Are We Consistently Biased?
Multidimensional Analysis of Biases in Distributional Word Vectors

Anne Lauscher and **Goran Glavaš**
Data and Web Science Research Group
University of Mannheim
Mannheim, Germany
{anne, goran}@informatik.uni-mannheim.de

Abstract

Word embeddings have recently been shown to reflect many of the pronounced societal biases (e.g., gender bias or racial bias). Existing studies are, however, limited in scope and do not investigate the consistency of biases across relevant dimensions like embedding models, types of texts, and different languages. In this work, we present a systematic study of biases encoded in distributional word vector spaces: we analyze how consistent the bias effects are across languages, corpora, and embedding models. Furthermore, we analyze the cross-lingual biases encoded in bilingual embedding spaces, indicative of the effects of bias transfer encompassed in cross-lingual transfer of NLP models. Our study yields some unexpected findings, e.g., that biases can be emphasized or downplayed by different embedding models or that user-generated content may be less biased than encyclopedic text. We hope our work catalyzes bias research in NLP and informs the development of bias reduction techniques.

1 Introduction

Recent work demonstrated that word embeddings induced from large text collections encode many human biases (e.g., Bolukbasi et al., 2016; Caliskan et al., 2017). This finding is not particularly surprising given that (1) we are likely project our biases in the text that we produce and (2) these biases in text are bound to be encoded in word vectors due to the distributional nature (Harris, 1954) of the word embedding models (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017). For illustration, consider the famous analogy-based gender bias example from Bolukbasi et al. (2016): *"Man is to computer programmer as woman is to homemaker"*. This bias will be reflected in the text (i.e., the word *man* will co-occur more often with words like *programmer* or *engineer*, whereas *woman* will more often appear next to *homemaker* or *nurse*), and will, in turn, be captured by word embeddings built from such biased texts. While biases encoded in word embeddings can be a useful data source for diachronic analyses of societal biases (e.g., Garg et al., 2018), they may cause ethical problems for many downstream applications and NLP models.

In order to measure the extent to which various societal biases are captured by word embeddings, Caliskan et al. (2017) proposed the *Word Embedding Association Test* (WEAT). WEAT measures semantic similarity, computed through word embeddings, between two sets of *target* words (e.g., insects vs. flowers) and two sets of *attribute* words (e.g., pleasant vs. unpleasant words). While they test a number of biases, the analysis is limited in scope to English as the only language, GloVe (Pennington et al., 2014) as the embedding model, and Common Crawl as the type of text. Following the same methodology, McCurdy and Serbetci (2017) extend the analysis to three more languages (German, Dutch, Spanish), but test only for gender bias.

In this work, we present the most comprehensive study of biases captured by distributional word vector to date. We create XWEAT, a collection of multilingual and cross-lingual versions of the WEAT dataset, by translating WEAT to six other languages and offer a comparative analysis of biases over seven diverse languages. Furthermore, we measure the consistency of WEAT biases across different embedding models and types of corpora. What is more, given the recent surge of models for inducing cross-lingual embedding spaces (Mikolov et al., 2013a; Hermann and Blunsom, 2014; Smith et al., 2017; Conneau et al., 2018; Artetxe et al., 2018; Hoshen and Wolf, 2018, *inter alia*) and their ubiquitous application in cross-lingual transfer of NLP models for downstream tasks, we investigate cross-lingual biases encoded in cross-lingual embedding spaces and compare them to bias effects present of corresponding monolingual embeddings.

*Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 85–91
Minneapolis, June 6–7, 2019. ©2019 Association for Computational Linguistics

Our analysis yields some interesting findings: biases do depend on the embedding model and, quite surprisingly, they seem to be less pronounced in embeddings trained on social media texts. Furthermore, we find that the effects (i.e., amount) of bias in cross-lingual embedding spaces can roughly be predicted from the bias effects of the corresponding monolingual embedding spaces.

2 Data for Measuring Biases

We first introduce the WEAT dataset (Caliskan et al., 2017) and then describe XWEAT, our multilingual and cross-lingual extension of WEAT designed for comparative bias analyses across languages and in cross-lingual embedding spaces.

2.1 WEAT

The Word Embedding Association Test (WEAT) (Caliskan et al., 2017) is an adaptation of the Implicit Association Test (IAT) (Nosek et al., 2002). Whereas IAT measures biases based on response times of human subjects to provided stimuli, WEAT quantifies the biases using semantic similarities between word embeddings of the same stimuli. For each bias test, WEAT specifies four stimuli sets: two sets of *target* words and two sets of *attribute* words. The sets of target words represent stimuli *between* which we want to measure the bias (e.g., for gender biases, one target set could contain male names and the other females names). The *attribute* words, on the other hand, represent stimuli *towards* which the bias should be measured (e.g., one list could contain pleasant stimuli like *health* and *love* and the other negative *war* and *death*). The WEAT dataset defines ten bias tests, each containing two target and two attribute sets.[1] Table 1 enumerates the WEAT tests and provides examples of the respective target and attribute words.

2.2 Multilingual and Cross-Lingual WEAT

We port the WEAT tests to the multilingual and cross-lingual settings by translating the test vocabularies consisting of attribute and target terms from English to six other languages: German (DE), Spanish (ES), Italian (IT), Russian (RU), Croatian (HR), and Turkish (TR). We first automatically translate the vocabularies and then let native speakers of the respective languages (also fluent in English) fix the incorrect automatic translations (or introduce better fitting ones). Our aim was to translate WEAT vocabularies to languages from diverse language families[2] for which we also had access to native speakers. Whenever the translation of an English term indicated the gender in a target language (e.g., *Freund* vs. *Freundin* in DE), we asked the translator to provide both male and female forms and included both forms in the respective test vocabularies. This helps avoiding artificially amplifying the gender bias stemming from the grammatically masculine or feminine word forms.

The monolingual tests in other languages are created by simply using the corresponding translations of target and attribute sets in those languages. For every two languages, L1 and L2 (e.g., DE and IT), we create two cross-lingual bias tests: we pair (1) target translations in L1 with L2 translations of attributes (e.g., for T2 we combine DE target sets $\{$*Klavier*, *Cello*, *Gitarre*, ...$\}$ and $\{$*Gewehr*, *Schwert*, *Schleuder*, ...$\}$ with IT attribute sets $\{$*salute*, *amore*, *pace*, ...$\}$ and $\{$*abuso*, *omicidio*, *tragedia*, ...$\}$), and vice versa, (2) target translations in L2 with attribute translations in L1 (e.g., for T2, IT target sets $\{$*pianoforte*, *violoncello*, *chitarra*, ...$\}$ and $\{$*fucile*, *spada*, *fionda*, ...$\}$ with DE attribute sets $\{$*Gesundheit*, *Liebe*, *Frieden*, ...$\}$ and $\{$*Missbrauch*, *Mord*, *Tragödie*, ...$\}$). We did not translate or modify proper names from WEAT sets 3–6. In our multilingual and cross-lingual experiments we, however, discard the (translations of) WEAT tests for which we cannot find more than 20% of words from some target or attribute set in the embedding vocabulary of the respective language. This strategy eliminates tests 3–5 and 10 which include proper American names, majority of which can not be found in distributional vocabularies of other languages. The exception to this is test 6, containing frequent English first names (e.g., *Paul*, *Lisa*), which we do find in distributional vocabularies of other languages as well. In summary, for languages other than EN and for cross-lingual settings, we execute six bias tests (T1, T2, T6–T9).

3 Methodology

We adopt the general bias-testing framework from Caliskan et al. (2017), but we span our study over multiple dimensions: (1) corpora – we analyze the

[1] Some of the target and attribute sets are shared across multiple tests.

[2] English and German from the Germanic branch of Indo-European languages, Italian and Spanish from the Romance branch, Russian and Croatian from the Slavic branch, and finally Turkish as a non-Indo-European language.

Test	Target Set #1	Target Set #2	Attribute Set #1	Attribute Set #2
T1	Flowers (e.g., *aster*, *tulip*)	Insects (e.g., *ant*, *flea*)	Pleasant (e.g., *health*, *love*)	Unpleasant (e.g., *abuse*)
T2	Instruments (e.g., *cello*, *guitar*)	Weapons (e.g., *gun*, *sword*)	Pleasant	Unpleasant
T3	Euro-American names (e.g., *Adam*)	Afro-American names (e.g., *Jamel*)	Pleasant (e.g., *caress*)	Unpleasant (e.g., *abuse*)
T4	Euro-American names (e.g., *Brad*)	Afro-American names (e.g., *Hakim*)	Pleasant	Unpleasant
T5	Euro-American names	Afro-American names	Pleasant (e.g., *joy*)	Unpleasant (e.g., *agony*)
T6	Male names (e.g., *John*)	Female names (e.g., *Lisa*)	Career (e.g. *management*)	Family (e.g., *children*)
T7	Math (e.g., *algebra*, *geometry*)	Arts (e.g., *poetry*, *dance*)	Male (e.g., *brother*, *son*)	Female (e.g., *woman*, *sister*)
T8	Science (e.g., *experiment*)	Arts	Male	Female
T9	Physical condition (e.g., *virus*)	Mental condition (e.g., *sad*)	Long-term (e.g., *always*)	Short-term (e.g., *occasional*)
T10	Older names (e.g., *Gertrude*)	Younger names (e.g., *Michelle*)	Pleasant	Unpleasant

Table 1: WEAT bias tests.

consistency of biases across distributional vectors induced from different types of text; (2) embedding models – we compare biases across distributional vectors induced by different embedding models (on the same corpora); and (3) languages – we measure biases for word embeddings of different languages, trained from comparable corpora. Furthermore, unlike Caliskan et al. (2017), we test whether biases depend on the selection of the similarity metric. Finally, given the ubiquitous adoption of cross-lingual embeddings (Ruder et al., 2017; Glavaš et al., 2019), we investigate biases in a variety of bilingual embedding spaces.

Bias-Testing Framework. We first describe the WEAT framework (Caliskan et al., 2017). Let X and Y be two sets of *targets*, and A and B two sets of *attributes* (see §2.1). The tested statistic is the difference between X and Y in average similarity of their terms with terms from A and B:

$$s(X,Y,A,B) = \sum_{x\in X} s(x,A,B) - \sum_{y\in Y} s(y,A,B)\,, \quad (1)$$

with association difference for term t computed as:

$$s(t,A,B) = \frac{1}{|A|}\sum_{a\in A} f(\mathbf{t},\mathbf{a}) - \frac{1}{|B|}\sum_{b\in B} f(\mathbf{t},\mathbf{b})\,, \quad (2)$$

where $\mathbf{t}$ is the distributional vector of term t and f is a similarity or distance metric, fixed to cosine similarity in the original work (Caliskan et al., 2017). The significance of the statistic is validated by comparing the score $s(X,Y,A,B)$ with the scores $s(X_i,Y_i,A,B)$ obtained for different equally sized partitions $\{X_i,Y_i\}_i$ of the set $X\cup Y$. The p-value of this permutation test is then measured as the probability of $s(X_i,Y_i,A,B) > s(X,Y,A,B)$ computed over all permutations $\{X_i,Y_i\}_i$.[3] The effect size, that is, the "amount of bias", is computed as the normalized measure of separation between association distributions:

$$\frac{\mu\left(\{s(x,A,B)\}_{x\in X}\right) - \mu\left(\{s(y,A,B)\}_{y\in Y}\right)}{\sigma\left(\{s(w,A,B)\}_{w\in X\cup Y}\right)}\,, \quad (3)$$

where μ denotes the mean and σ standard deviation.

Dimensions of Bias Analysis. We analyze the bias effects across multiple dimensions. First, we analyze the effect that different embedding models have: we compare biases of distributional spaces induced from English Wikipedia, using CBOW (Mikolov et al., 2013b), GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2017), and Dict2Vec algorithms (Tissier et al., 2017). Secondly, we investigate the effects of biases in different corpora: we compare biases between embeddings trained on the Common Crawl, Wikipedia, and a corpus of tweets. Finally, and (arguably) most interestingly, we test the consistency of biases across seven languages (see §2.2). To this end, we test for biases in seven monolingual fastText spaces trained on Wikipedia dumps of the respective languages.

Biases in Cross-Lingual Embeddings. Cross-lingual embeddings (CLEs) are widely used in multilingual NLP and cross-lingual transfer of NLP models. Despite the ubiquitous usage of CLEs, the biases they potentially encode have not been analyzed so far. We analyze projection-based CLEs (Glavaš et al., 2019), induced through post-hoc linear projections between monolingual embedding spaces (Mikolov et al., 2013a; Artetxe et al., 2016; Smith et al., 2017). The projection is commonly learned through supervision with few thousand word translation pairs. Most recently, however, a number of models have been proposed that learn the projection without any bilingual signal (Artetxe et al., 2018; Conneau et al., 2018; Hoshen and Wolf, 2018; Alvarez-Melis and Jaakkola, 2018, *inter alia*). Let $\mathbf{X}$ and $\mathbf{Y}$ be, respectively, the distributional spaces of the source (S) and target (T) language and let $D = \{w^i_S, w^i_T\}_i$ be the word translation dictionary. Let $(\mathbf{X}_S, \mathbf{X}_T)$ be the aligned subsets of monolingual embeddings, corresponding to word-aligned pairs from D. We

[3] If f is a distance rather than a similarity metric, we measure the probability of $s(X_i,Y_i,A,B) < s(X,Y,A,B)$.

Metric	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Cos	1.7	1.6	-0.1*	-0.2*	-0.2*	1.8	1.3	1.3	1.7	-0.6*
Euc	1.7	1.6	-0.1*	-0.2*	-0.1*	1.8	1.3	1.3	1.7	-0.7*

Table 2: WEAT bias effects (EN fastText embeddings trained on Wikipedia) for cosine similarity and Euclidean distance. Asterisks indicate bias effects that are insignificant at $\alpha < 0.05$.

then compute the orthogonal matrix $\mathbf{W}$ that minimizes the Euclidean distance between $\mathbf{X_S W}$ and $\mathbf{X_T}$ (Smith et al., 2017): $\mathbf{W} = \mathbf{UV}^\top$, where $\mathbf{U \Sigma V}^\top = SVD(\mathbf{X}_T \mathbf{X}_S{}^\top)$. We create comparable bilingual dictionaries D by translating 5K most frequent EN words to other six languages and induce a bilingual space for all 21 language pairs.

4 Findings

Here, we report and discuss the results of our multi-dimensional analysis. Table 2 shows the effect sizes for WEAT T1–T10 based on Euclidean or cosine similarity between word vector representations trained on the EN Wikipedia using fastText. We observe the highest bias effects for T6 (Male/Female – Career/Family), T9 (Physical/Mental deseases – Long-term/Short-term), and T1 (Insects/Flowera – Positive/Negative). Importantly, the results show that biases do not depend on the similarity metric. We observe nearly identical effects for cosine similarity and Euclidean distance for all WEAT tests. In the following experiments we thus analyze biases only for cosine similarity.

Word Embedding Models. Table 3 compares biases in embedding spaces induced with different models: CBOW, GloVe, fastText, and Dict2Vec. While the first three embedding methods are trained on Wikipedia only, Dict2Vec employs definitions from dictionaries (e.g., Oxford dictionary) as additional resources for identifying strongly related terms.[4] We only report WEAT test results T1, T2, and T7–T9 for Dict2Vec, as the Dict2Vec's vocabulary does not cover most of the proper names from the remaining tests.

Somewhat surprisingly, the bias effects seem to vary greatly across embedding models. While GloVe embeddings are biased according to all tests,[5] fastText and especially CBOW exhibit significant biases only for a subset of tests. We hypothesize that the bias effects reflected in the distributional space depend on the preprocessing steps of the embedding model. fastText, for instance, relies on embedding subword information, in order to avoid issues with representations of out-of-vocabulary and underrepresented terms: additional reliance on morpho-syntactic signal may make fastText more resilient to biases stemming from distributional signal (i.e., word co-occurrences). The fact that the embedding space induced with Dict2Vec exhibits larger bias effects may seem counterintuitive at first, since the dictionaries used for vector training should be more objective and therefore less biased than encyclopedic text. We believe, however, that the additional dictionary-based training objective only propagates the distributional biases across definitionally related words. Generally, we find these results to be important as they indicate that embedding models may accentuate or diminish biases expressed in text.

WEAT	**CBOW**	**GloVe**	**fastText**	**Dict2Vec**
T1	1.20	1.41	1.67	1.35
T2	1.38	1.45	1.55	1.66
T3	−0.28*	1.16	−0.09*	–
T4	−0.35*	1.36	−0.17*	–
T5	−0.36*	1.40	−0.18*	–
T6	1.78	1.75	1.83	–
T7	1.28	1.16	1.30	1.48
T8	0.39*	1.28*	1.30	1.30
T9	1.55	1.35	1.72	1.69
T10	0.09*	1.17	−0.61*	–

Table 3: WEAT bias effects for spaces induced (on EN Wikipedia) with different embedding models: CBOW, GloVe, fastText, and Dict2Vec methods. Asterisks indicate bias effects that are insignificant at $\alpha < 0.05$.

Corpus	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Wiki	1.4	1.5	1.2	1.4	1.4	1.8	1.2	1.3	1.3	1.2
CC	1.5	1.6	1.5	1.6	1.4	1.9	1.1	1.3	1.4	1.3
Tweets	1.2	1.0	1.1	1.2	1.2	1.2	−0.2*	0.6*	0.7*	0.8*

Table 4: WEAT bias effects for GloVe embeddings trained on different corpora: Wikipedia (Wiki), Common Crawl (CC), and corpus of tweets (Tweets). Asterisks indicate bias effects that are insignificant at $\alpha < 0.05$.

Corpora. In Table 4 we compare the biases of embeddings trained with the same model (GloVe) but on different corpora: Common Crawl (i.e., noisy web content), Wikipedia (i.e., encyclopedic text) and a corpus of tweets (i.e., user-generated content). Expectedly, the biases are slightly more pronounced for embeddings trained on the

[4] Two terms A and B are strongly related if B appears in the definition of A and vice versa (Tissier et al., 2017).

[5] This is consistent with the original results obtained by Caliskan et al. (2017).

XW	EN	DE	ES	IT	HR	RU	TR
T1	1.67	1.36	1.47	1.28	1.45	1.28	1.21
T2	1.55	1.25	1.47	1.36	1.10	1.46	0.83
T6	1.83	1.59	1.67	1.72	1.83	1.87	1.85
T7	1.30	0.46*	1.47	1.00	0.72*	0.59*	-0.88
T8	1.30	0.05*	1.16	0.10*	0.13*	0.37*	1.72
T9	1.72	0.82*	1.71	1.57	-0.40^*	1.73	1.09*
Avg_{all}	1.56	0.92	1.49	1.17	0.81	1.22	0.88
Avg_{sig}	1.68	1.4	1.54	1.45	1.46	1.54	1.30

Table 5: XWEAT effects across languages (FASTTEXT embeddings trained on Wikipedias). Avg_{all}: average of effects over all tests; Avg_{sig}: average over the subset of tests yielding significant biases for all languages. Asterisks indicate bias effects that are insignificant at $\alpha < 0.05$.

XW	EN	DE	ES	IT	HR	RU	TR
EN	–	1.09*	1.58	1.49	0.72*	1.17*	1.20*
DE	1.53	–	1.50	1.45	0.55*	1.35	1.07*
ES	1.52	0.79*	–	1.38*	0.60*	1.37*	1.09*
IT	1.33*	0.69*	1.27	–	0.53*	0.82*	0.80*
HR	1.47	1.30*	1.29	1.18*	–	1.14*	1.11*
RU	1.47	0.72*	1.35	1.35	0.77*	–	0.80*
TR	1.41	0.90*	1.37*	1.45	0.29*	0.64*	–

Table 6: XWEAT bias effects (aggregated over all six tests) for cross-lingual word embedding spaces. Rows: *targets* language; columns: *attributes* language. Asterisks indicate the inclusion of bias effects sizes in the aggregation that were insignificant at $\alpha < 0.05$.

Common Crawl than for those obtained on encyclopedic texts (Wikipedia). Countering our intuition, the corpus of tweets seems to be consistently less biased (across all tests) than Wikipedia. In fact, the biases covered by tests T7–T10 are not even significantly present in the vectors trained on tweets. This finding is indeed surprising and the phenomenon warrants further investigation.

Multilingual Comparison. Table 5 compares the bias effects across the seven different languages. Whereas many of the biases are significant in all languages, DE, HR, and TR consistently display smaller effect sizes. Intuitively, the amount of bias should be proportional to the size of the corpus.[6] Wikipedias in TR and HR are the two smallest ones – thus they are expected to contain least biased statements. DE Wikipedia, on the other hand, is the second largest and low bias effects here suggest that German texts are indeed less biased than texts in other languages. Additionally, for (X)WEAT T2, which defines a universally accepted bias (Instruments vs. Weapons), TR and HR exhibit the smallest effect sizes, while the highest bias is observed for EN and IT. We measure the highest gender bias, according to (X)WEAT T6, for TR and RU, and the lowest for DE.

Biases in Cross-Lingual Embeddings. We report bias effects for all 21 bilingual embedding spaces in Table 6. For brevity, here we report the bias effects averaged over all six XWEAT tests (we provide results detailing bias effects for each of the tests separately in the supplementary materials). Generally, the bias effects of bilingual spaces are in between the bias effects of the two corresponding monolingual spaces (cf. Table 5): this means that we can roughly predict the amount of bias in a cross-lingual embedding space from the same bias effects of corresponding monolingual spaces. For example, effects in cross-lingual spaces increase over monolingual effects for low-bias languages (HR and TR), and decrease for high-bias languages (EN and ES).

5 Conclusion

In this paper, we have presented the largest study to date on biases encoded in distributional word vector spaces. To this end, we have extended previous analyses based on the WEAT test (Caliskan et al., 2017; McCurdy and Serbetci, 2017) in multiple dimensions: across seven languages, four embedding models, and three different types of text. We find that different models may produce embeddings with very different biases, which stresses the importance of embedding model selection when fair text representations are to be created. Surprisingly, we find that the user-generated texts, such as tweets, may be less biased than redacted content. Furthermore, we have investigated the bias effects in cross-lingual embedding spaces and have shown that they may be predicted from the biases of corresponding monolingual embeddings. We make the XWEAT dataset and the testing code publicly available,[7] hoping to fuel further research on biases encoded in word representations.

Acknowledgments

We thank our six native speakers for manually correcting and improving (i.e., post-editing) the automatic translations of the WEAT dataset.

[6] The larger the corpus the larger is the overall number of contexts in which some bias may be expressed.

[7] At: https://github.com/umanlp/XWEAT.

References

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of EMNLP*, pages 1881–1890.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of EMNLP*, pages 2289–2294.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL*, pages 789–798.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of NIPS*, pages 4356–4364, USA. Curran Associates Inc.

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

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of ICLR*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *arXiv preprint arXiv:1902.00508*.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of ACL*, pages 58–68.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of EMNLP*, pages 469–478.

Katherine McCurdy and Oguz Serbetci. 2017. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. In *Proceedings of WiNLP*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR, abs/1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceesings of NIPS*, pages 3111–3119.

Brian A. Nosek, Anthony G. Greenwald, and Mahzarin R. Banaji. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics*, 6:101–115.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*.

Samuel L. Smith, David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

Julien Tissier, Christophe Gravier, and Amaury Habrard. 2017. Dict2vec : Learning word embeddings using lexical dictionaries. In *Proceedings of EMNLP*, pages 254–263.

XW1	EN	DE	ES	IT	HR	RU	TR
EN	–	1.28	1.63	1.62	1.59	1.49	1.32
DE	1.55	–	1.28	1.45	1.41	1.03	1.29
ES	1.45	1.25	–	1.28	1.21	1.31	1.09
IT	1.18	1.10	1.28	–	1.29	0.61	1.09
HR	1.57	1.62	1.59	1.62	–	1.62	1.63
RU	1.41	1.12	1.20	1.38	1.46	–	1.29
TR	1.23	1.21	1.06	1.26	1.24	1.04	–

Table 7: XWEAT T1 effect sizes for cross-lingual embedding spaces. Rows denote the target set language, column the attribute set language.

XW2	EN	DE	ES	IT	HR	RU	TR
EN	–	1.35	1.51	1.48	1.60	1.56	1.15
DE	1.37	–	1.25	1.19	1.31	1.47	1.16
ES	1.55	1.50	–	1.53	1.50	1.57	1.22
IT	1.54	1.37	1.28	–	1.47	1.39	1.27
HR	1.19	1.25	0.72	1.09	–	1.26	0.81
RU	1.46	1.26	1.23	1.08	1.13	–	0.71
TR	1.29	1.44	1.21	1.4	1.25	1.57	–

Table 8: XWEAT T2 effect sizes for cross-lingual embedding spaces. Rows denote the target set language, column the attribute set language.

A Additional Results

For completeness, we report detailed results on bias effects for each of the six XWEAT tests and bilingual word embedding spaces for all 21 language pairs. Tables 7 to 12 show bias effects for XWEAT tests T1, T2, and T6–T9.

XW6	EN	DE	ES	IT	HR	RU	TR
EN	–	1.77	1.81	1.88	1.83	1.78	1.89
DE	1.82	–	1.77	1.85	1.84	1.74	1.86
ES	1.71	0.95	–	1.81	1.80	1.61	1.50
IT	1.76	1.58	1.703	–	1.72	1.77	1.76
HR	1.68	1.65	1.66	1.43	–	1.74	1.73
RU	1.86	1.74	1.74	1.82	1.86	–	1.80
TR	1.90	1.66	1.77	1.82	1.77	1.55	–

Table 9: XWEAT T6 effect sizes for cross-lingual embedding spaces. Rows denote the target set language, column the attribute set language.

XW7	EN	DE	ES	IT	HR	RU	TR
EN	–	0.34^*	1.36	1.33	0.26^*	0.46^*	0.49^*
DE	1.51	–	1.60	1.42	0.23^*	1.33	-0.62^*
ES	1.63	0.24^*	–	1.26	0.60^*	1.29	1.55
IT	1.12	0.65^*	1.01	–	0.51^*	-0.20^*	-1.08
HR	1.46	0.94	0.95	1.27	–	0.62^*	0.00^*
RU	1.19	-0.51^*	1.30	1.09	0.81^*	–	-0.79^*
TR	1.22	0.07^*	0.81^*	1.30	-0.23^*	-0.48^*	–

Table 10: XWEAT T7 effect sizes for cross-lingual embedding spaces. Rows denote the target set language, column the attribute set language.

XW8	EN	DE	ES	IT	HR	RU	TR
EN	–	0.68^*	1.49	1.01	-0.38^*	-0.06^*	0.71^*
DE	1.17	–	1.43	1.10	-0.09^*	1.06	1.16
ES	1.13	-0.69^*	–	0.61^*	-0.19^*	0.67^*	-0.18^*
IT	0.75^*	-0.76^*	0.87	–	-0.18^*	-0.52^*	0.04^*
HR	1.36	0.42^*	0.92	0.76^*	–	-0.16^*	0.90
RU	1.09	-0.84^*	0.96	0.99	0.19^*	–	1.00
TR	0.93	0.06^*	1.49	1.21	-0.47^*	-0.43^*	–

Table 11: XWEAT T8 effect sizes for cross-lingual embedding spaces. Rows denote the target set language, column the attribute set language.

XW9	EN	DE	ES	IT	HR	RU	TR
EN	–	1.12	1.66	1.61	-0.59^*	1.76	1.65
DE	1.74	–	1.68	1.66	-1.39	1.46	1.57
ES	1.64	1.48	–	1.79	-1.34	1.75	1.37
IT	1.62	0.19^*	1.47	–	-1.63	1.87	1.74
HR	1.54	1.89	1.87	0.96^*	–	1.73	1.59
RU	1.82	1.54	1.64	1.72	-0.84^*	–	0.80^*
TR	1.88	0.98^*	1.88	1.70	-1.80	0.58^*	–

Table 12: XWEAT T9 effect sizes for cross-lingual embedding spaces. Rows denote the target set language, column the attribute set language.