Models

Cristina España-Bonet  
DFKI GmbH

Language in the Human-Machine Era Workshop

Budapest, Hungary

28th August 2023

## The (Undesired) Attenuation of Human Biases by Multilinguality

Cristina España-Bonet, Alberto Barrón-Cedeño

### Abstract

Some human preferences are universal. The odor of vanilla is perceived as pleasant all around the world. We expect neural models trained on human texts to exhibit these kind of preferences, i.e. biases, but we show that this is not always the case. We explore 16 static and contextual embedding models in 9 languages and, when possible, compare them under similar training conditions. We introduce and release CA-WEAT, multilingual cultural aware tests to quantify biases, and compare them to previous English-centric tests. Our experiments confirm that monolingual static embeddings do exhibit human biases, but values differ across languages, being far from universal. Biases are less evident in contextual models, to the point that the original human association might be reversed. Multilinguality proves to be another variable that attenuates and even reverses the effect of the bias, specially in contextual multilingual models. In order to explain this variance among models and languages, we examine the effect of asymmetries in the training corpus, departures from isomorphism in multilingual embedding spaces and discrepancies in the testing measures between languages.

**Anthology ID:** 2022.emnlp-main.133

**Volume:** [Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing](#)

**Month:** December

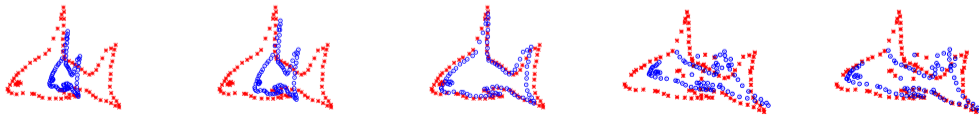
**Year:** 2022

# Motivation

## What does Multilinguality Involve in Machine Learning Models?

Most multilingual models just use a combination of monolingual corpora for training.

Are **we** distorting semantics?



[[https://en.wikipedia.org/wiki/Point-set\\_registration](https://en.wikipedia.org/wiki/Point-set_registration)]

Most multilingual models just use a combination of monolingual corpora for training.

Are **we** distorting semantics?

- We need something that is language and cultural independent
- We chose non-social (human) biases for this

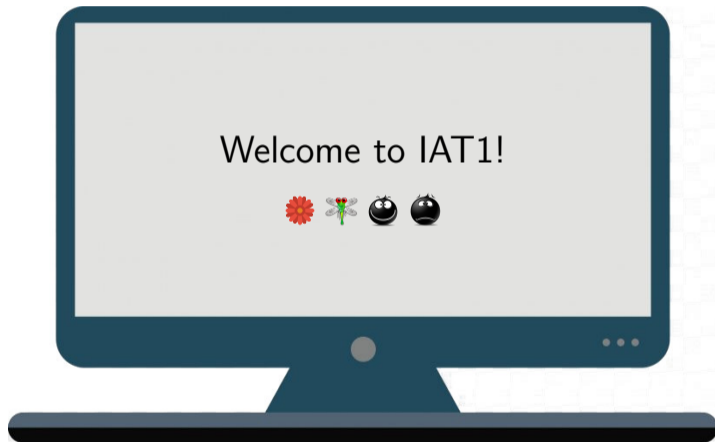
# The (Undesired) Attenuation of Human Biases by ML

## Outline

- 1 Measuring Biases
- 2 Multilinguality and Cultural-Aware WEAT (CA-WEAT)
- 3 Experiments
- 4 Conclusions

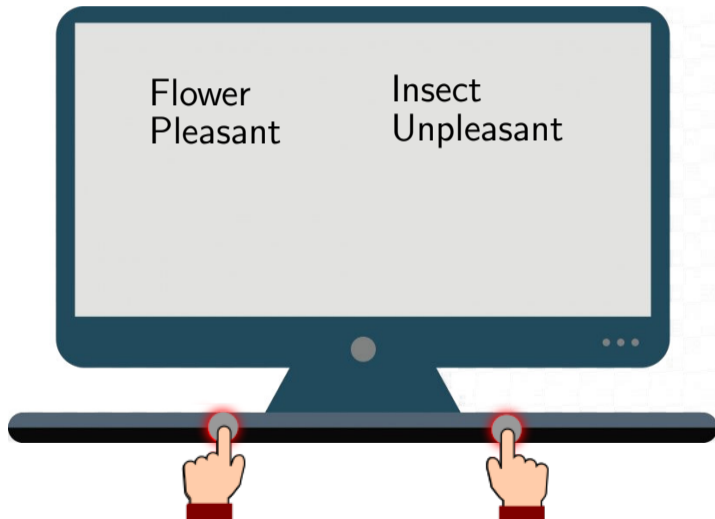
# IAT: Implicit Association Tests

## Non-Social Human Biases



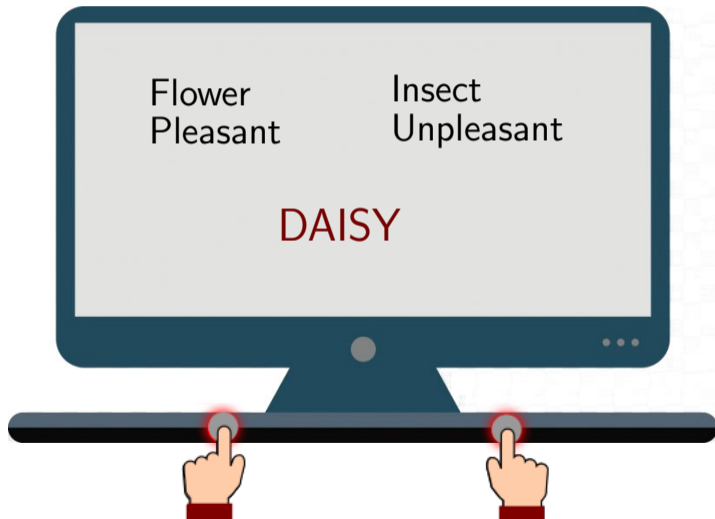
# IAT: Implicit Association Tests

## Non-Social Human Biases



# IAT: Implicit Association Tests

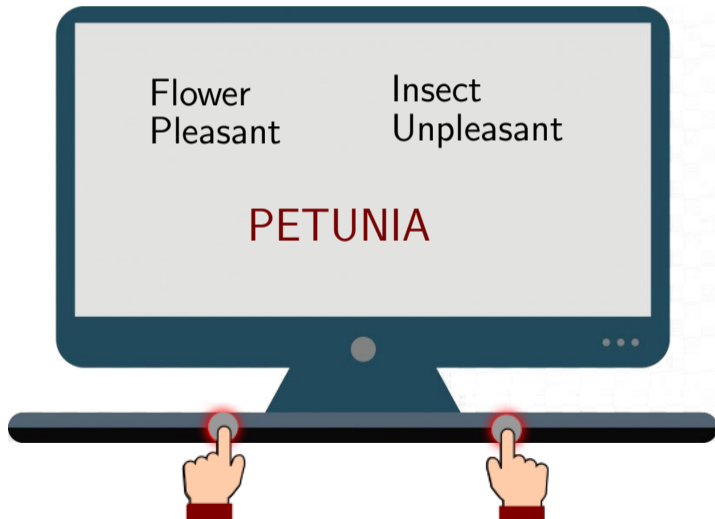
## Non-Social Human Biases





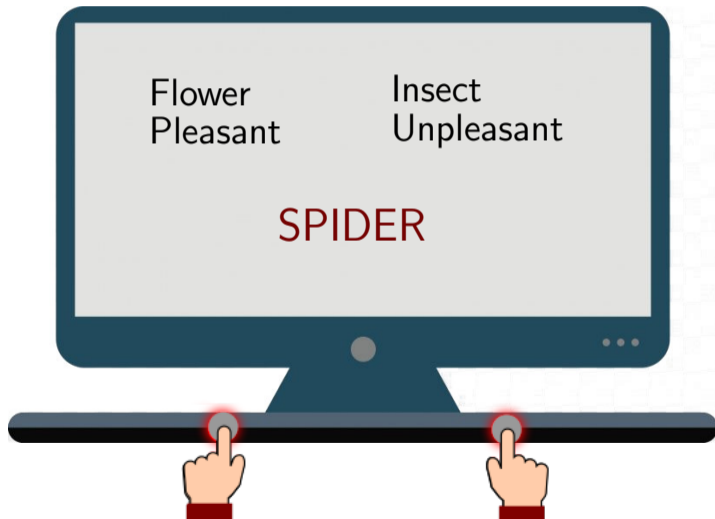
# IAT: Implicit Association Tests

## Non-Social Human Biases



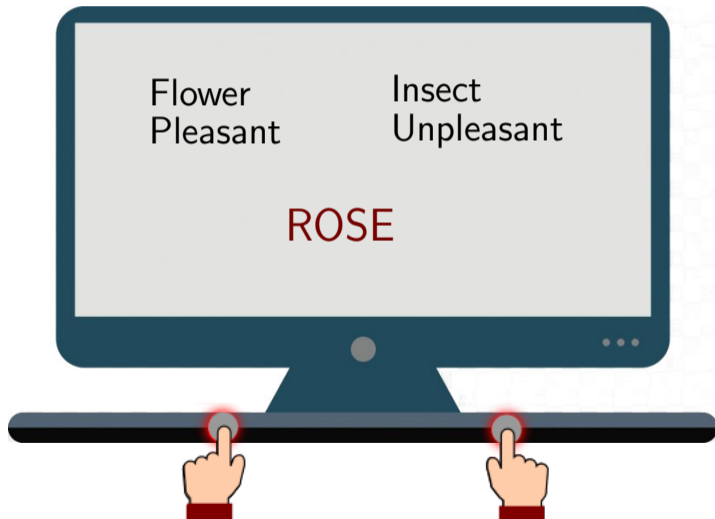
# IAT: Implicit Association Tests

## Non-Social Human Biases



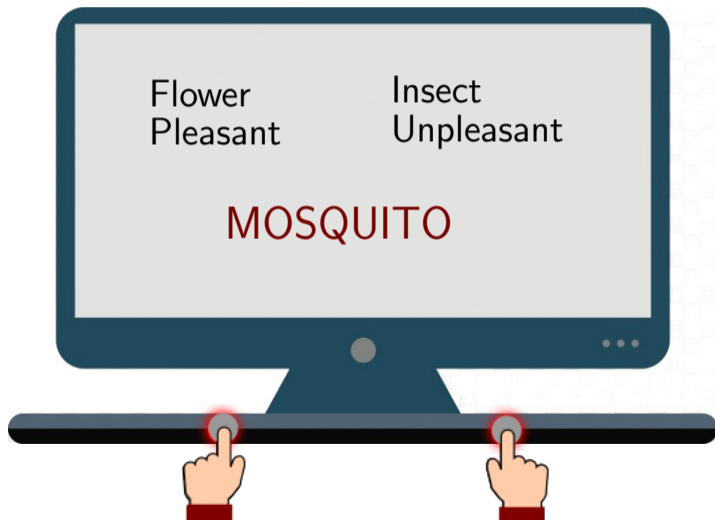
# IAT: Implicit Association Tests

## Non-Social Human Biases



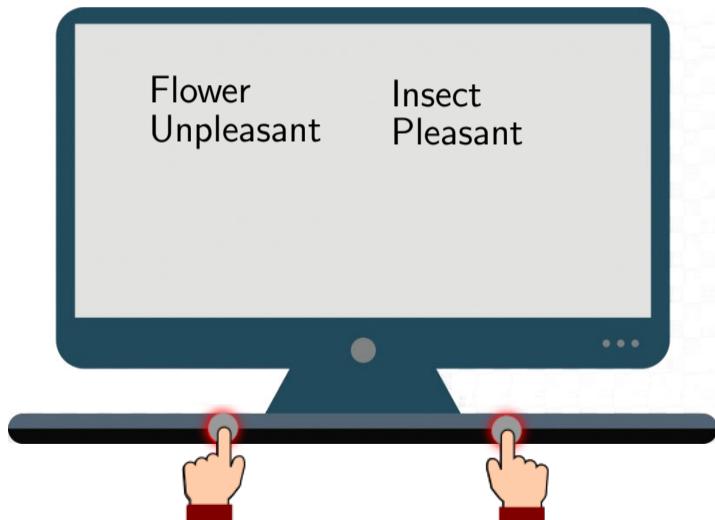
# IAT: Implicit Association Tests

## Non-Social Human Biases



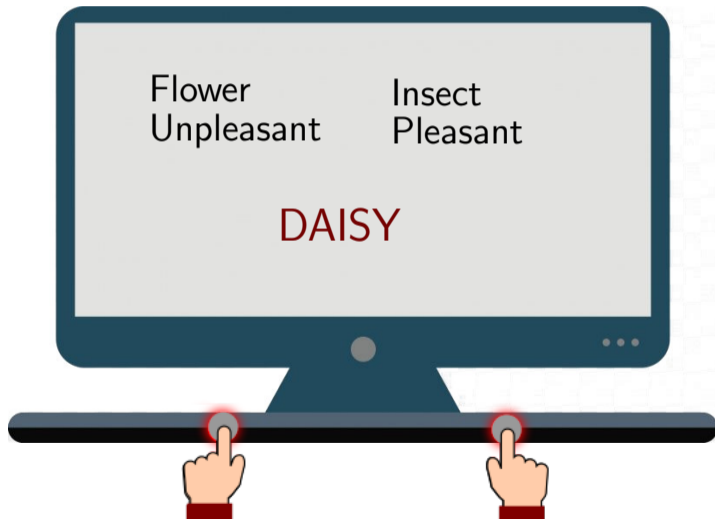
# IAT: Implicit Association Tests

## Non-Social Human Biases



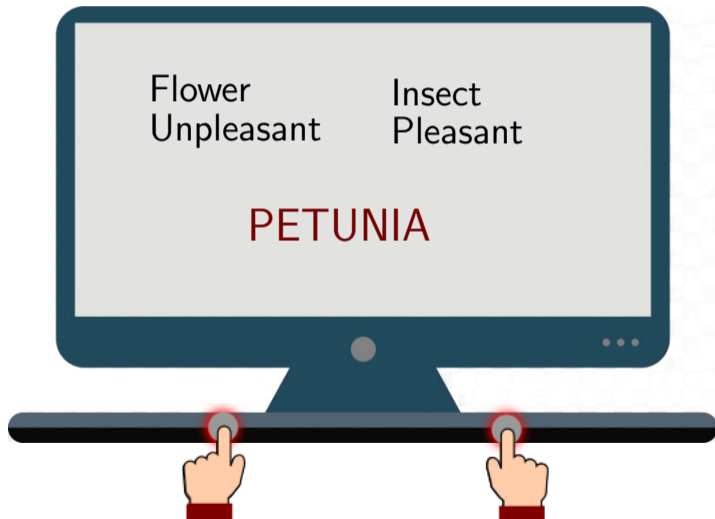
# IAT: Implicit Association Tests

## Non-Social Human Biases



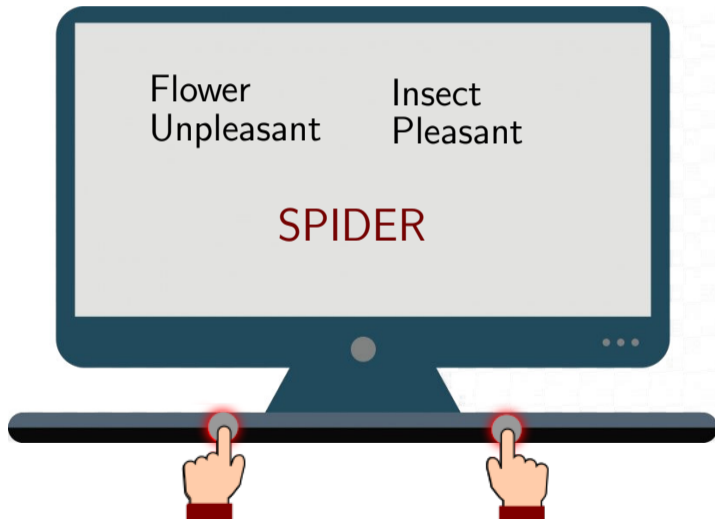
# IAT: Implicit Association Tests

## Non-Social Human Biases



# IAT: Implicit Association Tests

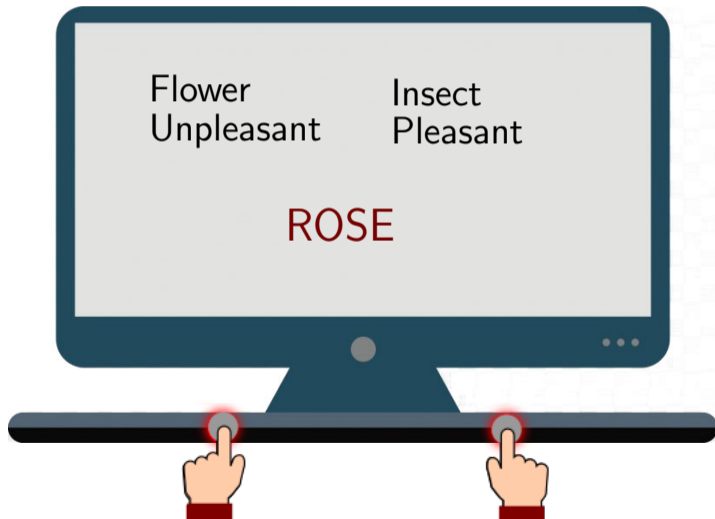
## Non-Social Human Biases





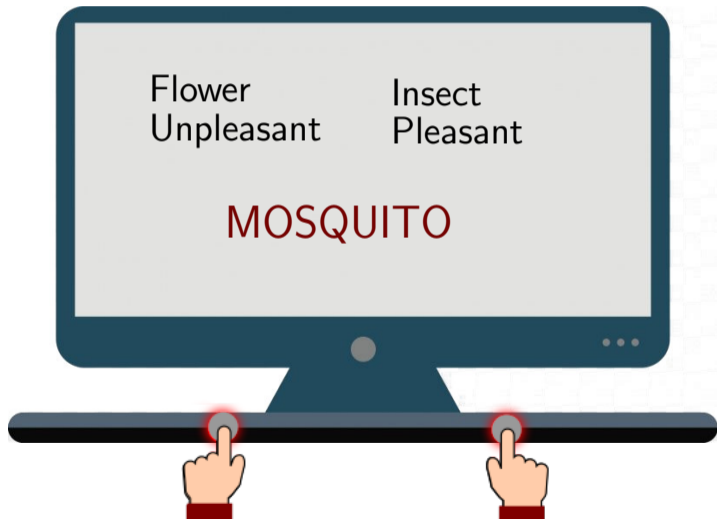
# IAT: Implicit Association Tests

## Non-Social Human Biases



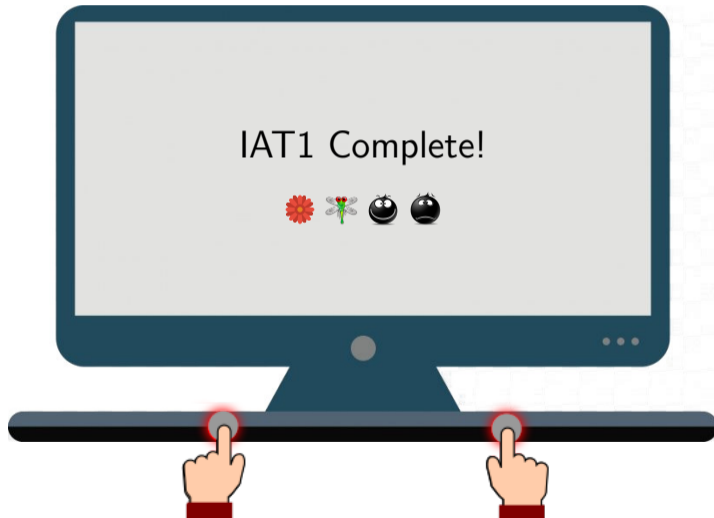
# IAT: Implicit Association Tests

## Non-Social Human Biases



# IAT: Implicit Association Tests

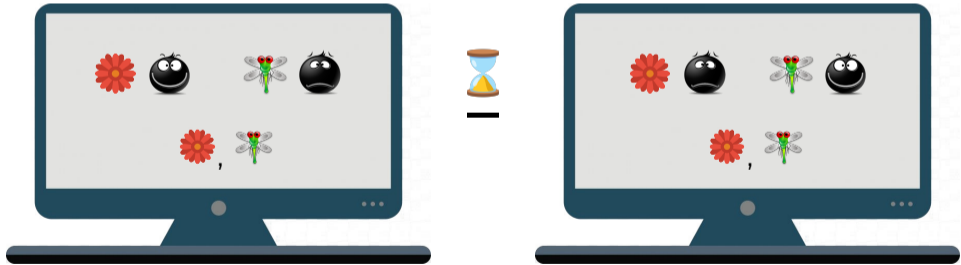
## Non-Social Human Biases



# IAT: Implicit Association Tests

## Non-Social Human Biases

IAT1: difference in response time  
(flowers & insects)



# IAT: Implicit Association Tests

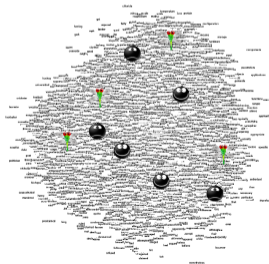
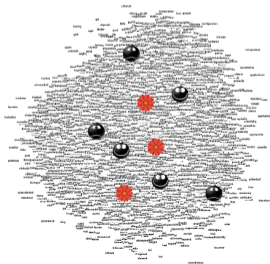
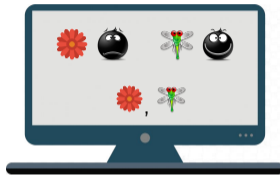
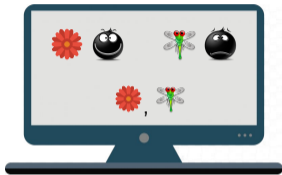
## Non-Social Human Biases

IAT2: difference in response time  
(musical instruments & weapons)



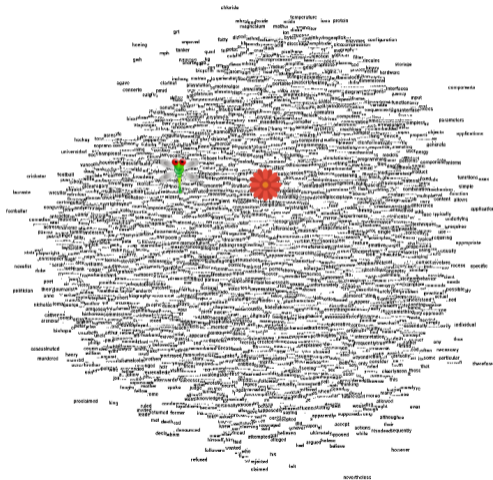
# WEAT: Association Tests in Word Embeddings

## WEAT, Intuition



# WEAT: Association Tests in Word Embeddings

Intuition, in our Embedding Space we can Measure Distances

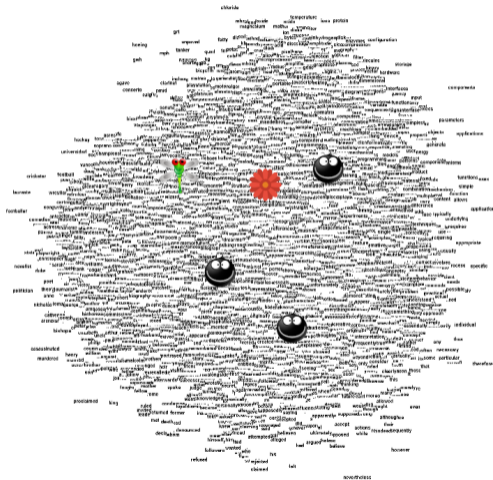






# WEAT: Association Tests in Word Embeddings

Intuition, in our Embedding Space we can Measure Distances

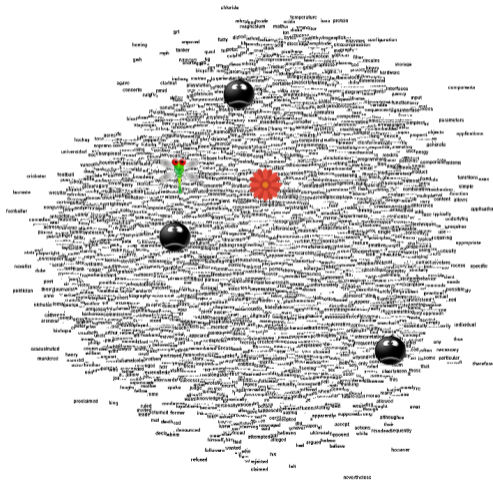


$$\frac{\sum_{\text{smiley} \in \vec{w}} \cos(\vec{w}, \text{smiley})}{|\vec{w}|}$$

$$\frac{\sum_{\text{smiley} \in \vec{w}} \cos(\vec{w}, \text{flower})}{|\vec{w}|}$$

# WEAT: Association Tests in Word Embeddings

Intuition, in our Embedding Space we can Measure Distances

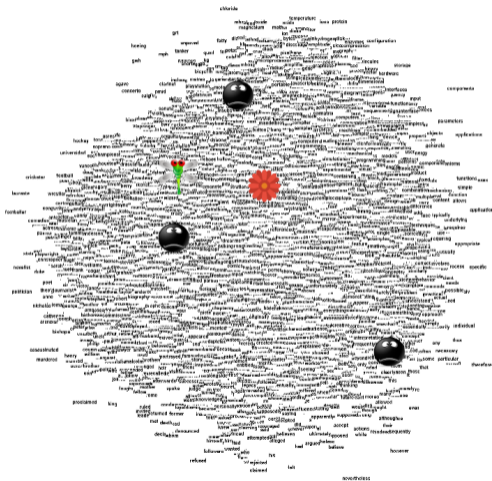


$$\frac{\sum_{\bullet \in \vec{b}} \cos(\bullet, \bullet)}{|\vec{b}|}$$

$$\frac{\sum_{\bullet \in \vec{b}} \cos(\bullet, \bullet)}{|\vec{b}|}$$

# WEAT: Association Tests in Word Embeddings

Intuition, in our Embedding Space we can Measure Distances



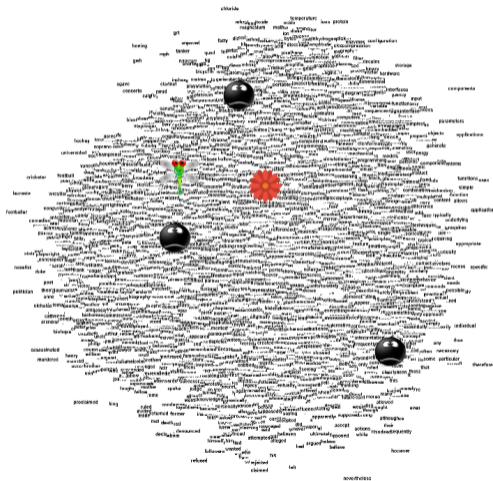
$$\frac{\sum_{\bullet \in \vec{a}} \cos(\vec{t}, \bullet)}{|\vec{a}|}$$

$$\frac{\sum_{\bullet \in \vec{a}} \cos(\vec{t}, \bullet)}{|\vec{a}|}$$

$$assoc(t, A) = \frac{\sum_{a \in A} \cos(\mathbf{t}, \mathbf{a})}{|A|}$$

# WEAT: Association Tests in Word Embeddings

Intuition, in our Embedding Space we can Measure Distances



$$\frac{\sum_{\text{bomb} \in \vec{t}} \cos(\vec{t}, \text{bomb})}{|\vec{t}|}$$

$$\frac{\sum_{\text{bomb} \in \vec{t}} \cos(\vec{t}, \text{flower})}{|\vec{t}|}$$

$$\text{assoc}(t, A) = \frac{\sum_{a \in A} \cos(\vec{t}, \vec{a})}{|A|}$$

$$\Delta_{\text{assoc}}(t, A, B) = \text{assoc}(t, A) - \text{assoc}(t, B)$$

# WEAT: Association Tests in Word Embeddings

## What do we Measure?

The difference in association for a term:

$$\Delta_{assoc}(t, A, B) = assoc(t, A) - assoc(t, B)$$

The statistic:

$$s(X, Y, A, B) = \sum_{x \in X} \Delta_{assoc}(x, A, B) - \sum_{y \in Y} \Delta_{assoc}(y, A, B)$$

$$s(\vec{\text{red flower}}, \vec{\text{green flower}}, \vec{\text{bomb}}, \vec{\text{bomb}}) = \sum_{\vec{\text{red flower}} \in \vec{\text{red flower}}} \Delta_{assoc}(\vec{\text{red flower}}, \vec{\text{bomb}}, \vec{\text{bomb}}) - \sum_{\vec{\text{green flower}} \in \vec{\text{green flower}}} \Delta_{assoc}(\vec{\text{green flower}}, \vec{\text{bomb}}, \vec{\text{bomb}})$$

# WEAT: Association Tests in Word Embeddings

What do we Measure?

$$s(\vec{w}_1, \vec{w}_2, \vec{w}_3, \vec{w}_4) = \sum_{\vec{w} \in \vec{w}_1} \Delta_{\text{assoc}}(\vec{w}, \vec{w}_3, \vec{w}_4) - \sum_{\vec{w} \in \vec{w}_2} \Delta_{\text{assoc}}(\vec{w}, \vec{w}_3, \vec{w}_4)$$



# WEAT: Association Tests in Word Embeddings

## What do we Measure?

The statistic:

$$s(\vec{w}_1, \vec{w}_2, \vec{w}_3, \vec{w}_4) = \sum_{\vec{w}_1 \in \vec{w}_1} \Delta_{assoc}(\vec{w}_1, \vec{w}_3, \vec{w}_4) - \sum_{\vec{w}_2 \in \vec{w}_2} \Delta_{assoc}(\vec{w}_2, \vec{w}_3, \vec{w}_4)$$

The size effect:

$$d(\vec{w}_1, \vec{w}_2, \vec{w}_3, \vec{w}_4) = \frac{\mu(\Delta_{assoc}(\vec{w}_1, \vec{w}_3, \vec{w}_4)_{\forall \vec{w}_1 \in \vec{w}_1}) - \mu(\Delta_{assoc}(\vec{w}_2, \vec{w}_3, \vec{w}_4)_{\forall \vec{w}_2 \in \vec{w}_2})}{\sigma(\Delta_{assoc}(\vec{w}_1, \vec{w}_3, \vec{w}_4)_{\forall \vec{w}_1 \in \vec{w}_1 \cup \vec{w}_2})}$$

# WEAT: Association Tests in Word Embeddings

## Do Word Embeddings Reflect Implicit Human Associations?

[Caliskan et al., Nature, 2017]

Semantics derived automatically from language corpora contain human-like biases:

- morally neutral as toward insects or flowers, —our non-social—
- problematic as toward race or gender,
- veridical, reflecting the status quo distribution of gender with respect to careers or first names.



# WEAT: Association Tests in Word Embeddings

## Do Word Embeddings Reflect Implicit Human Associations?

[Caliskan et al., Nature, 2017]

Semantics derived automatically from language corpora contain human-like biases:

- morally neutral as toward insects or flowers, —our non-social—
- problematic as toward race or gender,
- veridical, reflecting the status quo distribution of gender with respect to careers or first names.

For multilinguality we need universals  $\Rightarrow$  non-social only

# Multilinguality and Cultural-Aware WEAT (CA-WEAT)

## Outline



- 1 Measuring Biases
  - IAT: Implicit Association Tests
  - WEAT: Association Tests in Word Embeddings
- 2 Multilinguality and Cultural-Aware WEAT (CA-WEAT)**
- 3 Experiments
  - Wide Overview
  - WEAT vs X-WEAT vs CA-WEAT
  - Data Asymmetries and Isomorphism
- 4 Conclusions

# Multilinguality and Cultural-Aware WEAT

## WEAT1 and WEAT2 Original Lists



---

### WEAT1 target items

	Flowers	aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia
	Insects	ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil



---

### WEAT2 target items

	Instruments	bagpipe, cello, guitar, lute, trombone, banjo, clarinet, harmonica, mandolin, trumpet, bassoon, drum, harp, oboe, tuba, bell, fiddle, harpsichord, piano, viola, bongo, flute, horn, saxophone, violin
	Weapons	arrow, club, gun, missile, spear, axe, dagger, harpoon, pistol, sword, blade, dynamite, hatchet, rifle, tank, bomb, firearm, knife, shotgun, teargas, cannon, grenade, mace, slingshot, whip

---

### WEAT1 and WEAT2 attributes

	Pleasant	caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation
	Unpleasant	abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison

---

# Multilinguality and Cultural-Aware WEAT

## Original and X-WEAT Lists

### Original version (WEAT1, WEAT2)

[Battig and Montague, 1969; Bellezza et al., 1986; Greenwald et al., 1998]

- Collected from college students in Eastern US
- Frequent terms
- Non-ambiguous terms

# Multilinguality and Cultural-Aware WEAT

## Original and X-WEAT Lists

### Original version (WEAT1, WEAT2)

[Battig and Montague, 1969; Bellezza et al., 1986; Greenwald et al., 1998]

- Collected from college students in Eastern US
- Frequent terms
- Non-ambiguous terms

### Multilingual version (X-WEAT)

[Lauscher and Glavaš, 2019; Lauscher et al., 2020]

- Literal translation
- Arabic, Croatian, German, Italian, Russian, Spanish and Turkish

# Multilinguality and Cultural-Aware WEAT

Features and Issues with WEAT and X-WEAT (the safe version :-)


- **WEAT**: American English, represents the culture of the (Eastern) US
- **X-WEAT**: Multilingual, but represents the culture of the (Eastern) US!  
—and this applies to all NLP using translation—
  - duplicates? (violin, fiddle → violín)
  - frequent terms? (gnat → jején)
  - non-ambiguous terms? (blade → hoja)
- **CA-WEAT**: Multilingual and culturally aware

# Multilinguality and Cultural-Aware WEAT

## CA-WEAT

Cultural Aware WEAT

Preguntes Respostes 86 Configuració



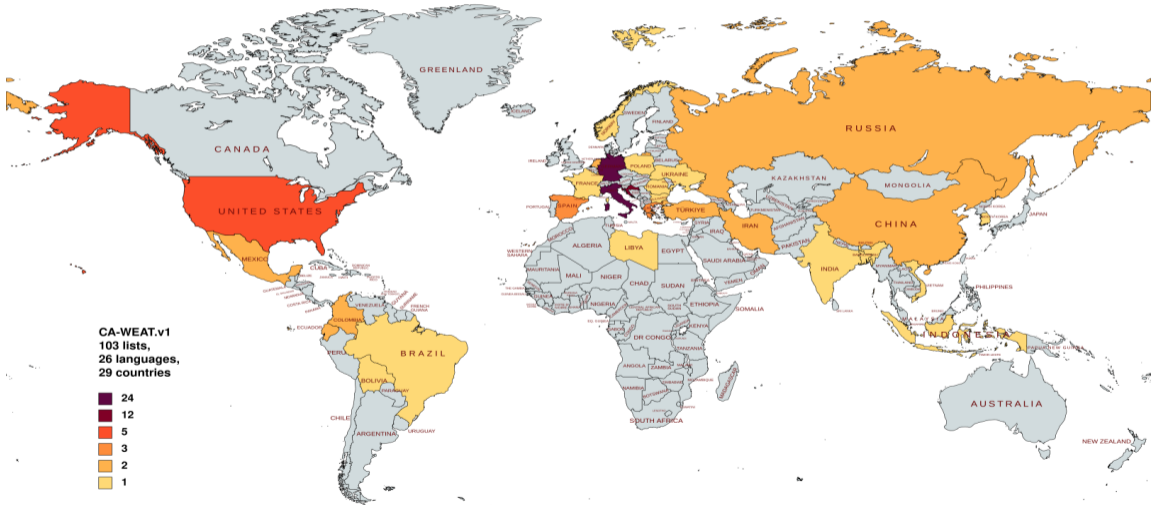
Secció 1 de 3

### Cultural Aware WEAT

The words we use when communicating are related to our culture and environment. There is less need for us to use words that reflect concepts that do not appear in our everyday life, and everyday life is different in each country. With this form, we try to collect lists of words from all around the world that reflect different cultures. You might have travelled a lot, either in person or through reading, but we'd like you to focus on your home only and list words that are relevant there.

# Multilinguality and Cultural-Aware WEAT

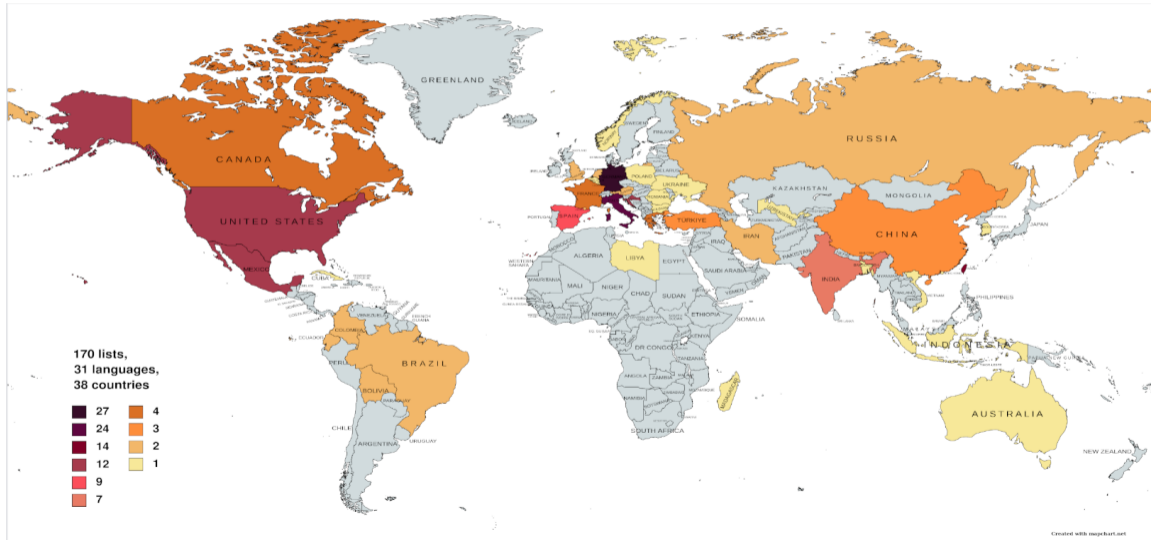
CA-WEATs per Country (not the best Distribution!)





# Multilinguality and Cultural-Aware WEAT

A few more Today!



# Multilinguality and Cultural-Aware WEAT

## Monolingual (English) lists



apple, pear, grape, strawberry, blackberry, blueberry, raspberry, plum, apricot, orange, tangerine, clementine, lemon, lime, watermelon, pepper, squash, pumpkin, tomato, banana, pineapple, fig, date, mango, papaya



apple, banana, orange, grape, cherry, strawberry, raspberry, blueberry, tangerine, mango, peach, nectarine, pineapple, plum, mandarin, kiwi, papaya, blackberry, blackcurrant, redcurrant, apricot, raisin, gooseberry, pear, melon



apple, orange, banana, kiwi, grape, lemon, cherry, pear, strawberry, blueberry, raspberry, blackberry, avocado, lime, mango, peach, plum, apricot, nectarine, pineapple, papaya, watermelon, lychee, longan, durian

# Multilinguality and Cultural-Aware WEAT

## Monolingual (English) lists



apple, pear, grape, strawberry, blackberry, blueberry, raspberry, plum, apricot, orange, tangerine, clementine, lemon, lime, watermelon, pepper, squash, pumpkin, tomato, banana, pineapple, fig, date, mango, papaya



apple, banana, orange, grape, cherry, strawberry, raspberry, blueberry, tangerine, mango, peach, nectarine, pineapple, plum, mandarin, kiwi, papaya, blackberry, blackcurrant, redcurrant, apricot, raisin, gooseberry, pear, melon



apple, orange, banana, kiwi, grape, lemon, cherry, pear, strawberry, blueberry, raspberry, blackberry, avocado, lime, mango, peach, plum, apricot, nectarine, pineapple, papaya, watermelon, lychee, longan, durian

# Multilinguality and Cultural-Aware WEAT

## Monolingual (English) lists



US English

apple, pear, grape, strawberry, blackberry, blueberry, raspberry, plum, apricot, orange, tangerine, clementine, lemon, lime, watermelon, pepper, squash, pumpkin, tomato, banana, pineapple, fig, date, mango, papaya



UK English

apple, banana, orange, grape, cherry, strawberry, raspberry, blueberry, tangerine, mango, peach, nectarine, pineapple, plum, mandarin, kiwi, papaya, blackberry, blackcurrant, redcurrant, apricot, raisin, gooseberry, pear, melon



AU English

apple, orange, banana, kiwi, grape, lemon, cherry, pear, strawberry, blueberry, raspberry, blackberry, avocado, lime, mango, peach, plum, apricot, nectarine, pineapple, papaya, watermelon, lychee, longan, durian

# Multilinguality and Cultural-Aware WEAT

## Multilingual Lists. Disclaimer: my Translation...



### US English

apple, pear, grape, **strawberry**, blackberry, blueberry, **raspberry**, plum, apricot, orange, **tangerine**, clementine, **lemon**, lime, **watermelon**, pepper, squash, pumpkin, tomato, **banana**, **pineapple**, fig, date, mango, papaya

### BR Portuguese

apple, **banana**, guava, **pineapple**, apricot, pear, **watermelon**, orange, **lemon**, cherry, **tangerine**, kiwi, pequi, açai, cashew, hog plum, soursop, **strawberry**, **raspberry**, blackberry, plum, peach, passion fruit, lychee, jabuticaba

maçã, banana, goiaba, abacaxi, damasco, pêra, melancia, laranja, limão, cereja, mexerica, kiwi, pequi, açai, caju, cajá, graviola, morango, framboesa, amora, ameixa, pêssego, maracujá, lichia, jabuticaba

### Traditional Chinese

**banana**, **apple**, **pineapple**, guava, orange, grape, peach, cherry, blueberry, Java apple, papaya, lychee, **strawberry**, tomato, cantaloupe, **tangerine**, **lemon**, lime, **raspberry**, Japanese banana, sugarcane, **watermelon**, durian, sugar apple, coconut

香蕉, 蘋果, 鳳梨, 芭樂, 柳丁, 葡萄, 水蜜桃, 櫻桃, 藍莓, 蓮霧, 木瓜, 荔枝, 草莓, 番茄, 哈密瓜, 橘子, 檸檬, 萊姆, 覆盆莓, 芭蕉, 甘蔗, 西瓜, 榴蓮, 釋迦, 椰子

# Multilinguality and Cultural-Aware WEAT

## Multilingual Lists. Disclaimer: my Translation...



### US English

apple, pear, grape, **strawberry**, blackberry, blueberry, **raspberry**, plum, apricot, **orange**, **tangerine**, clementine, **lemon**, lime, **watermelon**, pepper, squash, pumpkin, tomato, **banana**, **pineapple**, fig, date, mango, papaya

### BR Portuguese

apple, **banana**, guava, **pineapple**, apricot, pear, **watermelon**, **orange**, **lemon**, cherry, **tangerine**, kiwi, **pequi**, **açaí**, cashew, **hog plum**, **soursop**, **strawberry**, **raspberry**, blackberry, plum, peach, **passion fruit**, lychee, **jabuticaba**

maçã, banana, goiaba, abacaxi, damasco, pêra, melancia, laranja, limão, cereja, mexerica, kiwi, pequi, açaí, caju, cajá, graviola, morango, framboesa, amora, ameixa, pêssego, maracujá, lichia, jabuticaba

### Traditional Chinese

**banana**, **apple**, **pineapple**, guava, **orange**, grape, peach, cherry, blueberry, **Java apple**, papaya, lychee, **strawberry**, tomato, cantaloupe, **tangerine**, **lemon**, lime, **raspberry**, **Japanese banana**, sugarcane, **watermelon**, **durian**, **sugar apple**, **coconut**

香蕉, 蘋果, 鳳梨, 芭樂, 柳丁, 葡萄, 水蜜桃, 櫻桃, 藍莓, 蓮霧, 木瓜, 荔枝, 草莓, 番茄, 哈密瓜, 橘子, 檸檬, 萊姆, 覆盆莓, 芭蕉, 甘蔗, 西瓜, 榴蓮, 釋迦, 椰子

# Multilinguality and Cultural-Aware WEAT

## (Cross-lingual) Cultural Biases in NLP

- Disclaimer: 1 list is just an example, 100 lists start saying something
- But CA-WEATs seem different to X-WEAT!
- Multilingual models trained on an asymmetric distribution of data
  - Most of it US English
  - Yes, also chatGPT :-)
- "Write an article about agriculture"

- 1 Measuring Biases
  - IAT: Implicit Association Tests
  - WEAT: Association Tests in Word Embeddings
- 2 Multilinguality and Cultural-Aware WEAT (CA-WEAT)
- 3 Experiments**
  - Wide Overview
  - WEAT vs X-WEAT vs CA-WEAT
  - Data Asymmetries and Isomorphism
- 4 Conclusions



# Experiments

## Embedding Models & Languages

❖ Pre-trained fastText word embeddings

WP

WPali

CCWP

❖ Comparable word embeddings with a subset of CC-100

CCe

CCeVMuns

CCeVMsup

CCe2langs

CCe9langs

❖ Word embeddings extracted from contextual models at different layers

BERT

mBERT

XLM

XGLM

# Experiments

## Embedding Models & Languages

❖ Pre-trained fastText word embeddings

WP

WPali

CCWP

❖ Comparable word embeddings with a subset of CC-100

CCe

CCeVMuns

CCeVMsup

CCe2langs

CCe9langs

❖ Word embeddings extracted from contextual models at different layers

BERT

mBERT

XLM

XGLM

🗨️ Languages

Arabic (ar), Catalan (ca), Croatian (hr), English (en), German (de), Italian (it),  
Russian (ru), Spanish (es) and Turkish (tr)

# Experiments

## What we Report here (More in the Paper!)

### ■ Size effect

$$d(\vec{r}_1, \vec{r}_2, \vec{r}_3, \vec{r}_4) = \frac{\mu\left(\Delta_{assoc}(\vec{r}_1, \vec{r}_2, \vec{r}_3)_{\forall \vec{r}_1 \in \vec{r}_1}\right) - \mu\left(\Delta_{assoc}(\vec{r}_2, \vec{r}_3, \vec{r}_4)_{\forall \vec{r}_2 \in \vec{r}_2}\right)}{\sigma\left(\Delta_{assoc}(\vec{r}_1, \vec{r}_2, \vec{r}_3)_{\forall \vec{r}_1 \in \vec{r}_1 \cup \vec{r}_2}\right)}$$

**Sawilowsky's scale:** very small ( $d < 0.01$ ), small ( $< 0.20$ ), medium ( $< 0.50$ ), large ( $< 0.80$ ), very large ( $< 1.20$ ), and huge ( $< 2.00$ )

- CA-WEAT: median and 95% CI with order statistics
- WEAT, CA-WEAT, X-WEAT: 5,000 bootstraps (median and 95% CI)

# Experiments

## What we Report here (More in the Paper!)

### ■ Size effect

$$d(\vec{r}, \vec{g}, \vec{b}, \vec{b}) = \frac{\mu\left(\Delta_{\text{assoc}}(\vec{r}, \vec{b}, \vec{b})_{\forall \vec{r} \in \vec{r}}\right) - \mu\left(\Delta_{\text{assoc}}(\vec{g}, \vec{b}, \vec{b})_{\forall \vec{g} \in \vec{g}}\right)}{\sigma\left(\Delta_{\text{assoc}}(\vec{r}, \vec{b}, \vec{b})_{\forall \vec{r} \in \vec{r} \cup \vec{g}}\right)}$$

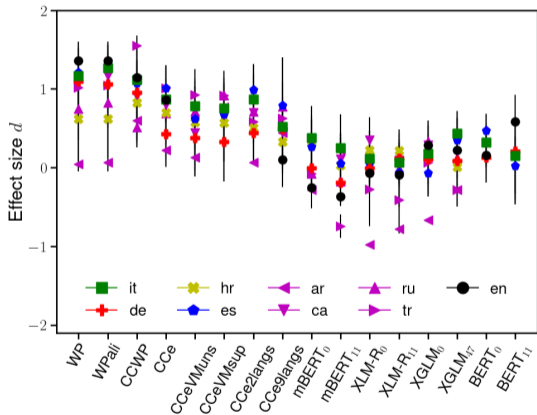
**Sawilowsky's scale:** very small ( $d < 0.01$ ), small ( $< 0.20$ ), medium ( $< 0.50$ ), large ( $< 0.80$ ), very large ( $< 1.20$ ), and huge ( $< 2.00$ )

- CA-WEAT: median and 95% CI with order statistics
- WEAT, CA-WEAT, X-WEAT: 5,000 bootstraps (median and 95% CI)
- IAT1 ( $\vec{r} \vec{g}$ ); IAT2 ( $\vec{r} \vec{g}$ ) is equivalent

Do our embeddings show (human) biases?  
All embedding models? All languages?

# Experiments

## Wide Overview (WEAT, CA-WEAT)

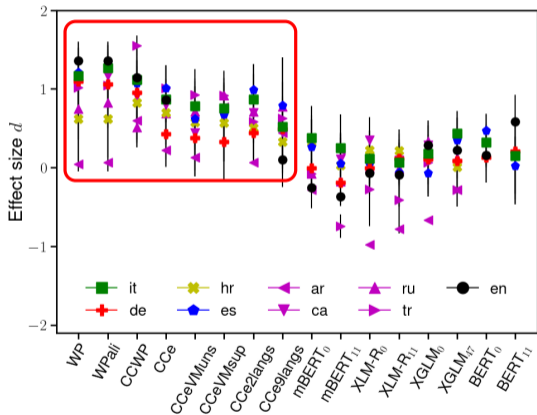


# Experiments

## Wide Overview (WEAT, CA-WEAT)

Word embeddings:

- All WE models have  $d > 0$

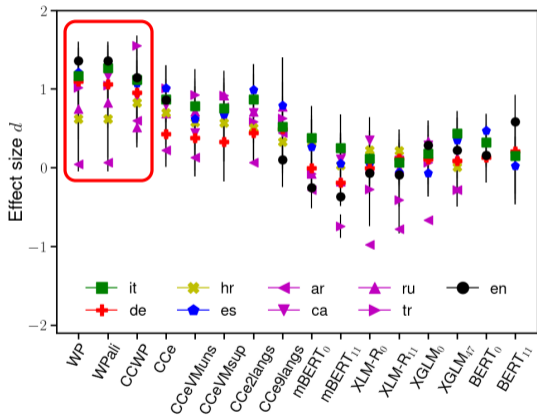


# Experiments

## Wide Overview (WEAT, CA-WEAT)

Word embeddings:

- All WE models have  $d > 0$
- Pre-trained models have higher  $\sigma$  across languages



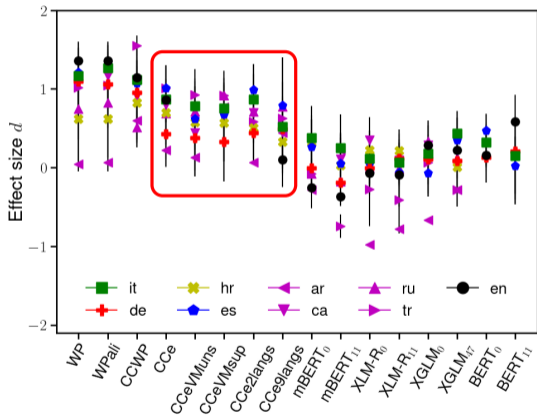


# Experiments

## Wide Overview (WEAT, CA-WEAT)

Word embeddings:

- All WE models have  $d > 0$
- Pre-trained models have higher  $\sigma$  across languages
- Equivalent projection methods

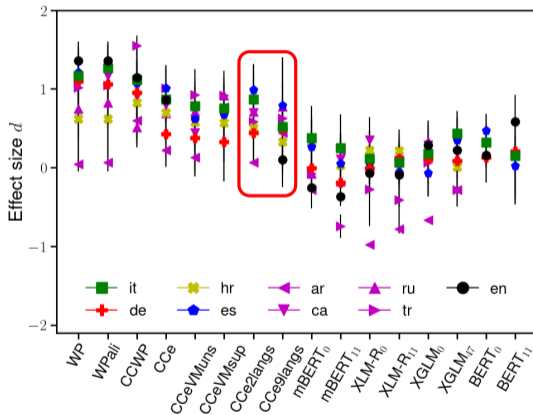


# Experiments

## Wide Overview (WEAT, CA-WEAT)

Word embeddings:

- All WE models have  $d > 0$
- Pre-trained models have higher  $\sigma$  across languages
- Equivalent projection methods
- Multilinguality attenuates

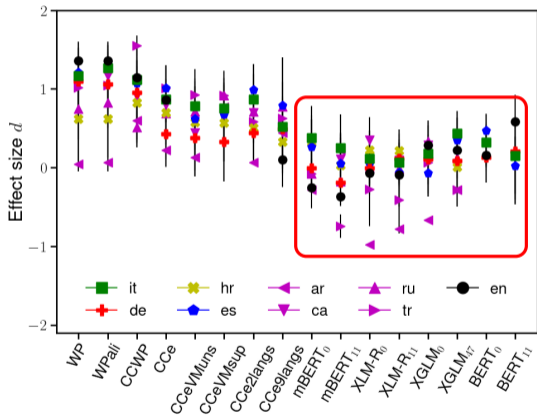


# Experiments

## Wide Overview (WEAT, CA-WEAT)

Contextual embeddings:

■  $d$  compatible with no bias

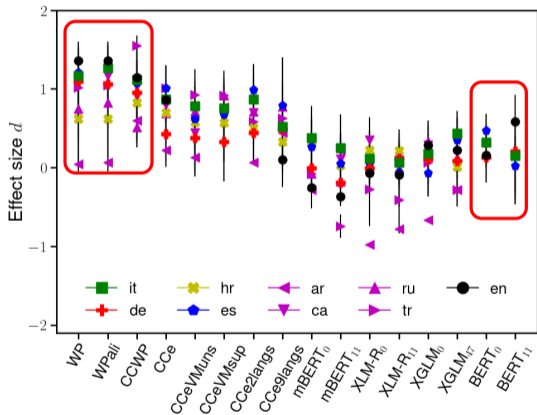


# Experiments

## Wide Overview (WEAT, CA-WEAT)

Contextual embeddings:

- $d$  compatible with no bias
- Effect of contextualisation

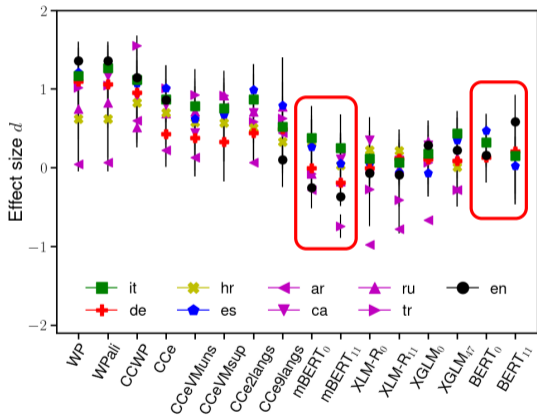


# Experiments

## Wide Overview (WEAT, CA-WEAT)

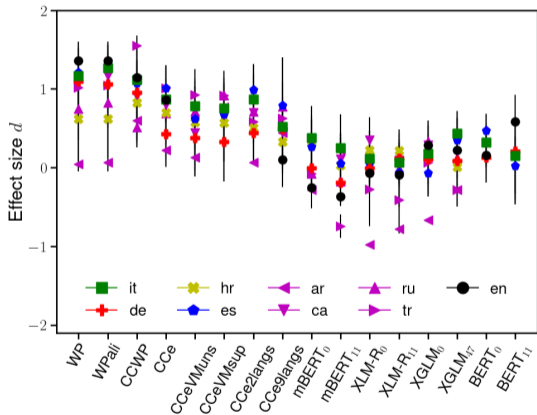
Contextual embeddings:

- $d$  compatible with no bias
- Effect of contextualisation
- But multilinguality attenuates further



# Experiments

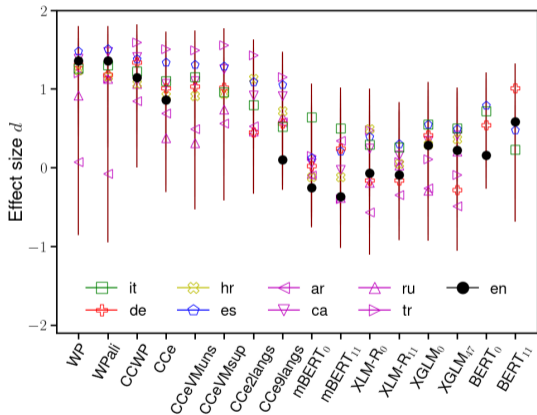
## Wide Overview (CA-WEAT vs X-WEAT)



# Experiments

## Wide Overview (CA-WEAT vs X-WEAT)

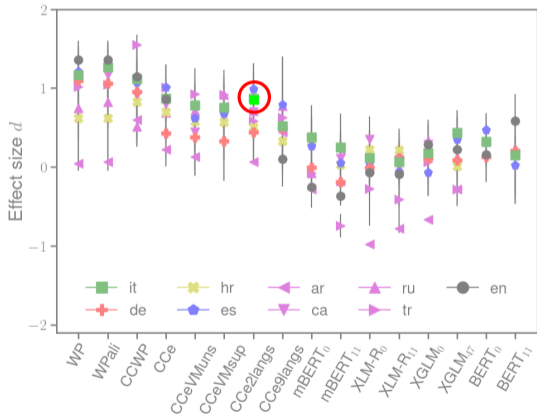
- X-WEAT shows similar trends as CA-WEAT
- **But!** With a higher dispersion across languages
- No universal  $d$



# Experiments

## Wide Overview (CA-WEAT vs X-WEAT)

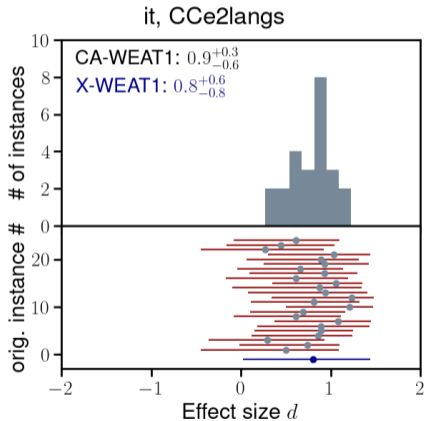
Let's focus!





# Experiments

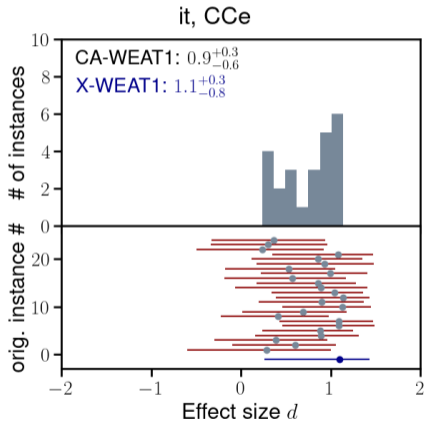
## WEAT vs X-WEAT vs CA-WEAT



- Lists show a high dispersion (bootstrapped and averaged)
- X-WEAT lies within CA-WEAT (close languages)

# Experiments

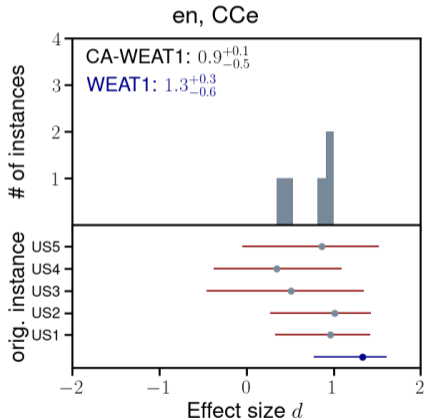
## WEAT vs X-WEAT vs CA-WEAT



- Lists show a high dispersion (bootstrapped and averaged)
- X-WEAT lies within CA-WEAT (close languages)
- Distributions non-normal (yet!)

# Experiments

## WEAT vs X-WEAT vs CA-WEAT



- Lists show a high dispersion (bootstrapped and averaged)
- X-WEAT lies within CA-WEAT (close languages)
- Distributions non-normal (yet!)
- English interesting for further study

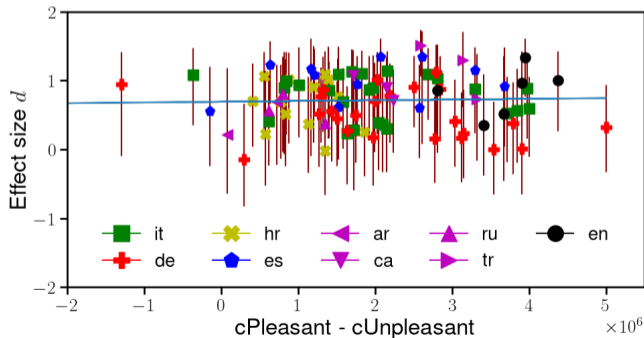
Why is  $d$  non-universal?

Is it data differences? Is it forcing multilinguality?  
Is it the dispersion?

# Experiments

## Why is $d$ non-universal?

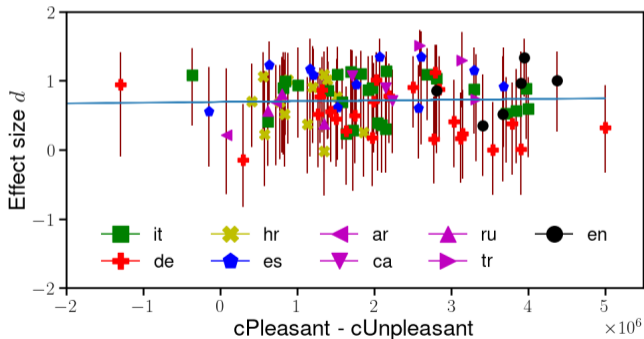
- Asymmetries in term frequencies are not a reason (Pleasant vs Unpleasant terms in CCe)  
 $\rho = 0.0$ ; explains 0% of the variance



# Experiments

## Comparison with Previous Work

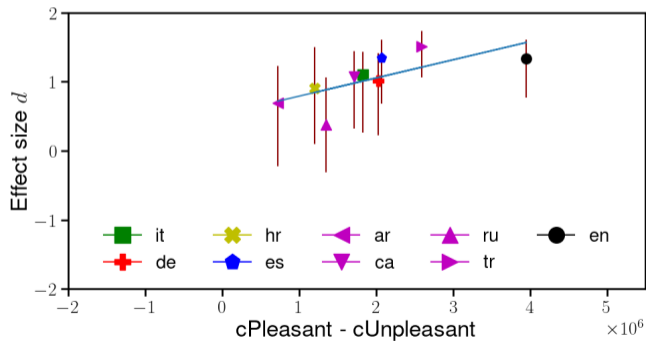
WEAT1+X-WEAT1+CA-WEAT1: no relation



# Experiments

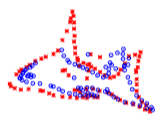
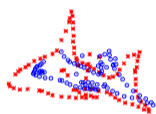
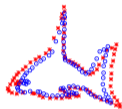
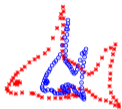
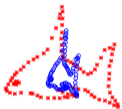
## Comparison with Previous Work

X-WEAT1: Simpson's paradox?



# Experiments

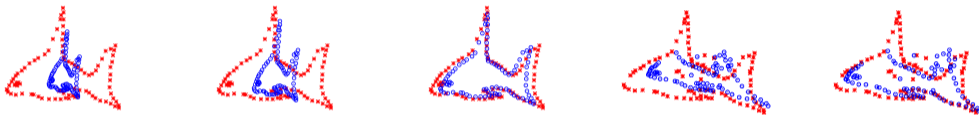
## Isomorphism





# Experiments

## Isomorphism



- Measures: Gromov-Hausdorff (GH) distance and Eigenvector similarity (EV)
- Isomorphism between a language (sub-)space and the English (sub-)space
- For contextual models we consider the vocab from CcE

# Experiments

## Isomorphism between a Language (sub-)Space and the English (sub-)Space

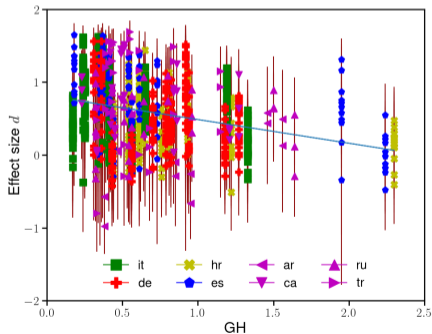
	ar		ca		de		es		hr		it		ru		tr	
	EV	GH	EV	GH	EV	GH	EV	GH	EV	GH	EV	GH	EV	GH	EV	GH
WP	106	0.47	12	0.49	12	0.31	10	0.18	42	0.54	21	0.24	16	0.43	49	0.39
WPali	143	0.55	22	0.51	22	0.36	16	0.37	46	0.61	19	0.34	30	0.32	36	0.44
CCWP	15	0.40	85	0.42	42	0.92	23	0.41	51	0.65	41	0.37	32	0.64	28	0.55
CCe	55	0.62	253	0.23	26	0.79	166	0.54	91	0.61	223	0.25	8	0.56	25	0.43
CCeVMuns	229	1.56	229	1.27	27	0.82	167	1.95	69	0.93	220	1.19	27	0.96	36	0.84
CCeVMsup	36	0.56	231	0.86	32	0.70	87	0.73	27	0.61	123	0.65	25	0.80	11	0.41
CCe2langs	93	0.53	8	0.43	19	0.94	72	0.35	33	0.81	51	0.41	39	0.51	64	0.61
CCe9langs	475	1.46	23	0.84	171	1.27	21	0.61	53	1.22	51	0.41	403	1.50	149	1.15
mBERT <sub>0</sub>	154	0.85	133	0.33	95	0.56	99	0.56	270	0.44	131	0.17	161	0.54	589	0.51
XLM-R <sub>0</sub>	54	0.38	74	0.45	59	0.43	150	0.44	58	0.54	113	0.56	111	0.32	277	0.33
XGLM <sub>0</sub>	67	0.95	88	1.21	144	1.18	135	2.24	*2584	*2.30	130	1.33	85	1.64	475	0.68

- No clear distinction between WE and CE wrt. isomorphism distances
- Language and embedding model effects are also mixed

# Experiments

## Why is $d$ non-universal?

- (Lack of) isomorphism between (sub-)spaces is not a reason either!  
 $\rho = -0.3$ ; explains 10% of the variance



# Conclusions

## Outline

- 1 Measuring Biases
  - IAT: Implicit Association Tests
  - WEAT: Association Tests in Word Embeddings
- 2 Multilinguality and Cultural-Aware WEAT (CA-WEAT)
- 3 Experiments
  - Wide Overview
  - WEAT vs X-WEAT vs CA-WEAT
  - Data Asymmetries and Isomorphism
- 4 Conclusions

# Conclusions

## Wrapping up

- Using (literal) translation in NLP does not in general preserve culture
- We therefore create CA-WEAT (in contrast to X-WEAT) to analyse desirable biases in embeddings across languages

# Conclusions

## Wrapping up

- Using (literal) translation in NLP does not in general preserve culture
- We therefore create CA-WEAT (in contrast to X-WEAT) to analyse desirable biases in embeddings across languages
- Monolingual and bilingual WE reproduce non-social human biases
- We do not observe a universal value even in the comparable setting
- Contextualisation and multilinguality attenuate biases, why?

- Using (literal) translation in NLP does not in general preserve culture
- We therefore create CA-WEAT (in contrast to X-WEAT) to analyse desirable biases in embeddings across languages
- Monolingual and bilingual WE reproduce non-social human biases
- We do not observe a universal value even in the comparable setting
- Contextualisation and multilinguality attenuate biases, why?
- Due to the large variability (models & languages) we want...

# Conclusions

## Future Work

- Better understanding of individual vs cultural differences
- Better understanding of intralanguage cultural differences
- Better understanding of language models

170 lists,  
31 languages,  
38 countries



...so, still collecting CA-WEATs!

<https://github.com/cristinae/CA-WEAT>





That's All Folks!

Thanks! And...

Questions?



# Extra Slides

## A Reviewer's Comment



There is a huge variability.

Shouldn't one use more (WEAT) tests?



How do we find more tests?!

We want universality...

[Arshamian et al., Current Biology, 2022]

- Culture plays a minimal role in the perception of odor pleasantness
- Individuals within cultures vary as to which odors they find pleasant
- Human olfactory perception is strongly constrained by universal principles

# Extra Slides

## The Perception of Odor Pleasantness is Shared Across Cultures

