The (Undesired) Attenuation of Human Biases by Multilinguality

Cristina España-Bonet
DFKI GmbH, Saarland Informatics Campus
Saarbrüken, Germany
cristinae@dfki.de

Alberto Barrón-Cedeño
Universitá di Bologna
Forlì, Italy
a.barron@unibo.it

Abstract

Some human preferences are universal. The odor of vanilla is perceived as pleasant all around the world. We expect neural models trained on human texts to exhibit these kind of preferences, i.e. biases, but we show that this is not always the case. We explore 16 static and contextual embedding models in 9 languages and, when possible, compare them under similar training conditions. We introduce and release CA-WEAT, multilingual cultural aware tests to quantify biases, and compare them to previous English-centric tests. Our experiments confirm that monolingual static embeddings do exhibit human biases, but values differ across languages, being far from universal. Biases are less evident in contextual models, to the point that the original human association might be reversed. Multilinguality proves to be another variable that attenuates and even reverses the effect of the bias, specially in contextual multilingual models. In order to explain this variance among models and languages, we examine the effect of asymmetries in the training corpus, departures from isomorphism in multilingual embedding spaces and discrepancies in the testing measures between languages.

1 Introduction

The perception of odor pleasantness has been shown to be universal. Even if variations across individuals exist, they do not depend on culture (Arshamian et al., 2022). We call this is a *human bias*: judging a phenomenon in terms of values that are inherent in human beings, regardless of their sociocultural background. On the contrary, a *cultural bias* is “the tendency to interpret and judge phenomena in terms of the distinctive values, beliefs, and other characteristics of the society or community to which one belongs”.[1] Racism, sexism, ageism, etc. are all *social cultural biases* and can be more or less present in distinct communities.

As long as neural systems are trained on general-domain texts written by humans, one would expect and desire human biases to be present in embedding models for all languages. One would also expect (but not always desire) cultural biases, but these would depend on the language or the culture behind it. Lots of work has been done on detecting social cultural biases and trying to mitigate them (Bolukbasi et al., 2016; Zhao et al., 2018; Gonen and Goldberg, 2019; Ravfogel et al., 2020; Dev et al., 2020; Schick et al., 2021; Zhou et al., 2022). Also recent work has investigated the presence of human biases in static word embeddings; first in English (Caliskan et al., 2017) and later in other languages (Lauscher and Glavaš, 2019; Lauscher et al., 2020a,b). These works make use of the Word Embedding Association Test (WEAT) for English —lists of concepts and attributes that hide implicit human associations as described in Section 2— and X-WEAT, WEAT’s translations, for other languages. To the best of our knowledge, research on language models and contextualised embeddings, has been done only in English, using extensions or variations of WEAT (May et al., 2019; Kurita et al., 2019; Guo and Caliskan, 2021).

Natural language processing aspires to multilinguality. When talking about embeddings, one achieves multilinguality by mapping two or more spaces into a single one, or by joint training with text in two or more languages. These are the two main approaches to crosslingual word embeddings (Ruder et al., 2019) and the approaches used to train language models (Devlin et al., 2019; Lin et al., 2021; Conneau et al., 2020b).

We go deeper into the understanding of multilingual embedding models and the effect, if any, of the different approaches used to build them. For this purpose, we *first* review results and some of the implicit assumptions made in previous works

[1] https://dictionary.apa.org/cultural-bias

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2056 - 2077
December 7-11, 2022 ©2022 Association for Computational Linguistics

by investigating (*i*) whether translations from the original WEAT English lists (X-WEAT) are fair tests for other languages and (*ii*) whether a single list, either original or translated, is representative for a language. *Second*, for the first time we define and collect cultural aware WEAT lists (CA-WEAT) to enable multilingual analyses completely independent of English. *Third*, we use WEAT, X-WEAT and CA-WEAT to study the effect of cross- and multilinguality in embedding models; that is, how they differ with respect to the monolingual English. Since differences do exist, we try to explain them in terms of the testing measure (*-WEAT), the nature of the training corpora, the method to achieve multilinguality and the final differences in the topology of the embedding spaces between languages. We perform a systematic study over 9 languages of 3 families in word embeddings from 16 static and contextual models. Our premise is that biases should be equally present, and we analyse departures from this premise with the focus on multilinguality.

2 Quantifying Bias with WEAT

The Word Embedding Association Test (WEAT) (Caliskan et al., 2017) is a bias measurement method for word embeddings. WEAT is inspired by the Implicit Association Test (IAT) for humans (Greenwald et al., 1998), which measures differences in response time when subjects are requested to pair items and attributes that they find similar and when pairing items and attributes that they find different. To give an example, subjects would be first asked to label the item *orchid* as *flower-pleasant* or *insect-unpleasant*. This would be repeated for several flowers and insects. In a second part, subjects would be asked to label the same list of items as *insect-pleasant* or *flower-unpleasant*. Experiments show that the response time for the first part is lower than for the second one. The cognitive effort for the latter is higher because the association flower-unpleasant is less expected than flower-pleasant. These results expose a human bias: flowers are more pleasant than insects and insects are more unpleasant than flowers. If this is a universal human bias, one would expect it to be present also in embeddings created from human texts. Flowers in a semantic space should be closer to pleasant attributes than insects, and insects closer to unpleasant attributes than flowers.

Many WEAT (and IAT) tests exist; most of them are designed to measure biases (implicit associations) towards racial groups, gender, sexuality, age, and religion. Only two are non-social and therefore we expect them to be culture- and language-independent: flowers/insects vs. pleasant/unpleasant (WEAT1) and musical instruments/weapons vs. pleasant/unpleasant (WEAT2). These are considered "universally accepted stereotypes" (Greenwald et al., 1998; Toney and Caliskan, 2021). Each attribute and concept in these tests has associated a list of 25 English terms, collected as described in Section 3 (listed in Appendix C).

The original WEAT measure (Caliskan et al., 2017) defines the association of each term t (e.g. *orchid*) as its average cosine similarity to the list of target attributes A (e.g. *pleasant* concepts):

$$assoc(t, A) = \frac{\sum_{a \in A} sim(\mathbf{t}, \mathbf{a})}{|A|}, \tag{1}$$

where $\mathbf{t}$ is the embedding for t and $\mathbf{a}$ is the embedding for an element $a \in A$.[2] The association difference Δ_{assoc} for a term t between attributes A (*pleasant*) and B (*unpleasant*) is then

$$\Delta_{assoc}(t, A, B) = assoc(t, A) - assoc(t, B). \tag{2}$$

Given two sets of target terms X and Y with n elements each (e.g., *flowers* and *insects*), the statistic s is the difference in average similarity of their terms with elements from A and B:

$$\begin{aligned} s(X, Y, A, B) = & \sum_{x \in X} \Delta_{assoc}(x, A, B) - \\ & \sum_{y \in Y} \Delta_{assoc}(y, A, B). \end{aligned} \tag{3}$$

We use Cohen's d to estimate the effect size (i.e. the strength of the bias). Cohen's d is a standardised measure of the effect defined as the difference between the two means divided by the standard deviation for all instances in X and Y:

$$d = \frac{\mu\left(\Delta_{assoc}(x, A, B)_{\forall x \in X}\right) - \mu\left(\Delta_{assoc}(y, A, B)_{\forall y \in Y}\right)}{\sigma\left(\Delta_{assoc}(w, A, B)_{\forall w \in X \cup Y}\right)}. \tag{4}$$

Sawilowsky (2009) defined the scale of magnitude for d as very small (< 0.01), small (< 0.20), medium (< 0.50), large (< 0.80), very large (< 1.20), and huge (< 2.00).

[2] Lauscher and Glavaš (2019) showed no significant difference in the results obtained with cosine similarity and Euclidean distance. We confirmed the results and only report those with cosine similarity.

3 Multilingual Aspects and CA-WEAT

Both IAT and WEAT have been traditionally created in the north east of the US and performed in English. The pleasant and unpleasant words used in WEAT1 and WEAT2 were selected from norms in Bellezza et al. (1986), where college students in Ohio rated a list of words for pleasantness. From this list, 25 elements were taken as pleasant words and 25 as unpleasant words by Greenwald et al. (1998). The lists for flowers, insects, musical instruments, and weapons were extracted from Battig and Montague (1969), where college students from Maryland and Illinois were given 30 seconds to write down as many objects within each category as possible. Greenwald et al. (1998) selected 25 *unambiguous* items that they thought their students would be *familiar* with. These two requests are relevant: they ensure taking frequent words in the language that have a single meaning.

The first experiment with word embeddings was done by Caliskan et al. (2017) who used pretrained English GloVe embeddings (Pennington et al., 2014). They observed that, according to their results on social biases, the training corpus "may be disproportionately American". Subsequent studies used crosslingual WEAT (X-WEAT) to go beyond analyses in English (Lauscher and Glavaš, 2019; Lauscher et al., 2020a). The original lists were translated into several languages and biases estimated using the translated items and attributes. They found differences in the biases obtained across languages and connected them to differences in the size of the training corpora. They also explored bilingual spaces and observed that the bias effects were in the middle of the two corresponding biases in the monolingual spaces.

As discussed, WEAT1 and WEAT2 originate in the US. Even if a concept (flower) might be considered pleasant in every culture, the items themselves (*orchid*, *broom*, etc.) can be different across cultures. This is most evident for elements that depend on geography (a flower that grows in the US or an insect that lives there might not be present in other locations), but it could happen for all the other items in WEAT1 and WEAT2. As a result, the representation of the original translated items in the training data in other languages might be smaller or, even worse, the distribution of the opposite attributes asymmetric. As we will show, there is already a variation in the terms and attributes used by different people within a common culture, but testing the existence of human biases with translations from American English might be inducing an additional cultural bias in the results. There is a second argument to avoid X-WEAT: translations are not perfect and one cannot assure that the requirements in Greenwald et al. (1998) (unambiguous and frequent words) hold. For example, the Spanish X-WEAT translates *blade* as *hoja* (the edge of a knife, but also a sheet of paper) and turns both *fiddle* and *violin* into *violín*. Whereas correct, translation introduces an ambiguous word lacking any association to (un)pleasant attributes in the former case, and reduces the size of the list in the latter.

To mitigate the problems introduced by translations and to estimate their impact in the analysis, we create CA-WEAT: a new collection of *cultural-aware* lists written by native speakers of different languages. We asked volunteers to create lists of flowers, insects, weapons and musical instruments, as well as both pleasant and unpleasant concepts with 25 elements each without any time constraint. The only requirement was that words needed to be common in their culture. Lists from different volunteers are semantically equivalent, since they characterise the same concepts, and can be seen as perturbations on a prototypical (or average) set.

We first conducted a pilot study with 14 volunteers from 11 nationalities to survey the difficulty of the task and prepare the guidelines of the experiment. Five of them failed to complete the task. After the pilot, we set up an online form with detailed instructions (see Appendix A) and distributed it to contacts in different countries.

We collected 112 CA-WEAT lists in 26 languages from which we discarded 9 after a quality check.[3] For the current experiments, we selected 9 of the 26 languages, for a total of 82 lists. The lists in the remaining 17 languages are provided in the CA-WEAT dataset but we do not use them in the subsequent analysis; statistics for all of them are reported in Appendix B. The languages considered here are chosen according to 3 criteria: high-quality embeddings could be obtained, the equivalent X-WEAT exists or could be created by a native speaker at hand, and different language families are covered. These constraints led to considering Arabic (ar), Catalan (ca), Croatian (hr), English (en), German (de), Italian (it), Russian (ru), Spanish (es) and Turkish (tr). The distribution among languages

[3] We excluded lists with less than 25 elements or filled with non-sensical words, and lists including stopwords.

is not even: we collected 24 lists in Italian and German, 12 in Croatian, 10 in Spanish, 5 in English, 2 in Catalan, Romanian and Turkish, and 1 in Arabic. Instead of aggregating the highest ranked/most frequent words per concept into a single list as WEAT does, CA-WEAT uses all the words and provides a list per subject. This allows us to study statistically the variations and the relevance of the test sets.

For X-WEAT, we use the translations provided by Lauscher and Glavaš (2019) and Lauscher et al. (2020b), after revising the Spanish ones and adding the Catalan translations.

4 Embedding Models

We consider 16 embedding models for the 9 selected languages. We select two kinds of models. On the left-hand side, widely-used out-of-the-box pre-trained models. On the right-hand side, models trained in-house. For the latter, we control both the amount and domain of the training data, as well as the approach used to reach multilinguality.

For static embeddings, we use pre-trained fastText word embeddings:[4]

WP: Monolingual models trained on Wikipedia using the skip-gram architecture with subword information, as described by Bojanowski et al. (2017).

WPali: *WP* aligned to English with the RCSLS method as described by Joulin et al. (2018).

CCWP: Models trained on Common Crawl and Wikipedia using CBOW with position weights and subword information (Grave et al., 2018).

We also build 5 static in-house word embeddings on Common Crawl using a subset of the CC-100 corpus (Conneau et al., 2020a; Wenzek et al., 2020). We enforce to have the same number of words for all 9 languages under study (CCe) by ceiling the size to that of the language with the smallest corpus: Catalan, with $1.7 \cdot 10^9$ words (see Appendix D). For comparison purposes, the training of the 5 models is done with the same architecture and hyperparameters as in *CCWP*: CBOW with position-weights, 300 dimensions, character 5-grams, a window of size 5 and 10 negatives. The five in-house embeddings are:

CCe Monolingual embeddings trained on CCe.

CCeVMuns *CCe* aligned to the English space using unsupervised VecMap (Artetxe et al., 2018b).[5]

CCeVMsup Supervised VecMap (Artetxe et al., 2018a) using the test part of the cross-lingual dictionaries in MUSE.[6]

CCe2langs Bilingual embeddings trained on the concatenation of CCe-en and CCe-L_i for one of the other 8 L_i languages.

CCe9langs Multilingual embeddings trained on the concatenation of the 9 CCe-L_i.

Purely static embeddings are compared to word embeddings extracted from 3 pre-trained contextual models:

mBERT$_0$ Static embeddings (layer 0) in multilingual BERT (Devlin et al., 2019); trained on 104 languages including the ones we analyse with a BPE vocabulary of 110k.

mBERT$_{11}$ Embeddings in the next-to-last layer (layer 11) of multilingual BERT.[7]

BERT$_0$ We use monolingual BERT for Arabic (Antoun et al., 2020), German,[8] Italian (Schweter, 2020b), Spanish (Cañete et al., 2020), Turkish (Schweter, 2020a) and English (Devlin et al., 2019). For the other languages, the model is not available or it finetunes *mBERT*.

BERT$_{11}$ Embeddings in the 11th layer of the same models as in *BERT*$_0$.

XLM-R$_0$ Static embeddings in XLM-RoBERTa (Conneau et al., 2020a); trained on 100 languages with a BPE vocabulary of 250k.

XLM-R$_{11}$ Embeddings in the next-to-last layer of XLM-RoBERTa.

XGLM$_0$ Static embeddings in XGLM (Lin et al., 2021); trained on 30 languages with a BPE vocabulary of 250k, excluding Croatian.

XGLM$_{47}$ Embeddings in the next-to-last layer (layer 47 in this case) of XGLM.

While static word2vec-like embeddings have 300 dimensions, BERT embeddings have 768, XLM-RoBERTa 1024 and XGLM 2048.

5 Experiments and Results

We calculate the statistic and the effect size (Cohen's d) for the 16 types of embeddings in the 9 languages for WEAT1, WEAT2, the 82 CA-WEAT1

[4] https://fasttext.cc/docs/en/pretrained-vectors.html

[5] https://github.com/artetxem/vecmap

[6] https://github.com/facebookresearch/MUSE#ground-truth-bilingual-dictionaries

[7] Jawahar et al. (2019) showed that the BERT layers encode different linguistic information: better semantics in the top and syntax in the middle layers. The top layer is the most adapted to the final task, so we use the next-to-last layer. Following this observation, we use the static layer and the next-to-last layer (instead of the last one) for our word embeddings extracted from contextual models.

[8] https://www.deepset.ai/german-bert

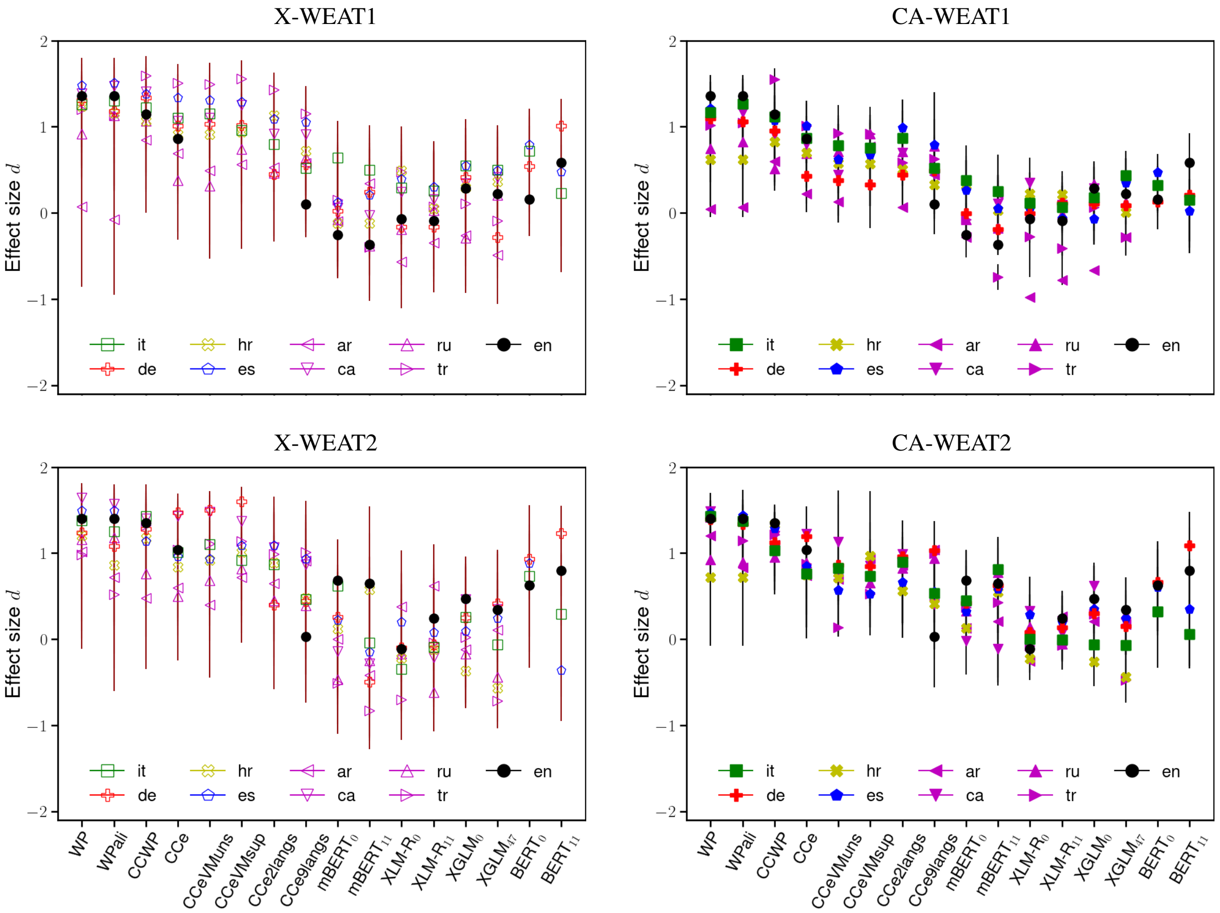


Figure 1: Effect sizes for X-WEAT and CA-WEAT tests for the 9 languages and 16 embedding models. The English entry corresponds to WEAT. Languages with less than 10 CA-WEATs are marked with magenta triangles.

and CA-WEAT2 and the 8 translated X-WEAT1 and X-WEAT2 lists.[9]

For the statistical analysis of the results, we estimate the uncertainty on the statistic and the effect size by (*i*) averaging the results with multiple lists for CA-WEAT and (*ii*) creating bootstrapped versions of a single instance in all cases: WEAT, X-WEAT and CA-WEAT. For the average version, we provide the median and 95% confidence intervals (CI) using order statistics given that the distributions are non-normal and contain few elements. For the bootstrapped version, we resample with replacement the four lists involved in an experiment to generate 5,000 synthetic sets per test.[10] Tables 6, 7, 8 and 9 in Appendix E show the detailed results.

WEAT vs X-WEAT vs CA-WEAT. First, we compare the conclusions one gets using each of the three testing alternatives. Figure 1 depicts the effect size estimations for X-WEAT and the median of the CA-WEAT tests. Biases measured with X-WEAT are in general higher than those with CA-WEAT, specially for pure static embeddings, but differences are not statistically significant at 95% level. The dispersion within a model and across languages is smaller for CA-WEAT than for X-WEAT; an indication that a single list is not representative for a language and the average of several lists helps to get closer to a *universal* effect size value.

The variation across lists of the same language is big. The top-row plots in Figure 2 compare the effect size of the original WEAT1 test and 5 CA-WEAT1 lists created by native speakers of American English, taking 3 models as representatives. The bottom part of each plot shows the effect size for the CA-WEAT1 tests (grey and red) and the WEAT1 test (blue). The top part shows the histogram of the CA-WEAT1 tests and reports the

[9] Both the source code and the data to reproduce our analysis, including the revised X- and the CA-WEAT lists, are available at `https://github.com/cristinae/CA-WEAT`.

[10] Previous work reported p-values in permutation tests. Since we are interested in variances between languages and within a language itself, we chose to report confidence intervals instead. We adapt the code in Lauscher and Glavaš (2019) for this purpose.

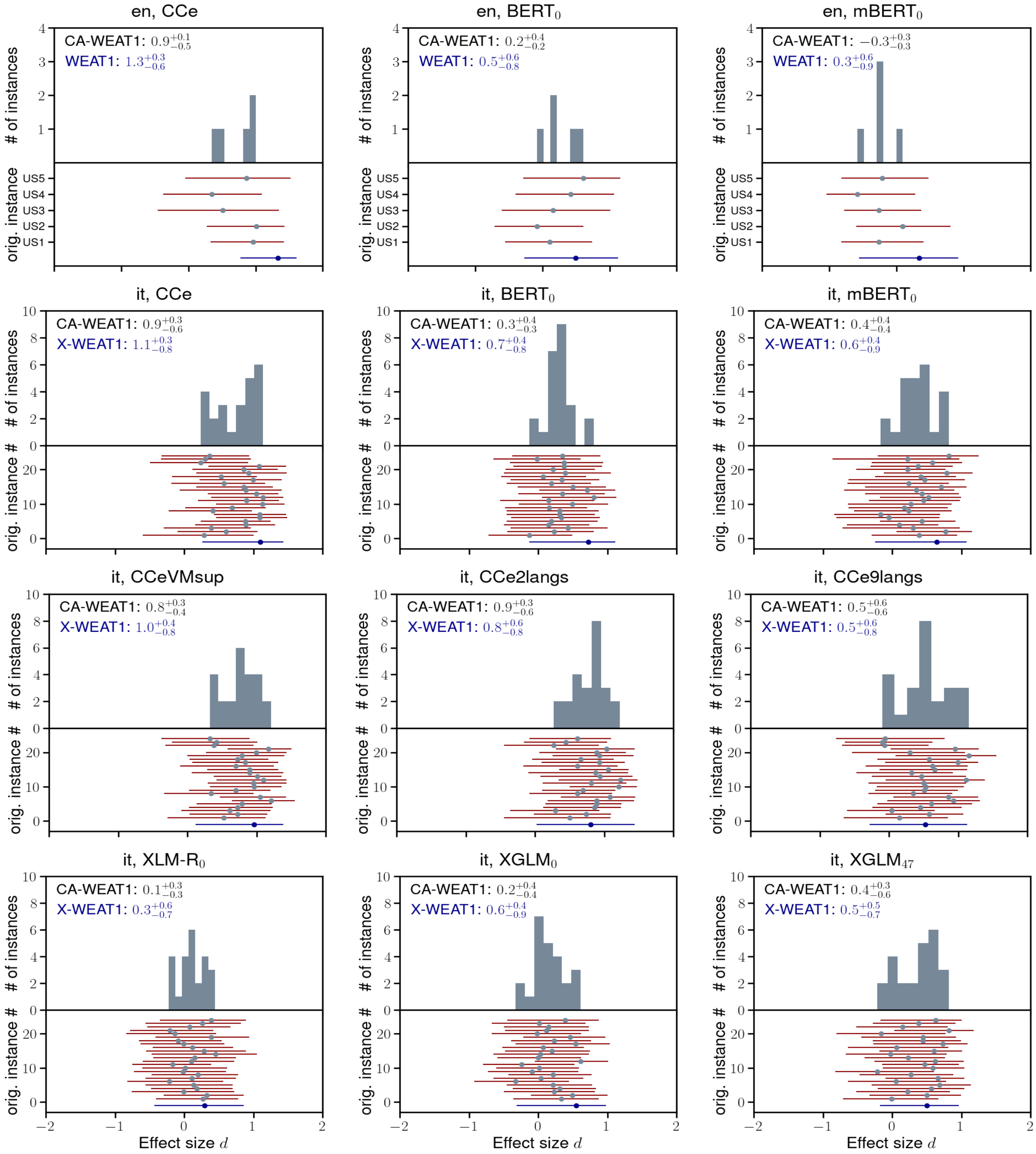


Figure 2: Effect sizes for WEAT1 and 5 CA-WEAT1 tests in American English (top row). The other plots correspond to Italian for 9 representative word embedding models.

numbers displayed in Figure 1. The five lists show very different behaviours across models, but also within a single model. For the monolingual model *CCe*, we see that for three of the lists the effect is compatible with no bias at 95% level. One would expect the average of the English CA-WEAT lists to be close to WEAT, as the latter was obtained by combining inputs from several subjects. However, the variation is huge and depends on the model. In order to be able to substitute WEAT with CA-WEAT, one needs multiple samples. X-WEAT can be considered as just one of these samples and this can explain why the variation across languages is larger with X-WEAT than with CA-WEAT.

The remaining plots in Figure 2 compare X- and CA-WEAT1 by looking at the variation in more samples, the Italian lists. The variation among CA-WEATs is large for all models, and X-WEAT results are within the CIs of CA-WEAT. According to Sawilowsky (2009)'s scale, the magnitude of the biases ranges from medium (d<0.5) to very large (<1.2) for X-WEAT and from small (<0.2) to large (<0.8) for CA-WEAT. Similar trends are seen for German, also with 24 CA-WEATs.

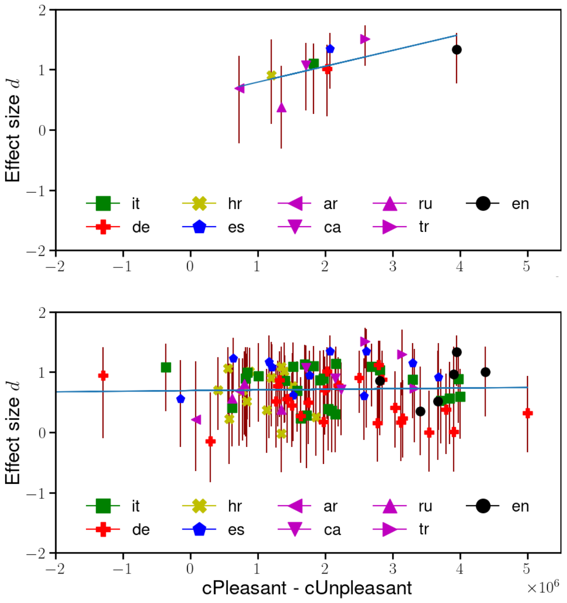


Figure 3: Effect sizes observed in the *CCe* model for only X-WEAT1 (top) and all *-WEAT1 tests (bottom) as a function of the asymmetry between counts of positive and negative attributes in the training corpus.

Comparable experiments with CCe. We compare the effect size d across languages and multilingual methods in the setting where all the embedding models are trained in a similar domain, using the same amount of data. The size of the corpus is not the only deciding factor though, as the number of times that the items in the lists appear in the corpus might also have an effect. Words in the lists are frequent enough (billions of occurrences) to get high quality embeddings, but an asymmetry between the number of positive and negative concepts might create an artificial bias.

We test this hypothesis by studying the correlation between the bias effect size and the difference between the counts of the positive (*pleasant*) and the negative (*unpleasant*) attributes.[11] Figure 3 shows the relation for the monolingual *CCe* embedding models. The correlation seems important for X-WEAT1 (top plot), which shows a positive trend with half of the variation in the effect size being explained by the number of counts R^2=0.493. However, this might be an effect of either having only 9 data points or X-WEAT and CA-WEAT coming from two different distributions. When we consider the 82 CA-WEAT1 tests and the 9 X-WEAT1 (bottom plot), we observe a flat slope where the variance is not explained by the counts (R^2=0.001). Results for X-/CA-WEAT2 are equivalent, with R^2=0.334 and R^2=0.008 respectively. CA-WEAT lists allow to see that the lack of trend is language independent. The average effect size for Spanish is higher than for German, the count difference is larger for English than for Croatian, but in none of the cases can the asymmetry counts explain the effect size variance.

Multilinguality changes the distribution of effect sizes and also the attribute counts in the training corpus of some models. While *CCeVMuns* and *CCeVMsup* project the pre-trained spaces into a common one, *CCe2langs* and *CCe9langs* train the joint embeddings on the concatenated corpus. In this case, the counts change as different languages can share the same surface token for some words. This might be a reason why differences in biases with respect to English in languages with different scripts are more relevant than differences with languages in different families. But it is not the only reason, as we observe the same trend in the projection methods. In general, the 4 multilingual approaches share the same conclusions as their monolingual counterparts: for X-WEAT one observes a positive correlation between d and the difference of counts but, when inspecting all the CA-WEAT lists, the correlation disappears. *CCe9langs* with the widest differences of counts shows almost no correlation for X-WEAT as well.

Multilingual models and isomorphism. Going from monolingual *CCe* to bilingual *CCeVMuns* implies a mitigation of the bias for CA-WEAT but biases remain close to constant for X-WEAT (see Figure 1). Contrary to the findings by Lauscher and Glavaš (2019), the bilingual embeddings do not show a bias halfway between that of the 2 languages. The effect of supervision (*CCeVMuns* vs *CCeVMsup*) is not consistent and results are in general equivalent. With few exceptions, the effect of *CCe2langs* and especially of *CCe9langs* is also a mitigation of the bias with respect to the one observed in the corresponding monolingual model *CCe*, but this is not statistically significant at 95% level. Arabic and Russian, both with non-Latin alphabets, have the lowest d (together at times with Turkish, the only other non-Indo-European language), but we cannot attribute it to multilinguality, because their effect size is also low for the monolingual embeddings.

To further investigate what is behind the variance

[11] The asymmetry could also come with the target items, but attributes are an order of magnitude more abundant and differences more significant (see counts in Appendix D).

	ar		ca		de		es		hr		it		ru		tr	
	EV	GH	EV	GH	EV	GH	EV	GH	EV	GH	EV	GH	EV	GH	EV	GH
WP	106	0.47	12	0.49	12	0.31	10	0.18	42	0.54	21	0.24	16	0.43	49	0.39
WPali	143	0.55	22	0.51	22	0.36	16	0.37	46	0.61	19	0.34	30	0.32	36	0.44
CCWP	15	0.40	85	0.42	42	0.92	23	0.41	51	0.65	41	0.37	32	0.64	28	0.55
CCe	55	0.62	253	0.23	26	0.79	166	0.54	91	0.61	223	0.25	8	0.56	25	0.43
CCeVMuns	229	1.56	229	1.27	27	0.82	167	1.95	69	0.93	220	1.19	27	0.96	36	0.84
CCeVMsup	36	0.56	231	0.86	32	0.70	87	0.73	27	0.61	123	0.65	25	0.80	11	0.41
CCe2langs	93	0.53	8	0.43	19	0.94	72	0.35	33	0.81	51	0.41	39	0.51	64	0.61
CCe9langs	475	1.46	23	0.84	171	1.27	21	0.61	53	1.22	51	0.41	403	1.50	149	1.15
mBERT$_0$	154	0.85	133	0.33	95	0.56	99	0.56	270	0.44	131	0.17	161	0.54	589	0.51
XLM-R$_0$	54	0.38	74	0.45	59	0.43	150	0.44	58	0.54	113	0.56	111	0.32	277	0.33
XGLM$_0$	67	0.95	88	1.21	144	1.18	135	2.24	*2584	*2.30	130	1.33	85	1.64	475	0.68

Table 1: Isomorphism measures (EV and GH) between the English (sub)space and the (sub)spaces for the other 8 languages for 11 representative embedding models. $^{(*)}$XGLM does not include hr data in training.

in the multilingual setting, we evaluate the isomorphism between spaces. Intuitively, if spaces are isomorphic, multilinguality should not alter the properties of the monolingual embeddings; if spaces are far from being isomorphic, a joint training might distort semantics and translations could lie further apart in projected spaces. Some differences in d could be therefore explained if spaces are not (close to) isomorphic. We use two well known measures: the Eigenvector similarity (EV) (Søgaard et al., 2018) and the Gromov-Hausdorff distance (GH) (Patra et al., 2019).[12] In both cases, lower values indicate more isomorphic spaces. For *WP*, *WPali*, *CCWP*, *CCe*, *CCeVMuns* and *CCeVMsup*, we calculate the values between English and each one of the other 8 languages L_i. For *CCe2langs*, *CCe9langs*, *mBERT*$_0$, *XLM-R*$_0$ and *XGLM*$_0$, we extract the subspaces for English and L_i using the vocabulary in the English CCe and in the L_i CCe and calculate the distance between the subspaces. Table 1 compiles the results. As noted by Patra et al. (2019) and Dutta Chowdhury et al. (2021), the metrics do not produce equivalent results; the correlation between EV and GH is $\rho = 0.47$ in our case. Interestingly, there is a systematic decrease of the effect size with increasing distances, both for EV and GH. The trend is more evident for GH and applies both to X-WEAT and CA-WEAT. Figure 4 shows the trend for *-WEAT1. The correlation between GH and the effect size is in this case -0.29 and GH describes only a 10% of the variance. Similar results are obtained for EV, GH, X-WEAT2 and CA-WEAT2, as detailed in Appendix F.

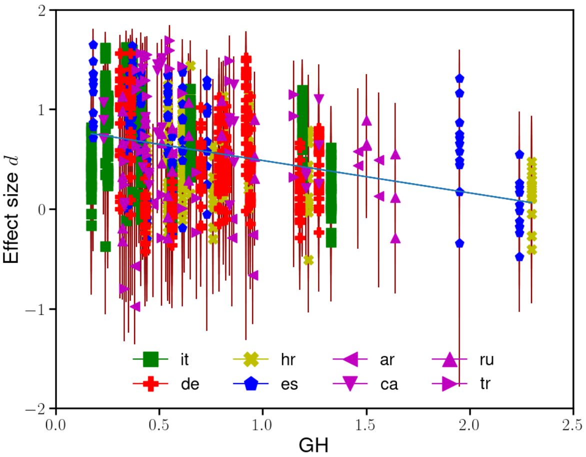


Figure 4: Effect size as a function of GH for X-/CA-WEAT1 in all the embedding models of Table 1.

Static embeddings in contextualised models. Previous work focused on simple English sentences. May et al. (2019) introduced the Sentence Encoder Association Test (SEAT). They generate templates such as "This is [TARGET]." or "[TARGET] are things.", were [TARGET] is substituted by words in WEAT tests. Kurita et al. (2019) also use templates ("[TARGET] is a [ATTRIBUTE]") for social biases and WEAT1 and compare the results from the standard WEAT measure and a new log probability bias score. Finally, Guo and Caliskan (2021) define yet another metric: the Contextualized Embedding Association Test (CEAT) which, instead of using a single template, collects sentences where WEAT terms appear in different contexts. The three methods have been evaluated on WEAT1 with BERT with May et al. (2019) ob-

[12] We use `https://github.com/cambridgeltl/iso-study` for EV and GH (Vulić et al., 2020) with the default most frequent 10k words for EV and 5k for GH.

taining an effect size of 0.22, Kurita et al. (2019) of 0.67 and Guo and Caliskan (2021) of 0.64.

When we move to the multilingual setting in our experiments, we do not consider sentences, but single words for which we extract the corresponding embedding. Words are built from the subunits in the language model vocabulary but, other than that, no context is considered. This allows a fair comparison across languages. As Figure 1 shows, static embeddings in contextual models show almost no bias and even negative effect sizes. The trend is specially confirmed for the languages for which more CA-WEAT lists are available (it, de, hr, es). Also notice that multilinguality further blurs the biases. In general, monolingual BERT models present less (desired) biases than *WP*, *WPali* and *CCWP* for all languages but more biases than the multilingual language models. This is not a consequence of a lower isomorphism between language subspaces, as Table 1 shows. We conjecture that building word vectors from subunits might have an impact in semantics at the word level. Differences observed on models coming from different layers are neither consistent nor significant.

6 Summary and Conclusions

Non-social human biases in embeddings are usually measured through WEAT association tests built in English, in the US. We hypothesise that this can be an issue when analysing embeddings in other languages or cultures. In order to address the question, we collected WEAT1 and WEAT2 lists written by natives of 9 languages, which we call cultural aware tests: CA-WEAT. We showed that different CA-WEATs produce a large variation in the biases and their effect size d. The values for WEAT and X-WEAT always lie within the CA-WEAT confidence intervals. This supports the idea that a single list (test) is not suitable for the analysis. Since we do not have a gold standard for the real bias we should expect in embeddings, one could argue that the *correct* bias is the one given by the WEAT test, as it has been carefully designed. However, this argument only holds for (American) English. For any other language, X-WEAT cannot assure the same properties as WEAT. CA-WEAT, the multilingual crowdsourced versions of WEAT, are the alternative to X-WEAT

We extend previous work to multilingual and language models taking this observation into account and perform in parallel the analysis for WEAT, X-WEAT and CA-WEAT. We confirm that monolingual static embeddings show signs of non-social human biases in all languages under study. When using the average CA-WEAT, the dispersion of d among languages for each model is smaller than the dispersion with X-WEAT. This is an indication that the average of several lists helps to get closer to the expected *universal* results across languages.

Multilinguality has the effect of mitigating biases. This is seen in static word embeddings but it is more evident in embeddings extracted from language models. On the other hand, word embeddings in language models already produce a huge mitigation with respect to their static counterparts, up to the point that effect sizes in multilingual language models can be negative. As a result, the trend can be inverse to the one observed in humans, being insects more pleasant than flowers in some languages.

Unexpectedly, the asymmetry between the amount of pleasant and unpleasant attributes in the training corpus is not responsible for the variance in the embeddings biases. Since CA-WEAT includes only frequent words in our training corpora, reliable representations are obtained, irrespective of any asymmetry. Differences in departures from isomorphism between languages in multilingual models describe up to a 10% of the variance. Even the trend is clear, this alone cannot explain the mitigation of the biases in either multilingual or contextualised models.

In the light of these outcomes, we expect to broaden the analysis to a more diverse set of languages extending the CA-WEAT tests, and design a fair multilingual setting for language models at sentence level.

Limitations

To the best of our knowledge, only 2 IAT tests are non-social, the other ones relate to other kinds of biases, such as gender or race, which are not pertinent for our study. We observe a large variability in the results between models but sometimes also between *-WEAT1 and *-WEAT2. More non-social psychologically-motivated IAT tests would be relevant to strengthen our conclusions.

When we defined CA-WEAT, we chose not to constrain the list of items and attributes to single words, multi-word terms were allowed for a wider coverage and also for a future reusability of the tests for sentence embeddings. As a result, some items in the lists can be out-of-vocabularies (OOV) in static word embeddings. This is relevant for Arabic, with 14 OOVs in both the *WP* embeddings and the *CCe* variants for X-WEAT1, and 6 and 7 OOVs for X-WEAT2 respectively. Turkish follows with 11 and 7 OOVs for X-WEAT1, and 5 and 3 for X-WEAT2. There are no OOVs in contextualised embeddings, where we sum the embeddings for all the subword units in a term.

Ethayarajh et al. (2019) showed that WEAT and the way how the effect size d is calculated causes a systematic overestimation of the biases. They also showed that in word embedding models that do some kind of matrix factorisation, such as skip-gram with negative sampling (it factorises a shifted word-context PMI matrix (Levy and Goldberg, 2014)) or GloVe (it factorises a logarithmic co-occurrence count matrix (Pennington et al., 2014)) having no bias is only possible if the positive (*pleasant* in our case) and negative (*unpleasant*) attributes occur with equal frequency in the corpus. Since the main motivation of our work is to study the effect of multilinguality and not the base models, we dodge the limitation for static embeddings by using CBOW, and we empirically show in Section 5 that the bias in our CBOW experiments does not correlate with the differences between pleasant and unpleasant counts in the corpus. However, the pre-trained *WP* and *WPali* used skip-gram and might be affected.

Acknowledgements

We thank the students of the Department of Interpreting and Translation at Università di Bologna, colleagues at Universität des Saarlandes, DFKI and several other locations around the world for producing the CA-WEAT dataset. CEB was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SFB 1102: Information Density and Linguistic Encoding.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Artin Arshamian, Richard C. Gerkin, Nicole Kruspe, Ewelina Wnuk, Simeon Floyd, Carolyn O'Meara, Gabriella Garrido Rodriguez, Johan N. Lundström, Joel D. Mainland, and Asifa Majid. 2022. The perception of odor pleasantness is shared across cultures. *Current Biology*, pages 1–6.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

William F. Battig and William Edward Montague. 1969. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of Experimental Psychology*, 80:1–46.

Francis Bellezza, Anthony Greenwald, and Mahzarin Banaji. 1986. Words high and low in pleasantness as rated by male and female college students. *Behavior Research Methods, Instruments, & Computers*, 18:299–303.

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

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. In *Advances in Neural Information Processing Systems*, volume 29, pages 4349–4357. 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.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *Practical ML for Developing Countries (PML4DC) at ICLR 2020.*

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7659–7666. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2021. Tracing source language interference in translation with graph-isomorphism measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 375–385, Held Online. INCOMA Ltd.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of personality and social psychology*, 74(6):1464–1480.

Wei Guo and Aylin Caliskan. 2021. *Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases*, page 122–133. Association for Computing Machinery, New York, NY, USA.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Anne Lauscher and Goran Glavaš. 2019. Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Anne Lauscher, Goran Glavas, Simone Paolo Ponzetto, and Ivan Vulic. 2020a. A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8131–8138. AAAI Press.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020b. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199, Barcelona, Spain (Online). Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot Learning with Multilingual Language Models. *arXiv preprint arXiv:2112.10668*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

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.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research (JAIR)*, 65(1):569–630.

Shlomo S. Sawilowsky. 2009. New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, 8:597–599.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Stefan Schweter. 2020a. BERTurk – BERT models for Turkish.

Stefan Schweter. 2020b. Italian BERT and ELECTRA models. *Zenodo*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Autumn Toney and Aylin Caliskan. 2021. ValNorm quantifies semantics to reveal consistent valence biases across languages and over centuries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7203–7218, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

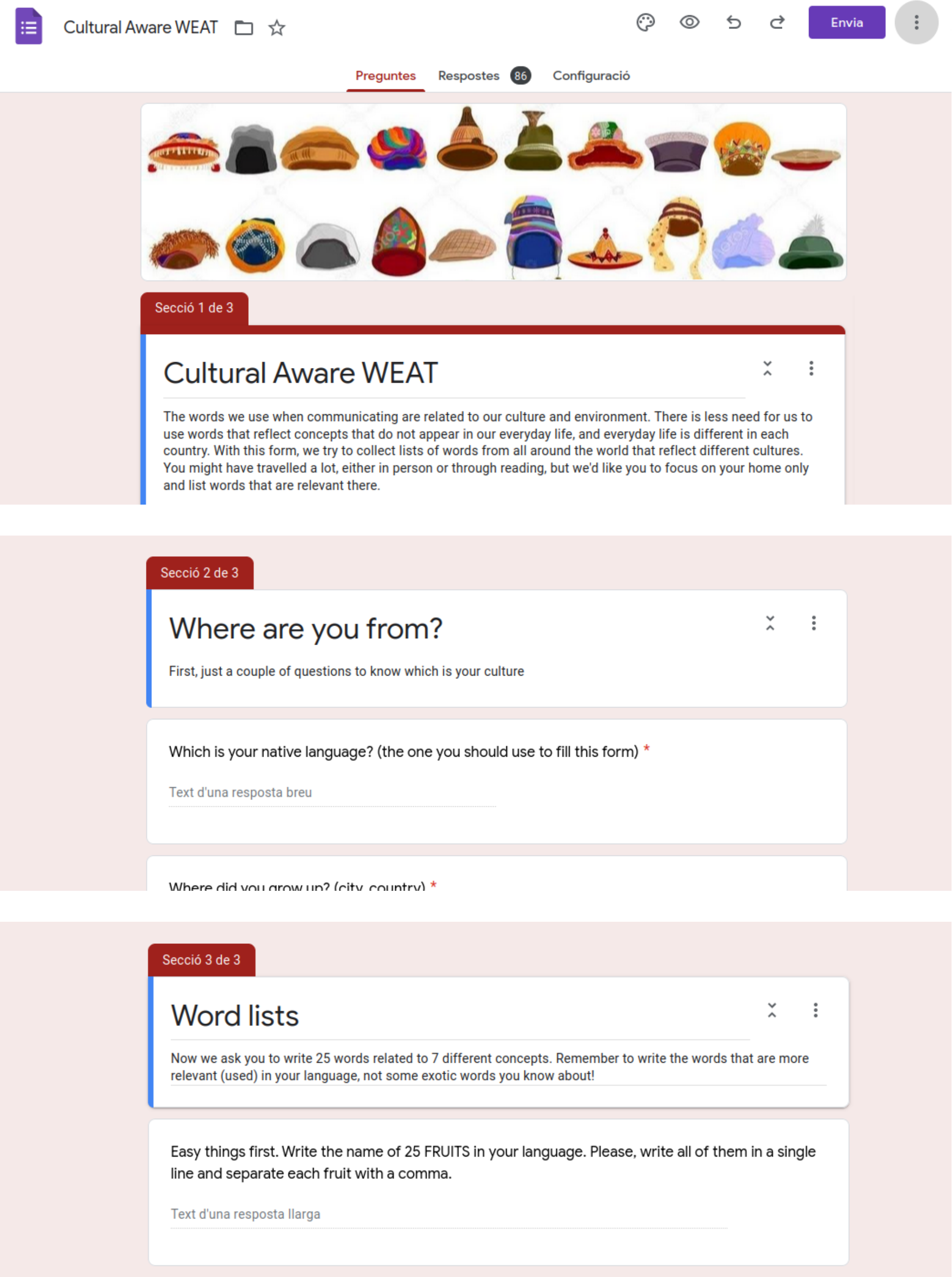
Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Yi Zhou, Masahiro Kaneko, and Danushka Bollegala. 2022. Sense Embeddings are also Biased–Evaluating Social Biases in Static and Contextualised Sense Embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages –. Association for Computational Linguistics.

A CA-WEAT Data Collection



Cultural Aware WEAT

The words we use when communicating are related to our culture and environment. There is less need for us to use words that reflect concepts that do not appear in our everyday life, and everyday life is different in each country. With this form, we try to collect lists of words from all around the world that reflect different cultures. You might have travelled a lot, either in person or through reading, but we'd like you to focus on your home only and list words that are relevant there.

Where are you from?

First, just a couple of questions to know which is your culture

Which is your native language? (the one you should use to fill this form) *

Text d'una resposta breu

Where did you grow up? (city, country) *

Word lists

Now we ask you to write 25 words related to 7 different concepts. Remember to write the words that are more relevant (used) in your language, not some exotic words you know about!

Easy things first. Write the name of 25 FRUITS in your language. Please, write all of them in a single line and separate each fruit with a comma.

Text d'una resposta llarga

Figure 5: The CA-WEAT lists were obtained thorough an online form. In the first section, the interface presents the guidelines for the task. The second part collects a minimal amount of personal information such as birth place and native language. Finally, the third part collects the list of words per concept. The form was created in English, Spanish, Catalan, French, Italian and German for a greater reach.

B The CA-WEAT Dataset

Language	ISO 639-1	# CA-WEATs	# X-WEATs
Arabic	ar	1	1
Bengali	bn	1	0
Bulgarian	bg	1	0
Catalan	ca	2	1
Chinese	zh	2	0
Croatian	hr	12	1
Dutch	nl	2	0
English	en	5	1
Farsi	fa	2	0
French	fr	1	0
German	de	24	1
Greek	el	3	0
Indonesian	id	1	0
Italian	it	24	1
Korean	ko	1	0
Luxembourgish	lb	1	0
Marathi	mr	1	0
Norwegian	no	1	0
Polish	po	1	0
Portuguese	pt	1	0
Romanian	ro	1	0
Russian	ru	2	1
Spanish	es	10	1
Turkish	tr	2	1
Ukrainian	uk	1	0
Vietnamese	vi	1	0

Table 2: Number of lists per language in the CA-WEAT.v1 data set. We include the languages with an available X-WEAT translation, which coincide with the languages used in the current work.

C WEAT1 and WEAT2 Original Lists

WEAT1 target items	
Flowers	aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia
Insects	ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil
WEAT2 target items	
Musical instruments	bagpipe, cello, guitar, lute, trombone, banjo, clarinet, harmonica, mandolin, trumpet, bassoon, drum, harp, oboe, tuba, bell, fiddle, harpsichord, piano, viola, bongo, flute, horn, saxophone, violin
Weapons	arrow, club, gun, missile, spear, axe, dagger, harpoon, pistol, sword, blade, dynamite, hatchet, rifle, tank, bomb, firearm, knife, shotgun, teargas, cannon, grenade, mace, slingshot, whip
WEAT1 and WEAT2 attributes	
Pleasant	caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation
Unpleasant	abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison

Table 3: List of words in the original English WEAT1 and WEAT2 tests.

D The CCe Corpus

	CC-100		CCe		
	lines	words	lines	words	tokens
ar	111,159,553	2,869,491,129	68,000,000	1,783,110,513	1,990,019,767
ca	77,707,813	1,752,301,188	70,697,475	1,752,301,188	2,073,242,760
de	348,047,236	10,297,244,661	62,000,000	1,784,586,244	2,077,747,370
en	1,857,736,518	55,607,824,084	67,000,000	1,792,874,585	2,087,270,544
es	318,730,600	9,374,385,063	65,000,000	1,784,255,267	2,024,552,573
hr	127,087,082	3,296,927,157	67,000,000	1,787,595,251	2,068,302,605
it	154,163,562	4,982,929,393	58,000,000	1,797,623,314	2,099,653,787
tr	109,279,716	2,736,027,827	70,000,000	1,751,214,765	2,151,242,947
ru	725,664,405	23,408,093,897	58,000,000	1,785,123,381	2,176,025,999

Table 4: Number of lines, words and tokens resulting after pre-processing for the CC-100 and CCe, the subset used to build our in-house embeddings (cf. Section 4).

	cFlowers		cInsects		cInstruments		cWeapons		cPleasant		cUnpleasant	
	X-	CA-WEAT	X-	CA-WEAT	X-	CA-WEAT	X-	CA-WEAT	X-	CA-WEAT	X-	CA-WEAT
en	0.13	0.12 ± 0.04	0.15	0.10 ± 0.05	0.17	0.21 ± 0.12	0.56	0.52 ± 0.29	5.32	4.94 ± 0.74	1.37	1.30 ± 0.39
ar	0.04	0.17 ± 0.00	0.02	0.03 ± 0.00	0.05	0.36 ± 0.00	0.36	0.19 ± 0.00	1.40	1.08 ± 0.00	0.68	1.00 ± 0.00
ca	0.35	0.30 ± 0.00	0.08	0.10 ± 0.03	0.22	0.54 ± 0.00	0.14	0.13 ± 0.03	2.92	3.92 ± 0.33	1.20	1.73 ± 0.38
de	0.04	0.03 ± 0.00	0.12	0.03 ± 0.01	0.11	0.10 ± 0.01	0.28	0.13 ± 0.10	2.76	3.35 ± 1.18	0.73	1.32 ± 1.32
es	0.20	0.14 ± 0.05	0.04	0.05 ± 0.00	0.10	0.64 ± 0.35	0.22	0.25 ± 0.12	3.65	3.00 ± 1.34	1.59	1.17 ± 0.66
hr	0.08	0.08 ± 0.01	0.06	0.05 ± 0.01	0.08	0.10 ± 0.06	0.53	0.18 ± 0.07	1.96	1.66 ± 0.42	0.76	0.60 ± 0.42
it	0.22	0.23 ± 0.02	0.08	0.11 ± 0.01	0.18	0.52 ± 0.22	0.43	0.22 ± 0.05	3.00	3.16 ± 1.09	1.18	1.11 ± 0.59
ru	0.01	0.07 ± 0.08	0.02	0.06 ± 0.06	0.03	0.56 ± 0.59	0.14	0.29 ± 0.23	1.69	1.16 ± 0.17	0.35	0.45 ± 0.05
tr	0.23	0.34 ± 0.06	0.08	0.43 ± 0.39	0.13	0.26 ± 0.15	0.51	0.56 ± 0.12	3.56	4.64 ± 0.22	0.97	1.41 ± 0.35

Table 5: Billions of words (counts) in the CCe corpus belonging to the X-/CA-WEAT1 and X-/CA-WEAT2 tests.

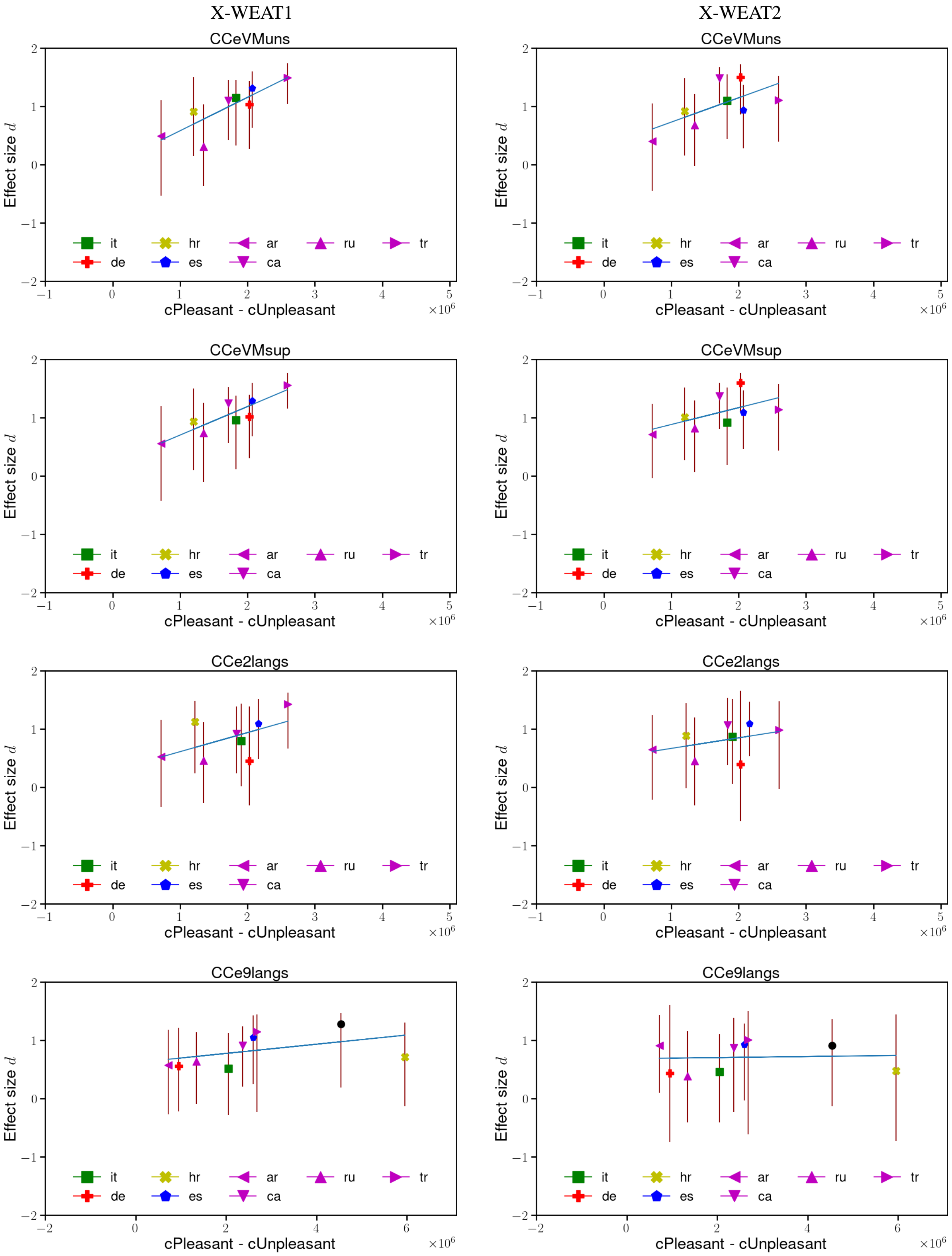


Figure 6: Effect sizes obtained with the X-WEAT tests in the *CCeVMsup*, *CCeVMuns*, *CCe2langs* and *CCe9langs* monolingual embedding models as a function of the difference in the number of pleasant and unpleasant attributes in the training corpus CCe.

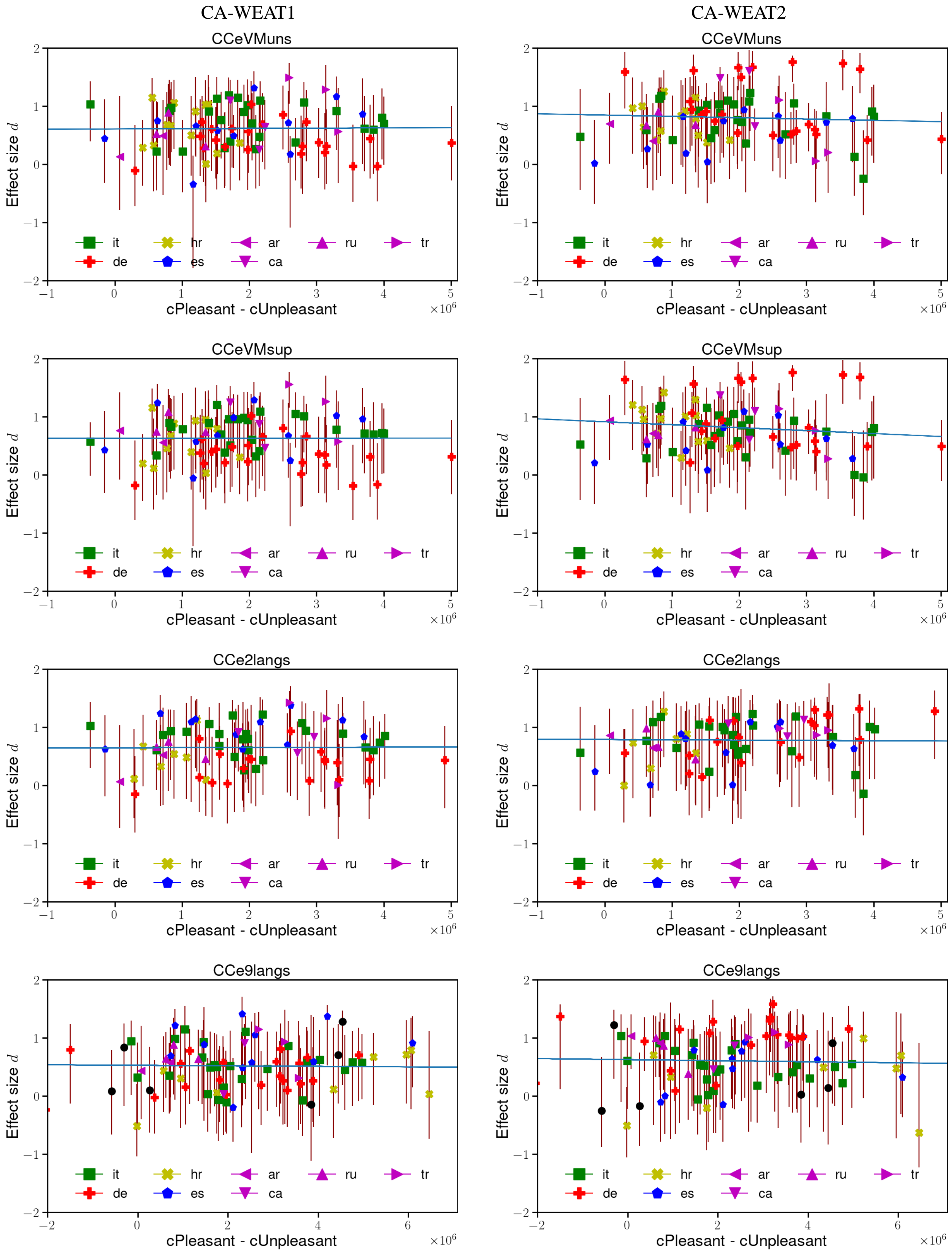


Figure 7: Effect sizes obtained with the CA-WEAT tests in the *CCeVMsup*, *CCeVMuns*, *CCe2langs* and *CCe9langs* monolingual embedding models as a function of the difference in the number of pleasant and unpleasant attributes in the training corpus CCe.

E Bias Statistic and Effect Size

	en$_{or}$	ar	ca	de	es	hr	it	ru	tr
WEAT 1: Flowers and insects									
WP	$1.5^{+0.6}_{-0.6}$	$0.0^{+0.4}_{-0.4}$	$1.1^{+0.6}_{-0.6}$	$1.0^{+0.7}_{-0.7}$	$1.3^{+0.7}_{-0.8}$	$0.8^{+0.5}_{-0.5}$	$0.9^{+0.6}_{-0.6}$	$0.6^{+0.6}_{-0.6}$	$0.7^{+0.6}_{-0.6}$
WPali	$1.5^{+0.6}_{-0.6}$	$-0.0^{+0.4}_{-0.4}$	$1.4^{+0.7}_{-0.7}$	$1.0^{+0.8}_{-0.8}$	$1.6^{+0.9}_{-0.9}$	$0.7^{+0.5}_{-0.5}$	$1.2^{+0.7}_{-0.7}$	$0.8^{+0.6}_{-0.6}$	$0.7^{+0.6}_{-0.6}$
CCWP	$1.5^{+0.8}_{-0.7}$	$0.5^{+0.5}_{-0.5}$	$1.0^{+0.5}_{-0.5}$	$0.6^{+0.4}_{-0.4}$	$1.2^{+0.7}_{-0.7}$	$0.8^{+0.5}_{-0.6}$	$0.8^{+0.7}_{-0.6}$	$0.9^{+0.7}_{-0.7}$	$1.5^{+0.7}_{-0.7}$
CCe	$1.7^{+0.8}_{-0.8}$	$0.5^{+0.6}_{-0.6}$	$1.0^{+0.7}_{-0.7}$	$0.7^{+0.5}_{-0.5}$	$1.5^{+0.7}_{-0.8}$	$1.2^{+1.0}_{-1.1}$	$0.8^{+0.7}_{-0.7}$	$0.4^{+0.6}_{-0.8}$	$1.7^{+0.6}_{-0.6}$
CCeVMuns	–	$0.5^{+1.3}_{-1.2}$	$1.8^{+1.2}_{-1.2}$	$1.4^{+1.0}_{-1.0}$	$2.1^{+1.1}_{-1.2}$	$1.8^{+1.4}_{-1.6}$	$1.5^{+1.1}_{-1.2}$	$0.4^{+0.9}_{-1.0}$	$2.9^{+1.1}_{-1.0}$
CCeVMsup	–	$0.6^{+1.0}_{-1.0}$	$1.7^{+1.0}_{-1.0}$	$1.4^{+1.0}_{-0.9}$	$2.2^{+1.1}_{-1.1}$	$1.4^{+1.1}_{-1.3}$	$1.3^{+1.1}_{-1.2}$	$0.7^{+0.8}_{-0.8}$	$2.9^{+1.0}_{-1.0}$
CCe2langs	–	$0.3^{+0.5}_{-0.4}$	$0.9^{+0.9}_{-0.7}$	$0.5^{+0.8}_{-1.0}$	$1.3^{+1.0}_{-0.8}$	$1.2^{+0.8}_{-0.8}$	$0.6^{+0.5}_{-0.6}$	$0.3^{+0.5}_{-0.6}$	$1.2^{+0.6}_{-0.6}$
CCe9langs	$0.7^{+0.7}_{-0.6}$	$0.2^{+0.3}_{-0.3}$	$0.6^{+0.6}_{-0.5}$	$0.3^{+0.5}_{-0.6}$	$0.7^{+0.6}_{-0.5}$	$0.5^{+0.6}_{-0.7}$	$0.2^{+0.4}_{-0.4}$	$0.2^{+0.2}_{-0.3}$	$0.6^{+0.7}_{-0.8}$
BERT$_0$	$0.2^{+0.4}_{-0.4}$	$0.2^{+0.4}_{-0.5}$	–	$0.2^{+0.3}_{-0.3}$	$0.3^{+0.3}_{-0.3}$	–	$0.2^{+0.2}_{-0.2}$	–	$0.4^{+0.4}_{-0.3}$
BERT$_{11}$	$-0.0^{+0.2}_{-0.2}$	$0.4^{+0.4}_{-0.4}$	–	$0.1^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	–	$0.0^{+0.3}_{-0.3}$	–	$0.4^{+0.4}_{-0.4}$
mBERT$_0$	$0.1^{+0.4}_{-0.4}$	$-0.0^{+0.4}_{-0.4}$	$0.0^{+0.4}_{-0.4}$	$0.0^{+0.4}_{-0.4}$	$0.0^{+0.3}_{-0.3}$	$-0.0^{+0.4}_{-0.4}$	$0.2^{+0.4}_{-0.3}$	$-0.0^{+0.3}_{-0.3}$	$0.0^{+0.4}_{-0.4}$
mBERT$_{11}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.1}$	$0.1^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.1}$
XLM-R$_0$	$0.0^{+0.4}_{-0.4}$	$-0.2^{+0.4}_{-0.5}$	$0.2^{+0.5}_{-0.5}$	$-0.1^{+0.4}_{-0.5}$	$0.2^{+0.4}_{-0.4}$	$0.2^{+0.4}_{-0.4}$	$0.1^{+0.4}_{-0.4}$	$-0.0^{+0.2}_{-0.2}$	$0.2^{+0.3}_{-0.3}$
XLM-R$_{11}$	$0.0^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.2}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$
XGLM$_0$	$0.1^{+0.2}_{-0.2}$	$-0.1^{+0.3}_{-0.3}$	$0.1^{+0.2}_{-0.2}$	$0.1^{+0.2}_{-0.2}$	$0.1^{+0.2}_{-0.2}$	$0.1^{+0.2}_{-0.2}$	$0.1^{+0.2}_{-0.1}$	$-0.1^{+0.2}_{-0.2}$	$0.0^{+0.2}_{-0.2}$
XGLM$_{47}$	$0.0^{+0.1}_{-0.1}$	$-0.1^{+0.3}_{-0.4}$	$0.1^{+0.2}_{-0.2}$	$-0.1^{+0.2}_{-0.4}$	$0.1^{+0.2}_{-0.1}$	$0.0^{+0.2}_{-0.1}$	$0.1^{+0.1}_{-0.1}$	$0.2^{+1.0}_{-0.8}$	$-0.0^{+0.5}_{-0.5}$
WEAT 2: Instruments and weapons									
WP	$1.8^{+0.8}_{-0.8}$	$0.6^{+0.5}_{-0.5}$	$1.5^{+0.8}_{-0.8}$	$1.1^{+0.8}_{-0.8}$	$1.3^{+0.7}_{-0.8}$	$0.8^{+0.8}_{-0.8}$	$1.0^{+0.7}_{-0.7}$	$0.8^{+0.7}_{-0.9}$	$0.6^{+0.6}_{-0.7}$
WPali	$1.8^{+0.8}_{-0.8}$	$0.3^{+0.5}_{-0.4}$	$1.7^{+0.9}_{-0.9}$	$0.9^{+0.8}_{-0.8}$	$1.7^{+0.8}_{-0.9}$	$0.5^{+0.7}_{-0.7}$	$1.0^{+0.8}_{-0.8}$	$0.9^{+0.8}_{-0.9}$	$0.3^{+0.6}_{-0.7}$
CCWP	$1.9^{+0.7}_{-0.7}$	$0.3^{+0.5}_{-0.5}$	$0.9^{+0.7}_{-0.7}$	$0.6^{+0.5}_{-0.5}$	$0.7^{+0.7}_{-0.6}$	$1.1^{+0.7}_{-0.7}$	$1.0^{+0.6}_{-0.6}$	$0.4^{+0.6}_{-0.7}$	$0.9^{+0.5}_{-0.5}$
CCe	$2.0^{+0.9}_{-0.8}$	$0.4^{+0.7}_{-0.6}$	$1.5^{+0.7}_{-0.7}$	$1.3^{+0.7}_{-0.8}$	$1.0^{+0.7}_{-0.8}$	$1.2^{+0.9}_{-1.3}$	$0.8^{+0.7}_{-0.6}$	$0.5^{+0.8}_{-0.9}$	$1.0^{+0.7}_{-0.7}$
CCeVMuns	–	$0.5^{+1.2}_{-1.0}$	$2.7^{+1.3}_{-1.2}$	$2.7^{+1.2}_{-1.3}$	$1.4^{+1.0}_{-1.0}$	$2.1^{+1.5}_{-1.9}$	$1.5^{+0.9}_{-0.9}$	$0.9^{+0.9}_{-1.0}$	$1.8^{+1.2}_{-1.1}$
CCeVMsup	–	$0.7^{+0.8}_{-0.7}$	$2.2^{+1.2}_{-1.1}$	$2.9^{+1.2}_{-1.2}$	$1.6^{+1.0}_{-1.0}$	$2.0^{+1.4}_{-1.5}$	$1.3^{+1.0}_{-1.0}$	$1.0^{+0.9}_{-1.0}$	$1.8^{+1.0}_{-1.1}$
CCe2langs	–	$0.4^{+0.5}_{-0.4}$	$1.2^{+0.8}_{-0.7}$	$0.4^{+1.2}_{-1.6}$	$1.3^{+1.0}_{-0.8}$	$1.0^{+0.9}_{-1.0}$	$0.6^{+0.5}_{-0.5}$	$0.4^{+0.6}_{-0.7}$	$0.8^{+0.8}_{-0.8}$
CCe9langs	$0.5^{+0.7}_{-0.6}$	$0.4^{+0.4}_{-0.3}$	$0.5^{+0.6}_{-0.7}$	$0.2^{+1.0}_{-1.2}$	$0.5^{+0.6}_{-0.5}$	$0.3^{+1.0}_{-1.1}$	$0.2^{+0.4}_{-0.4}$	$0.2^{+0.4}_{-0.4}$	$0.5^{+1.0}_{-1.0}$
BERT$_0$	$0.8^{+0.5}_{-0.5}$	$0.1^{+0.4}_{-0.3}$	–	$0.4^{+0.4}_{-0.4}$	$0.3^{+0.4}_{-0.4}$	–	$0.3^{+0.4}_{-0.4}$	–	$0.2^{+0.4}_{-0.4}$
BERT$_{11}$	$0.4^{+0.3}_{-0.2}$	$0.1^{+0.3}_{-0.3}$	–	$0.2^{+0.2}_{-0.1}$	$-0.0^{+0.1}_{-0.2}$	–	$0.0^{+0.3}_{-0.3}$	–	$0.4^{+0.5}_{-0.4}$
mBERT$_0$	$0.1^{+0.4}_{-0.5}$	$-0.0^{+0.4}_{-0.4}$	$-0.0^{+0.4}_{-0.4}$	$0.1^{+0.4}_{-0.4}$	$0.1^{+0.5}_{-0.4}$	$0.0^{+0.4}_{-0.4}$	$0.2^{+0.4}_{-0.5}$	$-0.2^{+0.4}_{-0.4}$	$-0.2^{+0.5}_{-0.4}$
mBERT$_{11}$	$0.1^{+0.2}_{-0.2}$	$-0.0^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.2}$	$-0.0^{+0.1}_{-0.2}$	$-0.0^{+0.1}_{-0.2}$	$0.0^{+0.1}_{-0.1}$	$-0.0^{+0.2}_{-0.2}$	$-0.0^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.1}$
XLM-R$_0$	$-0.2^{+0.5}_{-0.5}$	$0.2^{+0.5}_{-0.4}$	$-0.1^{+0.4}_{-0.5}$	$-0.0^{+0.4}_{-0.5}$	$0.1^{+0.3}_{-0.4}$	$-0.1^{+0.4}_{-0.4}$	$-0.2^{+0.5}_{-0.5}$	$-0.0^{+0.3}_{-0.4}$	$-0.3^{+0.6}_{-0.6}$
XLM-R$_{11}$	$0.0^{+0.2}_{-0.2}$	$0.1^{+0.2}_{-0.1}$	$-0.0^{+0.1}_{-0.2}$	$-0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.1}$	$-0.1^{+0.2}_{-0.2}$	$-0.0^{+0.2}_{-0.2}$
XGLM$_0$	$0.1^{+0.2}_{-0.2}$	$-0.0^{+0.3}_{-0.3}$	$0.1^{+0.2}_{-0.2}$	$0.0^{+0.2}_{-0.2}$	$0.0^{+0.2}_{-0.2}$	$-0.2^{+0.4}_{-0.6}$	$0.0^{+0.2}_{-0.2}$	$-0.0^{+0.2}_{-0.3}$	$0.0^{+0.2}_{-0.2}$
XGLM$_{47}$	$0.1^{+0.2}_{-0.1}$	$0.0^{+0.2}_{-0.2}$	$0.1^{+0.2}_{-0.1}$	$0.1^{+0.2}_{-0.2}$	$0.0^{+0.1}_{-0.1}$	$-0.1^{+0.3}_{-0.4}$	$-0.0^{+0.2}_{-0.1}$	$-0.3^{+1.3}_{-1.6}$	$-0.1^{+0.8}_{-0.7}$

Table 6: Statistic for X-WEAT tests for 8 languages plus the original English WEAT (en$_{or}$) test. 95% confidence intervals obtained by bootstrap resampling of the lists.

	en_{or}	$\text{en}_{(5)}$	$\text{ar}_{(1)}$	$\text{ca}_{(2)}$	$\text{de}_{(24)}$	$\text{es}_{(10)}$	$\text{hr}_{(12)}$	$\text{it}_{(24)}$	$\text{ru}_{(2)}$	$\text{tr}_{(2)}$
WEAT 1: Flowers and insects										
WP	1.5	$0.8^{+0.1}_{-0.2}$	0.0	$0.8^{+0.4}_{-0.4}$	$0.7^{+0.5}_{-0.7}$	$0.8^{+0.6}_{-0.4}$	$0.3^{+0.3}_{-0.3}$	$0.7^{+0.5}_{-0.5}$	$0.9^{+0.0}_{-0.0}$	$0.5^{+0.2}_{-0.2}$
WPali	1.5	$0.8^{+0.1}_{-0.2}$	0.0	$1.0^{+0.4}_{-0.4}$	$0.8^{+0.5}_{-0.7}$	$1.0^{+0.8}_{-0.4}$	$0.3^{+0.3}_{-0.3}$	$1.1^{+0.5}_{-0.8}$	$1.0^{+0.1}_{-0.1}$	$0.5^{+0.3}_{-0.3}$
CCWP	1.5	$1.0^{+0.1}_{-0.7}$	0.4	$0.7^{+0.4}_{-0.4}$	$0.3^{+0.3}_{-0.2}$	$0.8^{+0.5}_{-0.3}$	$0.5^{+0.4}_{-0.3}$	$0.9^{+0.4}_{-0.8}$	$0.4^{+0.2}_{-0.2}$	$1.2^{+0.3}_{-0.3}$
CCe	1.7	$0.7^{+0.5}_{-0.3}$	0.2	$0.7^{+0.2}_{-0.2}$	$0.4^{+0.5}_{-0.4}$	$1.1^{+0.2}_{-0.6}$	$0.7^{+0.4}_{-0.6}$	$0.6^{+0.4}_{-0.4}$	$0.7^{+0.1}_{-0.1}$	$0.9^{+0.1}_{-0.1}$
CCeVMuns	–	–	0.2	$1.4^{+0.4}_{-0.4}$	$0.8^{+1.2}_{-0.9}$	$1.7^{+0.6}_{-1.0}$	$1.3^{+0.9}_{-1.0}$	$1.1^{+0.6}_{-0.8}$	$0.9^{+0.2}_{-0.2}$	$1.4^{+0.3}_{-0.3}$
CCeVMsup	–	–	0.9	$1.0^{+0.4}_{-0.4}$	$0.6^{+1.1}_{-1.0}$	$1.6^{+0.5}_{-0.8}$	$1.0^{+0.6}_{-0.9}$	$1.0^{+0.7}_{-0.5}$	$1.2^{+0.2}_{-0.2}$	$1.2^{+0.3}_{-0.3}$
CCe2langs	–	–	0.0	$0.5^{+0.1}_{-0.1}$	$0.4^{+0.4}_{-0.3}$	$0.7^{+0.5}_{-0.3}$	$0.6^{+0.4}_{-0.4}$	$0.4^{+0.3}_{-0.2}$	$0.5^{+0.0}_{-0.0}$	$0.4^{+0.3}_{-0.3}$
CCe9langs	0.7	$0.0^{+0.3}_{-0.1}$	0.2	$0.3^{+0.2}_{-0.2}$	$0.3^{+0.2}_{-0.3}$	$0.5^{+0.7}_{-0.5}$	$0.2^{+0.3}_{-0.5}$	$0.2^{+0.1}_{-0.3}$	$0.3^{+0.0}_{-0.0}$	$0.3^{+0.0}_{-0.0}$
BERT_0	0.2	$0.1^{+0.2}_{-0.1}$	0.2	–	$0.1^{+0.1}_{-0.2}$	$0.2^{+0.1}_{-0.2}$	–	$0.1^{+0.2}_{-0.1}$	–	$0.0^{+0.1}_{-0.1}$
BERT_{11}	-0.0	$0.1^{+0.1}_{-0.2}$	0.2	–	$0.0^{+0.0}_{-0.1}$	$0.0^{+0.0}_{-0.1}$	–	$0.0^{+0.1}_{-0.1}$	–	$0.1^{+0.2}_{-0.2}$
mBERT_0	0.1	$-0.1^{+0.1}_{-0.1}$	-0.2	$0.1^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.2}$	$0.1^{+0.1}_{-0.2}$	$0.1^{+0.2}_{-0.2}$	$0.2^{+0.3}_{-0.2}$	$-0.0^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.1}$
mBERT_{11}	0.0	$-0.0^{+0.1}_{-0.0}$	0.0	$0.0^{+0.0}_{-0.0}$	$-0.0^{+0.1}_{-0.0}$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.1}_{-0.0}$	$-0.0^{+0.0}_{-0.0}$	$-0.1^{+0.0}_{-0.0}$
XLM-R_0	0.0	$-0.0^{+0.3}_{-0.5}$	-0.4	$0.2^{+0.0}_{-0.0}$	$0.0^{+0.2}_{-0.2}$	$0.0^{+0.3}_{-0.2}$	$0.1^{+0.2}_{-0.2}$	$0.0^{+0.2}_{-0.2}$	$0.0^{+0.1}_{-0.1}$	$-0.2^{+0.3}_{-0.3}$
XLM-R_{11}	0.0	$-0.0^{+0.0}_{-0.1}$	-0.1	$0.1^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.0}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$-0.0^{+0.0}_{-0.0}$	$-0.1^{+0.1}_{-0.1}$
XGLM_0	0.1	$0.1^{+0.1}_{-0.1}$	-0.3	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.1}_{-0.0}$	$-0.0^{+0.1}_{-0.0}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$0.1^{+0.1}_{-0.1}$	$0.0^{+0.0}_{-0.0}$
XGLM_{47}	0.0	$0.0^{+0.0}_{-0.1}$	-0.2	$0.1^{+0.0}_{-0.0}$	$0.0^{+0.0}_{-0.1}$	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.1}_{-0.1}$	$0.1^{+0.0}_{-0.1}$	$0.3^{+0.3}_{-0.3}$	$-0.1^{+0.1}_{-0.1}$
WEAT 2: Instruments and weapons										
WP	1.8	$0.9^{+0.3}_{-0.4}$	0.7	$1.1^{+0.2}_{-0.2}$	$1.1^{+0.8}_{-0.6}$	$1.3^{+0.5}_{-0.6}$	$0.3^{+0.4}_{-0.3}$	$0.9^{+0.4}_{-0.4}$	$1.1^{+0.2}_{-0.2}$	$0.9^{+0.1}_{-0.1}$
WPali	1.8	$0.9^{+0.3}_{-0.4}$	0.4	$1.1^{+0.2}_{-0.2}$	$1.2^{+0.7}_{-0.8}$	$1.6^{+0.6}_{-0.6}$	$0.3^{+0.4}_{-0.3}$	$1.2^{+0.5}_{-0.7}$	$1.1^{+0.0}_{-0.0}$	$0.6^{+0.1}_{-0.1}$
CCWP	1.9	$1.0^{+0.4}_{-0.4}$	0.7	$0.8^{+0.1}_{-0.1}$	$0.4^{+0.5}_{-0.2}$	$1.1^{+0.2}_{-0.7}$	$0.9^{+0.1}_{-0.4}$	$0.7^{+0.4}_{-0.4}$	$0.9^{+0.1}_{-0.1}$	$0.9^{+0.0}_{-0.0}$
CCe	2.0	$0.8^{+0.9}_{-0.3}$	0.8	$1.3^{+0.2}_{-0.2}$	$1.1^{+0.6}_{-0.4}$	$1.0^{+0.3}_{-0.8}$	$0.7^{+0.5}_{-0.3}$	$0.6^{+0.4}_{-0.6}$	$1.1^{+0.4}_{-0.4}$	$1.0^{+0.1}_{-0.1}$
CCeVMuns	–	–	1.2	$2.6^{+0.4}_{-0.4}$	$2.5^{+1.3}_{-0.9}$	$1.3^{+0.7}_{-1.3}$	$1.5^{+0.7}_{-0.8}$	$1.2^{+0.5}_{-1.0}$	$1.9^{+0.8}_{-0.8}$	$1.7^{+0.1}_{-0.1}$
CCeVMsup	–	–	1.2	$2.1^{+0.4}_{-0.4}$	$2.4^{+1.3}_{-0.6}$	$1.6^{+0.6}_{-1.3}$	$1.6^{+0.9}_{-0.8}$	$1.1^{+0.6}_{-1.0}$	$1.6^{+0.8}_{-0.8}$	$1.6^{+0.2}_{-0.2}$
CCe2langs	–	–	0.7	$0.8^{+0.1}_{-0.1}$	$0.9^{+0.9}_{-0.6}$	$0.7^{+0.1}_{-0.7}$	$0.5^{+0.4}_{-0.5}$	$0.5^{+0.2}_{-0.4}$	$0.7^{+0.3}_{-0.3}$	$1.1^{+0.2}_{-0.2}$
CCe9langs	0.5	$0.0^{+0.5}_{-0.2}$	0.6	$0.2^{+0.0}_{-0.0}$	$0.7^{+0.7}_{-0.6}$	$0.3^{+0.2}_{-0.4}$	$0.2^{+0.3}_{-0.8}$	$0.3^{+0.1}_{-0.2}$	$0.5^{+0.1}_{-0.1}$	$0.7^{+0.1}_{-0.1}$
BERT_0	0.8	$0.3^{+0.4}_{-0.0}$	0.2	–	$0.3^{+0.2}_{-0.3}$	$0.3^{+0.2}_{-0.2}$	–	$0.1^{+0.2}_{-0.3}$	–	$0.0^{+0.0}_{-0.0}$
BERT_{11}	0.4	$0.2^{+0.1}_{-0.0}$	0.3	–	$0.2^{+0.2}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	–	$0.0^{+0.1}_{-0.2}$	–	$0.1^{+0.0}_{-0.0}$
mBERT_0	0.1	$0.4^{+0.4}_{-0.3}$	0.1	$-0.0^{+0.0}_{-0.0}$	$0.2^{+0.2}_{-0.2}$	$0.2^{+0.1}_{-0.2}$	$0.0^{+0.4}_{-0.2}$	$0.3^{+0.3}_{-0.2}$	$0.2^{+0.2}_{-0.2}$	$0.2^{+0.0}_{-0.0}$
mBERT_{11}	0.1	$0.1^{+0.1}_{-0.1}$	0.0	$-0.0^{+0.0}_{-0.0}$	$0.1^{+0.2}_{-0.1}$	$0.1^{+0.1}_{-0.1}$	$0.1^{+0.1}_{-0.1}$	$0.1^{+0.2}_{-0.2}$	$0.1^{+0.0}_{-0.0}$	$0.1^{+0.0}_{-0.0}$
XLM-R_0	-0.2	$-0.1^{+0.3}_{-0.1}$	-0.1	$0.2^{+0.1}_{-0.1}$	$0.0^{+0.2}_{-0.1}$	$0.2^{+0.4}_{-0.4}$	$-0.1^{+0.2}_{-0.1}$	$0.0^{+0.3}_{-0.2}$	$0.1^{+0.2}_{-0.2}$	$-0.0^{+0.1}_{-0.1}$
XLM-R_{11}	0.0	$0.0^{+0.1}_{-0.0}$	0.0	$0.1^{+0.2}_{-0.2}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.1}_{-0.0}$	$0.0^{+0.1}_{-0.1}$	$-0.0^{+0.1}_{-0.1}$	$-0.0^{+0.0}_{-0.0}$
XGLM_0	0.1	$0.1^{+0.1}_{-0.1}$	0.1	$0.1^{+0.0}_{-0.0}$	$0.1^{+0.1}_{-0.1}$	$0.1^{+0.1}_{-0.1}$	$-0.1^{+0.2}_{-0.1}$	$-0.0^{+0.1}_{-0.1}$	$0.2^{+0.1}_{-0.1}$	$0.1^{+0.0}_{-0.0}$
XGLM_{47}	0.1	$0.0^{+0.1}_{-0.1}$	0.1	$0.0^{+0.0}_{-0.0}$	$0.0^{+0.1}_{-0.1}$	$0.0^{+0.0}_{-0.0}$	$-0.1^{+0.1}_{-0.1}$	$-0.0^{+0.0}_{-0.0}$	$0.4^{+0.2}_{-0.2}$	$-0.4^{+0.3}_{-0.3}$

Table 7: Statistic for CA-WEAT tests with lists created for 9 languages; the original English WEAT (en_{or}) test is shown for comparison. The number of lists per language is shown as subindex of the language. We report the median and 95% confidence intervals.

	en$_{or}$	ar	ca	de	es	hr	it	ru	tr
WEAT 1: Flowers and insects									
WP	$1.7^{+0.1}_{-0.4}$	$0.1^{+0.8}_{-0.9}$	$1.4^{+0.2}_{-0.6}$	$1.3^{+0.3}_{-0.8}$	$1.5^{+0.2}_{-0.8}$	$1.2^{+0.3}_{-0.8}$	$1.3^{+0.3}_{-0.7}$	$0.9^{+0.5}_{-0.9}$	$1.2^{+0.4}_{-0.8}$
WPali	$1.7^{+0.1}_{-0.4}$	$-0.1^{+0.8}_{-0.9}$	$1.5^{+0.2}_{-0.6}$	$1.2^{+0.4}_{-0.9}$	$1.5^{+0.2}_{-0.7}$	$1.2^{+0.4}_{-0.8}$	$1.3^{+0.3}_{-0.7}$	$1.1^{+0.4}_{-0.9}$	$1.1^{+0.4}_{-0.9}$
CCWP	$1.3^{+0.3}_{-0.6}$	$0.8^{+0.5}_{-0.8}$	$1.4^{+0.2}_{-0.6}$	$1.3^{+0.4}_{-0.9}$	$1.4^{+0.3}_{-0.7}$	$1.1^{+0.4}_{-0.8}$	$1.2^{+0.4}_{-0.8}$	$1.1^{+0.5}_{-0.8}$	$1.6^{+0.2}_{-0.5}$
CCe	$1.3^{+0.3}_{-0.6}$	$0.7^{+0.5}_{-0.9}$	$1.1^{+0.4}_{-0.8}$	$1.0^{+0.4}_{-0.8}$	$1.3^{+0.3}_{-0.7}$	$0.9^{+0.6}_{-0.8}$	$1.1^{+0.3}_{-0.8}$	$0.4^{+0.7}_{-0.7}$	$1.5^{+0.2}_{-0.4}$
CCeVMuns	–	$0.5^{+0.6}_{-1.0}$	$1.1^{+0.4}_{-0.7}$	$1.0^{+0.4}_{-0.8}$	$1.3^{+0.3}_{-0.7}$	$0.9^{+0.6}_{-0.8}$	$1.2^{+0.3}_{-0.8}$	$0.3^{+0.7}_{-0.7}$	$1.5^{+0.2}_{-0.4}$
CCeVMsup	–	$0.6^{+0.6}_{-1.0}$	$1.2^{+0.3}_{-0.7}$	$1.0^{+0.4}_{-0.7}$	$1.3^{+0.3}_{-0.6}$	$0.9^{+0.6}_{-0.8}$	$1.0^{+0.4}_{-0.8}$	$0.7^{+0.5}_{-0.8}$	$1.6^{+0.2}_{-0.4}$
CCe2langs	–	$0.5^{+0.6}_{-0.9}$	$0.9^{+0.5}_{-0.7}$	$0.4^{+0.9}_{-0.8}$	$1.1^{+0.4}_{-0.6}$	$1.1^{+0.4}_{-0.9}$	$0.8^{+0.6}_{-0.8}$	$0.5^{+0.7}_{-0.7}$	$1.4^{+0.2}_{-0.8}$
CCe9langs	$1.3^{+0.2}_{-1.1}$	$0.6^{+0.6}_{-0.8}$	$0.9^{+0.3}_{-0.7}$	$0.6^{+0.7}_{-0.8}$	$1.0^{+0.4}_{-0.8}$	$0.7^{+0.6}_{-0.8}$	$0.5^{+0.6}_{-0.8}$	$0.6^{+0.5}_{-0.7}$	$1.2^{+0.3}_{-1.4}$
BERT$_0$	$0.5^{+0.6}_{-0.8}$	$0.4^{+0.7}_{-0.8}$	–	$0.5^{+0.5}_{-0.8}$	$0.8^{+0.4}_{-0.8}$	–	$0.7^{+0.4}_{-0.8}$	–	$0.8^{+0.4}_{-0.7}$
BERT$_{11}$	$-0.1^{+0.9}_{-0.6}$	$1.1^{+0.3}_{-1.1}$	–	$1.0^{+0.3}_{-0.8}$	$0.5^{+0.6}_{-0.9}$	–	$0.2^{+0.6}_{-0.9}$	–	$0.7^{+0.5}_{-0.6}$
mBERT$_0$	$0.3^{+0.6}_{-0.9}$	$-0.1^{+0.7}_{-0.7}$	$0.1^{+0.7}_{-0.6}$	$0.0^{+0.8}_{-0.7}$	$0.1^{+0.6}_{-0.8}$	$-0.1^{+0.7}_{-0.6}$	$0.6^{+0.4}_{-0.9}$	$-0.1^{+0.7}_{-0.7}$	$0.2^{+0.6}_{-0.8}$
mBERT$_{11}$	$0.5^{+0.6}_{-0.9}$	$0.3^{+0.6}_{-0.8}$	$-0.0^{+0.7}_{-0.7}$	$0.2^{+0.6}_{-0.8}$	$0.2^{+0.6}_{-0.8}$	$-0.1^{+0.9}_{-0.6}$	$0.5^{+0.4}_{-0.8}$	$-0.4^{+0.8}_{-0.6}$	$-0.4^{+1.2}_{-0.6}$
XLM-R$_0$	$0.0^{+0.6}_{-0.7}$	$-0.6^{+1.0}_{-0.5}$	$0.2^{+0.6}_{-0.6}$	$-0.2^{+0.9}_{-0.6}$	$0.4^{+0.5}_{-0.7}$	$0.5^{+0.5}_{-0.8}$	$0.3^{+0.6}_{-0.7}$	$-0.2^{+0.7}_{-0.6}$	$0.5^{+0.5}_{-0.8}$
XLM-R$_{11}$	$0.2^{+0.5}_{-0.6}$	$-0.4^{+0.7}_{-0.6}$	$0.2^{+0.6}_{-0.6}$	$-0.2^{+0.7}_{-0.6}$	$0.3^{+0.5}_{-0.6}$	$0.0^{+0.7}_{-0.7}$	$0.3^{+0.6}_{-0.6}$	$0.0^{+0.7}_{-0.8}$	$0.1^{+0.6}_{-0.7}$
XGLM$_0$	$0.5^{+0.6}_{-0.8}$	$-0.3^{+0.7}_{-0.7}$	$0.3^{+0.6}_{-0.7}$	$0.4^{+0.6}_{-0.8}$	$0.6^{+0.4}_{-0.7}$	$0.3^{+0.5}_{-0.7}$	$0.6^{+0.4}_{-0.9}$	$-0.3^{+0.7}_{-0.6}$	$0.1^{+0.6}_{-0.7}$
XGLM$_{47}$	$0.2^{+0.8}_{-0.7}$	$-0.5^{+1.4}_{-0.6}$	$0.4^{+0.6}_{-0.6}$	$-0.3^{+0.8}_{-0.5}$	$0.5^{+0.5}_{-0.7}$	$0.4^{+0.5}_{-0.8}$	$0.5^{+0.5}_{-0.7}$	$0.2^{+0.6}_{-0.7}$	$-0.1^{+0.7}_{-0.5}$
WEAT 2: Instruments and weapons									
WP	$1.6^{+0.2}_{-0.4}$	$1.0^{+0.5}_{-0.8}$	$1.6^{+0.2}_{-0.6}$	$1.2^{+0.4}_{-0.7}$	$1.5^{+0.3}_{-0.7}$	$1.2^{+0.4}_{-1.1}$	$1.4^{+0.3}_{-0.9}$	$1.2^{+0.5}_{-1.2}$	$1.0^{+0.5}_{-1.1}$
WPali	$1.6^{+0.2}_{-0.4}$	$0.7^{+0.6}_{-0.9}$	$1.6^{+0.2}_{-0.6}$	$1.1^{+0.5}_{-0.9}$	$1.5^{+0.3}_{-0.7}$	$0.9^{+0.6}_{-1.2}$	$1.2^{+0.4}_{-1.1}$	$1.2^{+0.5}_{-1.1}$	$0.5^{+0.7}_{-1.1}$
CCWP	$1.6^{+0.2}_{-0.4}$	$0.5^{+0.7}_{-0.8}$	$1.4^{+0.3}_{-1.0}$	$1.3^{+0.4}_{-0.9}$	$1.1^{+0.4}_{-1.1}$	$1.2^{+0.4}_{-0.8}$	$1.4^{+0.3}_{-0.7}$	$0.8^{+0.7}_{-1.0}$	$1.3^{+0.3}_{-0.8}$
CCe	$1.3^{+0.3}_{-0.5}$	$0.6^{+0.6}_{-0.8}$	$1.4^{+0.2}_{-0.5}$	$1.5^{+0.2}_{-0.7}$	$1.0^{+0.4}_{-0.8}$	$0.8^{+0.7}_{-0.8}$	$1.0^{+0.5}_{-0.8}$	$0.5^{+0.7}_{-0.8}$	$1.0^{+0.4}_{-0.8}$
CCeVMuns	–	$0.4^{+0.6}_{-0.8}$	$1.5^{+0.2}_{-0.4}$	$1.5^{+0.2}_{-0.6}$	$0.9^{+0.4}_{-0.6}$	$0.9^{+0.6}_{-0.8}$	$1.1^{+0.4}_{-0.7}$	$0.7^{+0.5}_{-0.7}$	$1.1^{+0.4}_{-0.7}$
CCeVMsup	–	$0.7^{+0.5}_{-0.8}$	$1.4^{+0.2}_{-0.6}$	$1.6^{+0.2}_{-0.5}$	$1.1^{+0.4}_{-0.6}$	$1.0^{+0.5}_{-0.7}$	$0.9^{+0.6}_{-0.7}$	$0.8^{+0.5}_{-0.8}$	$1.1^{+0.4}_{-0.7}$
CCe2langs	–	$0.6^{+0.6}_{-0.9}$	$1.1^{+0.5}_{-0.7}$	$0.4^{+1.3}_{-1.0}$	$1.1^{+0.4}_{-0.6}$	$0.9^{+0.6}_{-0.9}$	$0.9^{+0.6}_{-0.8}$	$0.4^{+0.8}_{-0.8}$	$1.0^{+0.5}_{-1.0}$
CCe9langs	$0.9^{+0.4}_{-1.0}$	$0.9^{+0.5}_{-0.8}$	$0.9^{+0.5}_{-1.1}$	$0.4^{+1.2}_{-1.2}$	$0.9^{+0.4}_{-1.0}$	$0.5^{+1.0}_{-1.2}$	$0.5^{+0.6}_{-0.9}$	$0.4^{+0.8}_{-0.8}$	$1.0^{+0.5}_{-1.6}$
BERT$_0$	$1.2^{+0.4}_{-0.8}$	$0.4^{+0.6}_{-0.9}$	–	$0.9^{+0.3}_{-0.9}$	$0.9^{+0.4}_{-1.0}$	–	$0.7^{+0.5}_{-1.1}$	–	$0.6^{+0.6}_{-0.9}$
BERT$_{11}$	$1.1^{+0.4}_{-0.6}$	$0.5^{+0.5}_{-0.9}$	–	$1.2^{+0.3}_{-1.1}$	$-0.4^{+1.3}_{-0.6}$	–	$0.3^{+0.7}_{-1.0}$	–	$0.8^{+0.4}_{-0.9}$
mBERT$_0$	$0.1^{+0.7}_{-0.8}$	$-0.0^{+0.8}_{-0.8}$	$-0.1^{+0.8}_{-0.6}$	$0.3^{+0.5}_{-0.7}$	$0.2^{+0.6}_{-0.8}$	$0.1^{+0.8}_{-0.8}$	$0.6^{+0.5}_{-1.0}$	$-0.5^{+0.7}_{-0.5}$	$-0.5^{+1.2}_{-0.6}$
mBERT$_{11}$	$1.1^{+0.4}_{-1.4}$	$-0.4^{+1.1}_{-0.7}$	$-0.3^{+1.2}_{-0.9}$	$-0.5^{+1.2}_{-0.6}$	$-0.2^{+1.0}_{-0.8}$	$0.6^{+0.6}_{-1.3}$	$-0.0^{+1.0}_{-0.8}$	$-0.2^{+0.8}_{-0.6}$	$-0.8^{+1.5}_{-0.4}$
XLM-R$_0$	$-0.3^{+0.7}_{-0.6}$	$0.4^{+0.6}_{-1.0}$	$-0.2^{+0.6}_{-0.6}$	$-0.1^{+0.9}_{-0.6}$	$0.2^{+0.6}_{-0.7}$	$-0.2^{+0.8}_{-0.6}$	$-0.4^{+0.9}_{-0.6}$	$-0.2^{+0.9}_{-0.7}$	$-0.7^{+1.4}_{-0.5}$
XLM-R$_{11}$	$0.2^{+0.6}_{-0.6}$	$0.6^{+0.5}_{-1.2}$	$-0.2^{+0.6}_{-0.6}$	$-0.1^{+0.7}_{-0.6}$	$0.1^{+0.6}_{-0.6}$	$-0.1^{+0.7}_{-0.6}$	$-0.1^{+0.8}_{-0.6}$	$-0.6^{+1.4}_{-0.4}$	$0.0^{+0.9}_{-0.8}$
XGLM$_0$	$0.4^{+0.6}_{-0.8}$	$-0.1^{+0.7}_{-0.6}$	$0.5^{+0.5}_{-0.8}$	$0.2^{+0.6}_{-0.7}$	$0.1^{+0.7}_{-0.7}$	$-0.4^{+0.7}_{-0.4}$	$0.3^{+0.6}_{-0.8}$	$-0.2^{+0.9}_{-0.6}$	$0.0^{+0.9}_{-0.8}$
XGLM$_{47}$	$0.5^{+0.5}_{-0.7}$	$0.1^{+0.6}_{-0.7}$	$0.4^{+0.4}_{-0.6}$	$0.4^{+0.6}_{-0.8}$	$0.2^{+0.5}_{-0.6}$	$-0.6^{+1.6}_{-0.5}$	$-0.1^{+0.6}_{-0.6}$	$-0.4^{+1.5}_{-0.5}$	$-0.7^{+1.4}_{-0.3}$

Table 8: Effect size for X-WEAT tests for 8 languages plus the original English WEAT (en$_{or}$) test. 95% confidence intervals obtained by bootstrap resampling of the lists.

	en_{or}	$\text{en}_{(5)}$	$\text{ar}_{(1)}$	$\text{ca}_{(2)}$	$\text{de}_{(24)}$	$\text{es}_{(10)}$	$\text{hr}_{(12)}$	$\text{it}_{(24)}$	$\text{ru}_{(2)}$	$\text{tr}_{(2)}$
WEAT 1: Flowers and insects										
WP	1.7	$1.4^{+0.1}_{-0.4}$	0.0	$1.1^{+0.3}_{-0.3}$	$1.1^{+0.3}_{-1.0}$	$1.2^{+0.3}_{-0.4}$	$0.6^{+0.5}_{-0.7}$	$1.2^{+0.4}_{-0.8}$	$0.7^{+0.1}_{-0.1}$	$1.0^{+0.3}_{-0.3}$
WPali	1.7	$1.4^{+0.1}_{-0.4}$	0.1	$1.2^{+0.3}_{-0.3}$	$1.1^{+0.3}_{-0.9}$	$1.3^{+0.2}_{-0.5}$	$0.6^{+0.5}_{-0.7}$	$1.3^{+0.3}_{-0.8}$	$0.8^{+0.1}_{-0.1}$	$1.0^{+0.4}_{-0.4}$
CCWP	1.3	$1.1^{+0.2}_{-0.8}$	0.6	$0.9^{+0.4}_{-0.4}$	$1.0^{+0.5}_{-0.6}$	$1.1^{+0.3}_{-0.4}$	$0.8^{+0.5}_{-0.5}$	$1.1^{+0.4}_{-0.9}$	$0.5^{+0.2}_{-0.2}$	$1.6^{+0.1}_{-0.1}$
CCe	1.3	$0.9^{+0.1}_{-0.5}$	0.2	$0.8^{+0.1}_{-0.1}$	$0.4^{+0.5}_{-0.4}$	$1.0^{+0.3}_{-0.4}$	$0.7^{+0.4}_{-0.6}$	$0.9^{+0.2}_{-0.6}$	$0.7^{+0.1}_{-0.1}$	$1.0^{+0.2}_{-0.2}$
CCeVMuns	–	–	0.1	$0.4^{+0.2}_{-0.2}$	$0.4^{+0.5}_{-0.4}$	$0.6^{+0.4}_{-0.7}$	$0.6^{+0.5}_{-0.5}$	$0.8^{+0.4}_{-0.6}$	$0.7^{+0.2}_{-0.2}$	$0.9^{+0.3}_{-0.3}$
CCeVMsup	–	–	0.8	$0.7^{+0.2}_{-0.2}$	$0.3^{+0.5}_{-0.5}$	$0.7^{+0.5}_{-0.6}$	$0.6^{+0.5}_{-0.5}$	$0.8^{+0.3}_{-0.4}$	$0.9^{+0.2}_{-0.2}$	$0.9^{+0.3}_{-0.3}$
CCe2langs	–	–	0.1	$0.7^{+0.1}_{-0.1}$	$0.4^{+0.5}_{-0.4}$	$1.0^{+0.3}_{-0.4}$	$0.5^{+0.3}_{-0.4}$	$0.9^{+0.3}_{-0.6}$	$0.7^{+0.0}_{-0.0}$	$0.6^{+0.5}_{-0.5}$
CCe9langs	1.3	$0.1^{+0.7}_{-0.2}$	0.4	$0.4^{+0.4}_{-0.4}$	$0.5^{+0.3}_{-0.5}$	$0.8^{+0.6}_{-0.7}$	$0.3^{+0.4}_{-0.6}$	$0.5^{+0.6}_{-0.6}$	$0.8^{+0.1}_{-0.1}$	$0.6^{+0.3}_{-0.3}$
BERT$_0$	0.5	$0.2^{+0.4}_{-0.2}$	0.6	–	$0.1^{+0.3}_{-0.3}$	$0.5^{+0.1}_{-0.4}$	–	$0.3^{+0.4}_{-0.3}$	–	$0.1^{+0.3}_{-0.3}$
BERT$_{11}$	−0.1	$0.6^{+0.3}_{-0.9}$	0.7	–	$0.2^{+0.4}_{-0.5}$	$0.0^{+0.4}_{-0.5}$	–	$0.1^{+0.5}_{-0.6}$	–	$0.3^{+0.5}_{-0.5}$
mBERT$_0$	0.3	$-0.3^{+0.3}_{-0.3}$	−0.3	$0.3^{+0.3}_{-0.3}$	$-0.0^{+0.3}_{-0.3}$	$0.3^{+0.4}_{-0.5}$	$0.3^{+0.3}_{-0.6}$	$0.4^{+0.4}_{-0.4}$	$-0.1^{+0.2}_{-0.2}$	$-0.1^{+0.2}_{-0.2}$
mBERT$_{11}$	0.5	$-0.4^{+0.8}_{-0.1}$	0.2	$0.1^{+0.3}_{-0.3}$	$-0.2^{+0.6}_{-0.3}$	$0.1^{+0.4}_{-0.2}$	$0.0^{+0.3}_{-0.3}$	$0.2^{+0.4}_{-0.3}$	$-0.2^{+0.2}_{-0.2}$	$-0.7^{+0.2}_{-0.2}$
XLM-R$_0$	0.0	$-0.1^{+0.5}_{-0.6}$	−1.0	$0.4^{+0.0}_{-0.0}$	$-0.0^{+0.3}_{-0.3}$	$0.0^{+0.4}_{-0.3}$	$0.2^{+0.4}_{-0.4}$	$0.1^{+0.3}_{-0.3}$	$0.0^{+0.3}_{-0.3}$	$-0.3^{+0.5}_{-0.5}$
XLM-R$_{11}$	0.2	$-0.1^{+0.2}_{-0.4}$	−0.8	$0.1^{+0.2}_{-0.2}$	$0.1^{+0.3}_{-0.3}$	$-0.1^{+0.4}_{-0.4}$	$0.2^{+0.3}_{-0.4}$	$0.1^{+0.3}_{-0.3}$	$-0.0^{+0.2}_{-0.2}$	$-0.4^{+0.4}_{-0.4}$
XGLM$_0$	0.5	$0.3^{+0.3}_{-0.4}$	−0.7	$0.3^{+0.1}_{-0.1}$	$0.1^{+0.5}_{-0.2}$	$-0.1^{+0.3}_{-0.3}$	$0.1^{+0.3}_{-0.5}$	$0.2^{+0.4}_{-0.4}$	$0.3^{+0.2}_{-0.2}$	$0.1^{+0.3}_{-0.3}$
XGLM$_{47}$	0.2	$0.2^{+0.2}_{-0.5}$	−0.3	$0.4^{+0.1}_{-0.1}$	$0.1^{+0.4}_{-0.4}$	$0.4^{+0.3}_{-0.3}$	$0.0^{+0.6}_{-0.5}$	$0.4^{+0.3}_{-0.6}$	$0.1^{+0.2}_{-0.2}$	$-0.3^{+0.2}_{-0.2}$
WEAT 2: Instruments and weapons										
WP	1.6	$1.4^{+0.1}_{-0.3}$	1.2	$1.5^{+0.1}_{-0.1}$	$1.4^{+0.3}_{-0.7}$	$1.5^{+0.2}_{-0.3}$	$0.7^{+0.6}_{-0.8}$	$1.4^{+0.1}_{-0.6}$	$0.9^{+0.0}_{-0.0}$	$1.4^{+0.0}_{-0.0}$
WPali	1.6	$1.4^{+0.1}_{-0.3}$	0.8	$1.3^{+0.1}_{-0.1}$	$1.3^{+0.4}_{-0.6}$	$1.4^{+0.2}_{-0.5}$	$0.7^{+0.6}_{-0.8}$	$1.4^{+0.2}_{-0.6}$	$0.9^{+0.1}_{-0.1}$	$1.1^{+0.1}_{-0.1}$
CCWP	1.6	$1.4^{+0.2}_{-0.2}$	1.1	$1.2^{+0.0}_{-0.0}$	$1.1^{+0.4}_{-0.4}$	$1.3^{+0.2}_{-0.7}$	$1.3^{+0.2}_{-0.6}$	$1.0^{+0.4}_{-0.5}$	$1.0^{+0.0}_{-0.0}$	$1.2^{+0.1}_{-0.1}$
CCe	1.3	$1.0^{+0.2}_{-0.6}$	0.7	$1.2^{+0.0}_{-0.0}$	$1.2^{+0.4}_{-0.3}$	$0.8^{+0.3}_{-0.7}$	$0.7^{+0.5}_{-0.5}$	$0.8^{+0.4}_{-0.8}$	$0.9^{+0.1}_{-0.1}$	$0.9^{+0.1}_{-0.1}$
CCeVMuns	–	–	0.7	$1.1^{+0.4}_{-0.4}$	$0.9^{+0.9}_{-0.4}$	$0.6^{+0.3}_{-0.5}$	$0.7^{+0.5}_{-0.3}$	$0.8^{+0.4}_{-0.7}$	$0.8^{+0.1}_{-0.1}$	$0.1^{+0.1}_{-0.1}$
CCeVMsup	–	–	0.9	$0.9^{+0.2}_{-0.2}$	$0.8^{+0.9}_{-0.4}$	$0.5^{+0.4}_{-0.4}$	$1.0^{+0.4}_{-0.6}$	$0.7^{+0.4}_{-0.7}$	$0.7^{+0.0}_{-0.0}$	$0.5^{+0.2}_{-0.2}$
CCe2langs	–	–	0.9	$1.0^{+0.1}_{-0.1}$	$1.0^{+0.4}_{-0.8}$	$0.7^{+0.4}_{-0.6}$	$0.6^{+0.7}_{-0.4}$	$0.9^{+0.3}_{-0.7}$	$0.8^{+0.1}_{-0.1}$	$0.9^{+0.0}_{-0.0}$
CCe9langs	0.9	$0.0^{+1.0}_{-0.3}$	1.0	$0.4^{+0.0}_{-0.0}$	$1.0^{+0.3}_{-0.8}$	$0.5^{+0.4}_{-0.7}$	$0.4^{+0.5}_{-1.0}$	$0.5^{+0.5}_{-0.5}$	$0.9^{+0.1}_{-0.1}$	$1.0^{+0.1}_{-0.1}$
BERT$_0$	1.2	$0.6^{+0.5}_{-0.1}$	0.8	–	$0.7^{+0.4}_{-0.6}$	$0.6^{+0.5}_{-0.5}$	–	$0.3^{+0.5}_{-0.6}$	–	$0.0^{+0.0}_{-0.0}$
BERT$_{11}$	1.1	$0.8^{+0.5}_{-0.4}$	1.0	–	$1.1^{+0.4}_{-0.8}$	$0.3^{+0.6}_{-0.7}$	–	$0.1^{+0.4}_{-0.4}$	–	$0.3^{+0.1}_{-0.1}$
mBERT$_0$	0.1	$0.7^{+0.4}_{-0.5}$	0.1	$-0.0^{+0.0}_{-0.0}$	$0.4^{+0.2}_{-0.4}$	$0.3^{+0.7}_{-0.3}$	$0.1^{+0.7}_{-0.5}$	$0.4^{+0.5}_{-0.2}$	$0.3^{+0.2}_{-0.2}$	$0.4^{+0.0}_{-0.0}$
mBERT$_{11}$	1.1	$0.6^{+0.4}_{-0.3}$	0.2	$-0.1^{+0.4}_{-0.4}$	$0.6^{+0.2}_{-1.1}$	$0.6^{+0.2}_{-0.5}$	$0.5^{+0.6}_{-1.0}$	$0.8^{+0.4}_{-0.8}$	$0.8^{+0.0}_{-0.0}$	$0.4^{+0.1}_{-0.1}$
XLM-R$_0$	−0.3	$-0.1^{+0.4}_{-0.1}$	−0.3	$0.3^{+0.1}_{-0.1}$	$0.1^{+0.3}_{-0.2}$	$0.3^{+0.4}_{-0.6}$	$-0.2^{+0.4}_{-0.2}$	$-0.0^{+0.5}_{-0.4}$	$0.1^{+0.4}_{-0.4}$	$-0.1^{+0.2}_{-0.2}$
XLM-R$_{11}$	0.2	$0.2^{+0.2}_{-0.1}$	0.3	$0.1^{+0.3}_{-0.3}$	$0.1^{+0.4}_{-0.2}$	$0.2^{+0.3}_{-0.5}$	$-0.0^{+0.6}_{-0.2}$	$-0.0^{+0.3}_{-0.4}$	$-0.1^{+0.3}_{-0.3}$	$-0.1^{+0.2}_{-0.2}$
XGLM$_0$	0.4	$0.5^{+0.4}_{-0.4}$	0.2	$0.6^{+0.1}_{-0.1}$	$0.3^{+0.3}_{-0.4}$	$0.3^{+0.5}_{-0.3}$	$-0.3^{+0.3}_{-0.3}$	$-0.1^{+0.6}_{-0.2}$	$0.4^{+0.2}_{-0.2}$	$0.3^{+0.1}_{-0.1}$
XGLM$_{47}$	0.5	$0.3^{+0.2}_{-0.8}$	0.2	$0.2^{+0.1}_{-0.1}$	$0.1^{+0.4}_{-0.3}$	$0.2^{+0.5}_{-0.4}$	$-0.4^{+0.7}_{-0.3}$	$-0.1^{+0.2}_{-0.4}$	$0.3^{+0.0}_{-0.0}$	$-0.5^{+0.2}_{-0.2}$

Table 9: Effect size for CA-WEAT tests with lists created for 9 languages; the original English WEAT (en_{or}) test is shown for comparison. The number of lists per language is shown as subindex of the language. We report the median and 95% confidence intervals.

F Isomorphism Analysis

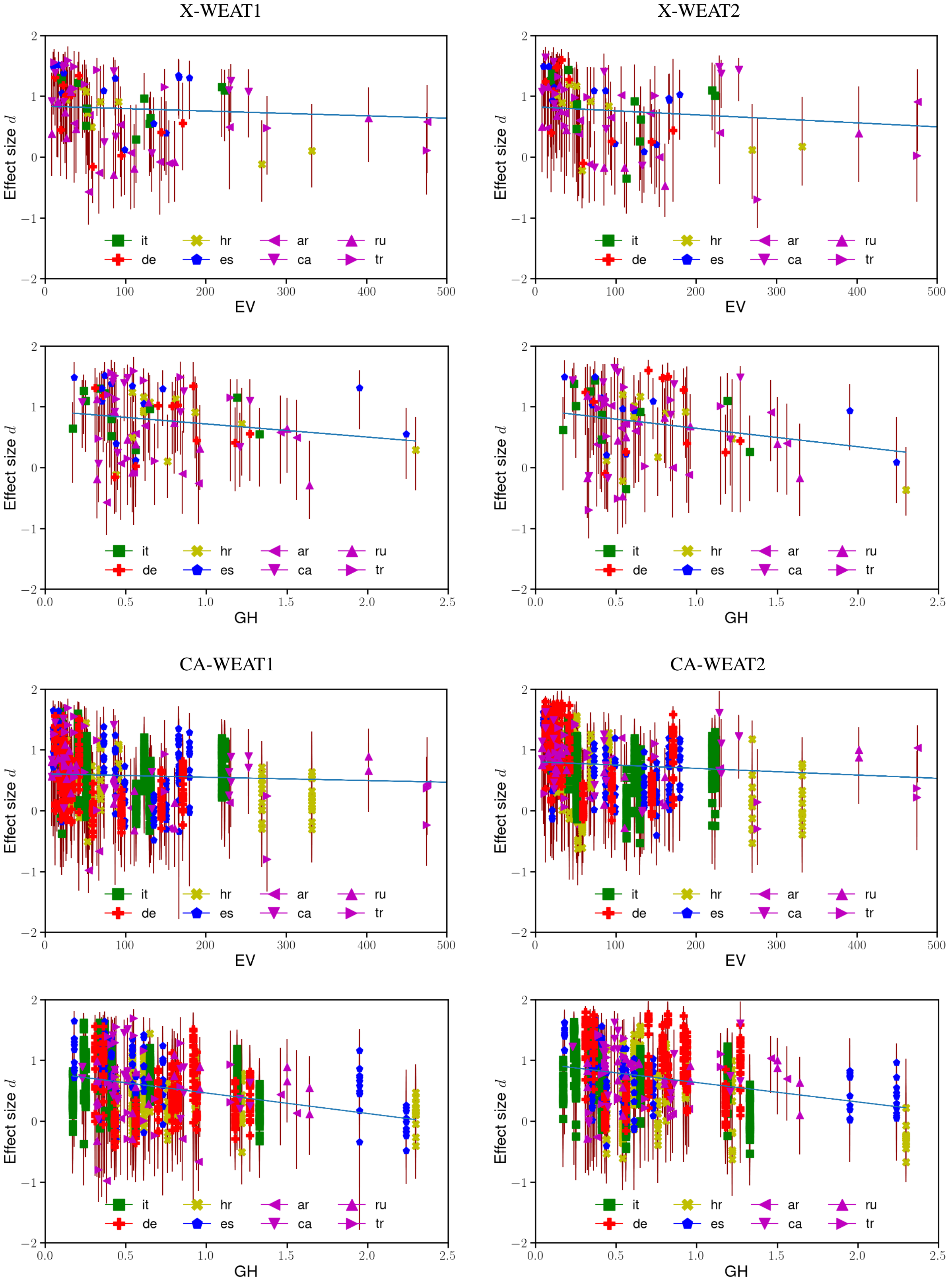


Figure 8: Effect size as a function of EV and GH between English and each of the other 8 languages. Considered models are *WP*, *WPali*, *CCWP*, *CCe*, *CCeVMuns*, *CCeVMsup*, *CCe2langs*, *CCe9langs*, *mBERT*$_0$, *XLM-R*$_0$ and *XGLM*$_0$. Table 1 in the main text reports the values of the measures for these models.