What an Elegant Bridge: Multilingual LLMs are Biased Similarly in Different Languages

Viktor Mihaylov **Aleksandar Shtedritski**

Oxford Artificial Intelligence Society, University of Oxford

Abstract

This paper investigates biases of Large Language Models (LLMs) through the lens of grammatical gender. Drawing inspiration from seminal works in psycholinguistics, particularly the study of gender's influence on language perception, we leverage multilingual LLMs to revisit and expand upon the foundational experiments of Boroditsky (2003). Employing LLMs as a novel method for examining psycholinguistic biases related to grammatical gender, we prompt a model to describe nouns with adjectives in various languages, focusing specifically on languages with grammatical gender. In particular, we look at adjective co-occurrences across gender and languages, and train a binary classifier to predict grammatical gender given adjectives an LLM uses to describe a noun. Surprisingly, we find that a simple classifier can not only predict noun gender above chance but also exhibit cross-language transferability. We find a strong social influence of language on the way multilingual LLMs reason.

1 Introduction

The way we perceive the world is not only affected by our culture (Oyserman and Lee, 2008; Masuda et al., 2008), but also the language we speak (Boroditsky et al., 2003; Boroditsky, 2001). The relationship between cognition and language has been of interest for a long time (Langacker, 1993), especially through the lens of gender (Boroditsky et al., 2003; Gygax et al., 2008). Recent advances in Large Language Models (LLMs), that match human performance on multiple tasks, provide an exciting opportunity to study the relationship between the psycholinguistic biases of humans and those of machines. While it is unclear whether the latter relationship exists, it would be a more scalable, affordable, and even ethical (Banyard and Flanagan, 2013) alternative to human studies.

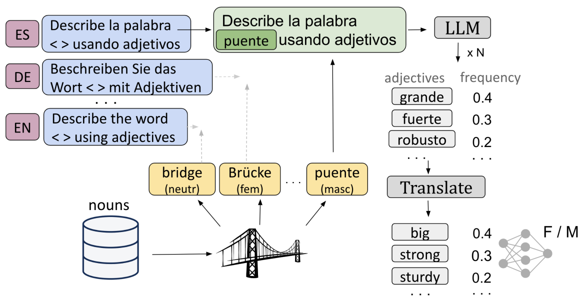

Figure 1: **Probing the bias of multilingual LLMs.** We prompt a LLM to describe gendered nouns using adjectives. This allows us to study psycholinguistic biases of LLMs. For example, if the generated adjectives are predictive of the nouns's gender, we can, by training a binary classifier, predict grammatical gender by only looking at the adjectives a LLM uses to describe a word.

In this work, we revisit the study of (Boroditsky et al., 2003) in the era of LLMs. To see how grammatical gender affects cognition, Boroditsky et al. (2003) ask speakers of languages with grammatical gender (where nouns have assigned genders) to describe various objects, finding that the language a person speaks affects the attribution of masculine or feminine characteristics to objects. For example, a Spanish speaker (where "bridge" is masculine) might describe a bridge with words like "strong" or "sturdy", while a German speaker (where "bridge" is feminine) might use terms like "elegant" or "beautiful". However, several subsequent studies fail to replicate such results (Haertlé; Mickan et al., 2014; Samuel et al., 2019), which is but a symptom of the replication crisis in psychology (Wiggins and Christopherson, 2019; Shrout and Rodgers, 2018; Maxwell et al., 2015). Similarly, studies in the field of NLP that examine the way gendered nouns are used in text corpora (Williams et al., 2021; Kann, 2019), find conflicting evidence on whether there is a relationship between grammatical gender and cognition.

The existence of gender bias has been well studied for word embeddings (Bolukbasi et al., 2016; Basta et al., 2019; Caliskan et al., 2017), as well as a range of NLP systems, such as ones for machine translation (Stanovsky et al., 2019; Vanmassenhove et al., 2018), image and video captioning (Tatman, 2017; Hall et al., 2023), or sentiment analysis (Kiritchenko and Mohammad, 2018). More recently, the social biases of LLMs have been studied (Kirk et al., 2021). While the multilingual capabilities of LLMs have been extensively evaluated, showing they perform well on machine translation (Hendy et al., 2023; Jiao et al., 2023; Wang et al., 2023) as well as various multilingual benchmarks (Ahuja et al., 2023; Bang et al., 2023), the evaluation of biases in the multilingual setting is less mature. Contrary to recent work showing that multilingual LLMs have different biases for different languages Mukherjee et al. (2023), we find that when it comes to gendered nouns, LLMs are biased in a similar way, as the biases are predictive of each other.

In this paper, we loosely follow the protocol of Boroditsky et al. (2003) and prompt LLMs to describe nouns using adjectives in different languages. Specifically, we focus on open-sourced LLMs (Llama-2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023)). We select 10 languages that have grammatical gender (e.g, German and Spanish), and use the LLMs to describe gendered nouns using adjectives. This allows us to see how adjectives co-occur across languages. Our most important findings are that (i) a simple classifier can predict the gender of a noun using the adjectives used to describe it, and (ii) such a classifier reliably transfers across languages, suggesting LLMs are biased similarly in different languages.

2 Method

In this work, we are interested in the adjectives a multilingual LLM uses to describe gendered nouns when asked in different languages. Here, we describe how we generate such adjectives, and how we examine whether they are predictive of the grammatical gender of the nouns.

2.1 Describing nouns in different languages

We show our pipeline for describing gendered nouns with adjectives in Figure 1. More formally, for a language l we have a database of K gendered nouns $\mathcal{N}^l = \{n_1^l, n_2^l, ..., n_K^l\}$, with corresponding grammatical genders $g(n_i^l) = \{\mathrm{f}, \mathrm{m}\}$ for feminine and masculine, respectively. We prompt the LLM to describe a noun n_k^l using adjectives, which we parse into a list of M adjectives $\mathcal{A}(n_k^l) = \{a_1^l, a_2^l, ..., a_M^l\}$. For every noun n, we repeat the prompting N times and compute the frequencies f with which the adjectives appear:

$$f(a_i) = \frac{\sum_{j=1}^{N} \mathbb{1}(a_i \in \mathcal{A}(n_j))}{N}. \tag{1}$$

Finally, we keep the adjectives with top-p frequencies. In practice, we use $N = 50$ and $p = 50$.

2.2 Predicting gender from descriptions

To examine to what extent the adjectives an LLM uses to describe a noun are predictive of its grammatical gender, we train a binary classifier Φ to predict grammatical gender:

$$\hat{g}(n_i^l) = \Phi\left(\sum_{i=1}^{p} f\left(a_i^l\right) e_g\left(a_i^l\right)\right),$$

where the input to the classifier are GloVe (Pennington et al., 2014) word embeddings e_g of the adjectives weighted by the adjectives frequencies f. In practice, we use a modified version of f, where $f' = -30/\log(f)$ to give us a better scaling. The classifier Φ is a 2-layer MLP and we train it with binary cross-entropy loss.

As shown in Figure 1, we first translate the generated adjectives to English. We do this for two reasons. Firstly, adjectives in some languages are also gendered and that would help the classifier learn this shortcut (e.g. *pretty* in Spanish is *bonito* and *bonita* for masculine and feminine, respectively). Adjectives in English are not gendered, so the classifier Φ has no way of inferring the gender of the noun from the grammatical form. Secondly, this allows for easy transfer of the classifier across languages – e.g. we can train Φ on words generated in Hindi, and evaluate on Italian.

3 Experiments

3.1 Implementation details

Languages We conduct experiments on the languages Bulgarian, Czech, French, German, Greek, Hindi, Italian, Latvian, Portuguese, and Spanish.

Nouns We automatically collect commonly used nouns from every language, and their corresponding grammatical gender. For details on the way we collect those nouns, and the number of nouns per language, please refer to the Appendix. We exclude neuter nouns as such nouns do not exist

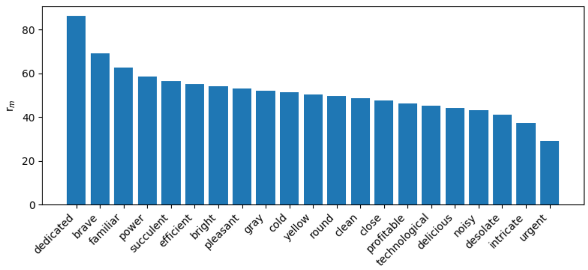

Figure 2: **Bias when describing gendered nouns.** Here we prompt an LLM in Spanish and for a random sample of adjectives, show the percentage of *masculine* nouns they were used for.

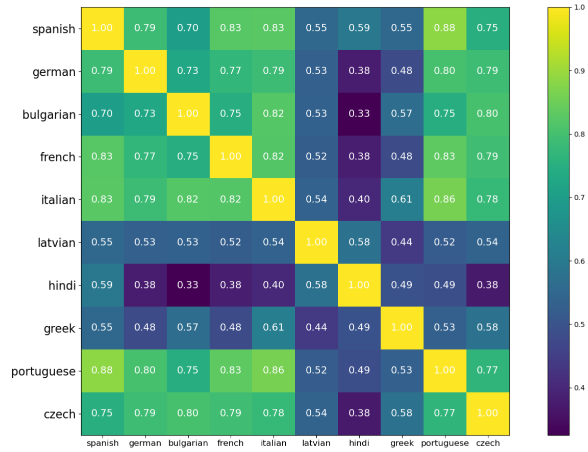

Figure 3: **Gendered adjective similarity sccores**.

in every language. We subsample the feminine or masculine nouns in each gender to ensure a uniform distribution for each language.

LLMs In our experiments we use the open-sourced Mistral-7B (Jiang et al., 2023) model, unless stated otherwise. We also repeat our experiments with Llama2-7B (Touvron et al., 2023).

Prompts We prompt the LLM to describe the given noun in the corresponding language using comma-separated adjectives. In practice, we use few-shot prompts, which we show in the Appendix.

Translation Where we translate nouns, adjectives, or prompts, we use Google Translate [1].

3.2 Bias in generated adjectives

First, we look at adjectives that commonly occur for masculine or feminine nouns.

For every adjective a_i, we look at the ratio r_m:

$$r_m(a_i) = \frac{\sum_{n \in \mathcal{N}, g(n)=\mathrm{m}} \mathbb{1}(a_i \in \mathcal{A}(n)))}{\sum_{n \in \mathcal{N}} \mathbb{1}(a_i \in \mathcal{A}(n)))}, \quad (2)$$

which shows the proportion of masculine words it was used to describe. We randomly sample adjectives and show their r_m in Figure 2. We see that adjectives like intricate and desolate are associated with feminine nouns, whereas adjectives like dedicated and brave are associated with masculine nouns. We show more examples for different languages in the Appendix.

3.3 Do languages show similar biases?

Next, we explore whether adjectives describing masculine and feminine nouns tend to co-occur in different languages. To this end, we compute a gendered-adjective similarity score S_{pq} for every language pair of languages l_p and l_q. We do that as follows. We take the set of N adjectives $a_1, a_2, ..., a_N$ that are used to describe at least 15 nouns in both l_p and l_q. Then for both languages, we construct a gendered-adjective score vector $\sigma \in \mathbb{R}^N$, where $\sigma[i] = r_m(a_i)$. Now, σ_p and σ_q contain the gender ratio for all N adjectives. Finally, we define the gendered-adjective similarity score S_{pq} as the cosine similarity between σ_p and σ_q.

In Figure 3 we show the score S for all language pairs. We see that in Romance languages (Spanish, Italian, French Portuguese), Slavic languages (Bulgarian, Czech), and Germanic languages (German), the LLM shows a high gendered-adjective similarity score, meaning that the adjectives in these languages tend to have similar value of r_m. On the other hand, Greek, Hindi and Latvian have a low score between themselves and others.

3.4 Predicting the gendered nouns

Can we predict the gender of a noun in some language given the adjectives used to describe it? Following Section 2.2, we train binary classifiers to predict the grammatical gender of a noun from the adjectives used to describe it (translated to English). We train a separate classifier for each language. As seen in Table 1, for all languages the classifier reliably does better than random – meaning that the adjectives are predictive of gender.

3.5 Transfer between languages

If we train a grammatical gender classifier, like in Section 3.4, can we predict the gender of a noun in an **unseen** language? To answer this, where we train grammatical gender classifiers on adjectives from 9 languages (translated to English), and eval-

[1] Google Translate, `https://translate.google.com/`

Language	F1	Accuracy Overall	Masc.	Fem.
Bulgarian	0.64	68.4%	72.4%	63.3%
Czech	0.52	59.0%	58.3%	60.2%
French	0.63	56.5%	55.8%	56.8%
German	0.60	60.0%	52.7%	69.4%
Greek	0.68	69.0%	62.7%	77.6%
Hindi	0.53	54.3%	57.5%	51.2%
Italian	0.46	68.2%	73.0%	54.3%
Latvian	0.64	62.6%	60.0%	65.0%
Portuguese	0.55	62.0%	62.7%	60.1 %
Spanish	0.62	63.3%	59.6%	68.0%

Table 1: **Predicting grammatical gender.** We train a classifier to predict the gender of nouns given the adjectives the LLM uses to describe them.

Language	F1	Accuracy Overall	Masc.	Fem.
Bulgarian	0.56	62.5%	64.4%	59.8%
Czech	0.45	60.6%	70.6%	43.5%
French	0.62	54.8%	50.3%	57.3%
German	0.54	58.6%	73.1%	46.0%
Greek	0.64	60.6%	47.8%	75.3%
Hindi	0.53	48.8%	37.9%	60.2%
Italian	0.40	60.1%	61.6%	55.6%
Latvian	0.41	51.7%	81.2%	29.7%
Portuguese	0.55	62.8%	63.0%	62.4%
Spanish	0.59	58.8%	56.7%	60.1%

Table 2: **Unseen Language Results.** We train on all other languages and predict the genders of nouns in the given language. We train a separate leave-one-out classifier for each language.

uate on the final language. As we see in Table 2, such classifiers can reliably predict gender across languages. Interestingly, they even work better than random for Greek, Hindi and Latvian, despite the results reported in Section 3.3. We suggest that although the LLM uses different adjectives to describe masculine and feminine nouns in different languages (hence low S_{pq}), they are semantically similar (hence high accuracy when evaluating the classifier on an unseen language).

4 Discussion

4.1 Reproducibility

Studying the phenomena relating cognition to grammatical gender in psychology has led to inconclusive results(Boroditsky, 2001; Haertlé; Mickan et al., 2014; Samuel et al., 2019). These could be explained by different experimental settings with speakers of different languages, which are difficult to control in a human study. Similarly, prior works that examine text corpora using NLP techniques show conflicting results (Williams et al., 2021; Kann, 2019). The results of these works heavily depend on the text corpora analyzed, and the methods used to identify adjective-noun pairs, which might be subpar for languages other than English. Our method presents more consistent results by ensuring consistent evaluation across languages.

LLM	Eval	F1	Accuracy Overall	Masc.	Fem.
Mistral-7B	Same	0.59	62.3%	61.5%	62.6%
Llama2-7B	Same	0.59	64.6%	67.9%	59.9%
Mistral-7B	Unseen	0.53	57.9%	60.7%	55.1%
Llama2-7B	Unseen	0.54	59.1%	62.6%	54.9%

Table 3: **Evaluating Llama-2.** We compare grammatical gender classifiers Llama-2 to Mistral when tested on the *same* language (as in Section 3.4), or an *unseen* language (as in Section 3.5). We show mean results over all 10 languages. We see that we observe a similar predictive performance on adjectives used by Llama-2 as those by Mistral.

4.2 Importance of our results

Our results are only valid for noun-adjective associations in LLMs. However, these associations have been learnt through co-occurences of these words in text corpora, which have been produced by speakers of the respective languages. Future work should study how well such biases in LLMs are predictive of biases of humans.

The results we present suggest a consistent bias that associates nouns with adjectives, depending on their grammatical gender. This could be important when LLMs are used to describe humans using objects, or vice versa (anthropomorphism, personification, metaphors, ...), where traits of these objects are transferred to the human. Furthermore, using LLMs to perform machine translation of such phrases could lead to a loss of meaning or unexpected biases.

5 Conclusion

In this work, we revisit the psycholinguistic experiments of Boroditsky et al. (2003), confirming the hypothesis of their work applies to LLMs, where different words are used to described masculine and feminine nouns. Our most surprising finding is that we can reliably zero-shot transfer a classifier that predicts grammatical gender across languages. This shows that while LLMs might think differently on different languages, they are biased similarly when it comes to grammatical gender. We hope this work inspires others to explore psycholonguistic experiments applied to LLMs, and to drive a discussion of whether such results can be useful to inform or motivate human experiments.

6 Limitations

We only conducted experiments and observed these effects for the opens-sourced Mistral-7B and Llama2-7B models. It is not clear if similar effects can be observed in larger LLMs, or commercial LLMs such as GPT-4. While we ensured to cover a wide range of languages, the ones we used are by no means exhaustive and only cover indo-european languages. Finally, we only explore the biases of general-purpose, multilingual LLMs. Looking into specialised LLMs, fine-tuned for the specific language, might be more representative of what models would be used in practice.

7 Acknowledgements

This work has been completed with the support of the Oxford AI student society, the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines Systems [EP/S024050/1] (A.S.) and compute resources from the AWS Responsible AI Grant.

References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Philip Banyard and Cara Flanagan. 2013. *Ethical issues in psychology*. Routledge.

Christine Basta, Marta R Costa-Jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Lera Boroditsky. 2001. Does language shape thought?: Mandarin and english speakers' conceptions of time. *Cognitive psychology*, 43(1):1–22.

Lera Boroditsky, Lauren A Schmidt, and Webb Phillips. 2003. Sex, syntax, and semantics. *Language in mind: Advances in the study of language and thought*, 22:61–79.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Pascal Gygax, Ute Gabriel, Oriane Sarrasin, Jane Oakhill, and Alan Garnham. 2008. Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. *Language and cognitive processes*, 23(3):464–485.

Izabella Haertlé. Does grammatical gender influence perception? a study of polish and french speakers. *Psychology of Language and Communication*, 21(1):386–407.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. 2023. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Katharina Kann. 2019. Grammatical gender, neowhorfianism, and word embeddings: A data-driven approach to linguistic relativity. *arXiv preprint arXiv:1910.09729*.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems*, 34:2611–2624.

Ronald W Langacker. 1993. Universals of construal. In *Annual Meeting of the Berkeley Linguistics Society*, volume 19, pages 447–463.

Takahiko Masuda, Mikako Akase, MH Radford, and Huaitang Wang. 2008. Effect of contextual factors on patterns of eye-movement: Comparing sensitivity to background information between japanese and westerners. *Shinrigaku Kenkyu: The Japanese Journal of Psychology*, 79(1):35–43.

Scott E Maxwell, Michael Y Lau, and George S Howard. 2015. Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, 70(6):487.

Anne Mickan, Maren Schiefke, and Anatol Stefanowitsch. 2014. Key is a llave is a schlüssel: A failure to replicate an experiment from boroditsky et al. 2003. *Yearbook of the German Cognitive Linguistics Association*, 2(1):39–50.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global voices, local biases: Socio-cultural prejudices across languages. *arXiv preprint arXiv:2310.17586*.

Daphna Oyserman and Spike WS Lee. 2008. Does culture influence what and how we think? effects of priming individualism and collectivism. *Psychological bulletin*, 134(2):311.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Steven Samuel, Geoff Cole, and Madeline J Eacott. 2019. Grammatical gender and linguistic relativity: A systematic review. *Psychonomic bulletin & review*, 26:1767–1786.

Patrick E Shrout and Joseph L Rodgers. 2018. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual review of psychology*, 69:487–510.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Bradford J Wiggins and Cody D Christopherson. 2019. The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, 39(4):202.

Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damián Blasi, and Hanna Wallach. 2021. On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *Transactions of the Association for Computational Linguistics*, 9:139–159.

Appendix

A Collecting nouns

We collect words in German [2] and Spanish [3] from a blog post that lists commonly used words in these languages, and shows their grammatical gender. For Bulgarian [4], Greek [5], Czech [6], French [7], Hindi [8], Italian [9], Latvian [10] and Portugese [11], we take a list of words and their grammatical gender from Wikipedia. Following that, we only select words whose English translation is in the list of commonly used words in either German or Spanish.

Language	Total	Masc.	Fem.
Bulgarian	1414	839	575
Czech	2383	1501	882
French	2763	996	1767
German	2031	952	1089
Greek	1257	670	587
Hindi	830	425	405
Italian	2919	2219	700
Latvian	1223	522	701
Portuguese	1766	1119	647
Spanish	1758	896	862

Table 4: **Dataset Statistics.** We present the number of masculine and feminine words we consider for all 10 languages. The languages are sorted alphabetically.

We show the number of collected nouns per language in Table 4. We use 90% of the nouns in each language for training, and 10% for testing.

B Excluding animate nouns

Following prior works that look into grammatical gender by looking at word co-occurrence in text corpora (Williams et al., 2021), we exclude animate nouns from our datasets in all languages (e.g. “uncle”, “cashier”, “engineer”, etc.). We repeat the experiments from Section 3.4 in Table 6, and see that the inclusion of animate nouns does not affect overall results.

LLM	F1	Accuracy Overall	Male	Female
Mistral-7B	0.57	55.0%	50.0%	60.0%
Llama2-7B	0.70	65.0%	50.0%	80.0%

Table 5: **Evaluating the agreement with native English.** We evaluate the agreement of our classifier trained on 10 gendered languages to the perceived grammatical gender of native English speakers, which we treat as ground truth.

Language	F1	Accuracy Overall	Masc.	Fem.
Bulgarian	0.70	71.1%	73.8%	68.3%
German	0.69	63.8%	63.1%	64.2%
Spanish	0.56	55.3%	56.2%	54.4%
Italian	0.51	65.2%	64.5%	67.1%
Czech	0.55	57.2%	54.3%	61.2%
Greek	0.68	69.5%	79.6%	60.1%
Portuguese	0.60	61.1%	56.7%	67.2%
Hindi	0.59	58.1%	67.7%	51.2%
Latvian	0.70	63.2%	60.0%	64.8%
French	0.60	57.0%	58.8%	55.8%

Table 6: **Gendered Nouns Predictions.** This table is for the filtered dictionaries, i.e. without jobs/mother/father etc.

C Gendered adjectives

We show more examples of adjectives that are predominantly used for masculine (or feminine) nouns in Figure 4, similarly to Section 3.2.

D Promps

The prompt we use in English is as follows:

```
***Question***: Describe the word “bottle” using comma-separated adjectives. ***Answer***: glass, sleek, thin, brittle, elegant, transparent, clear, tall, fragile, shiny
***Question***: Describe the word “stone” using comma-separated adjectives. ***Answer***: round, old, strong, cold, solid, ancient, sturdy, dense, natural, durable
***Question***: Describe the word <> using comma-separated adjectives. ***Answer***:
```

For the other languages we translate the prompt, e.g. in Spanish we use:

```
***Pregunta***: Describe la palabra “botella” usando adjetivos separados por comas. ***Respuesta***: vidrio, liso, delgado, quebradizo, elegante, transparente, claro, alto, frágil, brillante
```

[2] https://frequencylists.blogspot.com/2016/01/the-2980-most-frequently-used-german.html
[3] https://frequencylists.blogspot.com/2015/12/the-2000-most-frequently-used-spanish.html
[4] https://en.wiktionary.org/wiki/Category:Bulgarian_nouns_by_gender
[5] https://en.wiktionary.org/wiki/Category:Greek_nouns_by_gender
[6] https://en.wiktionary.org/wiki/Category:Czech_nouns_by_gender
[7] https://en.wiktionary.org/wiki/Category:French_nouns_by_gender
[8] https://en.wiktionary.org/wiki/Category:Hindi_nouns_by_gender
[9] https://en.wiktionary.org/wiki/Category:Italian_nouns_by_gender
[10] https://en.wiktionary.org/wiki/Category:Latvian_nouns_by_gender
[11] https://en.wiktionary.org/wiki/Category:Portuguese_nouns_by_gender

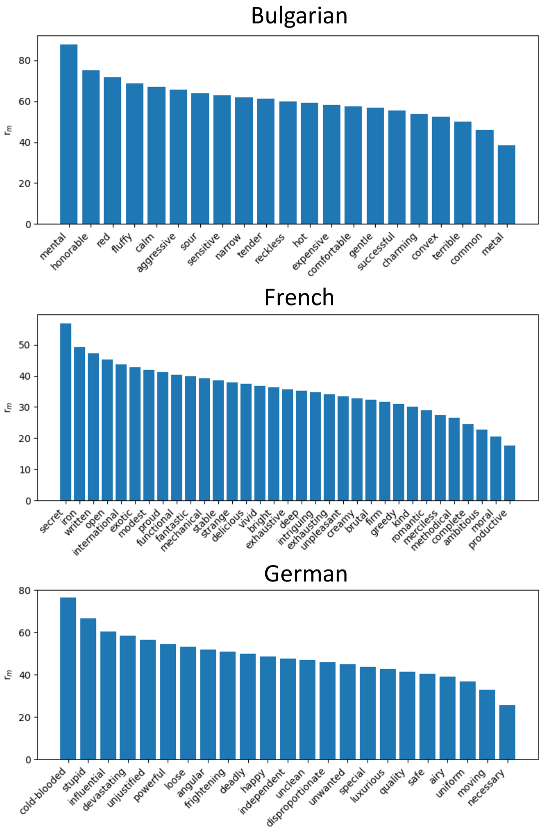

Figure 4: **Bias when describing gendered nouns**. Here we prompt an LLM in Bulgarian, French, and German and for a random sample of adjectives, show the percentage of masculine nouns they were used for.

```
***Pregunta***: Describe la palabra “piedra” usando adjetivos separados por comas. ***Respuesta***: redondo, viejo, fuerte, frío, sólido, antiguo, robusto, denso, natural, duradero
***Pregunta***: Describe la palabra <> usando adjetivos separados por comas. ***Respuesta***:
```