UNIVERSITY OF ICELAND

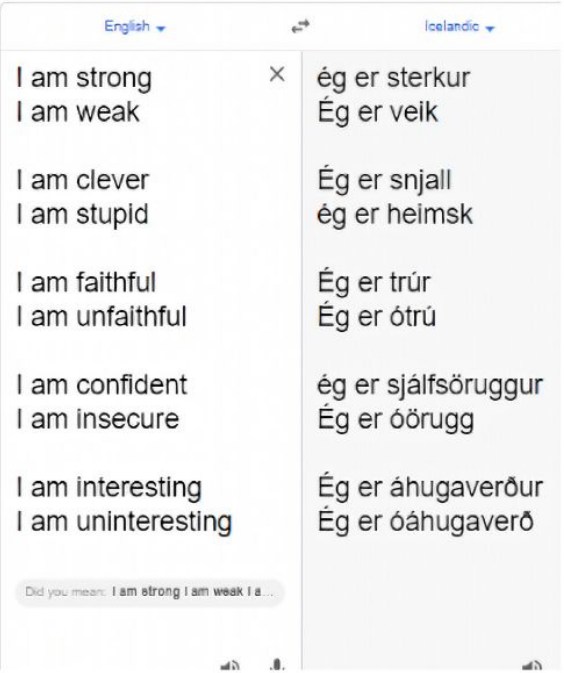
# Gendered Grammar or Ingrained Bias?

## Exploring Gender Bias in Icelandic Language Models

STEINUNN RUT FRIÐRIKSDÓTTIR
HAFSTEINN EINARSSON

# Biases in large language models

- Large language models are trained on a myriad of data
- More data is usually linked to better outputs
- Can result in the models absorbing all sorts of prejudices
- Sólmundsdóttir et al. (2022) showed pronounced gender bias in Icelandic machine translation models




Sólmundsdóttir, A., Guðmundsdóttir, D., Stefánsdóttir, L. B., & Ingason, A. (2022). Mean Machine Translations: On Gender Bias in Icelandic Machine Translations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3113-3121).

# What is *gender bias*?

- The tendency of LMMs to generate or perpetuate gender stereotypes
- Can lead to various types of harm
  - Reinforcement of harmful societal norms
  - Dismissal of individuals that fall outside of the norms
  - Feelings of exclusion
- Can cause direct harm when used for downstream tasks

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv preprint arXiv:2005.14050*.

# The objective

- To investigate the presence of gender bias within language models trained on Icelandic
- By focusing specifically on occupation-related terms, assess whether these models mirror the gender distributions observed in the Icelandic job market
- To achieve this, we cross-reference our findings with distribution data obtained from Statistics Iceland

UNIVERSITY OF ICELAND

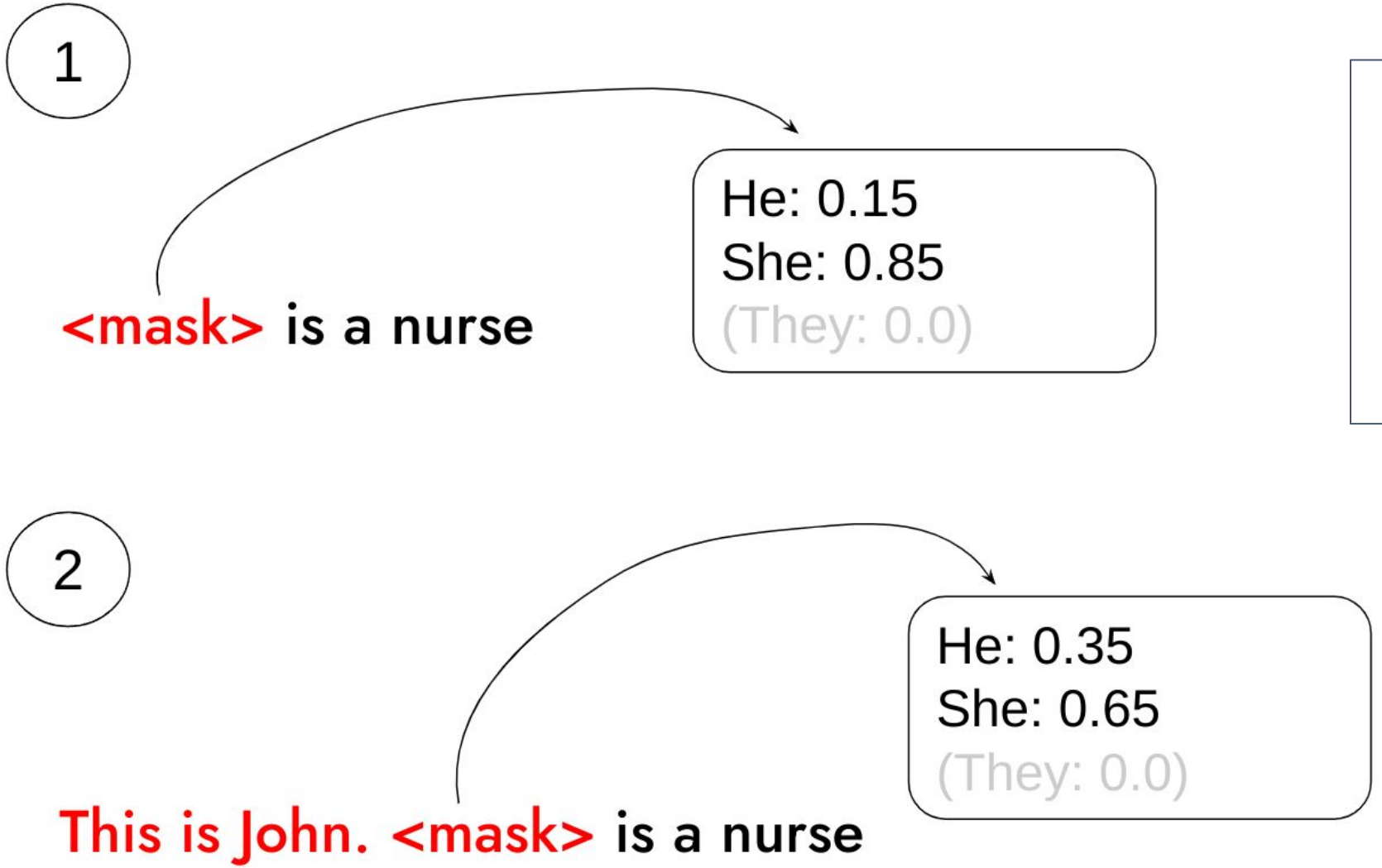
Statistics Iceland

# Icelandic and the generic masculine

- Icelandic is highly gendered
  - All words that inflect by case also inflect by gender
  - Doesn't reflect societal biases, just morphological agreement
- Icelandic favors the masculine when referring to a group of mixed-gendered people
  - Most occupational terms are grammatically masculine
    - 381 out of 394 in this case
    - A debate between two feminist movements
  - How does the language affect the results?

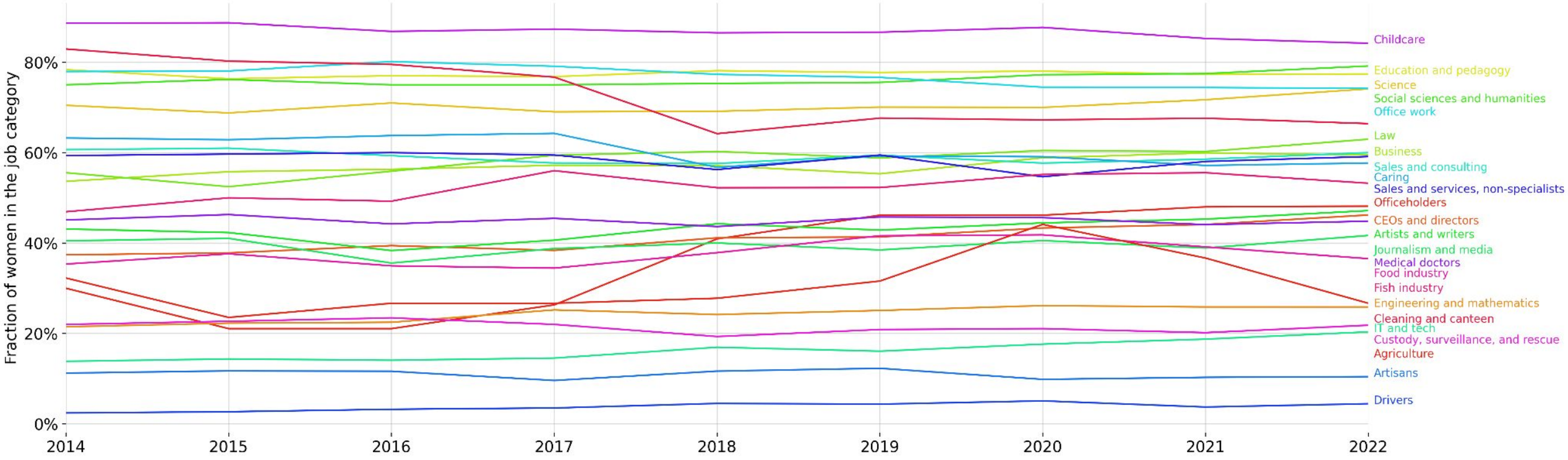
| Number | Karlkyn (MASCULINE) | Kvenkyn (FEMININE) | Hvorugkyn (NEUTER) |
|---|---|---|---|
| 1 | einn | ein | eitt |
| 2 | tveir | tvær | tvö |
| 3 | þrír | þrjár | þrjú |
| 4 | fjórir | fjórar | fjögur |

# Method

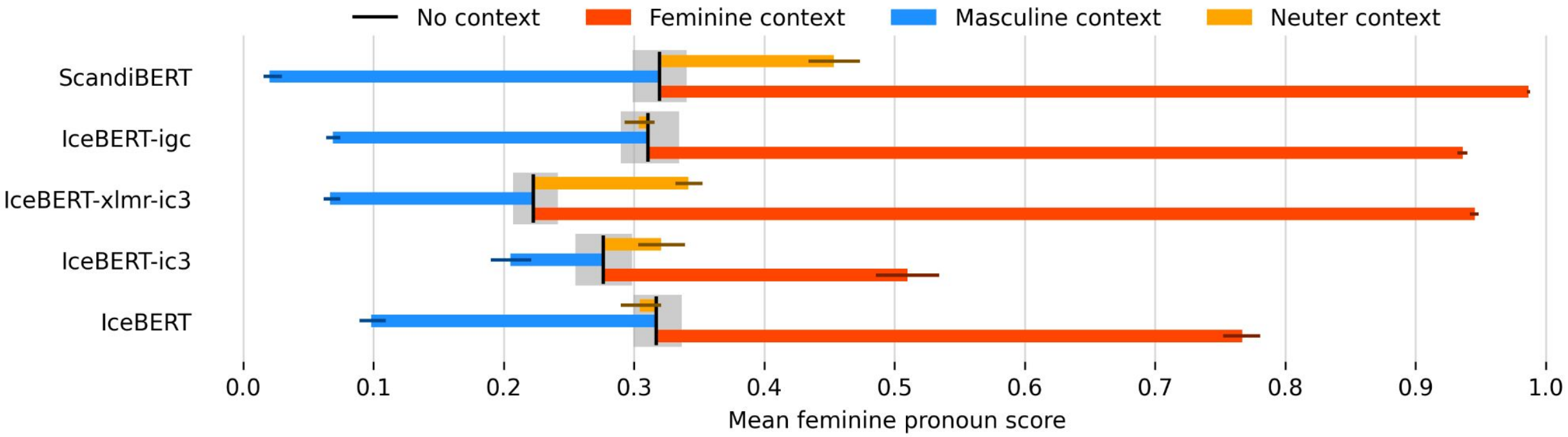



- IceBERT
- IceBERT-igc
- IceBERT-ic3
- IceBERT-xlmr-ic3
- ScandiBERT

# Information from Statistics Iceland



# Overall trends and the influence of context

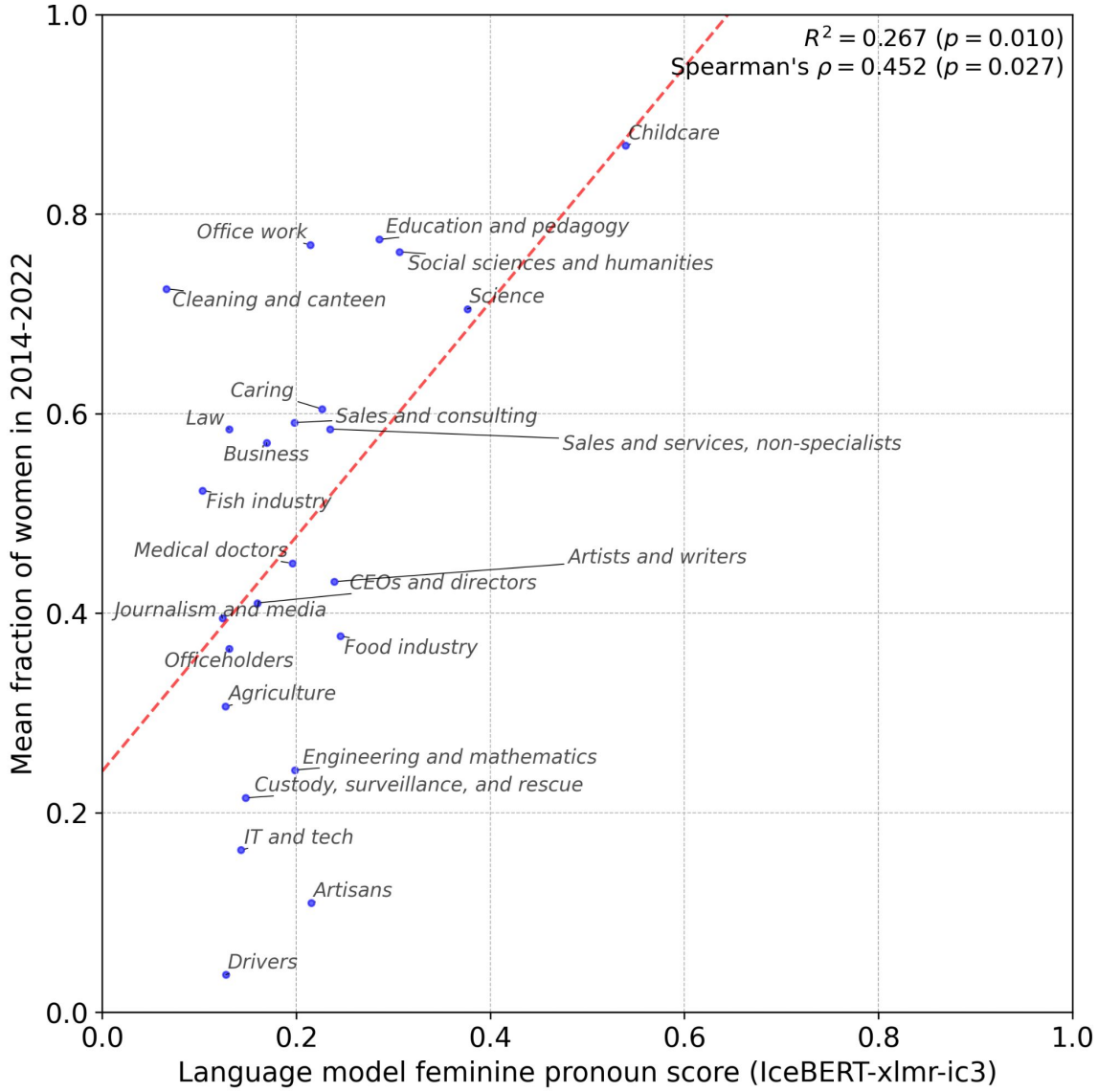



**But note!**

| | Masc | Fem | Neut |
|---|---|---|---|
| Proper names | 52,4% | 30% | 17,6% |
| Pronouns singular | 29% | 14,5% | 56,5% |
| Pronouns plural | 55,3% | 16,6% | 28,1% |

Table 3: Grammatical gender distribution in the Icelandic Gigaword Corpus, used to train three of our models.

# Statistical analyses




| Model | $r$ | $\rho$ |
|---|---|---|
| IceBERT | 0.40 (0.056) | 0.36 (0.086) |
| IceBERT-ic3 | 0.46 (0.025) | 0.42 (0.041) |
| IceB.-xlmr-ic3 | 0.52 (0.010) | 0.45 (0.027) |
| IceBERT-igc | 0.50 (0.012) | 0.33 (0.110) |
| ScandiBERT | 0.49 (0.016) | 0.58 (0.003) |

Table 2: The correlation coefficients ($r$), Spearman's rank correlation coefficients ($\rho$) and significance levels in brackets for different models when evaluating their relationship to job market data from Statistics Iceland.

# Notable results

- Overall masculine bias without context
- Female-centric occupations are not overpredicted
- Compounds ending with *maður* (e. *man*), *smiður* (e. *smith*) and *meistari* (e. *master*) are more likely to be masculine
- Lower-ranking university positions are more likely to be feminine but *prófessor* (e. *full professor*) is >80% likely to be masculine
- Kindergarten, elementary and secondary school teachers are more likely to be feminine but university teachers masculine

## Summary

- All models have a masculine bias but most can be mended with context
- Complex interplay of societal, linguistic and data selection biases
- We only have job categories, not individual occupations
- Would be interesting to compare our results to those of a language without a grammatical gender
- The dangers of aging data: Society evolves but the models might not be aware of it!