Visual Gender Biases in Wikipedia

A Systematic Evaluation across the Ten Most Spoken Languages

Presented at Wikimedia Research/Showcase – April 2023

Beytía, Pablo | Agarwal, Pushkal | Redi, Miriam | Singh, Vivek K.

Context and problem

- Wikipedia tends to register **systematic disparities** in multiple content dimensions, including gender (Redi et al. 2021).

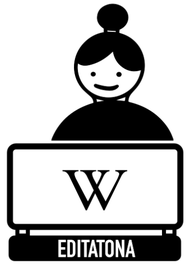
- **Less than 20%** of the biographies are about women (Wikidata 2020)**.**

- Previous research shows several **gender asymmetries** in content (e.g., in coverage, topics, lexicon, sources, and hypertext) (e.g., Graells-Garrido et al. 2015, Wagner et al. 2015, 2016)**.**

Context and problem

- Various **initiatives** have emerged to close this gender gap.

Women in red



- Few studies analyze the **joint and overlapping manifestation** of multiple gender biases. (e.g., Graells-Garrido et al. 2015, Wagner et al. 2015, 2016).

- Little research on **visual** gender biases (Young et. Al. 2016; Singh et al. 2020; Zagovora et al. 2017).

Our goal

Evaluate the content gender gap comprehensively and systematically

1

Considering the internal diversity of biographies

Different **Wikipedia versions**, different biases
(e.g. Overall & Rüger 2011).

Different **occupations**, different biases
(e.g. Singh et al. 2020; Zagovora et al. 2017).

2

Expanding the types of content analyzed

Analyzing written and **visual content** (multimodal approach).

Investigating content created in **multiple stages** of production.

Stages of visibility production in Wikipedia (Beytía & Wagner 2022)

BUILDING

SELECTION

POSITIONING

Stages of visibility production in Wikipedia (Beytía & Wagner 2022)

STAGE	*Editing process*	*Framing mode*	*Asymmetry*
POSITIONING	Association (classification / hyperlinks) Multilingual placement	Structural placement	Classification Network position Multilingual notability
BUILDING	Collaborative editing Discussion (talk pages)	Characterization	Writing length Lexical Topical Source
SELECTION	Topic Suggestions Acceptance	Representation	Article coverage Deletion

So, in what sense is this a **systematic approach**?

Research on visual gender biases in Wikipedia

PREVIOUS STUDIES
(Young et. Al. 2016; Singh et al. 2020; Zagovora et al. 2017)

- Monolingual focus (English or German only).
- A small number of biographies analyzed (1,000 articles at most).
- One typical "stage": the building of articles.
- One typical metric: the number of images.

OUR RESEARCH

- **Multilingual** approach: the ten most spoken languages in the world.
- **All biographical articles** on each Wikipedia version.
- Examining the **three editing "stages"**: selection, building, and positioning.
- **Multimodal** approach (textual and visual metrics)

(Beytía, Agarwal, Redi & Singh 2022)

Methods

Methods

- We **compiled from Wikidata** the list of the 6.22 million people with biography in some language (February 2020), with their gender, the main (most referenced) occupation, and birthplace.

- We **classified** genders (male / female / non-binary) and occupations (10 categories).

- We calculated the **quality of the articles** with an automatic classifier (Wikimedia's Language Agnostic Quality Classifier) based on the structure of the articles (e.g., number of sources, and length).

- In the 10 languages with the most speakers, we **collected all the images** available in the biographies.

- We estimated the **quality of each image** by training an Image Quality classifier (with 150k high-quality images and 150k random images from Wikimedia Commons).

(Beytía, Agarwal, Redi & Singh 2022)

Methods

- We calculated **8 metrics of gender asymmetries** in content:

	SELECTION	BUILDING		POSITIONING
Textual	Number of biographies (f/m*)	Average number of characters (f/m*)	Average article quality (f/m*)	Average number of languajes (f/m)
Visual	Number of biographies with images (f/m*)	Average number of images (f/m*)	Average image quality (f/m*)	Average number of languajes with images (f/m*)

* f/m = ratio between female and male biographies

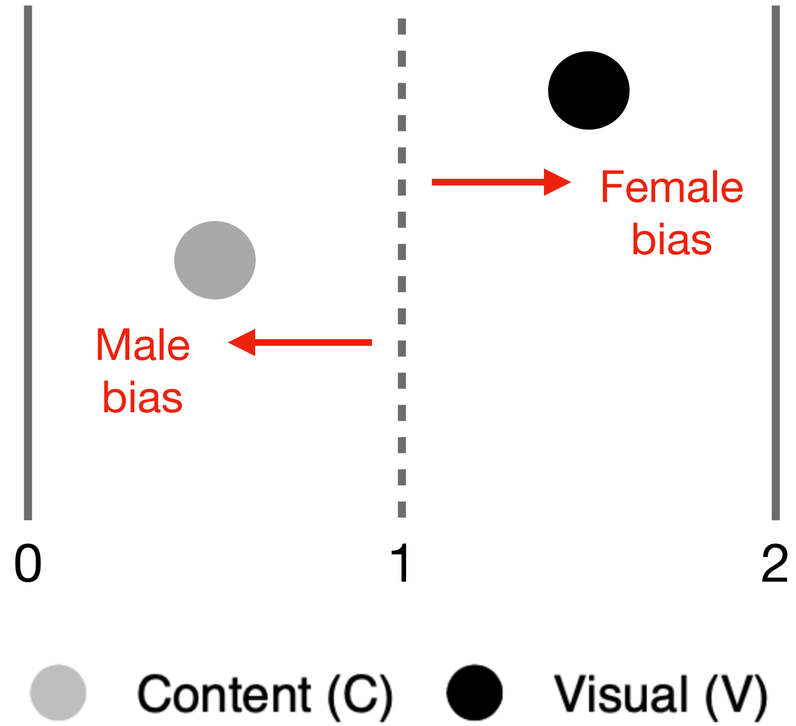
(Beytía, Agarwal, Redi & Singh 2022)

Results

Results

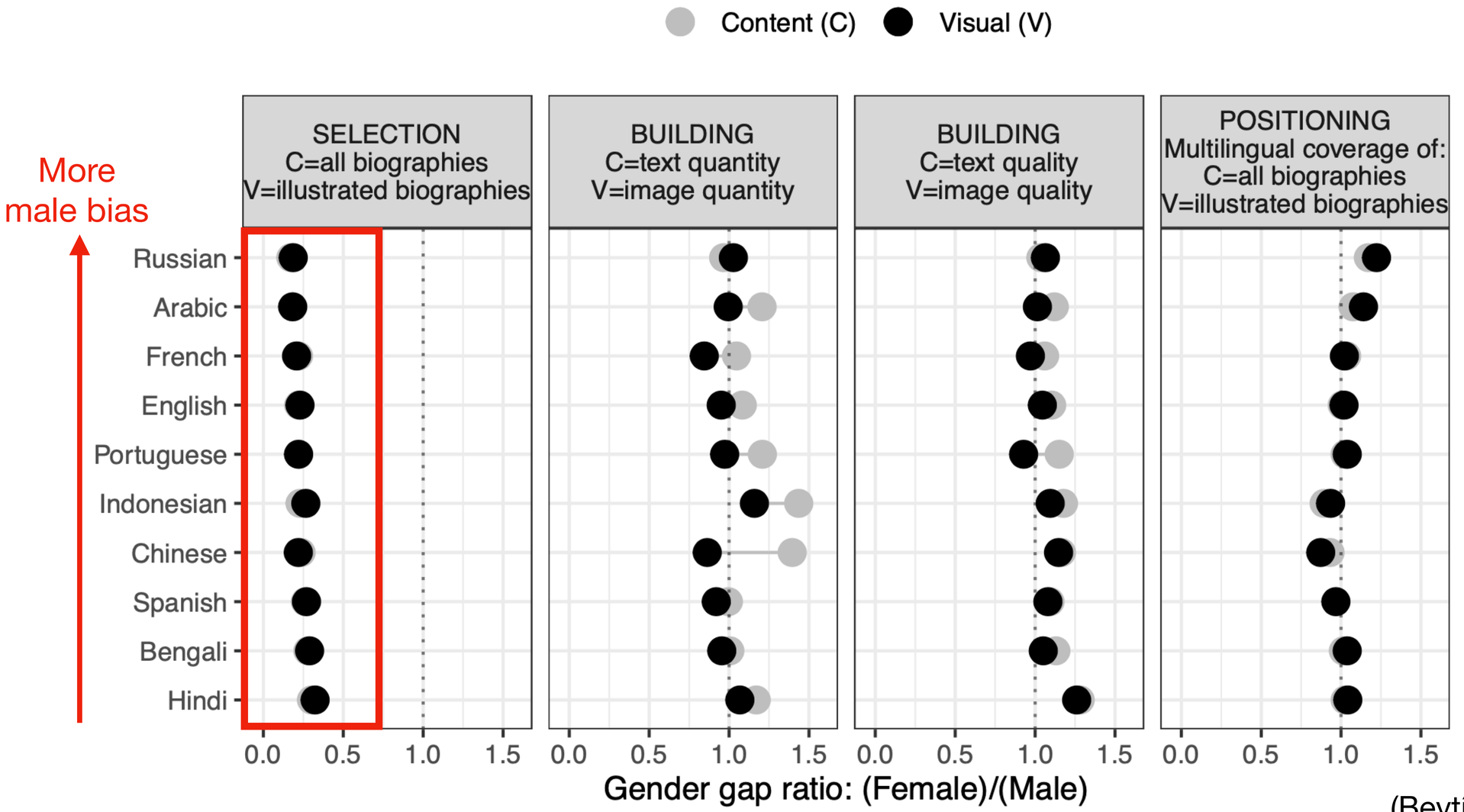
How to interpret our charts?

We use ratios: female content / male content



Results: multilingual analysis

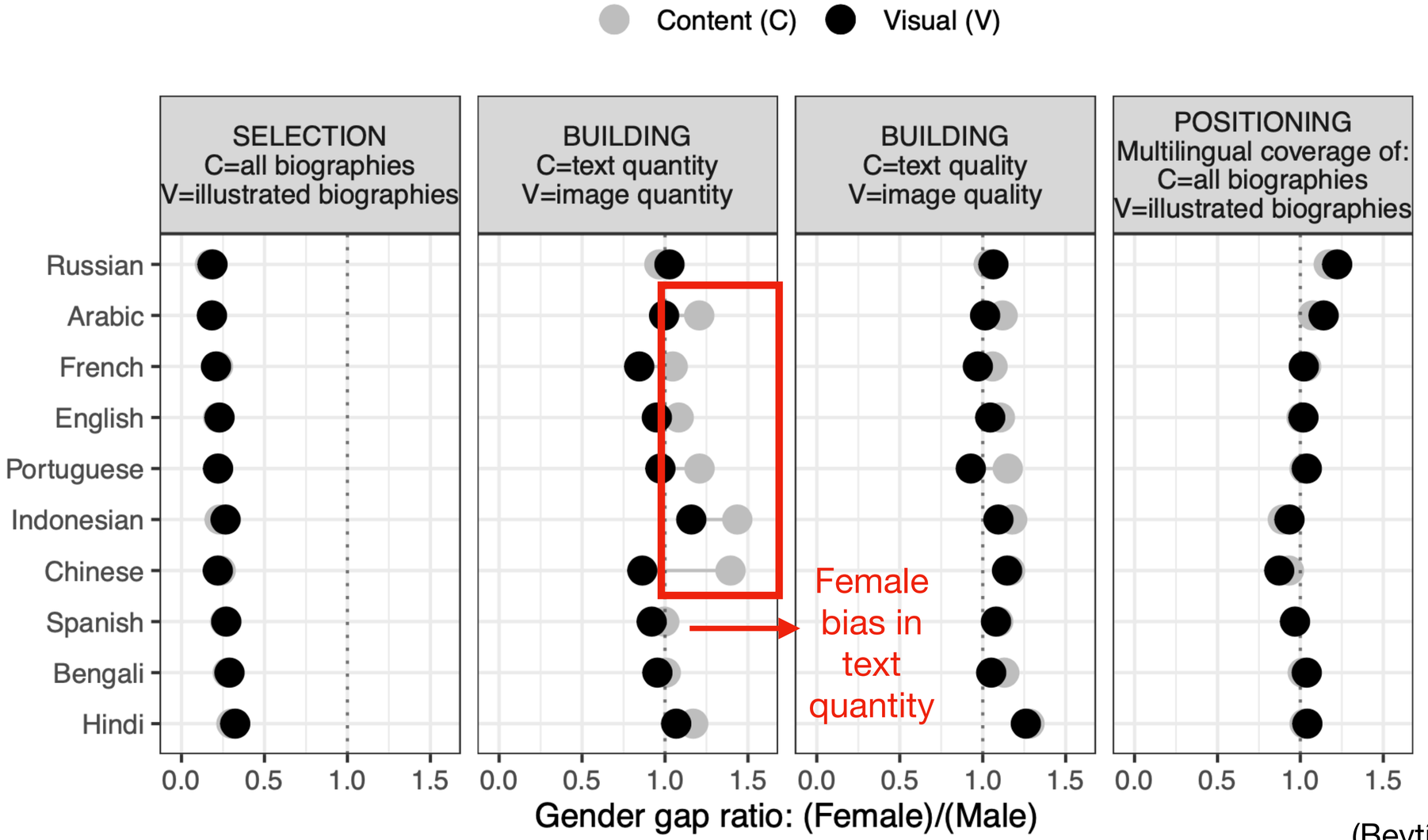
Visual and non-visual gender gaps by Wikipedia language version



(Beytía, Agarwal, Redi & Singh 2022)

Results: multilingual analysis

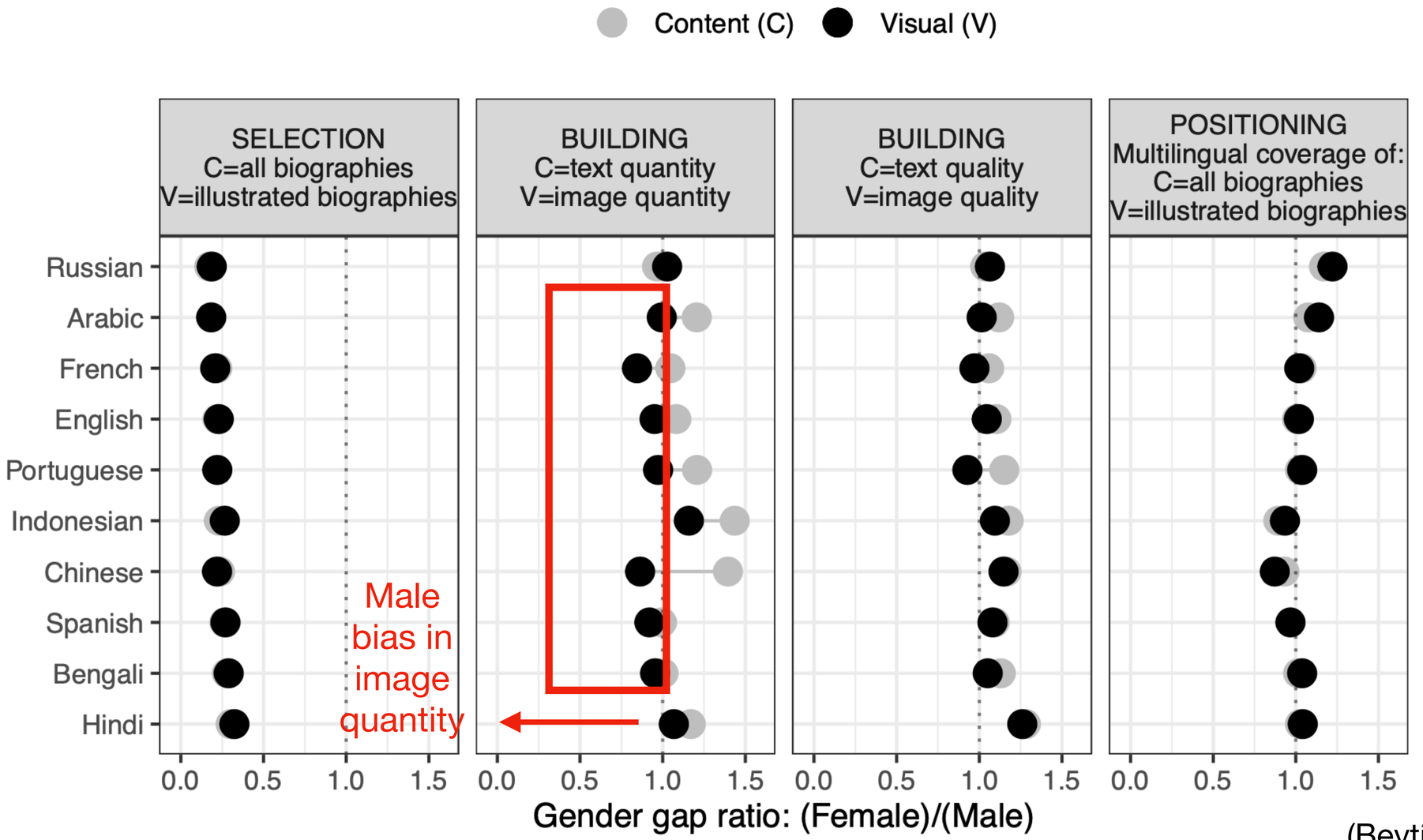
Visual and non-visual gender gaps by Wikipedia language version



(Beytía, Agarwal, Redi & Singh 2022)

Results: multilingual analysis

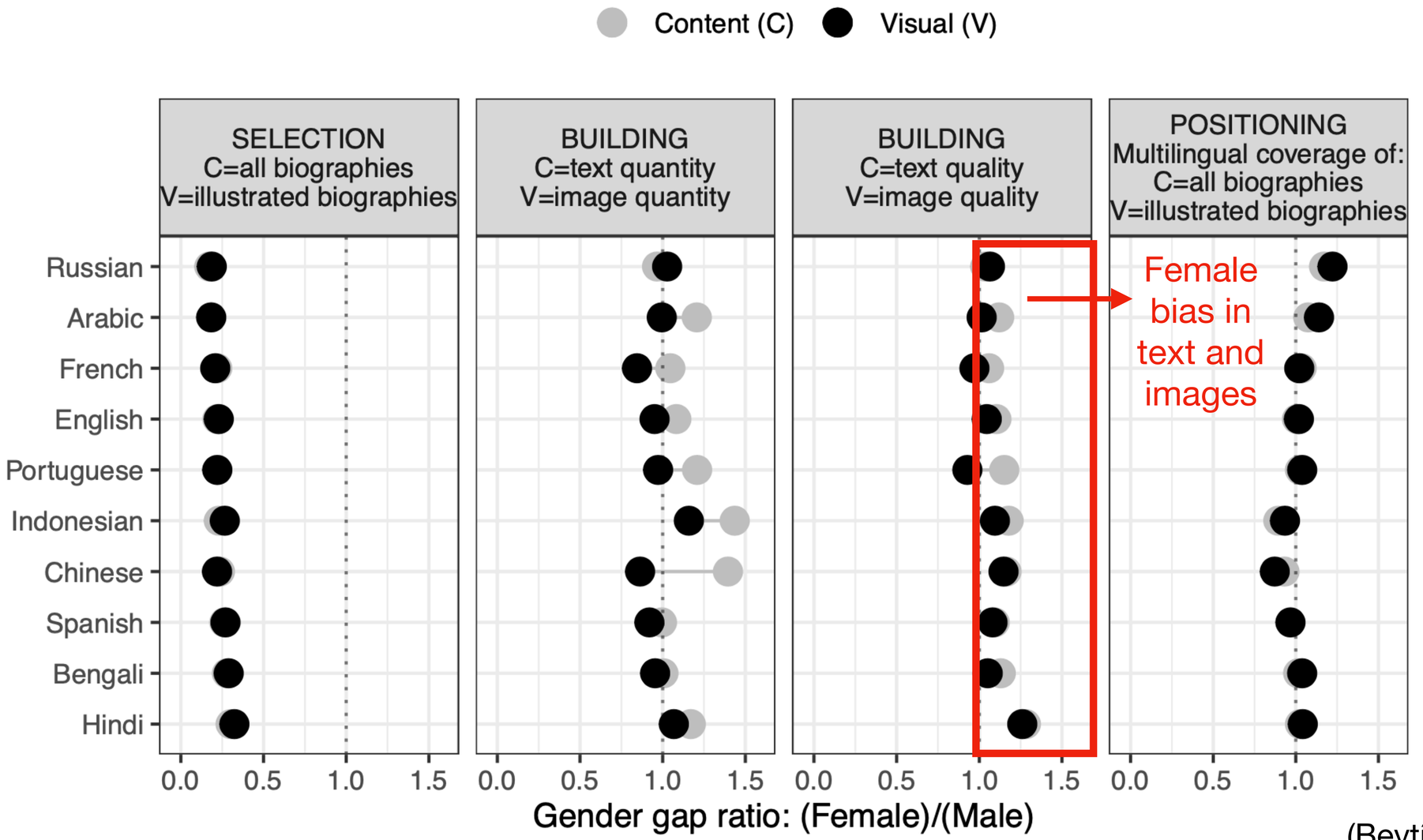
Visual and non-visual gender gaps by Wikipedia language version



(Beytía, Agarwal, Redi & Singh 2022)

Results: multilingual analysis

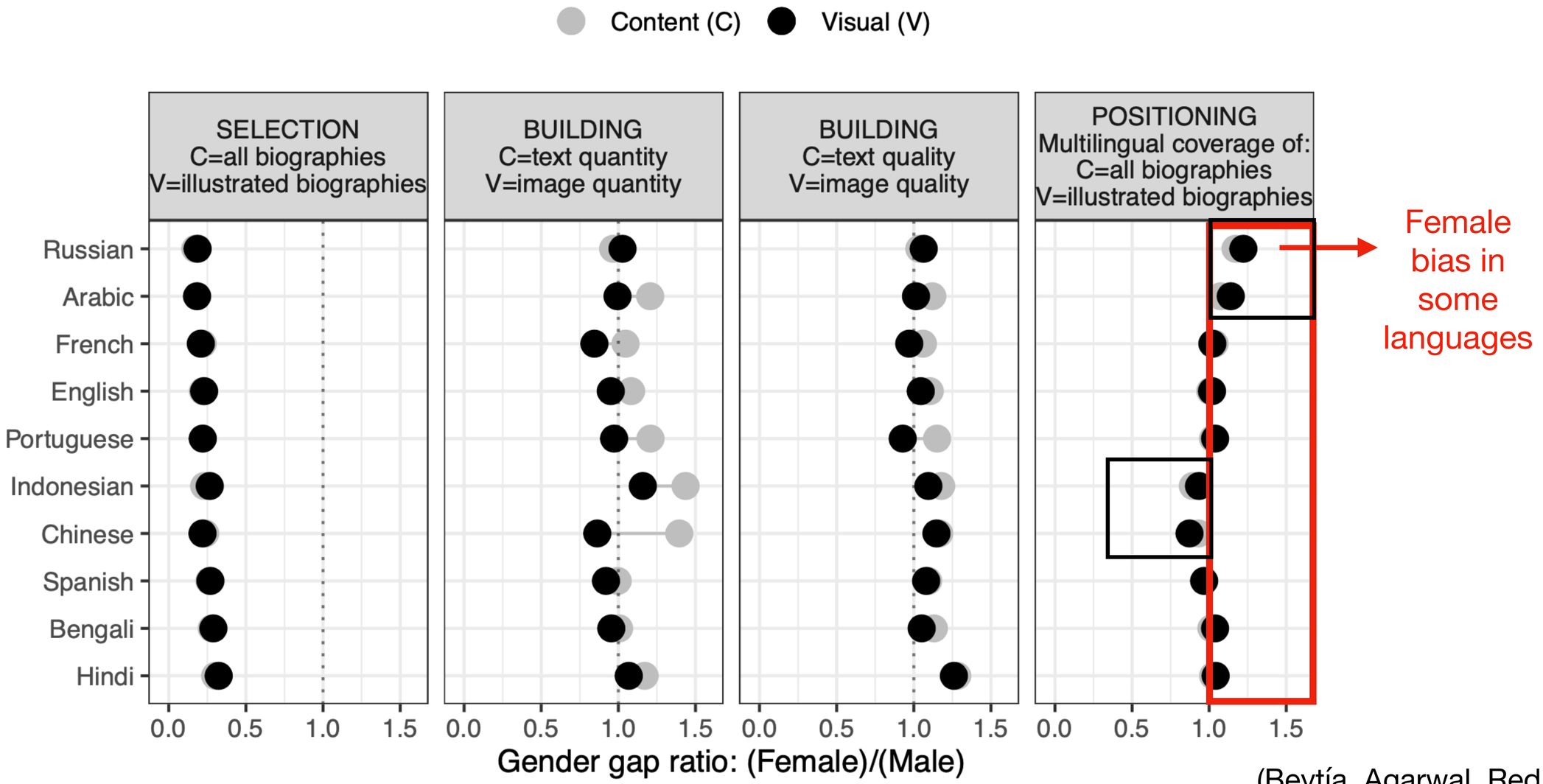
Visual and non-visual gender gaps by Wikipedia language version



(Beytía, Agarwal, Redi & Singh 2022)

Results: multilingual analysis

Visual and non-visual gender gaps by Wikipedia language version



(Beytía, Agarwal, Redi & Singh 2022)

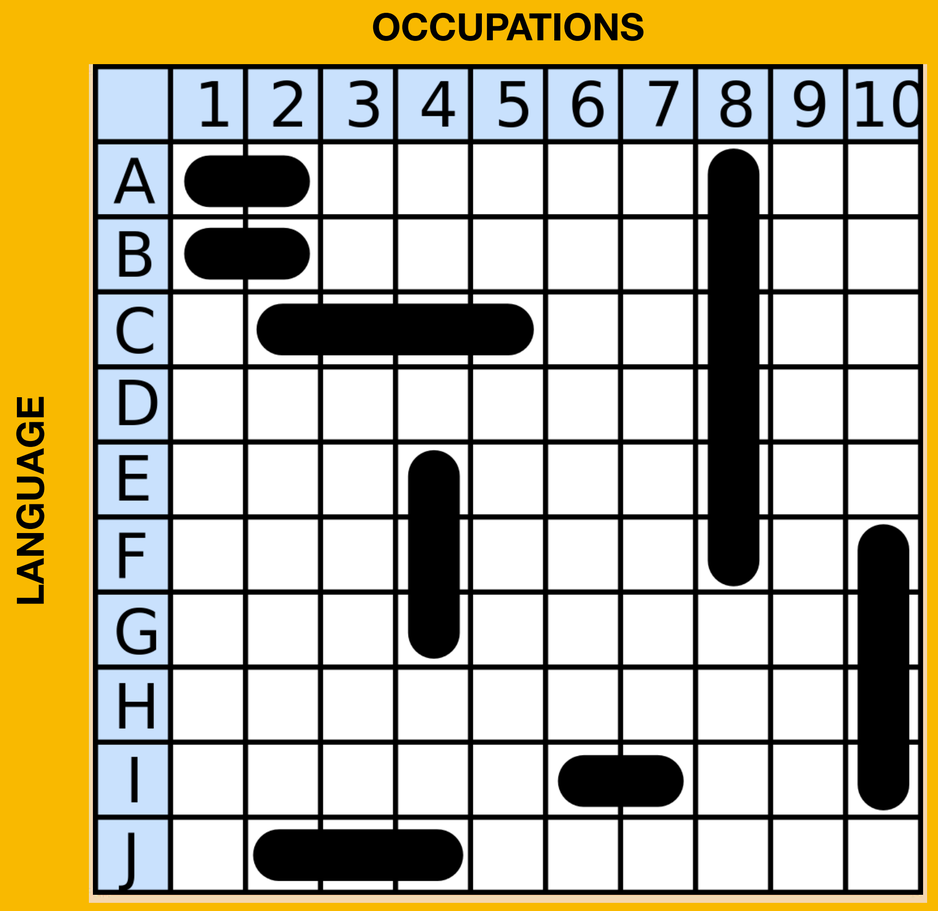
Summarizing

- The most salient male biases appear when editors select which personalities should have a Wikipedia page.
- The trends in written and visual content are dissimilar.
- Male biographies tend to have more images across languages.
- Female biographies have better visual quality on average.

(Beytía, Agarwal, Redi & Singh 2022)

How can we use these data to combat biases?

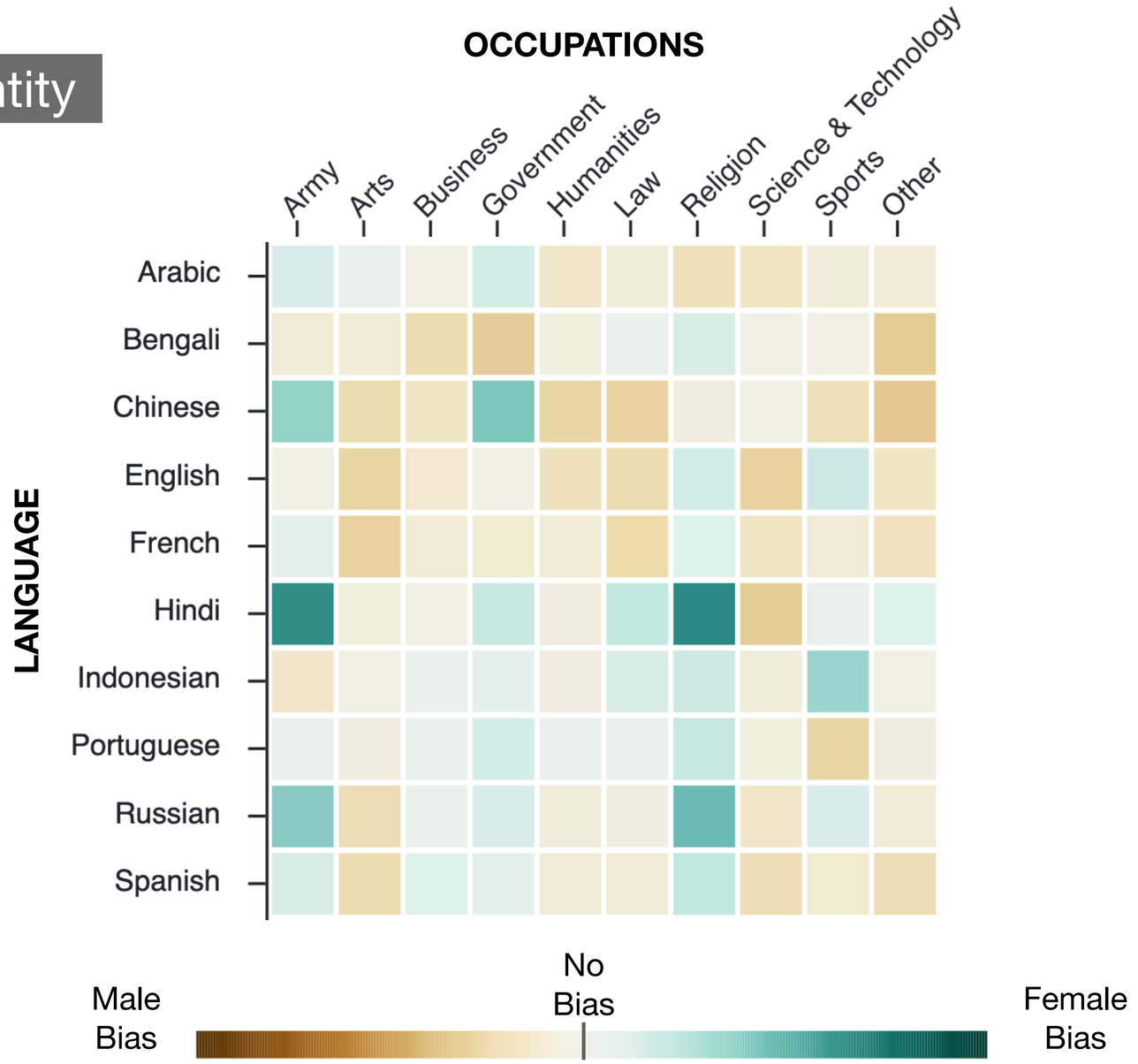
Battleship!



Source: Wikimedia commons
Author: Magasjukur2

Bias matrices: occupations by languages

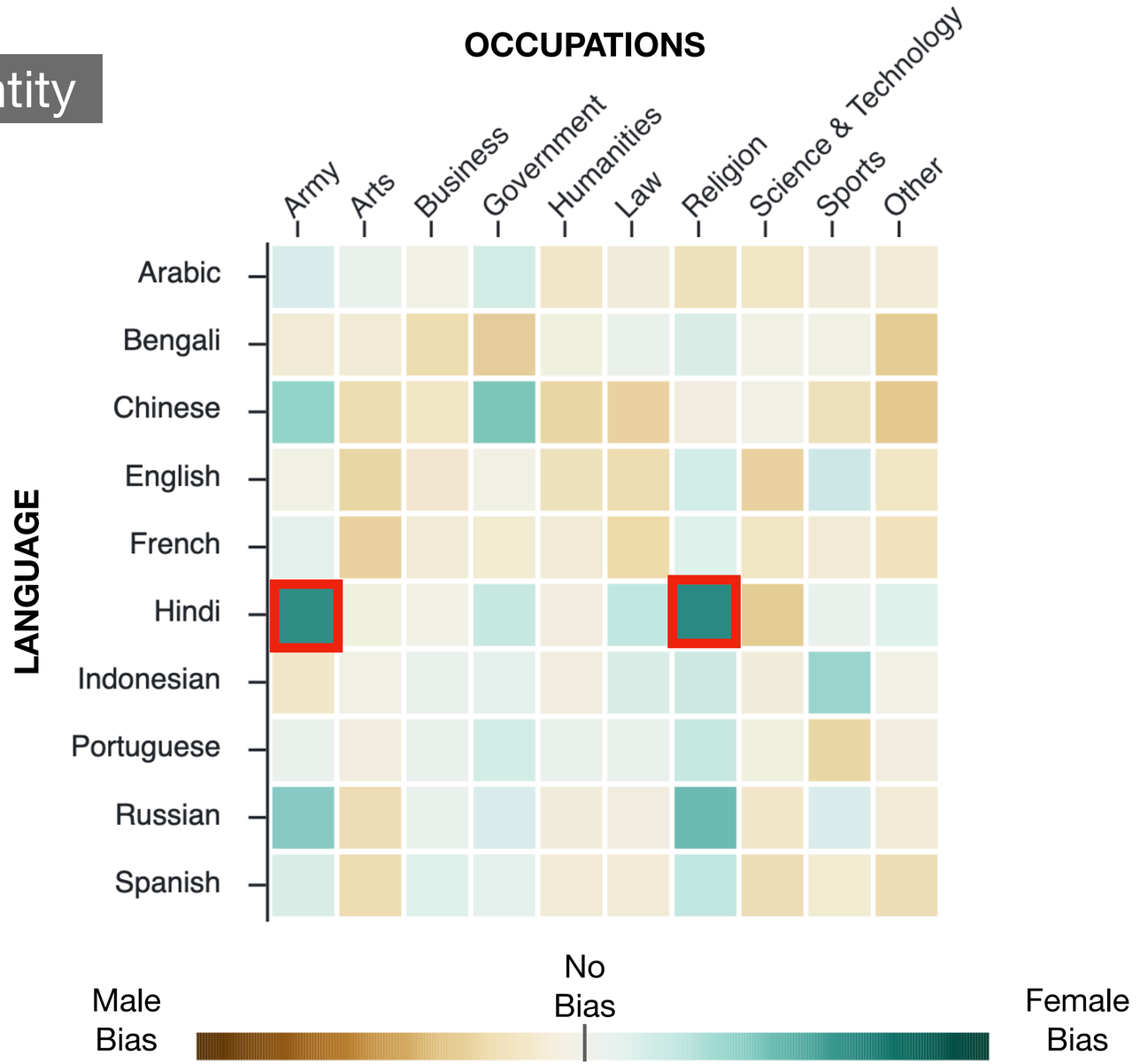
Image quantity



(Beytía 2023)

Bias matrices: occupations by languages

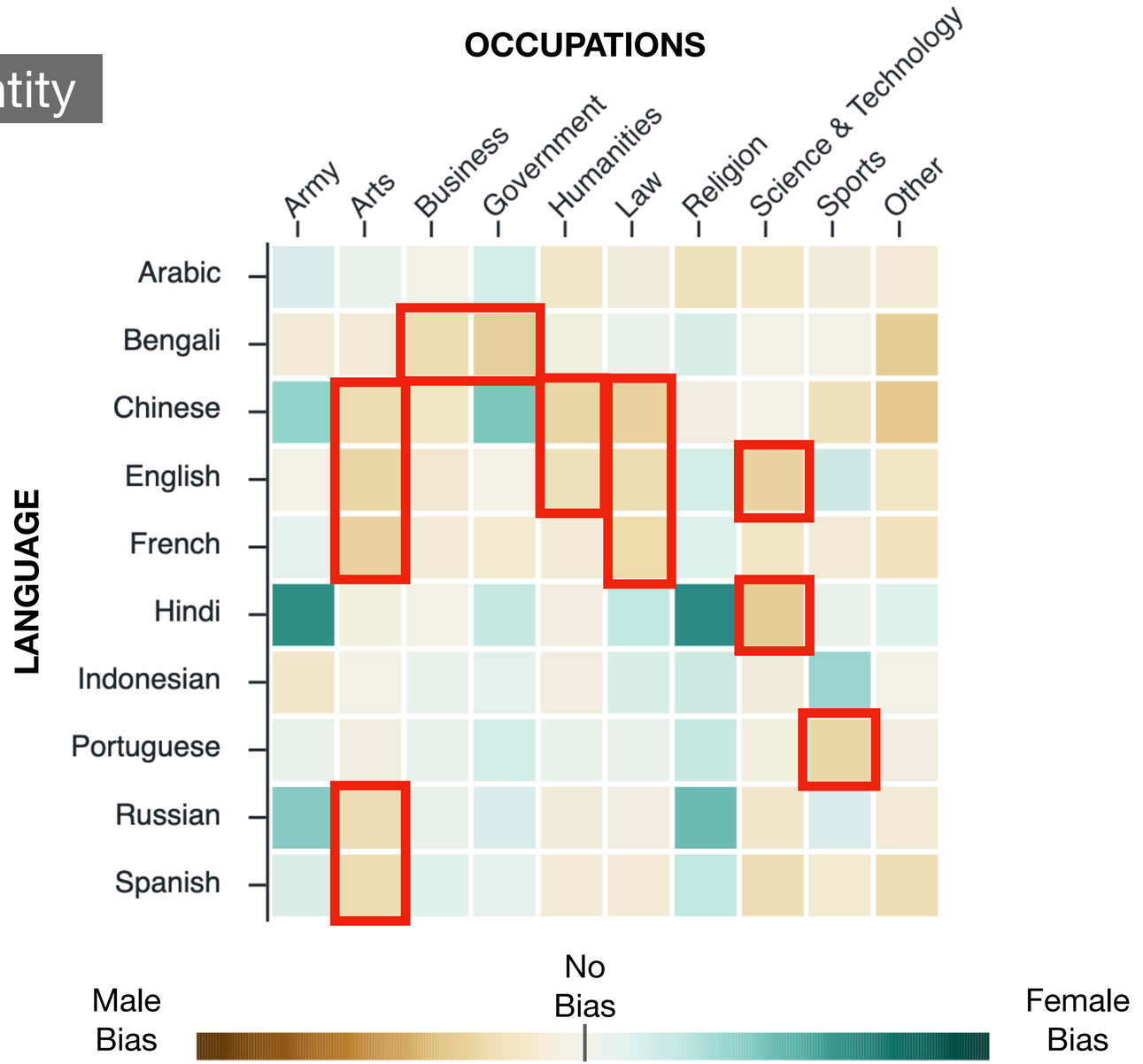
Image quantity



(Beytía 2023)

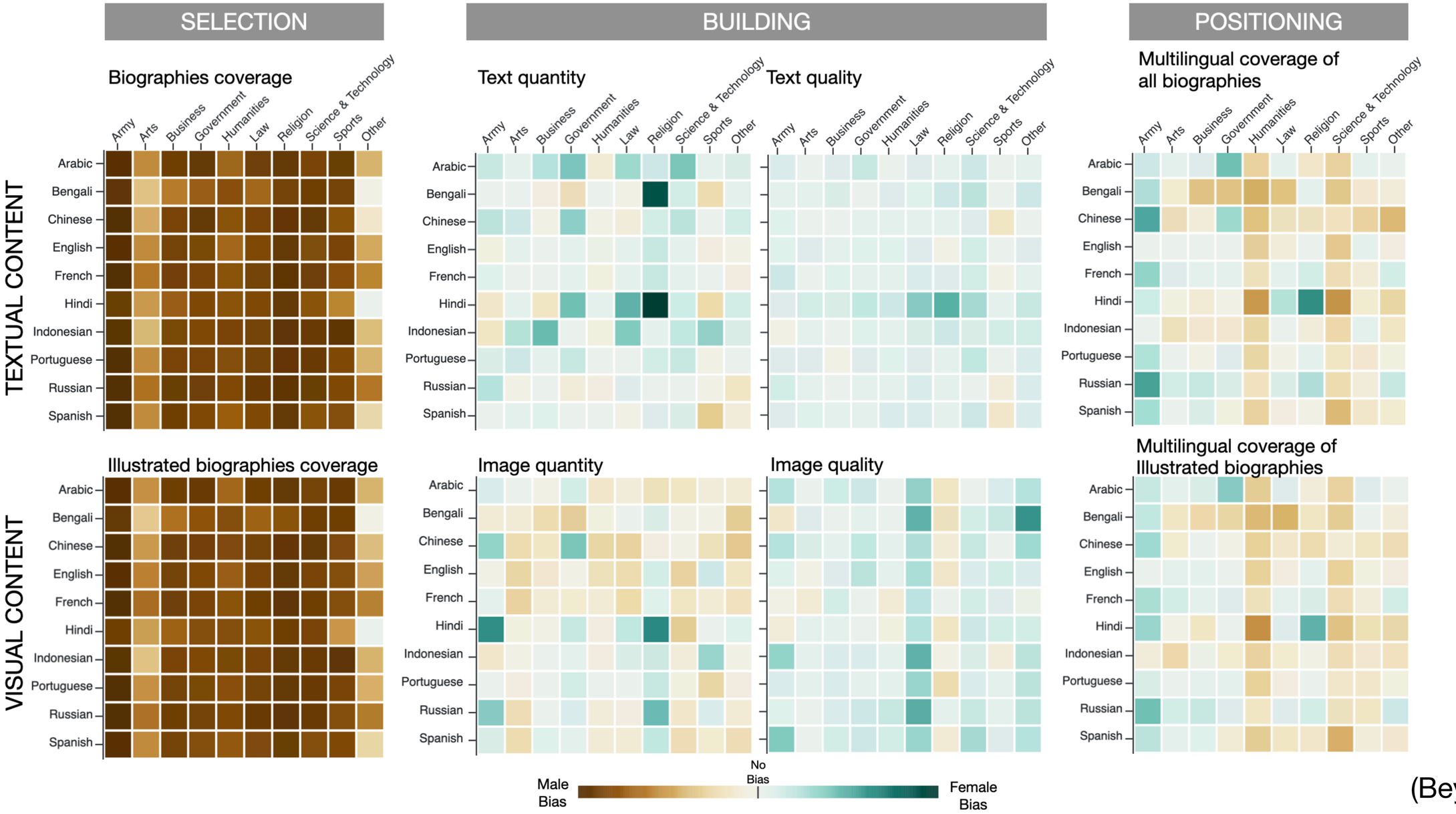
Bias matrices: occupations by languages

Image quantity



(Beytía 2023)

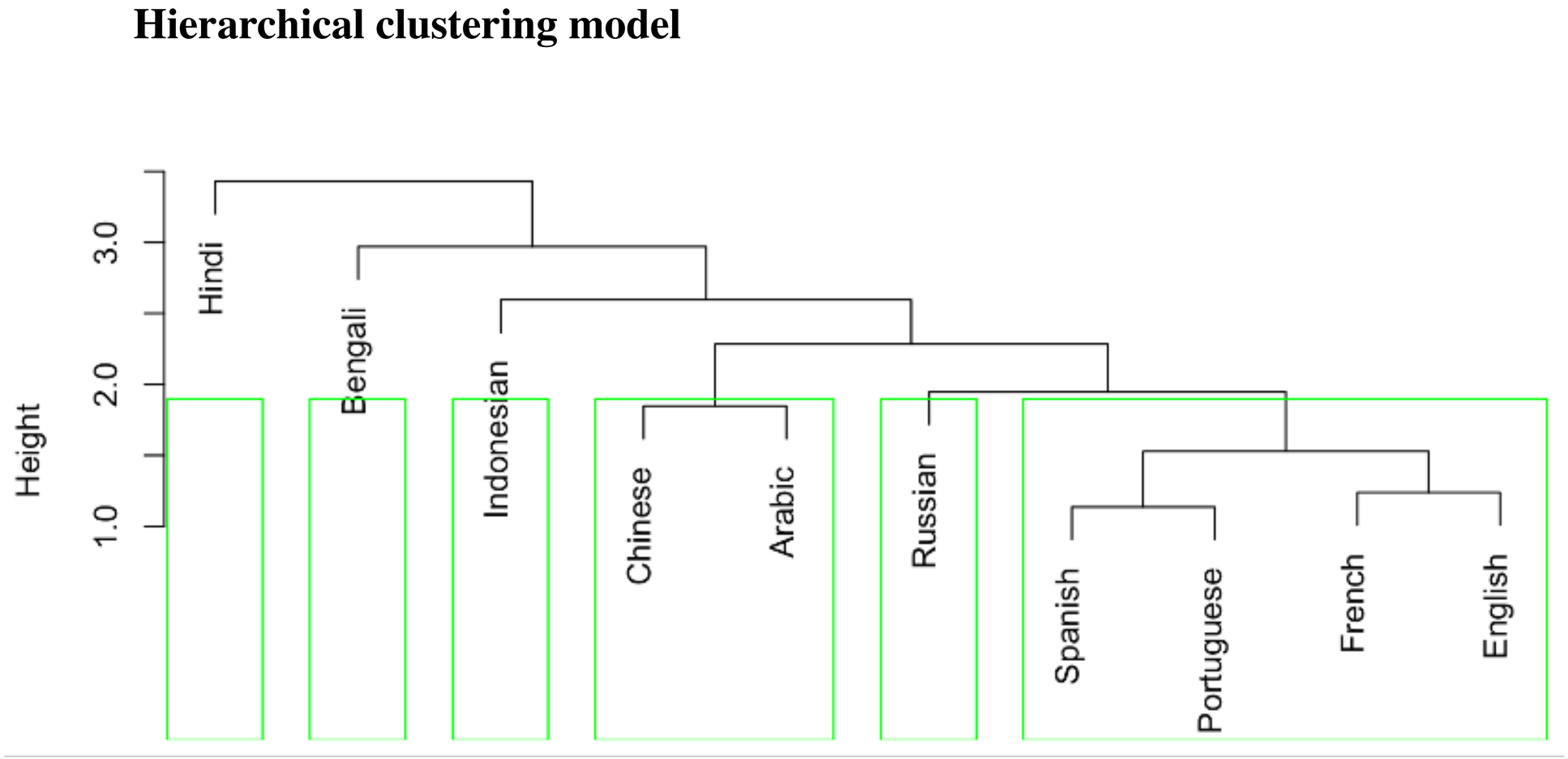
Bias matrices: occupations by languages



(Beytía 2023)

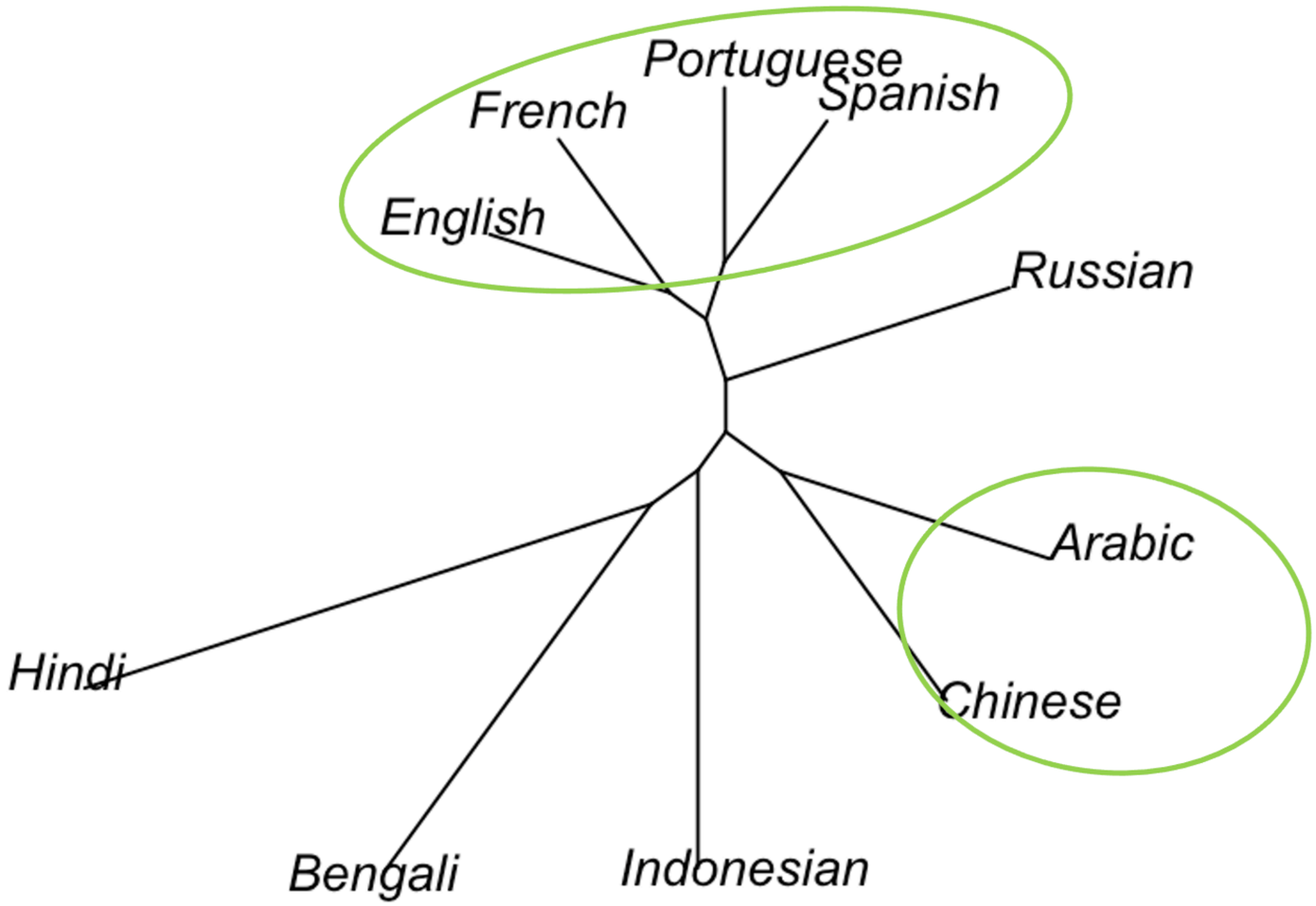
Which versions are closest in terms of their gender bias composition?

The similarity between languages (based on 80 gender ratios)



(Beytía 2023)

The similarity between languages (based on 80 gender ratios)



(Beytía 2023)

References

- Beytía, P., Agarwal, P., Redi, M., & Singh, V. K. (2022). Visual Gender Biases in Wikipedia: A Systematic Evaluation across the Ten Most Spoken Languages. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1), 43-54. https://doi.org/10.1609/icwsm.v16i1.19271

- Beytía, P. & Wagner, C. (2022). Visibility layers: a framework for systematizing the gender gap in Wikipedia content. *Internet Policy Review*, 11(1). https://doi.org/10.14763/2022.1.1621

- Beytía, P. (April 2023). A Digital Setting of Human History. Social Memory and Discursive Power in the Biographical Storage of Wikipedia. (*Draft of doctoral dissertation*).

Visual Gender Biases in Wikipedia

A Systematic Evaluation across the Ten Most Spoken Languages

Presented at Wikimedia Research/Showcase – April 2023

Beytía, Pablo | Agarwal, Pushkal | Redi, Miriam | Singh, Vivek K.