# The Emergence of Semantic Units in Massively Multilingual Models

**Andrea de Varda, Marco Marelli**
University of Milano – Bicocca
Department of Psychology
a.devarda@campus.unimib.it, marco.marelli@unimib.it

**Abstract**

Massively multilingual models can process text in several languages relying on a shared set of parameters; however, little is known about the encoding of multilingual information in single network units. In this work, we study how two semantic variables, namely valence and arousal, are processed in the latent dimensions of mBERT and XLM-R across 13 languages. We report a significant cross-lingual overlap in the individual neurons processing affective information, which is more pronounced when considering XLM-R vis-à-vis mBERT. Furthermore, we uncover a positive relationship between cross-lingual alignment and performance, where the languages that rely more heavily on a shared cross-lingual neural substrate achieve higher performance scores in semantic encoding.

**Keywords:** Massively multilingual models, neuron-level analysis, affective variables

## 1. Introduction

Recently, NLP research has seen a rapid surge of massively multilingual models (MMMs) such as mBERT (Devlin et al., 2019), XLM (Lample and Conneau, 2019), XLM-R (Conneau et al., 2020a), XGLM (Lin et al., 2021), BLOOM (Scao et al., 2022), and mGPT (Shliazhko et al., 2022). MMMs are usually derived from monolingual models based on the Transformer architecture (Vaswani et al., 2017), trained with a (masked) language modelling objective on non-aligned multilingual text in several languages (up to 104 in mBERT), without being exposed to any cross-lingual signal during training. MMMs reach impressive performance levels in zero-shot cross-lingual transfer, enabling the training of a model on supervised data in a source language and its application to a different target language, with no additional training. Crucially, cross-lingual transfer has been documented across a range of languages and tasks (Pires et al., 2019; Wu and Dredze, 2019; Dufter and Schütze, 2020; Liu et al., 2020; Lauscher et al., 2020; Winata et al., 2022; see Doddapaneni et al., 2021 for a review), even when source and target language share very few (Karthikeyan et al., 2020) or no (Karthikeyan et al., 2019; Conneau et al., 2020b) items in their vocabulary.

From a practical perspective, MMMs show several desirable properties: they require less resource and maintenance with respect to multiple monolingual models (Dufter and Schütze, 2020), and provide reasonably good representations for low-resource languages, for which it would be impossible to construct well-performing monolingual models due to data scarcity. From a theoretical perspective, MMMs open the way for a promising research question: Is it possible to develop an *interlingua* from text data alone, where both syntactic and semantic representations are encoded in a shared format across languages?

## 2. Related work

A rather limited body of findings is contributing to the question of whether MMMs develop cross-lingual internal representations, showing that representational similarity between matching sentences in different languages increases in the intermediate layers of the networks (Del and Fishel, 2021; Muller et al., 2021, but see Singh et al., 2019). Similar conclusions have been reached with different approaches, such as employing the network's intermediate subspaces to perform machine translation (Pires et al., 2019; Libovický et al., 2020), word alignment (Libovický et al., 2020), or to reconstruct cross-lingual syntactic trees (Chi et al., 2020). These studies, however, considered layer-wise representations as a whole, overlooking the role played by the individual neural units in the embeddings. There have been, however, other research efforts with a greater degree of granularity, which analyzed the cross-lingual consistency in the individual units processing linguistic properties in MMMs. In a fine-grained neuron-level study, Antverg and Belinkov (2021) reported that mBERT and XLM-R encode morphosyntactic features such as number, tense, and gender in an overlapping set of dimensions across languages (see also Stanczak et al., 2022). Similar results were obtained in a strictly syntactic probing setting, showing that the processes underpinning number agreement computations across languages could be ascribed to an overlapping set of latent dimensions in the structural embeddings of the models (de Varda and Marelli, 2023).

While most of the experiments centered at the layer-wise representational level considered



*LREC-COLING 2024*, pages 15910–15921

sentence representations as a whole, conflating semantic and syntactic properties (Singh et al., 2019; Del and Fishel, 2021), the more fine-grained neuron-level studies focused on morphological and syntactic features (Antverg and Belinkov, 2021; Stanczak et al., 2022). Thus, to our knowledge, no study has studied whether MMMs respond to purely semantic features in the same neural units across languages. This leaves the broad question of this study – i.e., whether MMMs develop an *interlingua* from text data alone – partially unanswered.

## 3. Aims

In this paper, we investigate the processes that support the generation of lexical meaning representations in mBERT and XLM-R across 13 languages. Instead of considering lexical representations as a whole, we increase the experimental control over our analyses by focusing on two specific lexical semantic features related to the emotional content of a word. The emotional connotation of a word plays a fundamental role in its processing, influencing recognition times (Kousta et al., 2009; Larsen et al., 2006), neurophysiological responses (Kissler et al., 2007; Kuchinke et al., 2005; Citron, 2012), and bodily responses such as facial muscular activity and heartbeat (Vergallito et al., 2019). The paramount importance of affective information in language processing is well motivated from an evolutionary perspective, given that emotion processing relies upon a phylogenetically ancient system aimed at survival (see for instance Van Berkum, 2010). Given the broad recognition of the foundational role played by emotion-related information in language, we considered affective content as a natural candidate feature that should be encoded in a consistent way in the MMMs parameters. The emotional content of a word (Kuperman et al., 2014) and emotions in general (Russell, 2003; Russell and Barrett, 1999) are typically operationalized along two axes, namely valence and arousal[1]. The former dimension reflects the degree to which a word is pleasant (e.g. *friendship*) or unpleasant (*nausea*), whereas the latter indicates the extent to which it is calming (*slumber*) or exciting (*fight*)[2]. In this study, we tested whether semantic knowledge about the affective properties of a word spontaneously emerged in the networks we consider. We approached the analysis with an atomistic perspective, aimed at identifying sub-layer clusters of neurons that are associated with the two affective variables we study. With respect to this first objective, Radford et al. (2017) have provided evidence that a single neuron in a multiplicative LSTM-based language model spontaneously learned to respond to sentiment and affective content, exemplifying a case of symbol emergence. The authors took this result as an indication that sentiment might serve as a high-level conditioning feature with strong predictive capability for language modelling. This observation, in conjunction with the psychological and neuroscientific evidence presented above, hints at the possibility that affective dimensions might be suitable candidate features to be considered in the search for an interlingua in MMMs. This rationale motivated the second objective of our study, namely the assessment of the cross-lingual congruence in the processing of affective information in MMMs. If single units encode emotional information in multilingual transformer models as they do in monolingual mLSTMs, are the units involved in this process coherent across languages? In this paper, we tested whether the encoding of these lexical properties taxed the same components within the networks' layers.

## 4. Methods and materials

In this study, we trained different probes on the mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a) representations to predict a lexical affective variable (valence and arousal) from the internal activation of the models across a sample of 13 languages. We then employed the learned weights of the probes to evaluate the relevance of the networks' individual units in encoding the affective variable, and ultimately assessed the cross-lingual consistency of such encoding.

The code is publicly available on GitHub[3].

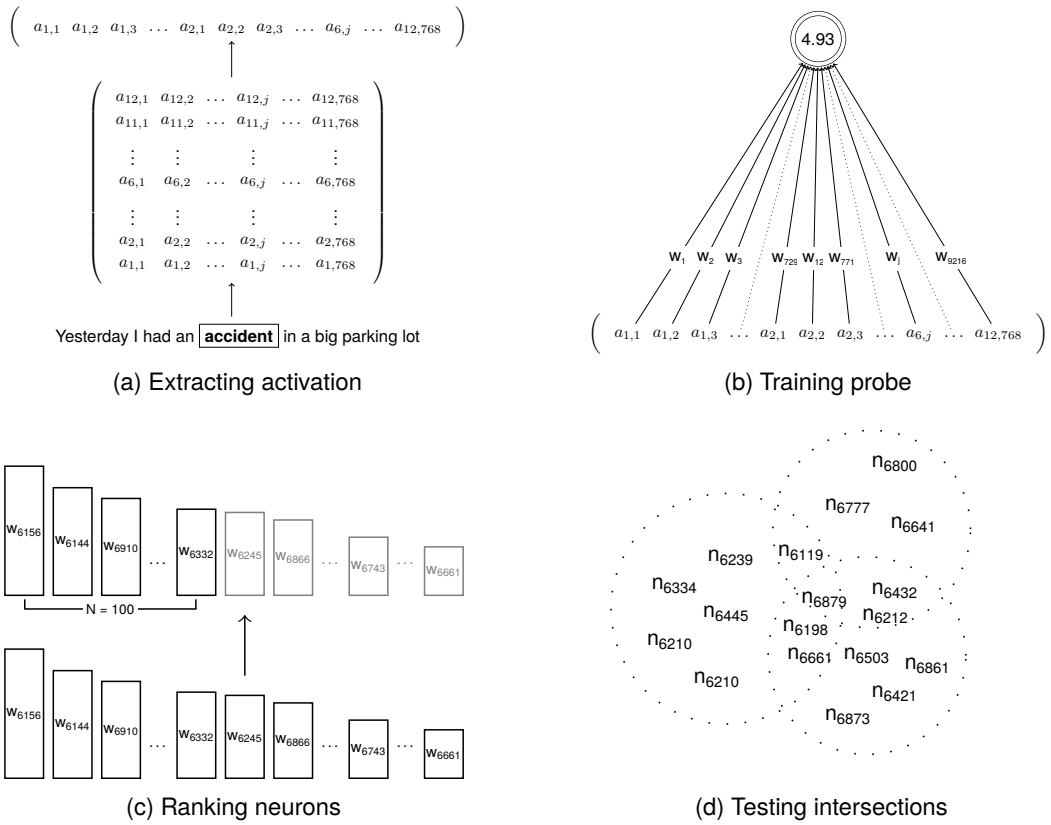
### 4.1. Data

We considered all the human-annotated affective norms available at the time of writing, except for the Finnish norms released by Eilola and Havelka (2010), which only include ratings for 210 items. Our language sample is summarized in Table 1; it includes 13 languages, covering 4 language families and 3 scripts (Latin in all cases but Greek and Chinese). For each word in our datasets, we scraped the Wikipedia data of the corresponding language to search for 15 sentences that contained that word. We chose not to present the models with

[1] A third semantic dimension, dominance, is sometimes considered when operationalizing emotional content. However, we did not include it in our analyses due to data scarcity, since many of the affective norms we employed in our study did not include it.

[2] These two dimensions are theoretically orthogonal (Carver and Scheier, 1990; Feldman Barrett and Russell, 1998), although different accounts of their empirical relationship have been proposed (see Kuppens et al., 2013, for an overview).

[3] https://github.com/Andrea-de-Varda/affective-interlingua




(a) Extracting activation

(b) Training probe

(c) Ranking neurons

(d) Testing intersections

Figure 1: Graphical summary of the experimental pipeline. First, we extract the network activation in response to a contextualized target word (1a). Then, we train a linear probe to predict the value of the affective variable (either valence or arousal) from the network activation (1b). We employ the absolute weights of the probe to sort the network units according to their relevance in detecting valence and arousal, and truncate the ranks at the 100$^{th}$ position (1c). Lastly, we assess the cross-lingual overlap in the neural units encoding affective information (1d).

| Authors | Language | Family | Items | Raters |
|---|---|---|---|---|
| Montefinese et al. (2013) | Italian | ine | 1,121 | 684 |
| Warriner et al. (2013) | English | ine | 13,915 | 1,827 |
| Moors et al. (2013) | Dutch | ine | 4,300 | 224 |
| Redondo et al. (2007) | Spanish | ine | 1,034 | 720 |
| Imbir (2016) | Polish | ine | 4,900 | 400 |
| Ćoso et al. (2019) | Croatian | ine | 3,022 | 933 |
| Monnier and Syssau (2014) | French | ine | 1,031 | 469 |
| Palogiannidi et al. (2016) | Greek | ine | 1,034 | 105 |
| Schmidtke et al. (2014) | German | ine | 1,003 | 65 |
| Soares et al. (2011) | Portuguese | ine | 1,034 | 958 |
| Kapucu et al. (2021) | Turkish | trk | 2,031 | 1,527 |
| Yao et al. (2017) | Chinese | sit | 1,100 | 960 |
| Sianipar et al. (2016) | Indonesian | map | 1,402 | 1,490 |

Table 1: Affective norms. The language family is indicated with the respective ISO 639-5 code (ine: Indoeuropean; trk: Turkic; sit: Sino-Tibetan; map: Austronesian).

single words without context since single words are an unnatural input to the pre-trained encoders, which rarely encountered them in isolation (Bommasani et al., 2020). In order to license meaningful comparisons across languages, all the data were downsampled to match the smallest dataset available, i.e., the Greek norms, for which we were able to find 15 sentences from Wikipedia data only for 520 words (7,800 sentences). Since the psycholinguistic norms were evaluated on different Likert scales, we min-max normalized them before training the probes.

## 4.2. Models

We carried out our analyses employing the original releases of mBERT and XLM-R. The networks were employed as out-of-the-box masked language models and did not undergo any fine-tuning or adaptation process. In particular, we relied on mBERT$_{\text{BASE}}$ (cased) and on XLM-R$_{\text{BASE}}$. The two models have a similar configuration ($L$ = 12, $H$ = 768, $A$ = 12), but while mBERT is jointly trained with a masked language modeling (MLM) and a next sentence prediction (NSP) objective, XLM-R is optimized only for MLM on more training data. We did not mask any words throughout this work.

## 4.3. Methods

Our procedure is divided into four steps (see Figure 1), described in the following subsections.

#### 4.3.1. Extracting activations

As a first step, we extracted the internal activation of the networks in response to the contextualized presentation of each word in our dataset (Figure 1a). In the case of multi-token words, we averaged the network activation for the various tokens, following Dalvi et al. (2019a). We standardized unit-wise the activation matrix, so that the activations of each neuron across sentences had $\overline{x} = 0$ and $s = 1$. Standardization was performed to favor the interpretability of the probe, so that the learned weights were not affected by the activation variance.

#### 4.3.2. Training the probe

In a subsequent step, we trained language-specific linear probes to predict arousal and valence values from the internal activation of the networks in response to the target token (Figure 1b). The probes were trained independently in each language, and there was no inherent bias towards learning cross-lingual patterns. Probing studies tend to consider categorical labels as outputs, and thus are generally based on probing classifiers (see for instance Belinkov et al., 2017; Dalvi et al., 2019a; Pires et al., 2019). However, since the affective variables we consider in this article are continuous in nature, we adopted a linear regression approach. The probes were trained with mean squared error loss and elastic net regularisation as additional loss term (Zou and Hastie, 2005), with $\lambda_1 = \lambda_2 = 0.001$. The probes were thus trained to minimize the following loss function:

$$\mathcal{L}_\theta = \frac{1}{m}\sum_{i=1}^{m}(y-\hat{y})^2 + \lambda_1 \sum_{j=1}^{n}|w_j| + \lambda_2 \sum_{j=1}^{n} w_j^2 \quad (1)$$

Where *m* is the number of observations and *n* the number of weights in $\theta$. The terms $\lambda_1 \sum_{j=1}^{n}|w_j|$ and $\lambda_2 \sum_{j=1}^{n} w_j^2$ correspond to L1 and L2 regularization, respectively. Their combination in elastic-net regularization balances the tendencies to identify few localized features (L1 in Lasso regression) versus distributed neurons (L2 in Ridge regression; Durrani et al., 2020). The probes were trained and tested on two subsets of the original data (80% train, N = 6,240; 20% test, N = 1,560). The training was performed with the `NeuroX` toolkit (Dalvi et al., 2019b), a Python package to perform interpretation and analysis of deep neural networks. In this step, we also identified the layers that displayed the strongest response to the affective variables we considered to restrict the following analyses to a relevant population of neural units[4].

[4]Note that if we had not restricted our analyses to a pre-specified layer, our estimates of cross-lingual overlap might have been overestimated, since similar processes with similar degrees of abstraction are likely to be processed in certain layers of the networks (e.g. semantic features in deeper layers, see Jawahar et al., 2019).

#### 4.3.3. Ranking neurons

The learned weights of the probe associated with the relevant layer were then employed to rank the neurons with respect to the task as a measure of the relevance of the corresponding units with respect to the linguistic property being investigated (namely, the semantic notions of valence and arousal). More precisely, neurons were ranked according to their absolute weight. The 100 units[5] with the highest absolute weight were independently selected for each language and affective variable (Figure 1c).

#### 4.3.4. Testing intersections

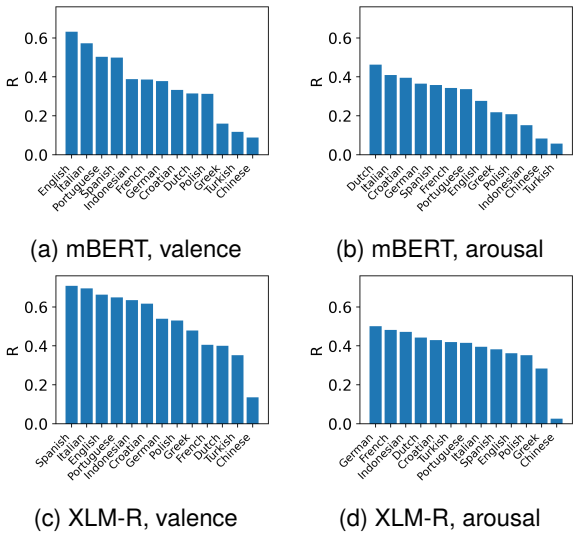
Once the neuron ranking was obtained, we empirically assessed the cross-lingual overlap of the sets of neurons responding to valence and arousal across languages in the pre-specified layer (Figure 1d). The Fold Enrichment (*FE*, i.e., the ratio between observed and expected overlap) and the statistical significance of the resulting intersections were calculated with the super exact test (Wang et al., 2015), a procedure for computing the distributions of multi-set intersections[6]. We considered the cross-lingual unit-level converge of information as quantified by the super-exact test as our operational definition of the theoretical construct of the *interlingua*.

## 5. Results

We hereby report the results obtained by the linear probes in each of the 13 languages considered in this study. As can be seen in Figure 2, the performance of the probes tends to be higher with XLM-R (Figures 2c, 2d) than with mBERT (Figures 2a, 2b), and when considering valence (Figures

[5]We considered 100 units as Antverg and Belinkov (2021) and de Varda and Marelli (2023).

[6]The super exact test computes the probability of obtaining a number of elements that overlap between two or more sets (i.e. intersection size) that is equal to or higher than the one observed. This probability (one-tailed) in binary intersections can be calculated by integrating over a hypergeometric function from the observed value to the maximum possible overlap size. The super exact test extends this analysis to multi-set intersections; however, since an analogous calculation of the probabilities of multi-set intersections would involve integrations over all possible hierarchical intersections across all the sets (thus with exponential growth of the operations), the super exact test optimizes the procedure through a forward algorithm which produces the same results with operation complexity $\mathcal{O}(m^2)$, where $m$ is the smallest set size.




(a) mBERT, valence (b) mBERT, arousal

(c) XLM-R, valence (d) XLM-R, arousal

Figure 2: Whole-network-based probe performances divided by model, affective variable, and language.

2a, 2c) compared to arousal (Figures 2b, 2d). Non-Indoeuropean languages – in particular Chinese and Turkish – and languages written in non-Latin script – Greek and Chinese – tend to be associated with lower performance levels.

The results of our layer-wise probes are summarized in Figure 3. The peak performance, measured as the Pearson correlation between probe predictions and target affective ratings, is reached across model types in the intermediate layers of the networks. More precisely, the highest performance scores are obtained in layer 7 by mBERT both in the case of valence and arousal and in the layers $[\![5, 6]\!]$ by XLM-R for valence and arousal, respectively. We also note that a second local *maximum* seems to be obtained in the deepest layers, and in particular in layer 11, across model types. Regarding model and measure differences, the layer-wise trends reveal higher performance scores for XLM-R with respect to mBERT and when employing valence as the dependent variable as opposed to arousal.

In light of the layer-wise patterns identified above, we restricted our multi-set intersection analyses to layer 7 for mBERT and layers $[\![5, 6]\!]$ for XLM-R. The pairwise intersection patterns document a general overrepresentation of units in the set intersections relative to what would be expected by chance[7]. In the case of mBERT, out of C(13, 2) = 78 combinations, 62 have more units than what would be expected assuming independence (79.49%), both when considering arousal and valence. This overrepresentation is significant with $p < .05$ in 13 cases when considering valence (16.67%) and 16 when considering arousal (20.51%). Overall, affective information displays a greater degree of cross-lingual convergence when considering XLM-R. The overlap in valence-responding units has $FE > 1$ in 77 combinations (98.71%), and the overlap in arousal-sensitive neurons has $FE > 1$ in all 78 pairs; in the case of valence, the results are significant with $p < .05$ in 66 cases (84.62%), while in the case of valence the number raises up to 74 (94.87%).

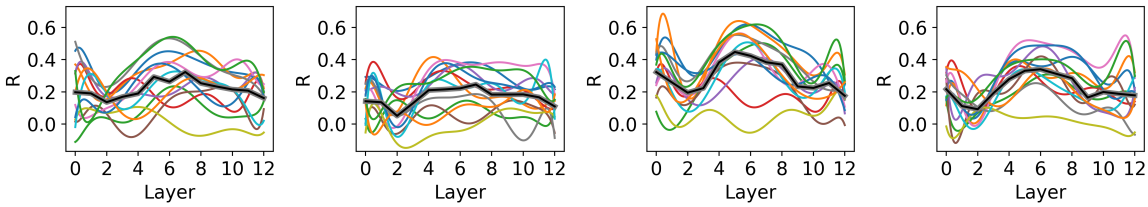
While we cannot comment exhaustively all the $N_{2\ldots13}$-way intersections, we report the highest degree intersections obtained for each combination of model and affective variable. Four neurons in mBERT occupied a position in the top 100 units responding to arousal in 7 languages, i.e. (1) Id[8] ∩ It ∩ Pt ∩ Hr ∩ Es ∩ Pl ∩ Tr; (2) En ∩ Pt ∩ Fr ∩ Hr ∩ Zh ∩ Es ∩ Pl; (3) En ∩ Nl ∩ Fr ∩ Hr ∩ Zh ∩ Es ∩ El; (4) Id ∩ En ∩ It ∩ Nl ∩ Pt ∩ Es ∩ De ($FE$ = 2,051.95, $p$ = 0.0004). When considering arousal, a singleton eighth-degree intersection was found in Pt ∩ Id ∩ El ∩ Hr ∩ Tr ∩ Pl ∩ It ∩ En ($FE$ = 15,758.99, $p = 6.34 \cdot 10^{-5}$). When considering the XLM-R model, we found two neurons involved in processing valence in 11 languages (Pt ∩ De ∩ Nl ∩ Id ∩ El ∩ Tr ∩ Hr ∩ Zh ∩ Es ∩ En ∩ It; Pt ∩ De ∩ Nl ∩ Fr ∩ Id ∩ El ∩ Hr ∩ Zh ∩ Pl ∩ Es ∩ It, with $FE$ = 7,138,586, $p = 1.40 \cdot 10^{-7}$), and one neuron involved in processing arousal in 10 languages (Pt ∩ De ∩ Nl ∩ Fr ∩ Id ∩ Tr ∩ Hr ∩ Pl ∩ Es ∩ It, $FE$ = 929,503.4, $p = 1.08 \cdot 10^{-6}$). In summary, we found several neurons responding to valence and arousal in most of the languages considered in the study; the language coverage of those units was generally bigger when considering XLM-R as opposed to mBERT, in accordance with what we found with the pairwise intersection analyses.

## 6. Generalization to an unseen language

In §5, we identified some units that encoded affective information across several languages (up to 11 in XLM-R). We considered those units as the implementational substrate that performs affective semantic computations across languages. In this section, we studied whether their individual activation patterns alone were sufficient to encode a significant amount of affective information. To increase the generalizability of our findings and further assess the cross-lingual generalizations of

[7]Note that the expected intersection size (i.e. $FE$ = 1) under the assumption of independence is 13.02.

[8]To increase readability, the set of the top 100 units selected for a given language is reported with the ISO 639-1 code of that language (Italian: It, Spanish: Es, Greek: El, Dutch: Nl, French: Fr, Turkish: Tr, German: De, English: En, Chinese: Zh, Polish: Pl, Croatian: Hr, Indonesian: Id, Portuguese: Pt).




(a) mBERT, valence (b) mBERT, arousal (c) XLM-R, valence (d) XLM-R, arousal

Figure 3: Layer-wise results of the linear probes for each language, divided by model (XLM-R and mBERT) and affective variable (valence, arousal). The layer 0 includes the embedding layer representations. The layer-wise scores are fit as 8th degree polynomials for readability purposes. The thick black/gray line indicates the scores averaged across languages (without polynomial fit).

those units, we tested the units in Finnish, a language that did not concur in the unit selection process. Note that Finnish belongs to the Uralic language family, and thus is typologically distant from all the languages considered in the previous experiments and on which the unit selection process was based (see Table 1).

The Finnish dataset comprised 210 words (Eilola and Havelka, 2010); following the same procedure as in §4.1, we searched for 15 sentences for each word in the Finnish Wikipedia; only 123 words had a sufficient number of occurrences, resulting in a corpus of 1,845 sentences. We then employed the output of each of the neurons identified in §5 taken singularly as a predictor in a linear regression model, with Eilola and Havelka's affective norms as dependent variables. As an additional baseline beyond chance level, we also randomly sampled 100 neural units from the corresponding layers in the two models (layer 7 for mBERT, and layers 5 and 6 for XLM-R), and compared their results with the ones obtained by the target neurons.

The results of the single-unit analyses in Finnish are summarized in Table 2. Overall, the units singled out in §5 were able to significantly predict arousal and valence values in Finnish in the majority of the cases (75%, i.e., six out of eight neurons), and to do so better than the randomized baseline (62.5%). A notable exception to this trend are the units $u_{5665}$ and $u_{6077}$ in the case of mBERT, and $u_{4313}$ in the case of XLM-R, which perform worse than randomly sampled units from the same layers. Predictably, the stronger results were obtained from the units that had been identified in higher-degree intersections ($u_{4324}$ and $u_{4747}$), and thus were encoding semantic content in more languages. These results show that some of the neural units that were responsible for encoding valence and arousal in a set of languages can successfully capture the same properties in Finnish, a different language that did not concur in the unit selection.

While the units identified in §5 outperformed the randomized baseline in most of the cases, it should be noted that, on average, the network units of the relevant layers achieve above-chance performance in the regression. This suggests that while some units have a preferential association with specific semantic content, the computation of affective meaning is not bounded to a small cluster of specialized units, but is instead processed with redundancy across the intermediate layers of the networks.

| | | | Target | | | Baseline | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Var** | $\mathbf{u_i}$ | $\hat{\mathbf{B}}$ | ***t*** | ***p*** | $\hat{\mathbf{B}}_\mathbf{b}$ | $t_b$ | $p_b$ |
| mBERT | v | 6078 | -0.027 | -4.066 | <.001 | 0.021 | 3.222 | <.001 |
| mBERT | v | 5665 | -0.001 | -0.111 | 0.911 | 0.021 | 3.222 | <.001 |
| mBERT | v | 6077 | -0.002 | -0.268 | 0.789 | 0.021 | 3.222 | <.001 |
| mBERT | v | 5685 | 0.028 | 4.189 | <.001 | 0.021 | 3.222 | <.001 |
| mBERT | a | 6038 | -0.045 | -6.733 | <.001 | 0.030 | 4.417 | <.001 |
| XLM-R | v | 4324 | 0.076 | 11.881 | <.001 | 0.030 | 4.538 | <.001 |
| XLM-R | v | 4313 | 0.015 | 2.175 | 0.030 | 0.030 | 4.538 | <.001 |
| XLM-R | a | 4747 | 0.083 | 12.780 | <.001 | 0.031 | 4.649 | <.001 |

Table 2: Results of the single neurons $u_i$ identified in §5 on the Finnish dataset grouped by model and affective variable (Var). The results of the target neurons ($\hat{B}$, $t$, $p$) are paired with a baseline that averages the results obtained from 100 randomly sampled units ($\hat{B}_b$, $t_b$, $p_b$).

## 7. Cross-lingual overlap and performance

In a second follow-up study, we inspected the relationship between the cross-lingual alignment of the representational subspace occupied by a language in a model and the model performance in that language. Theoretically, two opposite patterns may link representational quality and alignment. First, if the cross-lingual encoding in a multilingual network is driven by competition for the finite parameter space (Dufter and Schütze, 2020), one may expect that high-resource languages, being largely represented in the training data, would occupy larger language-specific subspaces. Consequently, low-resource languages would tax more strongly the

language-neutral components of the network, unable to seize exclusive network subspaces. Alternatively, if cross-lingual alignment is not a by-product of compression but rather an inherent tendency in multilingual representation learning, we would expect performance to be directly proportional to cross-lingual alignment. To test these hypotheses, we defined a metric of cross-lingual alignment of a language $l_i$ as the mean pairwise overlap (MPO) of the intersection with all the other $l \in L \setminus \{l_i\}$ languages (see Eq. 2).

$$MPO(l_i) = \frac{1}{n-1} \sum_{l \in L \setminus \{l_i\}}^{n-1} |\, l \cap l_i \,| \qquad (2)$$

Where $n$ is the number of $L$ languages. We then employed the MPO measure to predict the performance obtained in encoding the affective variables in each language, defined once again as the Pearson correlation coefficient between predictions and target affective values obtained by the probes in §5. Our choice to measure performance using the Pearson scores from the probes (thus, focusing specifically on encoding valence and arousal) was deliberate: since our primary goal was to examine the overlap in how models encode these semantic dimensions, we chose a measure reflecting the encoding of such dimensions, rather than a more general language model performance metric. To account for the hierarchical nature of the data, where observations are grouped by language, model type (mBERT and XLM-R) and output variable (valence and arousal), we fit a linear mixed-effects regression with uncorrelated random slopes and intercepts for model type and output variable, and random intercepts for languages. We found a significant, positive relationship between cross-lingual alignment and performance ($\hat{B}$ = 0.0188, SE = 0.0047, $t$ = 4.033, $p$ = 0.012), which we represented graphically in Figure 4. This result empirically supports the idea that representational quality is positively associated with MPO, challenging the competitive account of cross-lingual alignment.

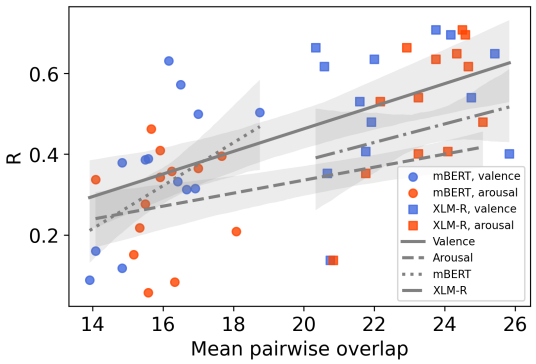



Figure 4: Relationship between mean pairwise intersection size and model performance, expressed as the Pearson correlation coefficient between predictions and targets.

## 8. Discussion

Our analyses showed signs of cross-lingual consistency in the encoding of affective semantic variables, both in terms of layer-wise flow of information and within-layer organization. At the layer level, affective information peaked in the intermediate layers of the networks in most languages, with a second local *maximum* in layer 11. From a neuron-level perspective, emotional information tended to converge towards the same units within a layer across languages. Most pairwise intersections showed an overrepresentation with respect to their expected size assuming independence, and some individual network units were implicated in affective content processing in several languages (up to 11), showing that multilingual training is indeed sufficient to develop a relative *interlingua*. Interestingly, by simply processing non-aligned text in several languages, mBERT and XLM-R developed implicit knowledge about specific semantic content; the representational format of this knowledge was sufficiently abstracted from the superficial features of the input to be encoded in partially overlapping subspaces across linguistic boundaries. Furthermore, the single units identified in our first analyses were generally able to capture affective content in Finnish, a language that was not considered in the unit selection process and belonged to a different language family. This result showed that single units can coherently encode semantic content in a quasi-symbolic format across a set of diverse languages. However, when the activation patterns of those neurons were compared with randomly sampled units from the same layers, it turned out that affective information was encoded with redundancy in the multilingual networks.

Our study also revealed a positive relationship between cross-lingual alignment and performance. The direction of this relationship suggests that cross-lingual alignment is not a by-product of the competition for a finite parameter space, but rather the result of a synergetic process that organizes encoded information according to language-invariant strategies. However, an important *caveat* must be made with respect to this observation. In §7, we reported correlational evidence that links performance and MPO, but we did not experimentally manipulate the two factors. Hence, it is problematic to infer the direction of the causal relationship that ties these two variables. On the one hand, cross-lingual alignment might be directly beneficial in terms of performance; under this account, sublexical representations in a given language might take

advantage of both *within*- and *across*-languages statistical relationships between subword tokens in the vocabulary. Languages that are on average more aligned with other languages might benefit more from the *across*-languages statistical information encoded in the parameters of the model. On the other hand, the causal relationship might follow a reverse pattern, where performance increases produce enhanced alignment. Under this account, the target linguistic encoding to be learned during pre-training might be a language-neutral objective, and the extent to which each language is encoded in accordance with this abstract, language-independent representation might determine the MPO of the representations produced by the network in that language. Adjudicating between these two competing accounts requires the experimental manipulation of MPO and performance, but it is not trivial to develop a methodology for manipulating one of the two variables without affecting the other, so we leave this matter for future research.

Across experiments and follow-up analyses, we consistently observed that cross-lingual convergence was more pronounced when considering XLM-R as opposed to mBERT, in accordance with what has been reported in the context of (morpho)syntactic probing (see Antverg and Belinkov, 2021; Stanczak et al., 2022; de Varda and Marelli, 2023); this corroborates the previous observation that the next sentence prediction objective is not a determining factor of cross-lingual alignment, and that the amount of pre-training data and, consequently, model performance, is associated with the development of stronger language-neutral internal components. This result nicely mirrors the positive relationship between representational quality and cross-lingual convergence at the highest level of analysis of this study, i.e. model comparison, with the best-performing model being also the one which shows the greater representational alignment across languages.

## 9. Conclusion

In this study, we presented the first neuron-level analysis targeting the cross-lingual encoding of specific semantic content in mBERT and XLM-R. Our results confirmed the observation that MMMs encode linguistic information in both language-dependent and language-independent subspaces (see also Doddapaneni et al., 2021; Gonen et al., 2022), but crucially revealed that the reliance on the latter is associated with enhanced performance. While previous research on MMMs has considered sentence representations as a whole or restricted its focus to (morpho)syntactic information, we narrowed our study to affective content in lexical representations, showing that MMMs tend to allocate a set of partially overlapping units to the construction of affective meaning. We additionally observed an example case of symbol emergence (Taniguchi et al., 2018), where semantic knowledge arose from the activation patterns of individual units across languages. We hope that our results will set the stage for future studies examining the behavior of single units in MMMs in relation to other facets of the semantic spectrum.

## Limitations

In this set of studies, we showed that a relatively specific expression of semantic content is encoded, at least in part, by language-neutral parameters in mBERT and XLM-R. However, we must acknowledge that valence and arousal constitute only two of the multitude of semantic dimensions that contribute to lexical meaning. While narrowing our focus to specific semantic information arguably increases the internal validity of our study, it negatively affects its external validity. Hence, the generalizability of our findings to other semantic dimensions should be assessed in future research. Furthermore, our study focuses on lexical semantic variables, but the extent to which our findings can be extended to sentence-level semantic properties needs to be carefully evaluated. A third shortcoming of our study consists in the language sample considered, which is dominated by Indoeuropean languages written in Latin scripts. It has been shown that both typological similarity (Pires et al., 2019) and script (Muller et al., 2020) are intervening factors in cross-lingual alignment; hence, it would be desirable to test our hypotheses on a more heterogeneous language sample. However, we also considered three non-Indoeuropean languages (Chinese, Turkish, and Indonesian), which are very informative for the aims of our study. Additionally, our zero-shot test presented in §6 is performed in Finnish, which belongs to the Uralic family. We believe that the results obtained on the more typologically diverse languages are particularly informative for the extent of the generalization abilities of massively multilingual models.

## 10. Bibliographical References

Omer Antverg and Yonatan Belinkov. 2021. On the pitfalls of analyzing individual neurons in language models. In *International Conference on Learning Representations*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about

morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Charles S Carver and Michael F Scheier. 1990. Origins and functions of positive and negative affect: A control-process view. *Psychological review*, 97(1):19.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577.

Francesca MM Citron. 2012. Neural correlates of written emotion word processing: a review of recent electrophysiological and hemodynamic neuroimaging studies. *Brain and language*, 122(3):211–226.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

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.

Bojana Ćoso, Marc Guasch, Pilar Ferré, and José Antonio Hinojosa. 2019. Affective and concreteness norms for 3,022 croatian words. *Quarterly Journal of Experimental Psychology*, 72(9):2302–2312.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019a. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317.

Fahim Dalvi, Avery Nortonsmith, Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. 2019b. Neurox: A toolkit for analyzing individual neurons in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9851–9852.

Andrea Gregor de Varda and Marco Marelli. 2023. Data-driven Cross-lingual Syntax: An Agreement Study with Massively Multilingual Models. *Computational Linguistics*, pages 1–39.

Maksym Del and Mark Fishel. 2021. Establishing interlingua in multilingual language models. *ArXiv*, abs/2109.01207.

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.

Sumanth Doddapaneni, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2021. A primer on pretrained multilingual language models. *arXiv preprint arXiv:2107.00676*.

Philipp Dufter and Hinrich Schütze. 2020. Identifying necessary elements for bert's multilinguality. *arXiv preprint arXiv:2005.00396*.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. *arXiv preprint arXiv:2010.02695*.

Tiina M Eilola and Jelena Havelka. 2010. Affective norms for 210 british english and finnish nouns. *Behavior Research Methods*, 42(1):134–140.

Lisa Feldman Barrett and James A Russell. 1998. Independence and bipolarity in the structure of current affect. *Journal of personality and social psychology*, 74(4):967.

Hila Gonen, Shauli Ravfogel, and Yoav Goldberg. 2022. Analyzing gender representation in multilingual models. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 67–77.

Kamil K. Imbir. 2016. Affective norms for 4900 polish words reload (anpw_r): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Frontiers in Psychology*, 7.

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.

Aycan Kapucu, Aslı Kılıç, Yıldız Özkılıç, and Bengisu Sarıbaz. 2021. Turkish emotional word norms for arousal, valence, and discrete emotion categories. *Psychological reports*, 124(1):188–209.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

K Karthikeyan, Wang Zihan, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Johanna Kissler, Cornelia Herbert, Peter Peyk, and Markus Junghofer. 2007. Buzzwords: early cortical responses to emotional words during reading. *Psychological science*, 18(6):475–480.

Stavroula-Thaleia Kousta, David P Vinson, and Gabriella Vigliocco. 2009. Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, 112(3):473–481.

Lars Kuchinke, Arthur M Jacobs, Claudia Grubich, Melissa L-H Vo, Markus Conrad, and Manfred Herrmann. 2005. Incidental effects of emotional valence in single word processing: an fmri study. *NeuroImage*, 28(4):1022–1032.

Victor Kuperman, Zachary Estes, Marc Brysbaert, and Amy Beth Warriner. 2014. Emotion and language: valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, 143(3):1065.

Peter Kuppens, Francis Tuerlinckx, James A Russell, and Lisa Feldman Barrett. 2013. The relation between valence and arousal in subjective experience. *Psychological bulletin*, 139(4):917.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Randy J Larsen, Kimberly A Mercer, and David A Balota. 2006. Lexical characteristics of words used in emotional stroop experiments. *Emotion*, 6(1):62.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

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*.

Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2020. Exploring fine-tuning techniques for pre-trained cross-lingual models via continual learning. *arXiv preprint arXiv:2004.14218*.

Catherine Monnier and Arielle Syssau. 2014. Affective norms for french words (fan). *Behavior research methods*, 46(4):1128–1137.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2013. The adaptation of the affective norms for english words (anew) for italian. *Behavior Research Methods*, 46.

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 dutch words. *Behavior research methods*, 45(1):169–177.

Benjamin Muller, Antonis Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2020. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. *arXiv preprint arXiv:2010.12858*.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Elisavet Palogiannidi, Polychronis Koutsakis, E Losif, and Alexandros Potamianos. 2016. Affective lexicon creation for the greek language.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT?

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3):600–605.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

James A Russell and Lisa Feldman Barrett. 1999. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

David S Schmidtke, Tobias Schröder, Arthur M Jacobs, and Markus Conrad. 2014. Angst: Affective norms for german sentiment terms, derived from the affective norms for english words. *Behavior research methods*, 46(4):1108–1118.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Agnes Sianipar, Pieter van Groenestijn, and Ton Dijkstra. 2016. Affective meaning, concreteness, and subjective frequency norms for indonesian words. *Frontiers in Psychology*, 7.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Ana Soares, Montserrat Comesaña, Ana Pinheiro, Alberto Simões, and Sofia Frade. 2011. The adaptation of the affective norms for english words (anew) for european portuguese. *Behavior research methods*, 44:256–69.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Tadahiro Taniguchi, Emre Ugur, Matej Hoffmann, Lorenzo Jamone, Takayuki Nagai, Benjamin Rosman, Toshihiko Matsuka, Naoto Iwahashi, Erhan Oztop, Justus Piater, et al. 2018. Symbol emergence in cognitive developmental systems: a survey. *IEEE transactions on Cognitive and Developmental Systems*, 11(4):494–516.

Jos JA Van Berkum. 2010. The brain is a prediction machine that cares about good and bad-any implications for neuropragmatics? *Italian Journal of Linguistics*, 22:181–208.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alessandra Vergallito, Marco Alessandro Petilli, Luigi Cattaneo, and Marco Marelli. 2019. Somatic and visceral effects of word valence, arousal and concreteness in a continuum lexical space. *Scientific reports*, 9(1):1–10.

Minghui Wang, Yongzhong Zhao, and Bin Zhang. 2015. Efficient test and visualization of multi-set intersections. *Scientific reports*, 5(1):1–12.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791, Online only. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

*Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Zhao Yao, Jia Wu, Yanyan Zhang, and Zhenhong Wang. 2017. Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 chinese words. *Behavior research methods*, 49(4):1374–1385.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.