GENDER BIAS IN DIFFERENT LANGUAGES

THIJS RAYMAKERS - CSE3000 - 2020-06-25

1. RESEARCH QUESTION

To what extent are word embeddings of different languages biased towards gender?

Word embedding are an important tool in natural language processing, but should be used with caution because they contain human biases. This research will look at the words *male* and *female* in word embedding of different languages in order to assess possible gender bias.

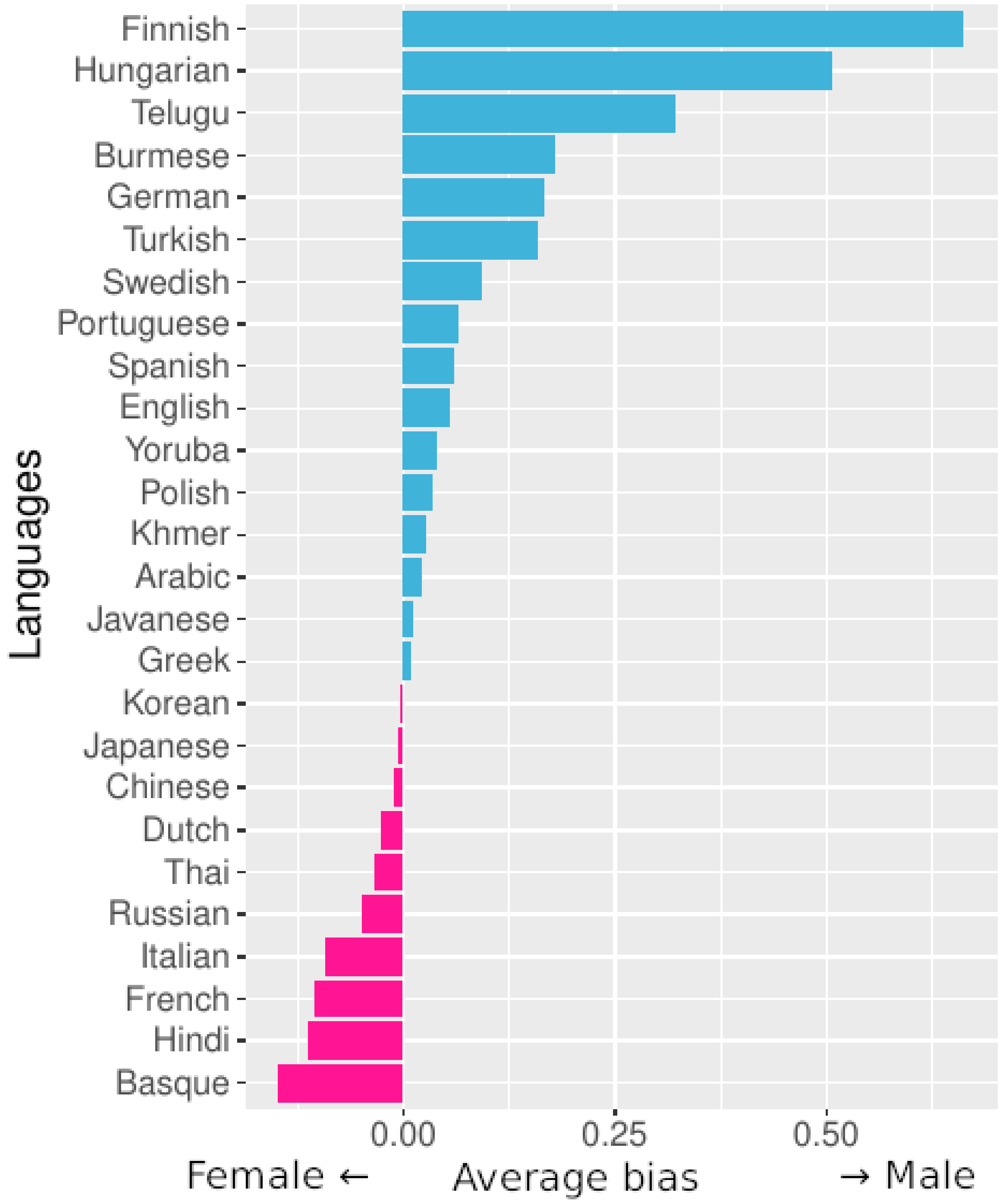
2. BACKGROUND

- Word embedding: math with words

$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$$

$$\overrightarrow{Paris} - \overrightarrow{France} + \overrightarrow{Germany} \approx \overrightarrow{Berlin}$$

- Previous research showed that English showed more bias in favor of *male*



3. METHOD

- 26 languages of 16 language families
- Measure cosine distance between attribute words (*male* and *female*) and target words (the words of a language)
- Average cosine distance and variance is used to calculate amount of bias per language

4. CONCLUSIONS

- Word embeddings of different languages are biased towards gender.
- Generally more biased in favor of *male* than *female*.
- Languages of the same language family often have a similar amount of bias (e.g. Finnish and Hungarian are both Uralic, Portugese and Spanish are both Iberian).
- Similar results when only the most used (instead of all) words are considered
- Possibly affirms linguistic relativity hypothesis (language influences how we think)

CONTACT INFO

Thijs Raymakers
Email: t.raijmakers@student.tudelft.nl
Professors: Marco Loog, David Tax
Supervised by:
Stavros Makrodimitris,
Arman Naseri Jahfari,
Tom Viering

TUDelft